

# Depth separations

Instructor: Lei Wu

PKU Summer School, 2021

# Outline

- One-dimensional cases:  $\text{poly}(1/\varepsilon)$  (shallow) v.s.  $\log(1/\varepsilon)$  (deep).
- High-dimensional cases: CoD (shallow nets) v.s. Free of CoD (deep nets)

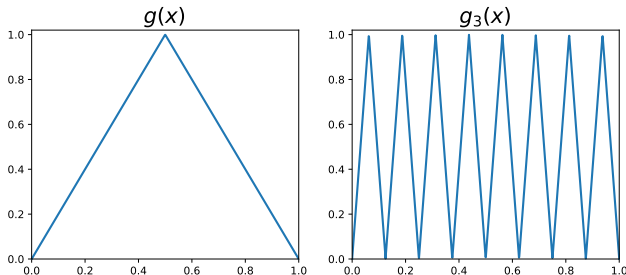
## Some useful facts

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

- Let  $g(x) = t(2x - 1)$  be the shift triangular function and

$$g_l(x) = \underbrace{g \circ g \circ \cdots \circ g}_l(x).$$

Obviously,  $g_l$  has  $2^l$  linear pieces.



- One challenge in examining the function composition is the change of input domains. The choice of  $g$  such that  $g : [0, 1] \mapsto [0, 1]$ , which dramatically simplifies the analysis of function composition.

# Approximating oscillated functions

## Theorem 1 (Telgarsky, 2016)

- On the one hand,  $g_l$  can be implemented as a  $O(l)$ -layer neural nets with the width less than 3.
- On the other hand, for any 2-layer ReLU net  $f_m(\cdot; \theta)$  with the width  $m = \text{poly}(l)$ , we have

$$\int_0^1 |f_m(x; \theta) - g_l(x)| dx \gtrsim 1.$$

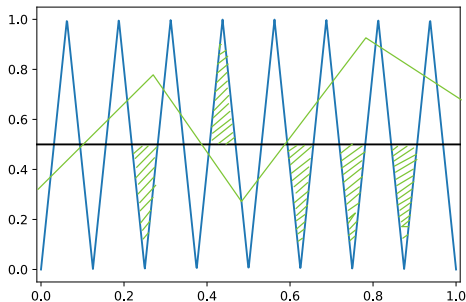
**Proof.** (1). First,  $g(x) = \text{ReLU}(2x) + \text{ReLU}(2x - 2) - 2 \text{ReLU}(2x - 1)$ , i.e.,  $g$  can be exactly represented as three neurons. Hence,  $g_l$  can be represented as  $2l$ -layer neural net with the width less than or equal to 3.

# Proof

## 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 the following figure. Draw the horizontal line  $y = 1/2$ .

$$\int_0^1 |f_m(x; \theta) - g_l(x)| dx \geq [\text{number of surviving triangles}] \cdot [\text{area of the triangle}]$$



## Proof (Cont'd)

Then, there are  $2^{l+1}$  (half) triangles.

$$\begin{aligned} \int_0^1 |f_m(x; \theta) - g_l(x)| dx &\geq [\text{number of surviving triangles}] \cdot [\text{area of the triangle}] \\ &\geq (2^{l+1} - 2^l - M) \cdot \left(\frac{1}{2} \cdot \frac{1}{2^{l+1}} \cdot \frac{1}{2}\right) \\ &\geq \frac{1}{8} - \frac{M}{2^{l+3}} \gtrsim 1. \end{aligned} \tag{0.1}$$

## Lemma 2

*For any fixed depth  $L \geq 1$  and width  $m \geq 1$ ,  $L$ -layer ReLU networks can only represent still piecewise linear functions. Moreover, the number of linear pieces is not greater than  $(m/L)^L$*

**Proof.** The case  $L = 1, 2$  is trivial. The proof of the general  $L > 0$  can be found in Section 6 of [Telgarsky's note].

- The number of linear pieces increases with the depth exponentially.
- For fixed-depth ReLU networks, approximating  $g_l$  needs  $(m/L)^L = 2^l$ . This implies that the number of total parameters:

$$Lm^2 = O(LL^2(2^l)^{2/L}) = O(L^3 2^{\frac{2l}{L}}),$$

which suffer from the curse of oscillation unless  $L \gtrsim l$ .

## Approximating $x^2$

Here, we consider the target function  $f(x) = x^2$ , which will be used to approximating general smooth functions.

### Lemma 3

Let  $S_M := (x_k)_{k=0}^M$  be the set of uniform grid points in  $[0, 1]$  with grid size  $h = 1/M$ . For any function  $f$ , let  $P_M f$  be the piecewise linear interpolation of  $f$  with the uniform grid points  $S_M$ :

$$P_M f(x) = \sum_{k=1}^M f(x_k) t\left(\frac{x - x_k}{h}\right), \quad (0.2)$$

where,  $t(\cdot)$  is the triangular function. Then,

$$\sup_{x \in [0,1]} |P_M f(x) - f^*(x)| \lesssim \frac{\sup_{x \in [0,1]} |f''(x)|}{m^2}$$



For each interval  $x_j + t \in [x_j, x_{j+1}]$ ,

$$|f(x_j + t) - P_M f(x_j + t)| = \left| f(x_j + t) - f(x_j) - \frac{f(x_{j+1}) - f(x_j)}{x_{j+1} - x_j} t \right| \quad (0.3)$$

$$= |f'(\xi_1)t - f'(\xi_2)t| \leq \max_x |f''(x)| t^2. \quad (0.4)$$

Piecewise linear approximators can only explore the second-order smoothness.

# Approximating $x^2$ with deep ReLU nets

## Proposition 1

Let  $f^*(x) = x^2$ . For any  $\varepsilon > 0$ , there exists a neural net  $\tilde{f}$ , whose depth and width is  $O(\log(1/\varepsilon))$  and  $O(1)$ , respectively, such that

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

# Proof I

**Proof.** One can show that  $l = 2, 3, \dots$ ,

$$P_{2^{l-1}} f^*(x) - P_{2^l} f^*(x) = \frac{g_l(x)}{2^{2l}}, \quad \forall x \in [0, 1]. \quad (0.5)$$

## Proof II

Construct a neural net as follows

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

Note that  $g$  can be implemented using 3 neurons. 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 3,

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

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

# Low bounds of approximating $x^2$ with piecewise linear functions

## Lemma 4

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

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

**Remark.** Similar results hold for the error measured by  $L^1$  norm.

**Proof.**

- First any interval  $[a, b] \subset [0, 1]$ , we have

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

# Proof

- 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  $c = o(|b - a|)$ ,  $d = o(|b - a|^2)$ . This results in the first term satisfies  $(b - a + c)^2 + d \gtrsim (|b - a|^2)$ . Combining them, we have  $I_{b,a} \gtrsim (b - a)^2$ .
- Since  $g \in \mathcal{G}_m$ , the number of piecewise linear parts of  $g$  is at most  $m$ . Hence, for any  $g \in \mathcal{G}$ , there must exist a domain  $[a, b]$  such that (1) there exist  $c, d$  such that  $g(x) = cx + d$  for  $x \in [a, b]$ ; (2)  $|b - a| \gtrsim 1/m$ . Then,

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

# Comparison

- For a depth- $L$  ReLU net, the number of pieces is at most  $(m/L)^L$ .
- For a target accuracy  $\varepsilon$ , the width needs to satisfy  $\frac{1}{(m/L)^{2L}} \leq \varepsilon$ , which yields  $m \geq L\varepsilon^{-1/(2L)}$ . Hence,

$$\text{Total parameters} \gtrsim m^2 L \gtrsim L^3 \varepsilon^{-1/L} = \text{poly}(1/\varepsilon).$$

- In a summary, for approximating  $x^2$  to reach the accuracy  $\varepsilon$ ,
  - deep ReLU network only need  $O(\log(1/\varepsilon))$  parameters;
  - shallow ReLU network needs at least  $O(\text{poly}(1/\varepsilon))$  parameters.

## 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:  $a_0 + a_1x + \dots + a_kx^k$ .
- Fast approximation of functions that can be efficiently approximated by polynomials, e.g., Sobolev space.



## Theorem 5 (Yarotsky, 2017)

Assume that  $f \in C^r([0, 1]^d)$  and  $\max_{|\alpha| \leq r} \text{ess sup}_{x \in [0, 1]^d} |D^\alpha f(x)| \leq 1$ . Then, there exists a neural net  $\tilde{f}$  of depth at most  $C(\log(1/\varepsilon) + 1)$  and width at most  $C\varepsilon^{-d/r}(\log(1/\varepsilon) + 1)$  such that

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

Here, the constant  $C$  depends on  $d, r$ .

The following theorem concerns the approximation rate for analytic target functions, which is given in (Wang and E, 2018).

**Theorem 2.6.** Let  $f$  be an analytic function over  $(-1, 1)^d$ . Assume that the power series  $f(\mathbf{x}) = \sum_{\mathbf{k} \in \mathbb{N}^d} a_{\mathbf{k}} \mathbf{x}^{\mathbf{k}}$  is absolutely convergent in  $[-1, 1]^d$ . Then for any  $\delta > 0$ , there exists a function  $\hat{f}$  that can be represented by a deep ReLU network with depth  $L$  and width  $d + 4$ , such that

$$|f(\mathbf{x}) - \hat{f}(\mathbf{x})| < 2 \sum_{\mathbf{k} \in \mathbb{N}^d} |a_{\mathbf{k}}| \cdot \exp(-d\delta(e^{-1}L^{\frac{1}{2d}} - 1)) \quad (2.7)$$

for all  $\mathbf{x} \in [-1 + \delta, 1 - \delta]^d$ .

- The preceding result only separate deep and fixed-depth nets for the *non-smooth* ReLU activation function.
- If considering smooth activation function, such as, Tanh, we may do not have this separation. In fact, (Maierov and Meir, 2000, Mhaskar, 1996) shows that for the Tanh activation function, depth-3 nets can achieve the same approximation rate as Theorem 5 (up to logarithmic terms).
- **Can we achieve the same rate for two-layer nets with some smooth activation functions?**

**Can we obtain separation results for high-dimensional functions?**

# Approximating radial functions

## Theorem 6 (Eladn, Shamir, ICML 2016)

Suppose  $|\sigma(z)| \lesssim (1 + |z|^\alpha)$  for all  $z \in \mathbb{R}$  and some constants  $\alpha > 0$ . Then, for  $d \gtrsim 1$ , there exists  $\mu \in \mathcal{P}(\mathbb{R}^d)$  and a radial function  $g(x) = g_0(\|x\|)$  such that

- $g_0(\|x\|)$  can be approximated with 3-layer neural network with  $\text{poly}(1/\varepsilon, d)$  parameters.
- For any two-layer net of width at most  $\exp(d)$ ,

$$\int |g_0(\|x\|) - f(x)|^2 d\mu(x) \gtrsim 1.$$

## Remarks

- The proof is rather intricate, which heavily utilizes the property of Fourier transform of two-layer neural nets. Moreover, this result is also unsatisfying in the sense that  $\mu$  is not explicit.
- (Daniely, COLT 2017) provides a more explicit construction:  $f : \mathbb{S}^{d-1} \otimes \mathbb{S}^{d-1} \mapsto \mathbb{R}$ ,

$$f(x, x') = \sin(d^3 \langle x, x' \rangle),$$

and the error is measured with respect to  $\mu = \text{Unif}(\mathbb{S}^{d-1} \otimes \mathbb{S}^{d-1})$ . However, this result needs to restrict the parameter magnitudes of two-layer net is not larger than  $2^d$ .

## Barrie of depth separation

The following theorem is from (Vardi, Shamir, 2020)

**Theorem 3.3.** *Let  $\mu$  be a density function on  $\mathbb{R}^d$  with an almost-bounded support and almost-bounded conditional density. Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be an approximately  $\text{poly}(d)$ -bounded function, and let  $k' > k \geq 4$  be constants, namely, independent of  $d$ . If  $f$  cannot be approximated by a neural network of depth  $k$  and width  $\text{poly}(d)$ , but can be approximated by a neural network of depth  $k'$  and width  $\text{poly}(d)$ , then there is a function that cannot be computed by a polynomial-sized threshold circuit of depth  $k - 2$ , but can be computed by a polynomial-sized threshold circuit of depth  $3k' + 1$ .*

The reduced problem is open problems and related to the natural-proof barrier in circuit complexity.

# Summary

- Increasing depth increases the ability to fit oscillations.
- Increasing depth increases the adaptivity to higher-order smoothness.
- $\exp(d)$  and  $\text{poly}(d)$  separation: approximating radial functions.
- Obtain general depth separation results is hard in the sense of circuit complexity.

There are some other separation results which shows that deep networks can be adaptive to some other notion of smoothness, i.e., Besov space. Please refer to (Suzuki, ICLR 2019) and (Bresler and Nagaraj, NeurIPS 2020).