Depth Separations in Neural Networks: What is Actually Being Separated?

Itay Safran ITAY.SAFRAN@WEIZMANN.AC.IL
Ronen Eldan RONEN.ELDAN@WEIZMANN.AC.IL
Ohad Shamir OHAD.SHAMIR@WEIZMANN.AC.IL
Weizmann Institute of Science

Editors: Alina Beygelzimer and Daniel Hsu

[1]**Abstract**

Existing depth separation results for constant-depth networks essentially show that certain radial functions in $\mathbb{R}^d$, which can be easily approximated with depth 3 networks, cannot be approximated by depth 2 networks, even up to constant accuracy, unless their size is exponential in d. However, the functions used to demonstrate this are rapidly oscillating, with a Lipschitz parameter scaling polynomially with the dimension d (or equivalently, by scaling the function, the hardness result applies to $\mathcal{O}(1)$-Lipschitz functions only when the target accuracy ϵ is at most $\text{poly}(1/d)$). In this paper, we study whether such depth separations might still hold in the natural setting of $\mathcal{O}(1)$-Lipschitz radial functions, when ϵ does not scale with d. Perhaps surprisingly, we show that the answer is negative: In contrast to the intuition suggested by previous work, it *is* possible to approximate $\mathcal{O}(1)$-Lipschitz radial functions with depth 2, size $\text{poly}(d)$ networks, for every constant ϵ. We complement it by showing that approximating such functions is also possible with depth 2, size $\text{poly}(1/\epsilon)$ networks, for every constant d. Finally, we show that it is not possible to have polynomial dependence in both $d, 1/\epsilon$ simultaneously. Overall, our results indicate that in order to show depth separations for expressing $\mathcal{O}(1)$-Lipschitz functions with constant accuracy – if at all possible – one would need fundamentally different techniques than existing ones in the literature.

1. Introduction

In the past few years, several works provided separation results between depth 2 and depth 3 networks: There are functions f and distributions μ on $\mathbb{R}^d$, which are

- Hard to approximate with a depth 2 network: $\mathbb{E}_{\mathbf{x}\sim\mu}[(N_2(\mathbf{x})-f(\mathbf{x}))^2] \geq c$ for some absolute $c > 0$, using any depth 2, width $\text{poly}(d)$ network $N_2(\mathbf{x}) := \sum_{i=1}^{\text{poly}(d)} u_i\sigma(\mathbf{w}_i^\top\mathbf{x} + b_i)$ (for some parameters $\{v_i, \mathbf{w}_i, b_i\}$ and univariate activation function σ).
- Easy to approximate with a depth 3 network: For any $\epsilon > 0$, it holds that $\mathbb{E}_{\mathbf{x}\sim\mu}[(N_3(\mathbf{x}) - f(\mathbf{x}))^2] \leq \epsilon$ (or sometimes even $\sup_{\mathbf{x}} |N_3(\mathbf{x}) - f(\mathbf{x})| \leq \epsilon$) for some depth 3, width $\text{poly}(d, 1/\epsilon)$ neural network network $N_3(\mathbf{x}) := \sum_{i=1}^{\text{poly}(d,1/\epsilon)} u_i\sigma\left(N_2^i(\mathbf{x}) + b_i\right)$ (where each N_2^i is a depth 2, width $\text{poly}(d, 1/\epsilon)$ network, and σ is a standard activation such as a ReLU).

1. Extended abstract. Full version appears as arXiv preprint 1904.06984 v2

© 2019 I. Safran, R. Eldan & O. Shamir.

Moreover, these "hard" functions have a simple form: They are essentially radial functions of the form $f(\mathbf{x}) = g(||\mathbf{x}||)$ for a univariate function g. Such radial functions are of interest in learning theory, since there are function classes that are essentially a mixture of radial functions (e.g. Gaussian kernels), and they are essential primitives in expressing functions which involve Euclidean distances. Overall, these results appear to provide a clear separation between the required widths of depth 2 and depth 3 networks, in terms of the dimension d.

However, a closer inspection of the constructions above reveals that in fact, this is not so clear. The reason is that the functions which are shown to be provably hard for depth 2 networks are rapidly oscillating, and require a Lipschitz constant (at least) polynomial in d to even approximate. Having such rapidly oscillating functions is not always a natural regime, since we are often interested in functions whose Lipschitz parameter is independent of the dimension. Nevertheless, previous work seems to suggest such separation results still hold when approximating $\mathcal{O}(1)$-Lipschitz functions.

Our Results. In this paper, we show that perhaps surprisingly, such separation results break when considering $\mathcal{O}(1)$-Lipschitz radial functions: For any constant ϵ, it *is* possible to approximate radial functions using poly(d)-width, depth 2 networks. We prove this result for networks employing any activation function $\sigma(\cdot)$ which satisfies the following mild assumption (taken from Eldan and Shamir (2016)), which implies that the activation can be used to approximate univariate functions well. This assumption is satisfied for all standard activations, such as ReLU and sigmoidal functions (see reference above for further discussion):

Assumption 1 *Given the activation function σ, there is a constant $c_\sigma \geq 1$ (depending only on σ) such that the following holds: For any L-Lipschitz function $f : \mathbb{R} \to \mathbb{R}$ which is constant outside a bounded interval $[-R, R]$, and for any δ, there exist scalars $a, \{\alpha_i, \beta_i, \gamma_i\}_{i=1}^{w}$, where $w \leq c_\sigma \frac{RL}{\delta}$, such that the function $h(x) = a + \sum_{i=1}^{w} \alpha_i \sigma(\beta_i x - \gamma_i)$ satisfies $\sup_{x \in \mathbb{R}} |f(x) - h(x)| \leq \delta$.*

We now formally state our main result:

Theorem 1 *Suppose $\sigma : \mathbb{R} \to \mathbb{R}$ satisfies Assumption 1. Then for any $\epsilon > 0$ and any 1-Lipschitz radial function $f(\mathbf{x}) = \varphi(||\mathbf{x}||)$, there exists a depth 2 neural network N with σ activations and width $\exp\left(\mathcal{O}\left(\epsilon^{-9}\log(d/\epsilon)\right)\right)$ satisfying $\sup_{\mathbf{x} \in B_d} |N(\mathbf{x}) - f(\mathbf{x})| \leq \epsilon$, where $B_d = \left\{\mathbf{x} \in \mathbb{R}^d : ||\mathbf{x}|| \leq 1\right\}$, and the big O notation hides a constant that depends solely on σ.*

We complement this by showing that for constant dimension d, approximation of any $\mathcal{O}(1)$-Lipschitz radial function is possible with poly$(1/\epsilon)$-width, depth 2 networks:

Theorem 2 *Suppose $f(\mathbf{x}) = \varphi(||\mathbf{x}||)$ is a 1-Lipschitz radial function on B_d. Then there exists a depth 2 ReLU neural network N of width $n = \exp\left(\mathcal{O}\left(d\log\left(1/\epsilon\right)\right)\right)$ such that $\sup_{\mathbf{x} \in B_d} |f(\mathbf{x}) - N(\mathbf{x})| < \epsilon$.*

Furthermore, we show that any even radial monomial, namely a radial function of the form $\mathbf{x} \mapsto ||\mathbf{x}||^{2k}$, for any fixed natural k, can be approximated to accuracy ϵ using a depth 2 network of width polynomial in both d and $1/\epsilon$:

Theorem 3 *Suppose $\sigma : \mathbb{R} \to \mathbb{R}$ satisfies Assumption 1. Then for any $\epsilon > 0$ and any natural $k \geq 1$, there exists a depth 2 neural network N with σ activations of width $\exp\left(\mathcal{O}\left(k^2 \log\left(d/\epsilon\right)\right)\right)$ satisfying $\sup_{\mathbf{x}\in B_d} \left|N(\mathbf{x}) - ||\mathbf{x}||^{2k}\right| \leq \epsilon$, where the big O notation hides a constant that depends solely on σ.*

Finally, we formally prove (using a reduction from Eldan and Shamir (2016); Daniely (2017), and using their assumptions) that it is impossible to obtain a general polynomial dependence on both d and $1/\epsilon$ in our setting. More formally, we have the following two theorems:

Theorem 4 *The following holds for some positive universal constants c_1, c_2, and any depth 2 network employing a ReLU activation function. Consider the 1-Lipschitz function $f(\mathbf{x}) = \frac{1}{2\pi d^3} \sin\left(2\pi d^3 ||\mathbf{x}||_2^2\right)$ on B_d. Suppose N is a depth 2 network of width $w(d, 1/\epsilon)$, with weights bounded by $\frac{2^{d+1}}{2\pi d^3}$, and satisfying $\sup_{\mathbf{x}\in B_d} |N(\mathbf{x}) - f(\mathbf{x})| \leq \epsilon$ for any $\epsilon > 0$ and any $d \geq 2$. Then for any $d > c_1$,*

$$w(d, 101\exp(2)\pi^3 d^3) \geq 2^{c_2 d \log d}.$$

In particular, depth 2 networks of width $\mathrm{poly}(d, 1/\epsilon)$ cannot approximate f to accuracy ϵ.

Theorem 5 *The following holds for some positive universal constants c_1, c_2, c_3, c_4, and any network employing an activation function satisfying Assumptions 1 and 2 in Eldan and Shamir (2016). Let $f(\mathbf{x}) = \max\{0, -||\mathbf{x}|| + 1\}$. For any $d > c_1$, there exists a continuous probability distribution γ on $\mathbb{R}^d$, such that for any $\epsilon > 0$, and any depth 2 neural network N satisfying $||N(\mathbf{x}) - f(\mathbf{x})||_{L_2(\gamma)} \leq \epsilon$ and having width $w(d, 1/\epsilon)$, it must hold that*

$$w(d, c_2 d^6) \geq c_3 \exp(c_4 d).$$

In particular, depth 2 networks of width $\mathrm{poly}(d, 1/\epsilon)$ cannot approximate f to accuracy ϵ.

Overall, these results show that to approximate radial functions with depth 2 networks, their width *can* be polynomial in either d or $1/\epsilon$, but generally not in both.

Acknowledgements

This research is supported in part by European Research Council (ERC) grant 754705.

References

Amit Daniely. Depth separation for neural networks. *arXiv preprint arXiv:1702.08489*, 2017.

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