Deep Neural Networks

Instructor: Lei Wu [1]

Topics in Deep Learning Theory (Spring 2024)

Peking University, Spring 2024

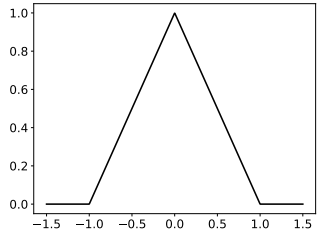
[1]School of Mathematical Sciences; Center for Machine Learning Research

Outline

- Deep Nets Have Stronger Adaptivity to (Anisotropic) Smoothness .
- Depth Separation via Dimension.

Piecewise Linear Approximation

Let $t(x) = \max(0, 1 - |x|)$ be the triangular function.



Theorem 1

Let $f : [0, 1] \mapsto \mathbb{R}$ with $\sup_{x\in[0,1]} |f''(x)| \leq M$. For any $n \in \mathbb{N}$, let $h = \frac{1}{n}$ and $x_i = \frac{ih}{n}$ the grid points. Consider the piecewise linear interpolation:

$$\mathcal{P}_n f = \sum_{i=1}^{n} f(x_i) t\left(\frac{x - x_i}{h}\right).$$

Then, we have

$$\|\mathcal{P}_n f - f\|_2 \lesssim \frac{M}{n^2}$$

Triangular map

- Consider the shift triangular map given by $g(x) = t(2x - 1)$ which $g : [0, 1] \mapsto \mathbb{R}$.
- $$g_l(x) = \underbrace{g \circ g \circ \cdots \circ g}_{l}(x).$$
- An illustration of g (i.e, g_1) and g_l is provided in Figure 1.

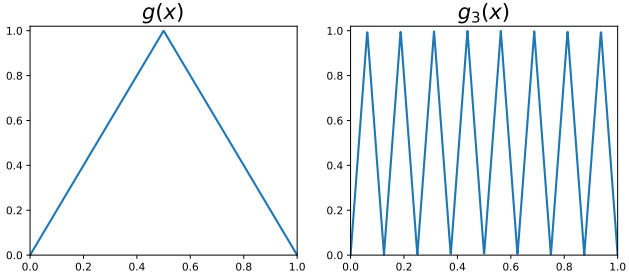


Figure 1: Illustration of the shift triangular function and the triangular wave.

Key Observations:

- g_l has 2^{l+1} linear pieces but can be implemented with a MLP using only l layers.

  First, $g(x) = \mathrm{ReLU}(2x) + \mathrm{ReLU}(2x-2) - 2\,\mathrm{ReLU}(2x-1)$, i.e., g can be exactly represented as three neurons. Hence, g_l can be represented as $2l$-layer neural net with the width less than or equal to 3.
- Two-layer ReLU nets of width m have at most m pieces.
- In summary, for multilayer ReLU nets, the number of pieces grows with the width polynomially but grows with depth exponentially.

This observation already gives us separation between deep and shallow ReLU nets?

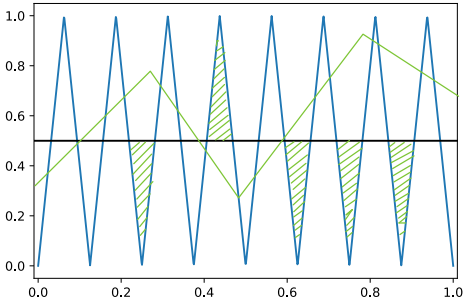
A Quantitative Result

Theorem 2 (Telgarsky, 2016)

Consider the target function $f^* = g_l$. Then, g_l can be implemented as a $c_1 l$-layer neural nets with the width less than c_2. For any 2-layer ReLU net $f_m(\cdot;\theta)$ with the width $m = \mathrm{poly}(l)$, we have

$$\int_0^1 |f_m(x;\theta) - g_l(x)|\,\mathrm{d}x \geq c_3.$$

Here c_1, c_2, c_3 are absolute constants.



Proof

- For a two-layer neural network of width m, let M denote its number of linear pieces. Obviously,

$$M \sim m.$$

- The proof of the lower bound proceeds by counting triangles as illustrated in Figure 2. Draw the horizontal line $y = 1/2$. Then, there are 2^{l+1} (half) triangles.

$$\begin{aligned}\int_0^1 |f_m(x;\theta) - g_l(x)|\,\mathrm{d}x &\geq [\text{number of surviving triangles}] \cdot [\text{area of the triangle}] \\ &\geq \frac{1}{2}(2^{l+1} - 2M) \cdot (\frac{1}{2} \cdot \frac{1}{2^{l+1}} \cdot \frac{1}{2}) \\ &\geq \frac{1}{2}\left(\frac{1}{4} - \frac{M}{2^{l+2}}\right) \geq \frac{1}{16}. \end{aligned} \tag{1}$$

Remark

- Why is it good to choose L^1 norm to measure the approximaiton error?
- Can you establish a similar separation result for L^∞ norm?
- **What is the implication?**
  - Deep nets are good at approximating high-frequency functions?
  - Deep nets are good at approximating non-smooth function?
- Can we claim we establish establish a separation between deep and shallow nets from a frequency perspective?

Approximating x^2 with Shallow Nets

Lemma 3

Let $\mathcal{G}_m$ denote the set of piecewise linear functions with the number of linear pieces less than or equal to m. Then, for any $g \in \mathcal{G}_m$, we have

$$\sup_{x \in [0,1]} |g(x) - x^2| \sim \frac{1}{m^2}.$$

Proof of Lemma 3

- Note that considering one piece is enough. For any $[a, b] \subset [0, 1]$, let

$$I_{a,b} := \min_{c,d\in\mathbb{R}} \max_{t\in[a,b]} |x^2 - cx - d|.$$

- Simplification:

$$\begin{aligned} I_{a,b} &= \min_{c,d\in\mathbb{R}} \max_{t\in[a,b]} |(x-a)^2 + 2ax + a^2 - cx - d| \\ &= \min_{c,d\in\mathbb{R}} \max_{t\in[0,b-a]} |x^2 - 2cx - d| \\ &= \min_{c,d\in\mathbb{R}} \max_{t\in[0,b-a]} |(x-c)^2 - d| \\ &= \min_{c,d} \max\{|(b-a-c)^2 - d|, |c^2 - d|, |d|\}. \end{aligned}$$

- We must have $I_{b,a} \gtrsim (b-a)^2$.
  - If $|c^2 - d| \gtrsim (b-a)^2$ or $|d| \gtrsim (b-a)^2$. Then, $I_{a,b} \gtrsim (b-a)^2$.
  - Otherwise, we must have $c = o(|b-a|), d = o(|b-a|^2)$, under which $(b-a+c)^2 - d \gtrsim |b-a|^2$

Proof of Lemma 3 (Cont'd)

- Since $g \in \mathcal{G}_m$, the number of piecewise linear parts of g is at most m. There must exist a piece $[a, b]$ such that $|b - a| \gtrsim 1/m$. Then,

$$\sup_{x \in [0,1]} |g(x) - x^2| \geq \sup_{x \in [a,b]} |x^2 - cx - d| \gtrsim |b - a|^2 \gtrsim \frac{1}{m^2}.$$

What about the square loss?

Lemma 4

Let $f(x) = ax^2 + bx + c$. If g is at most m pieces. Then

$$\|f - g\|_{L^2([0,1])} \gtrsim \frac{a}{m^2}$$

Proof: First, consider the one-piece case:

$$I_{\alpha,\beta}(f) := \inf_{u,w} \int_\alpha^\beta (ax^2 + bx + c - (ux + w))^2 \,\mathrm{d}x = \inf_{b,c} \int_\alpha^\beta (ax^2 + bx + c)^2 \,\mathrm{d}x$$
$$= \inf_{b,c} \int_{-1}^1 (\beta - \alpha)^4 (az^2 + bz + c)^2 \,\mathrm{d}z = Q_a (\beta - \alpha)^5,$$

where $Q_a = \inf_{b,c} \int_{-1}^1 (az^2 + bz + c)^2 \,\mathrm{d}z$.

What about the square loss?

Lemma 4

Let $f(x) = ax^2 + bx + c$. If g is at most m pieces. Then

$$\|f - g\|_{L^2([0,1])} \gtrsim \frac{a}{m^2}$$

Proof: First, consider the one-piece case:

$$I_{\alpha,\beta}(f) := \inf_{u,w} \int_\alpha^\beta (ax^2 + bx + c - (ux + w))^2 \,\mathrm{d}x = \inf_{b,c} \int_\alpha^\beta (ax^2 + bx + c)^2 \,\mathrm{d}x$$
$$= \inf_{b,c} \int_{-1}^1 (\beta - \alpha)^4 (az^2 + bz + c)^2 \,\mathrm{d}z = Q_a (\beta - \alpha)^5,$$

where $Q_a = \inf_{b,c} \int_{-1}^1 (az^2 + bz + c)^2 \,\mathrm{d}z$. For the general case, denote by $0 = z_0 < z_1 < \cdots < z_m = 1$ the knots. Then,

$$\|f - g\|^2_{L^2([0,1])} = \sum_{j=1}^m I_{z_{j-1},z_j}(f) \geq Q_a \sum_{j=1}^m (z_j - z_{j-1})^5 \gtrsim \frac{Q_a}{m^4}.$$

Compute Q_a

This can be done by using orthogonal polynomials. Let $\{h_n\}_{n=0}^{\infty}$ be the Legendre polynomials, which are orthonormal wrt in $L^2([-1,1])$:

$$h_0(x) = 1, h_1(x) = x, h_2(x) = \frac{3x^2 - 1}{2}, \cdots .$$

For any $f \in L^2([-1,1])$, we have

$$\inf_{c_0,c_1} \|f - c_1 h_1 - c_0 h_0\|_2^2 = \sum_{j=2}^{\infty} \langle f, h_j \rangle^2.$$

Let $f_a = ax^2$. Then,

$$Q_a = \inf_{c_0,c_1} \|f_a - c_1 h_1 - c_0 h_0\|_2^2 = \langle f_a, h_2 \rangle^2 \sim a^2.$$

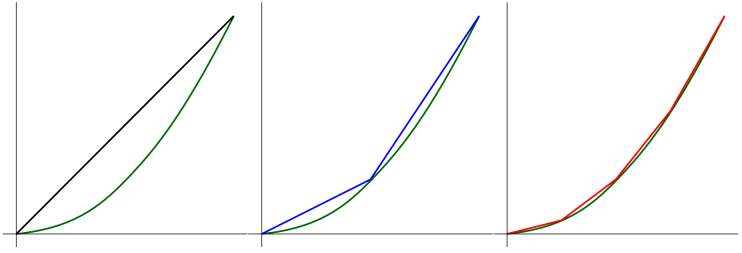
Approximating x^2 with Deep Nets

Proposition 5

For any $\varepsilon > 0$, there exits a neural net $\tilde{f}$, whose depth and width is $O(\log(1/\varepsilon))$ and $O(1)$, respectively, such that

$$\sup_{x \in [0,1]} |\tilde{f}(x) - x^2| \leq \varepsilon.$$

An illustration of the approximation scheme:



Proof

- First, we can show that $l = 2, 3, \ldots,$

$$P_{2^{l-1}} f^*(x) - P_{2^l} f^*(x) = \frac{g_l(x)}{2^{2l}}, \qquad \forall x \in [0, 1]. \tag{2}$$

Proof

- First, we can show that $l = 2, 3, \ldots,$

$$P_{2^{l-1}} f^*(x) - P_{2^l} f^*(x) = \frac{g_l(x)}{2^{2l}}, \qquad \forall x \in [0, 1]. \tag{2}$$

- Construct a neural net as follows

$$\begin{aligned} y_0 &= x \\ y_l &= g(y_{l-1}) \\ \tilde{f}(x) &= \sum_{l=1}^{L} \frac{y_l}{2^l}. \end{aligned}$$

Proof

- First, we can show that $l = 2, 3, \ldots,$

$$P_{2^{l-1}} f^*(x) - P_{2^l} f^*(x) = \frac{g_l(x)}{2^{2l}}, \qquad \forall x \in [0, 1]. \tag{2}$$

- Construct a neural net as follows

$$\begin{aligned} y_0 &= x \\ y_l &= g(y_{l-1}) \\ \tilde{f}(x) &= \sum_{l=1}^{L} \frac{y_l}{2^l}. \end{aligned}$$

- The last step introduces skip connections from each layer to the output layer. So, the depth and width of this net is $O(L)$ and $O(1)$, respectively.

Proof

- First, we can show that $l = 2, 3, \ldots,$

$$P_{2^{l-1}} f^*(x) - P_{2^l} f^*(x) = \frac{g_l(x)}{2^{2l}}, \qquad \forall x \in [0, 1]. \tag{2}$$

- Construct a neural net as follows

$$\begin{aligned} y_0 &= x \\ y_l &= g(y_{l-1}) \\ \tilde{f}(x) &= \sum_{l=1}^{L} \frac{y_l}{2^l}. \end{aligned}$$

- The last step introduces skip connections from each layer to the output layer. So, the depth and width of this net is $O(L)$ and $O(1)$, respectively.

Proof

- First, we can show that $l = 2, 3, \ldots,$

$$P_{2^{l-1}} f^*(x) - P_{2^l} f^*(x) = \frac{g_l(x)}{2^{2l}}, \qquad \forall x \in [0,1]. \tag{2}$$

- Construct a neural net as follows

$$\begin{aligned} y_0 &= x \\ y_l &= g(y_{l-1}) \\ \tilde{f}(x) &= \sum_{l=1}^{L} \frac{y_l}{2^l}. \end{aligned}$$

- The last step introduces skip connections from each layer to the output layer. So, the depth and width of this net is $O(L)$ and $O(1)$, respectively.

By Lemma 1,

$$\sup_{x\in[0,1]} |\tilde{f} - f(x)| = \sup_{x\in[0,1]} |P_{2^L} f(x) - f(x)| \lesssim \frac{1}{4^L}.$$

Taking $1/(4^L) = \varepsilon$, we complete the proof.

Why is approximating x^2 interesting?

From the approximation of $f(x) = x^2$, we can get many other results.

- Fast approximation of the *multiplication* $(x, y) \mapsto xy$ using

$$xy = \frac{(x+y)^2 - x^2 - y^2}{2}.$$

- Fast approximation of any monomials: x^k.
- Fast approximation of polynomials: $a_0 + a_1 x + \cdots + a_k x^k$.
- Fast approximation of functions that can be efficiently approximated by polynomials, e.g., Sobolev spaces.

Remark: The above argument implies that for achieving precison ϵ, deep ReLU nets with $L = \log(1/\epsilon)$ performs as well as polynomials.

Approximating Sobolev Spaces with Deep ReLU Nets

Theorem 6 (Yarotsky (2017))

Assume that $\|f\|_{W^{k,\infty}} := \max_{|\alpha|\le k} \operatorname{ess\,sup}_{x\in[0,1]^d} |D^\alpha f(x)| \le 1$. *Then, there exists a ReLU* $\tilde{f}$ *of depth at most* $O(\log(1/\varepsilon)+1)$ *and width at most* $O(\varepsilon^{-d/k}(\log(1/\varepsilon)+1))$ *such that*

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

Here, the constant C *depends on* d, k.

Remark:

- The result only separates deep and shallow nets for the *non-smooth* ReLU activation.
- If considering smooth activation function, no such separation exists.

More Thoughts

- Can we separate deep nets from shallow nets by expoliting "smoothness"?
- Can deep nets learn less-smooth functions ? See (Bresler and Nagaraj, 2020).
- Can deep nets adapt to anisotropic smoothness? See (Suzuki and Nitanda, 2021)
- How about target functions defined over a compact manifold with $\dim \mathcal{X} \ll d$?

Depth Separation in High Dimension

Depth Separation via Dimension

There exist functions $f_d : \mathbb{R}^d \mapsto \mathbb{R}$ such that approximating with deep nets require only $\mathrm{poly}(d)$ parameters but shallow networks require at least $\exp(d)$ parameters.

Depth Separation via Dimension

There exist functions $f_d : \mathbb{R}^d \mapsto \mathbb{R}$ such that approximating with deep nets require only $\mathrm{poly}(d)$ parameters but shallow networks require at least $\exp(d)$ parameters.

Relevant works:

- [Daniely, 2017] Depth separation for neural networks, COLT 2017 (only 6 pages).
- [Eldan and Shamir, 2016] The power of depth for feedforward neural network, COLT 2016.
- [Luca et al., 2021] *Depth separation beyond radial functions*, JMLR 2021.

The Daniely's Result

Theorem 7 (Daniely 2017)

Let $\mathcal{X} = \mathbb{S}^{d-1} \otimes \mathbb{S}^{d-1}$ *and input distribition* $\rho = \mathrm{Unif}(\mathbb{S}^{d-1})$. *Consider target function* $f(x, y) = h(x^\top y)$ *for* $h(z) := \sin\left(\pi d^3 z\right)$. *Then, we have*

- ϵ-*approximable by* **depth-3** ReLU *network of poly* $(d, 1/\epsilon)$ *width and weight sizes*
- *Not* $\Omega(1)$-*approximable by any* **depth-2 ReLU** *network of* $\exp(o(d \log d))$ *width and* $\mathcal{O}(\exp(d))$-*sized weights.*

More generally: the separation holds for any $h : [-1, 1] \mapsto \mathbb{R}$ which is inapproximable with $O\left(d^{1+\epsilon}\right)$-degree polynomial

Proof idea

- 2LNN implements the sum of m separable functions:

$$\sum_{j=1}^{m} a_j \sigma(w_j^\top x + v_j^\top y + b_j) = \sum_{j=1}^{m} a_j \varphi_j(w_j^\top x, v_j^\top y).$$

Proof idea

- 2LNN implements the sum of m separable functions:

$$\sum_{j=1}^{m} a_j \sigma(w_j^\top x + v_j^\top y + b_j) = \sum_{j=1}^{m} a_j \varphi_j(w_j^\top x, v_j^\top y).$$

- $(x, y) \mapsto h(x^\top y)$ are nearly orthogonal to any separable function $(x, y) \mapsto \psi(w^\top x, v^\top y)$.
- What is the intuition behind?

Harmonic Analysis on $\mathbb{S}^{d-1}$

- A multivariate polynomial p is said to be harmonic if $\Delta p = 0$. For instance, the harmonic polynomials up to degree 3 is given

$$1, \quad x, y, \quad xy, x^2 - y^2, \quad y^3 - 3x^2y, x^3 - 3xy^2.$$

Harmonic Analysis on $\mathbb{S}^{d-1}$

- A multivariate polynomial p is said to be harmonic if $\Delta p = 0$. For instance, the harmonic polynomials up to degree 3 is given

$$1, \quad x, y, \quad xy, x^2 - y^2, \quad y^3 - 3x^2y, x^3 - 3xy^2.$$

- Let $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ and $\tau_{d-1} = \mathrm{Unif}(\mathbb{S}^{d-1})$.

Harmonic Analysis on $\mathbb{S}^{d-1}$

- A multivariate polynomial p is said to be harmonic if $\Delta p = 0$. For instance, the harmonic polynomials up to degree 3 is given

$$1, \quad x, y, \quad xy, x^2 - y^2, \quad y^3 - 3x^2y, x^3 - 3xy^2.$$

- Let $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ and $\tau_{d-1} = \mathrm{Unif}(\mathbb{S}^{d-1})$.
- **Spherical harmonics:** Let $\mathcal{Y}_k^d$ be the space of all homogeneous harmonic polynomials of degree k in d dimensions restricted on $\mathbb{S}^{d-1}$; the dimension of the space $\mathcal{Y}_k^d$ is

$$N(d,k) = \frac{2k+d-2}{k}\binom{k+d-3}{d-2}.$$

Harmonic Analysis on $\mathbb{S}^{d-1}$

- A multivariate polynomial p is said to be harmonic if $\Delta p = 0$. For instance, the harmonic polynomials up to degree 3 is given

$$1, \quad x, y, \quad xy, x^2 - y^2, \quad y^3 - 3x^2y, x^3 - 3xy^2.$$

- Let $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ and $\tau_{d-1} = \mathrm{Unif}(\mathbb{S}^{d-1})$.
- **Spherical harmonics:** Let $\mathcal{Y}_k^d$ be the space of all homogeneous harmonic polynomials of degree k in d dimensions restricted on $\mathbb{S}^{d-1}$; the dimension of the space $\mathcal{Y}_k^d$ is

$$N(d,k) = \frac{2k+d-2}{k}\binom{k+d-3}{d-2}.$$

- Let $\{Y_{j,k}\}_{1 \le j \le N(d,k)}$ be an orthogonal basis of $\mathcal{Y}_k^d$ in $L^2(\tau_{d-1})$.

Harmonic Analysis on $\mathbb{S}^{d-1}$

- A multivariate polynomial p is said to be harmonic if $\Delta p = 0$. For instance, the harmonic polynomials up to degree 3 is given

$$1, \quad x, y, \quad xy, x^2 - y^2, \quad y^3 - 3x^2y, x^3 - 3xy^2.$$

- Let $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ and $\tau_{d-1} = \mathrm{Unif}(\mathbb{S}^{d-1})$.
- **Spherical harmonics:** Let $\mathcal{Y}_k^d$ be the space of all homogeneous harmonic polynomials of degree k in d dimensions restricted on $\mathbb{S}^{d-1}$; the dimension of the space $\mathcal{Y}_k^d$ is

$$N(d,k) = \frac{2k+d-2}{k}\binom{k+d-3}{d-2}.$$

- Let $\{Y_{j,k}\}_{1 \le j \le N(d,k)}$ be an orthogonal basis of $\mathcal{Y}_k^d$ in $L^2(\tau_{d-1})$.
- Then $\{Y_{j,k}\}_{k \in \mathbb{N}, 1 \le j \le N(d,k)}$ forms an orthogonal basis of $L^2(\tau_{d-1})$.

Legendre Polynomials

- Let $\pi_d \in \mathcal{P}([-1,1])$ be the distribution of x_1 for $x = (x_1, \ldots, x_d) \sim \tau_{d-1}$, whose support is $[-1,1]$ with density given by

$$\pi_d(z) = \frac{(1-z^2)^{\frac{d-3}{2}}}{B\left(\frac{1}{2}, \frac{d-1}{2}\right)}$$

[2] P_k should depends on d. We omit this dependence for notation brevity.

Legendre Polynomials

- Let $\pi_d \in \mathcal{P}([-1,1])$ be the distribution of x_1 for $x = (x_1, \ldots, x_d) \sim \tau_{d-1}$, whose support is $[-1,1]$ with density given by

$$\pi_d(z) = \frac{(1-z^2)^{\frac{d-3}{2}}}{B\left(\frac{1}{2}, \frac{d-1}{2}\right)}$$

- Legendre poynomials $\{P_k\}_{k=0}^{\infty}$ are the orthogonal polynomials [2] with respect to $L^2(\pi_d)$:

$$\langle P_k, P_j \rangle_{\pi_d} = \frac{\delta_{jk}}{N(d,k)}. \tag{3}$$

We shall use $p_k = \sqrt{N(d,k)} P_k$ to denote the normalized Legendre polynomial.

[2] P_k should depends on d. We omit this dependence for notation brevity.

Properties of Legendre poynomials

- P_k satisfies the following recursive formula

$$\begin{aligned} &P_0(t) = 0,\ P_1(t) = t, \\ &P_k(t) = \frac{2k+d-4}{k+d-3} t P_{k-1}(t) - \frac{k-1}{k+d-3} P_{k-2}(t),\ k \geq 2. \end{aligned} \tag{4}$$

- The Rodrigues's formula gives a closed-form expression of P_k:

$$P_k(t) = \left(-\frac{1}{2}\right)^k \frac{\Gamma((d-1)/2)}{\Gamma(k+(d-1)/2)} \left(1-t^2\right)^{(3-d)/2} \left(\frac{d}{dt}\right)^k \left(1-t^2\right)^{k+(d-3)/2}. \tag{5}$$

  The polynomial P_k is even (resp. odd) when k is even (resp. odd).

- Additionally, $|P_k(z)| \leq P(0) = 1$.

A Visualization of Legendre Polynomials

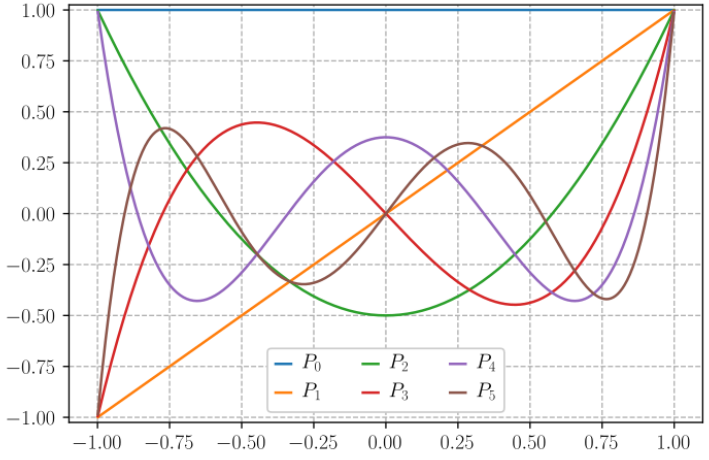


Figure 3: The Legendre polynomials with respect $\pi_d = \mathrm{Unif}([-1, 1])$, i.e., $d = 3$. This figure is taken from wikipedia.

Important Facts

- The spherical harmonics is related to the Legendre polynomials:

$$\frac{1}{N(d,k)^{1/2}} \sum_{j=1}^{N(d,k)} Y_{k,j}(x) Y_{k,j}(y) = p_k(x^\top y). \tag{6}$$

Important Facts

- The spherical harmonics is related to the Legendre polynomials:

$$\frac{1}{N(d,k)^{1/2}} \sum_{j=1}^{N(d,k)} Y_{k,j}(x) Y_{k,j}(y) = p_k(x^\top y). \tag{6}$$

- **The Hecke-Funk formula:** For any $h : [-1, 1] \mapsto \mathbb{R}$, $x \in \mathbb{S}^{d-1}$ and $Y_k \in \mathcal{Y}_k^d$, we have

$$\mathbb{E}_y[h(x^\top y) Y_k(y)] = \frac{1}{N(d,k)^{1/2}} \langle h, p_k \rangle_{\pi_d} Y_k(x). \tag{7}$$

  This implies that spherical harmonics are the eigenfunctions of integral operators induced by inner-product functions.

Important Facts

- The spherical harmonics is related to the Legendre polynomials:

$$\frac{1}{N(d,k)^{1/2}} \sum_{j=1}^{N(d,k)} Y_{k,j}(x) Y_{k,j}(y) = p_k(x^\top y). \tag{6}$$

- **The Hecke-Funk formula:** For any $h : [-1, 1] \mapsto \mathbb{R}$, $x \in \mathbb{S}^{d-1}$ and $Y_k \in \mathcal{Y}_k^d$, we have

$$\mathbb{E}_y[h(x^\top y) Y_k(y)] = \frac{1}{N(d,k)^{1/2}} \langle h, p_k \rangle_{\pi_d} Y_k(x). \tag{7}$$

  This implies that spherical harmonics are the eigenfunctions of integral operators induced by inner-product functions.

- Let $p_i^u(x) := p_i(u^\top x)$. Then, (6) and (7) gives

$$\langle p_i^u, p_j^v \rangle = \mathbb{E}_{x \sim \tau_{d-1}}[p_i(u^\top x) p_j(v^\top x)] = \frac{\delta_{i,j}}{N(d,j)^{1/2}} p_j(u^\top v)$$

Inner-product Functions

- Let $f(x, y) = h(x^\top y)$ with $x, y \in \mathbb{S}^{d-1}$ and $h \in L^2(\pi_d)$.
- Let $h(z) = \sum_{k=0}^{\infty} \hat{h}_k p_k(z)$ with $\hat{h}_k = \langle h, p_k \rangle_{\pi_d}$.
- Then, we can decompose an inner-product function via spherical harmonics [3]

$$h(x^\top y) = \sum_{k=0}^{\infty} \hat{h}_k p_k(x^\top y) = \sum_{k=0}^{\infty} \frac{\hat{h}_k}{N(d,k)^{1/2}} \sum_{j=1}^{N(d,k)} Y_{k,j}(x) Y_{k,j}(y)$$

[3] This decomposition implies an inner-product kernel is positive definite iff the "Fourier coefficient" $\{\hat{h}_k\}_k$ are non-negative.

The Separable Function

- Let $(p_i \otimes p_j)(s,t) = p_i(s)p_j(t)$. Then, $\{p_i \otimes p_j\}_{i,j=0}^{\infty}$ form an orthonormal basis of $L^2(\pi_d \otimes \pi_d)$.

The Separable Function

- Let $(p_i \otimes p_j)(s,t) = p_i(s)p_j(t)$. Then, $\{p_i \otimes p_j\}_{i,j=0}^{\infty}$ form an orthonormal basis of $L^2(\pi_d \otimes \pi_d)$.
- The expansion of a separable function is given by

$$\varphi(u^\top x, v^\top y) = \sum_{i,j=0}^{\infty} \hat{\varphi}_{i,j} p_i(u^\top x) p_j(v^\top y),$$

where $\hat{\varphi}_{i,j} = \mathbb{E}_{s,t\sim\pi_d}[\varphi(s,t)p_i(s)p_j(t)]$.

The Separable Function

- Let $(p_i \otimes p_j)(s,t) = p_i(s)p_j(t)$. Then, $\{p_i \otimes p_j\}_{i,j=0}^{\infty}$ form an orthonormal basis of $L^2(\pi_d \otimes \pi_d)$.
- The expansion of a separable function is given by

$$\varphi(u^\top x, v^\top y) = \sum_{i,j=0}^{\infty} \hat{\varphi}_{i,j} p_i(u^\top x) p_j(v^\top y),$$

  where $\hat{\varphi}_{i,j} = \mathbb{E}_{s,t\sim\pi_d}[\varphi(s,t)p_i(s)p_j(t)]$.
- The key observation: Let $\tilde{p}_k(x,y) := p_k(x^\top y)$, $p_k^u(x) = p_k(u^\top x)$ for any $u \in \mathbb{S}^{d-1}$. Then,

$$\langle \tilde{p}_k, p_i^u \otimes p_j^v \rangle = \begin{cases} \frac{1}{N(d,k)} p_k(u^\top v) & \text{if } i = j = k \\ 0 & \text{otherwise .} \end{cases}$$

The Separable Function

- Let $(p_i \otimes p_j)(s,t) = p_i(s)p_j(t)$. Then, $\{p_i \otimes p_j\}_{i,j=0}^{\infty}$ form an orthonormal basis of $L^2(\pi_d \otimes \pi_d)$.
- The expansion of a separable function is given by

$$\varphi(u^\top x, v^\top y) = \sum_{i,j=0}^{\infty} \hat{\varphi}_{i,j} p_i(u^\top x) p_j(v^\top y),$$

  where $\hat{\varphi}_{i,j} = \mathbb{E}_{s,t\sim\pi_d}[\varphi(s,t)p_i(s)p_j(t)]$.
- The key observation: Let $\tilde{p}_k(x,y) := p_k(x^\top y)$, $p_k^u(x) = p_k(u^\top x)$ for any $u \in \mathbb{S}^{d-1}$. Then,

$$\langle \tilde{p}_k, p_i^u \otimes p_j^v \rangle = \begin{cases} \frac{1}{N(d,k)} p_k(u^\top v) & \text{if } i = j = k \\ 0 & \text{otherwise .} \end{cases}$$

- This leads to

$$\langle \tilde{p}_k, \varphi(u^\top \cdot, v^\top \cdot) \rangle = \frac{\hat{\varphi}_{k,k}}{N(d,k)^{1/2}} P_k(u^\top v).$$

  This justifies why inner-product functions are nearly orthogonal to any separable functions.

The Main Result

Given $h : [-1, 1] \mapsto \mathbb{R}$, let

$$A_{n,d}(h) = \inf_{\text{q is a } n\text{-order polynomial}} \|q - h\|_{L^2(\pi_d)}.$$

The Main Result

Given $h : [-1, 1] \mapsto \mathbb{R}$, let

$$A_{n,d}(h) = \inf_{\text{q is a } n\text{-order polynomial}} \|q - h\|_{L^2(\pi_d)}.$$

Theorem 8

Let $g_1, \ldots, g_m$ be r arbitrary separable functions. Then, for any $n \in \mathbb{N}$, it holds that

$$\left\| f - \sum_{r=1}^{m} g_r \right\|^2 \geq A_{n,d}(h) \left(A_{n,d}(h) - \frac{2 \sum_{r=1}^{m} \|g_r\|}{\sqrt{N(d,n)}} \right).$$

The Main Result

Given $h : [-1, 1] \mapsto \mathbb{R}$, let

$$A_{n,d}(h) = \inf_{\text{q is a } n\text{-order polynomial}} \|q - h\|_{L^2(\pi_d)}.$$

Theorem 8

Let $g_1, \ldots, g_m$ be r arbitrary separable functions. Then, for any $n \in \mathbb{N}$, it holds that

$$\left\| f - \sum_{r=1}^{m} g_r \right\|^2 \geq A_{n,d}(h) \left(A_{n,d}(h) - \frac{2 \sum_{r=1}^{m} \|g_r\|}{\sqrt{N(d,n)}} \right).$$

It recovers Theorem 7 by

- taking $n = n_d = \Omega(d)$ and thus $N(d, n) = \exp(d)$;
- taking h such that $A_{n_d,d}(h) = \Omega(1)$.

This explains why we can take $h(z) = \sin(d^3 \pi z)$.

Proof of Theorem 8

- Recall $f(x,y) = \sum_{k=0}^{\infty} \hat{h}_k \tilde{p}_k(x,y)$ and
  $g_t(x,y) = \varphi^{(t)}(u_t^\top x, v_t^\top y) = \sum_{i,j=0}^{\infty} \hat{\varphi}_{i,j}^{(t)} p_i(u_t^\top x) p_j(v_t^\top y) = \sum_{i,j=0}^{\infty} \hat{\varphi}_{i,j}^{(t)} p_i^{u_t} \otimes p_j^{v_t}$

Proof of Theorem 8

- Recall $f(x,y) = \sum_{k=0}^{\infty} \hat{h}_k \tilde{p}_k(x,y)$ and $g_t(x,y) = \varphi^{(t)}(u_t^\top x, v_t^\top y) = \sum_{i,j=0}^{\infty} \hat{\varphi}_{i,j}^{(t)} p_i(u_t^\top x) p_j(v_t^\top y) = \sum_{i,j=0}^{\infty} \hat{\varphi}_{i,j}^{(t)} p_i^{u_t} \otimes p_j^{v_t}$
- Then, we have

$$\|f - \sum_{t=1}^{r} g_t\|^2 =$$

Proof of Theorem 8

- Recall $f(x,y) = \sum_{k=0}^{\infty} \hat{h}_k \tilde{p}_k(x,y)$ and
  $g_t(x,y) = \varphi^{(t)}(u_t^\top x, v_t^\top y) = \sum_{i,j=0}^{\infty} \hat{\varphi}_{i,j}^{(t)} p_i(u_t^\top x) p_j(v_t^\top y) = \sum_{i,j=0}^{\infty} \hat{\varphi}_{i,j}^{(t)} p_i^{u_t} \otimes p_j^{v_t}$
- Then, we have

$$\|f - \sum_{t=1}^{r} g_t\|^2 = \|\sum_{k=0}^{\infty} \hat{h}_k \tilde{p}_k - \sum_{t=1}^{r} \sum_{i,j=0}^{\infty} \hat{\varphi}_{i,j}^{(t)} p_i^{u_t} \otimes p_j^{v_t}\|^2$$

Proof of Theorem 8

- Recall $f(x,y) = \sum_{k=0}^{\infty} \hat{h}_k \tilde{p}_k(x,y)$ and
  $g_t(x,y) = \varphi^{(t)}(u_t^\top x, v_t^\top y) = \sum_{i,j=0}^{\infty} \hat{\varphi}_{i,j}^{(t)} p_i(u_t^\top x) p_j(v_t^\top y) = \sum_{i,j=0}^{\infty} \hat{\varphi}_{i,j}^{(t)} p_i^{u_t} \otimes p_j^{v_t}$
- Then, we have

$$\begin{aligned}\|f - \sum_{t=1}^{r} g_t\|^2 &= \|\sum_{k=0}^{\infty} \hat{h}_k \tilde{p}_k - \sum_{t=1}^{r} \sum_{i,j=0}^{\infty} \hat{\varphi}_{i,j}^{(t)} p_i^{u_t} \otimes p_j^{v_t}\|^2 \\ &= \sum_{k=0}^{\infty} \|\hat{h}_k \tilde{p}_k - \sum_{t=1}^{r} \hat{\varphi}_{k,k}^{(t)} p_k^{u_t} \otimes p_k^{v_t}\|^2\end{aligned}$$

All the cross terms disappear!!!

Proof of Theorem 8

- Recall $f(x,y) = \sum_{k=0}^{\infty} \hat{h}_k \tilde{p}_k(x,y)$ and
  $g_t(x,y) = \varphi^{(t)}(u_t^\top x, v_t^\top y) = \sum_{i,j=0}^{\infty} \hat{\varphi}_{i,j}^{(t)} p_i(u_t^\top x) p_j(v_t^\top y) = \sum_{i,j=0}^{\infty} \hat{\varphi}_{i,j}^{(t)} p_i^{u_t} \otimes p_j^{v_t}$
- Then, we have

$$\begin{aligned}\|f - \sum_{t=1}^{r} g_t\|^2 &= \|\sum_{k=0}^{\infty} \hat{h}_k \tilde{p}_k - \sum_{t=1}^{r} \sum_{i,j=0}^{\infty} \hat{\varphi}_{i,j}^{(t)} p_i^{u_t} \otimes p_j^{v_t}\|^2 \\ &= \sum_{k=0}^{\infty} \|\hat{h}_k \tilde{p}_k - \sum_{t=1}^{r} \hat{\varphi}_{k,k}^{(t)} p_k^{u_t} \otimes p_k^{v_t}\|^2\end{aligned}$$

All the cross terms disappear!!!

Proof of Theorem 8 (Cont'd)

$$\|f - \sum_{t=1}^{r} g_t\|^2 = \sum_{k=0}^{\infty} \|\hat{h}_k \tilde{p}_k - \sum_{t=1}^{r} \hat{\varphi}_{k,k}^{(t)} p_k^{u_t} \otimes p_k^{v_t}\|^2$$

$$\geq \sum_{k=0}^{\infty} \left(\hat{h}_k^2 - 2 \frac{\hat{h}_k}{N(d,k)^{1/2}} \sum_{t=1}^{r} \hat{\varphi}_{k,k}^{(t)} P_k(u_t^\top v_t) \right)$$

Proof of Theorem 8 (Cont'd)

$$\|f - \sum_{t=1}^{r} g_t\|^2 = \sum_{k=0}^{\infty} \|\hat{h}_k \tilde{p}_k - \sum_{t=1}^{r} \hat{\varphi}_{k,k}^{(t)} p_k^{u_t} \otimes p_k^{v_t}\|^2$$

$$\geq \sum_{k=0}^{\infty} \left(\hat{h}_k^2 - 2 \frac{\hat{h}_k}{N(d,k)^{1/2}} \sum_{t=1}^{r} \hat{\varphi}_{k,k}^{(t)} P_k(u_t^\top v_t) \right)$$

$$\geq \sum_{k=0}^{\infty} \left(\hat{h}_k^2 - 2 \frac{|\hat{h}_k|}{N(d,k)^{1/2}} \sum_{t=1}^{r} |\hat{\varphi}_{k,k}^{(t)}| \right)$$

Proof of Theorem 8 (Cont'd)

$$\|f - \sum_{t=1}^{r} g_t\|^2 = \sum_{k=0}^{\infty} \|\hat{h}_k \tilde{p}_k - \sum_{t=1}^{r} \hat{\varphi}_{k,k}^{(t)} p_k^{u_t} \otimes p_k^{v_t}\|^2$$

$$\begin{aligned}
&\geq \sum_{k=0}^{\infty} \left(\hat{h}_k^2 - 2 \frac{\hat{h}_k}{N(d,k)^{1/2}} \sum_{t=1}^{r} \hat{\varphi}_{k,k}^{(t)} P_k(u_t^\top v_t) \right) \\
&\geq \sum_{k=0}^{\infty} \left(\hat{h}_k^2 - 2 \frac{|\hat{h}_k|}{N(d,k)^{1/2}} \sum_{t=1}^{r} |\hat{\varphi}_{k,k}^{(t)}| \right) \\
&\geq \sum_{k=n}^{\infty} \hat{h}_k^2 - \frac{1}{N(d,n)^{1/2}} \sum_{r=1}^{m} \sum_{k=n}^{\infty} |\hat{h}_k| |\hat{\varphi}_{k,k}^{(t)}| \\
&\geq A_{n,d}(h)^2 - \frac{2 A_{n,d}(h) \sum_{r=1}^{m} \|g_r\|}{N(n,d)^{1/2}}.
\end{aligned}$$

The Eland and Shamir's Result

- Result assume that weights are not too large. Really necessary?
- (Eldan and Shamir, 2016) shows that the weight restriction is not necessary.

The Eland and Shamir's Result

- Result assume that weights are not too large. Really necessary?
- (Eldan and Shamir, 2016) shows that the weight restriction is not necessary.

Theorem 9

Assume σ is measuable and satisfies $|\sigma(t)| \leq C(1 + |t|^{\alpha})$ for all $t \in \mathbb{R}$ and some constants $C, \alpha > 0$. Then, there exists a radial function

$$f(x) = g(\|x\|_2)$$

such that

- **3-layer** *MLPs can approximate with* $\mathrm{poly}(d, 1/\epsilon)$ *parameter.*
- *Not* $o(1)$ *approximate by any* 2-**layer MLP** *with* $\exp(o(d))$-*wide.*

The Eland and Shamir's Result

- Result assume that weights are not too large. Really necessary?
- (Eldan and Shamir, 2016) shows that the weight restriction is not necessary.

Theorem 9

Assume σ is measuable and satisfies $|\sigma(t)| \leq C(1 + |t|^{\alpha})$ for all $t \in \mathbb{R}$ and some constants $C, \alpha > 0$. Then, there exists a radial function

$$f(x) = g(\|x\|_2)$$

such that

- **3-layer** *MLPs can approximate with* $\mathrm{poly}(d, 1/\epsilon)$ *parameter.*
- *Not* $o(1)$ *approximate by any* 2-**layer MLP** *with* $\exp(o(d))$-*wide.*

Remark

- Theorem 9 needs impose restriction on the activation functions. Can be obtain separation without any restriction on the weight size and activation functions?

The answer is NO, per the Kolmogorov-Arnold representation theorem, which solved the Hilbert's 13th problem.

Theorem 10

For any $f \in C([0,1]^d)$, there exists $\Phi_j : \mathbb{R} \mapsto \mathbb{R}$ and $\psi_{i,j} : \mathbb{R} \mapsto \mathbb{R}$ such that

$$f(x_1, x_2, \ldots, x_d) = \sum_{i=0}^{2d} \Phi_i \left(\sum_{j=1}^{d} \psi_{i,j}(x_j) \right)$$

Moreover, it can be further simplied as

$$f(x_1, x_2, \ldots, x_d) = \sum_{i=0}^{2d} \Phi \left(\sum_{j=1}^{d} \lambda_i \psi(x_j + \eta i) + i \right)$$

The Intuition via Fourier Analysis

- Consider the Fourier transform $\hat{f}(\xi) = \int f(x) e^{-2\pi i \xi^\top x} \, \mathrm{d}x$.

The Intuition via Fourier Analysis

- Consider the Fourier transform $\hat{f}(\xi) = \int f(x) e^{-2\pi i \xi^\top x} \, \mathrm{d}x$.
- A 2-layer MLP takes the form $N_m(x) = \sum_{j=1}^m n_j(w_j^\top x)$ and thus

$$\hat{N}_m(\xi) = \sum_{j=1}^m \hat{n}_j(w_j^\top \xi) \prod_{i=2}^d \delta(V_j^\top \xi - \cdot),$$

where $V_j \in \mathbb{R}^{d \times (d-1)}$ denotes the orthogonal complement of w_j.

The Intuition via Fourier Analysis

- Consider the Fourier transform $\hat{f}(\xi) = \int f(x) e^{-2\pi i \xi^\top x} \, \mathrm{d}x$.
- A 2-layer MLP takes the form $N_m(x) = \sum_{j=1}^m n_j(w_j^\top x)$ and thus

$$\hat{N}_m(\xi) = \sum_{j=1}^m \hat{n}_j(w_j^\top \xi) \prod_{i=2}^d \delta(V_j^\top \xi - \cdot),$$

  where $V_j \in \mathbb{R}^{d \times (d-1)}$ denotes the orthogonal complement of w_j.
- Thus,

$$\mathrm{supp}(\hat{N}_m) = \cup_{j=1}^m \{y = w_j^\top \xi : \xi \in \mathbb{R}^d\}.$$

The Intuition via Fourier Analysis

- Consider the Fourier transform $\hat{f}(\xi) = \int f(x) e^{-2\pi i \xi^\top x} \, \mathrm{d}x$.
- A 2-layer MLP takes the form $N_m(x) = \sum_{j=1}^m n_j(w_j^\top x)$ and thus

$$\hat{N}_m(\xi) = \sum_{j=1}^m \hat{n}_j(w_j^\top \xi) \prod_{i=2}^d \delta(V_j^\top \xi - \cdot),$$

  where $V_j \in \mathbb{R}^{d \times (d-1)}$ denotes the orthogonal complement of w_j.
- Thus,

$$\mathrm{supp}(\hat{N}_m) = \cup_{j=1}^m \{y = w_j^\top \xi : \xi \in \mathbb{R}^d\}.$$

- If f^* is radial, then $\hat{f}^*$ is still radial.

The Intuition via Fourier Analysis (Cont'd)

- Let $\rho \in \mathcal{P}(\mathbb{R}^d)$ be the input distribution. Let $\rho(x) = \varphi(x)^2$. Then,

$$\int (N_m(x) - f(x))^2 \rho(x)\,\mathrm{d}x = \|N_m\varphi - f\varphi\|^2_{L^2(\mathbb{R}^d)}$$
$$= \|\sum_{j=1}^{m} \hat{n}_{j,w_j} * \hat{\varphi} - \hat{f} * \hat{\varphi}\|^2_{L^2(\mathbb{R}^d)}$$

- Taking f and φ to be radial!!! and f or φ to be highly oscillated!!

$\hat{n}_{i,\mathbf{w}_i}(\boldsymbol{\xi}) * \hat{\varphi}(\boldsymbol{\xi})$ $\qquad$ $\hat{f}(\boldsymbol{\xi}) * \hat{\varphi}(\boldsymbol{\xi})$

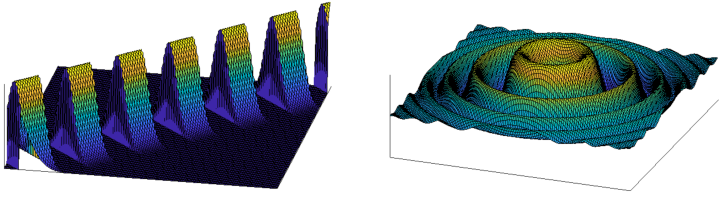


Figure 4: Intuition: Can't approximate "fat" function with few "thin" functions in high dimension. This figure is taken from Shamir's slide.

Summary

Depth separations for approximating some functions.

- $g_l = g \circ g \circ \cdots \circ g$: highly-oscillating

Summary

Depth separations for approximating some functions.

- $g_l = g \circ g \circ \cdots \circ g$: highly-oscillating
- $x \mapsto x^2$: bit-extraction (highly-oscillating)

Summary

Depth separations for approximating some functions.

- $g_l = g \circ g \circ \cdots \circ g$: highly-oscillating
- $x \mapsto x^2$: bit-extraction (highly-oscillating)
- $(x, y) \mapsto \sin(\pi d^3 \langle x, y \rangle)$: highly-oscillating

Summary

Depth separations for approximating some functions.

- $g_l = g \circ g \circ \cdots \circ g$: highly-oscillating
- $x \mapsto x^2$: bit-extraction (highly-oscillating)
- $(x, y) \mapsto \sin(\pi d^3 \langle x, y \rangle)$: highly-oscillating
- $x \mapsto f(\|x\|_2)$: f highly-oscillating

Summary

Depth separations for approximating some functions.

- $g_l = g \circ g \circ \cdots \circ g$: highly-oscillating
- $x \mapsto x^2$: bit-extraction (highly-oscillating)
- $(x, y) \mapsto \sin(\pi d^3 \langle x, y \rangle)$: highly-oscillating
- $x \mapsto f(\|x\|_2)$: f highly-oscillating

Summary

Depth separations for approximating some functions.

- $g_l = g \circ g \circ \cdots \circ g$: highly-oscillating
- $x \mapsto x^2$: bit-extraction (highly-oscillating)
- $(x, y) \mapsto \sin(\pi d^3 \langle x, y \rangle)$: highly-oscillating
- $x \mapsto f(\|x\|_2)$: f highly-oscillating

Other results:

- There are some depth separation result from circuit complexity perspective. (Venturi et al., JMLR 2021) provides an example "slightly" beyond radial functions.
- We refer interested readers to Shamir's slide https://users.cs.duke.edu/~rongge/stoc2018ml/Shamir_depthfordeep_STOC.pdf.

Open Questions

- What is the ‘big picture” beyond some specific examples?

Open Questions

- What is the ‘big picture” beyond some specific examples?
- Depth separation in dimension for depths > 3?

Open Questions

- What is the 'big picture" beyond some specific examples?
- Depth separation in dimension for depths > 3?
- Depth separation for estimation and optimization?