Depth Separation in Learning via Representation Costs

Suzanna Parkinson* Greg Ongie** Rebecca Willett* Ohad Shamir† Nathan Srebro‡

*University of Chicago **Marquette University †Weizmann Institute of Science ‡Toyota Technical Institute at Chicago

Are **depth-2** or **depth-3** neural networks better at **learning**?

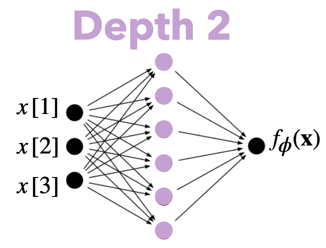
Depth-L Neural Networks

$$\phi = (\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_{L-1}, \mathbf{a})$$

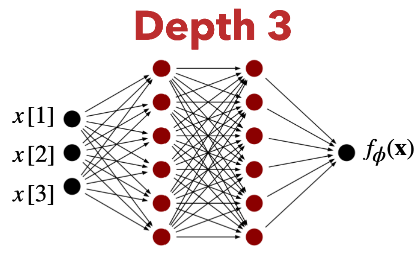
$$\sigma(x) = \mathrm{ReLU}(x) = \max(0, x)$$

$$f_\phi(\mathbf{x}) = \mathbf{a}^\top \sigma\left(\mathbf{W}_{L-1} \cdot \sigma\left(\cdots \sigma\left(\mathbf{W}_2 \sigma\left(\mathbf{W}_1 \mathbf{x}\right)\right)\right)\right)$$

Depth 2



Depth 3



PAC Learning

- The output of a learning rule $\mathscr{A}$ trained with m samples is (ε, δ)**-Probably Approximately Correct** if with probability $1-\delta$ over the training samples $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, the **generalization error** is less than ε:

$$\mathscr{L}_{\mathscr{D}}(\mathscr{A}(S)) := \mathbb{E}_{\mathbf{x}\sim\mathscr{D}}\left[\left(\mathscr{A}(S)(\mathbf{x}) - f(\mathbf{x})\right)^2\right] < \varepsilon\,.$$

- If our learning rule $\mathscr{A}$ gives a model that is (ε, δ)-Probably Approximately Correct using $m(\varepsilon, \delta)$ samples, then we say that we can **learn** with **sample complexity** $m(\varepsilon, \delta)$.

Controlling Generalization Error

- We often end up with error bounds like this:

$$\underbrace{\mathscr{L}_{\mathscr{D}}(\mathscr{A}(S))}_{\text{Generalization Error}} \leq \underbrace{\inf_{g\in\mathscr{G}} \mathscr{L}_{\mathscr{D}}(g)}_{\text{Approximation Error}} + \underbrace{2\sup_{g\in\mathscr{G}} |\mathscr{L}_S(g) - \mathscr{L}_{\mathscr{D}}(g)|}_{\text{Estimation Error}}$$

- **Approximation error:** Need existence of **one** good approximator $g \in \mathscr{G}$.[1,2] Both depth **2** and **3** networks of arbitrary width are universal approximations of continuous functions.
- **Estimation error:** Controlled using **size** of $\mathscr{G}$, here analyzed in terms of **Rademacher complexity**.[3,4] Naively, depth **3** networks have more parameters and so form a bigger model class

What if we measure model **size** in terms of **norm** of parameters instead of **number** of parameters?[4,5]

Weight Decay & Representation Cost

$$\hat{\phi}_S \in \arg\min_{\phi} \; \mathscr{L}_S(f_\phi) + \lambda C_L(\phi) \text{ where } \underbrace{C_L(\phi) = \frac{1}{L}\left(\sum_{\ell=1}^{L-1} \|\mathbf{W}_\ell\|_F^2 + \|\mathbf{a}\|_2^2\right)}_{\text{Weight Decay Cost}}$$

⬇

$$\mathscr{A}_L(S) \in \arg\min_{g\in\mathscr{N}_L} \mathscr{L}_S(g) + \lambda R_L(g) \text{ where } \underbrace{R_L(g) = \inf_{\phi} C_L(\phi) \text{ s.t. } f_\phi = g}_{\text{Representation Cost}}$$

Understanding **representation costs** across different depths helps us understand gaps in **learning** capabilities

Depth Separation in **Approximation**

$\exists f$ that requires **exponential width** (in dimension) with depth **2** but only **polynomial width** with depth **3** to be **approximated**.[6,7,8]

Depth Separation in **Learning**

- $\mathbf{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1} \times \mathbb{S}^{d-1})$, $f(\mathbf{x}) \in [-1, 1]$
- Depth-**2** vs. Depth **3** learning rules:

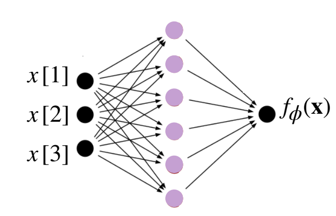
$$\mathscr{A}_2(S) \in \arg\min_{g\in\mathscr{N}_2} \mathscr{L}_S(g) + \lambda_2 R_2(g) \quad \text{vs.}$$

$$\mathscr{A}_3(S) \in \arg\min_{g\in\mathscr{N}_3} \mathscr{L}_S(g) + \lambda_3 R_3(g)$$

$\exists f$ that requires **exponential sample complexity** with depth **2** but only **polynomial sample complexity** with depth **3** to **learn**.

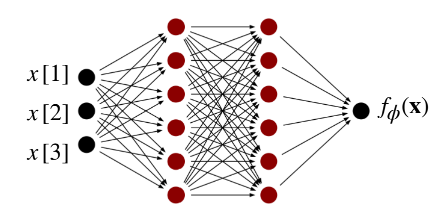
Key Idea: Choose f so that…

Large **representation cost** with **Depth 2**



Expensive

Small **representation cost** with **Depth 3**

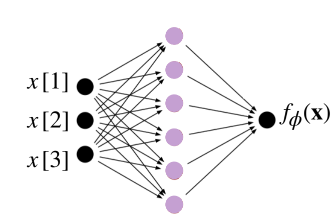


Cheap

$\forall f$ that can be **learned** with **polynomial sample complexity** with depth **2** can also be **learned** with **polynomial sample complexity** with depth **3**.

Key Idea:

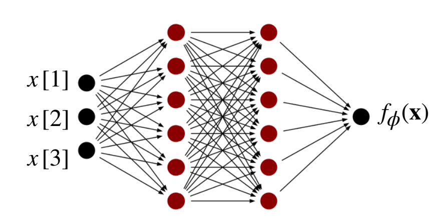
Small **representation cost** with **Depth 2**



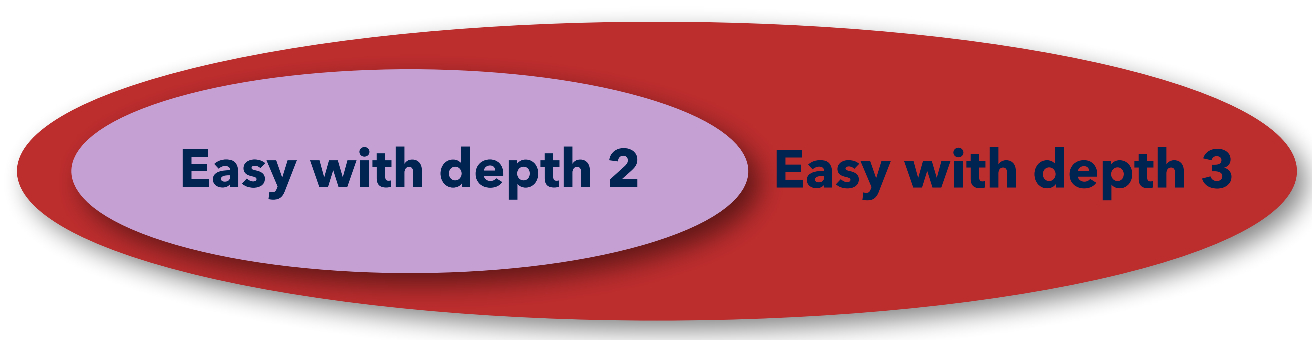
Cheap

$\Longrightarrow$

Small **representation cost** with **Depth 3**



Cheap



Open Questions & Extensions

- Depth separation between **other depths**?
- **Other architectures** beyond MLPs?
- We've implicitly assumed that we're **close to global minima** of our objective. How does **optimization** and the **loss-landscape** affect learning at different depths?

[1] Hornik (1991)
[2] Shen et al. (2022)
[3] Bartlett & Mendelson (2001)
[4] Neyshabur et al. (2015)
[5] Bartlett (1996)
[6] Eldan & Shamir (2016)
[7] Daniely (2017)
[8] Safran et al. (2021)

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 2140001. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

More details at: https://arxiv.org/abs/2402.08808