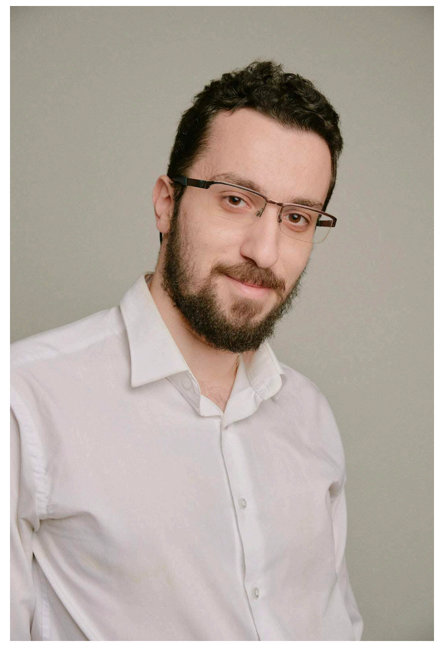
Approximation Powers and Limitations of Neural Networks

Clayton Sanford
Columbia CS



Based on work with Vaggos Chatziafratis, Daniel Hsu, Rocco Servedio, and Manolis Vlatakis

Many unanswered questions about NN theory

- Why does gradient descent attain near zero training loss? (Optimization)
- Why do models attain low test error despite overfitting and having more parameters than samples? (Benign overfitting)
- What are the properties of functions that gradient descent tends to converge to and how do they relate to generalization? (Inductive bias)
- How do neural networks provably learn hierarchical functions layer-by-layer? (Feature learning)
- **How do representational capabilities and limitations very among NN architectures? (Approximation)**

Core approximation theory question

- **Separation:** What functions can be represented by one model, but not by another?
- Classical example: Perceptron vs XOR
  - Perceptron: $x \mapsto \operatorname{sign}(w^T x - b)$
  - No perceptron can represent XOR function
  - But, feature expansions or two Perceptrons can

Vol. VI, No. 2, Summer 1958

research trends

CORNELL AERONAUTICAL LABORATORY, INC., BUFFALO 21, NEW YORK

The Design of an Intelligent AUTOMATON

by FRANK ROSENBLATT

Introducing the perceptron — A machine which senses, recognizes, remembers, and responds like the human mind.

STORIES about the creation of machines having human qualities have long been a fascinating province in the realm of science fiction. Yet we are now about to witness the birth of such a machine — a machine capable of perceiving, recognizing, and identifying its surroundings without any human training or control.

Development of that machine has stemmed from a search for an understanding of the physical mechanisms which underlie human experience and intelligence. The question of the nature of these processes is at least as ancient as any other question in western science and philosophy, and, indeed, ranks as one of the greatest scientific challenges of our time.

Our understanding of this problem has gone perhaps as far as had the development of physics before Newton. We have some excellent descriptions of the phenomena to be explained, a number of interesting hypotheses, and a little detailed knowledge about events in the nervous system. But we lack agreement on any integrated set of principles by which the functioning of the nervous system can be understood.

We believe now that this ancient problem is about to yield to our theoretical investigation for three reasons:

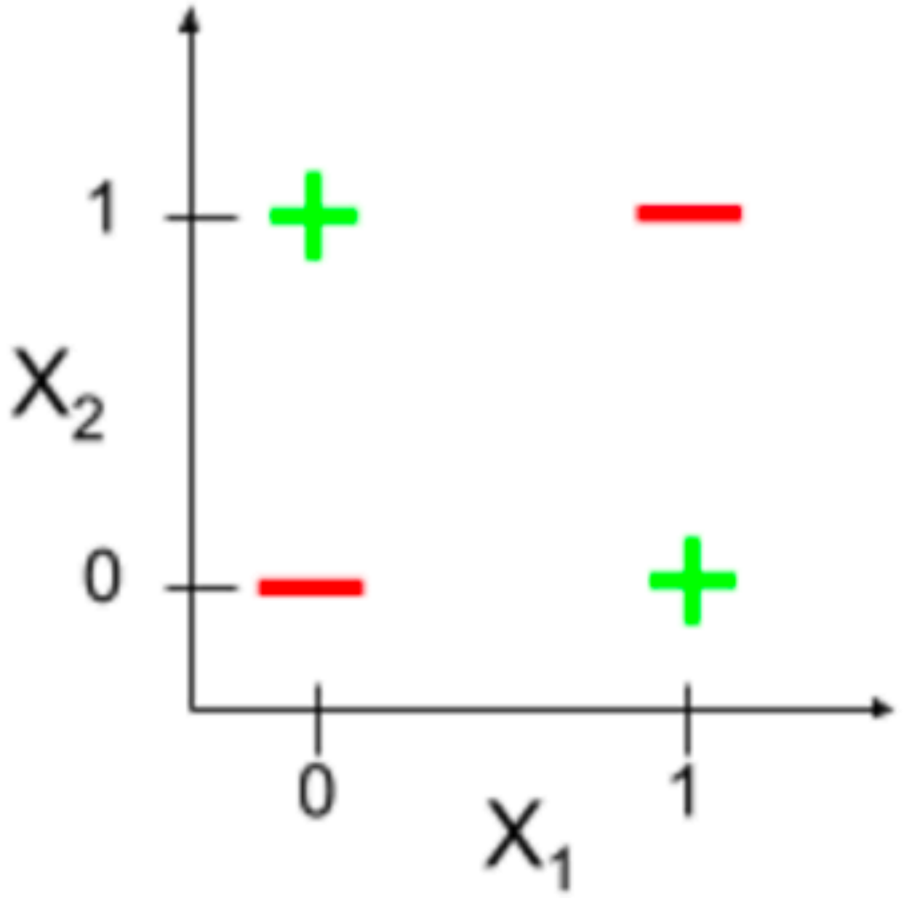
First, in recent years our knowledge of the functioning of individual cells in the central nervous system has vastly increased.

Second, large numbers of engineers and mathematicians are, for the first time, undertaking serious study of the mathematical basis for thinking, perception, and the handling of information by the central nervous system, thus providing the hope that these problems may be within our intellectual grasp.

Third, recent developments in probability theory and in the mathematics of random processes provide new tools for the study of events in the nervous system, where only the gross statistical organization is known and the precise cell-by-cell "wiring diagram" may never be obtained.

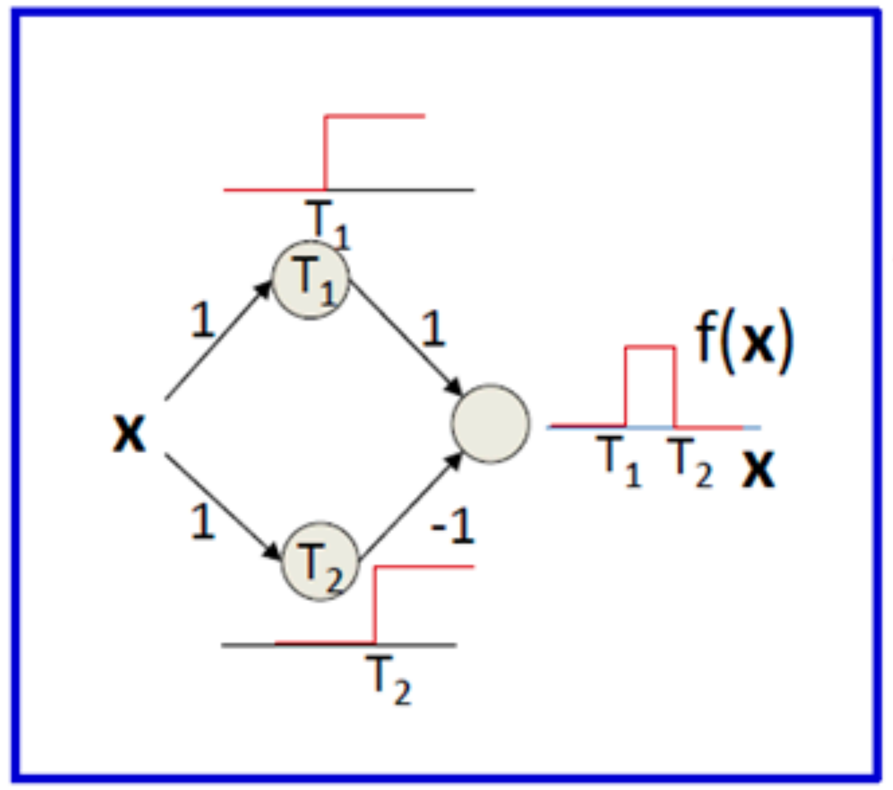
Receives Navy Support

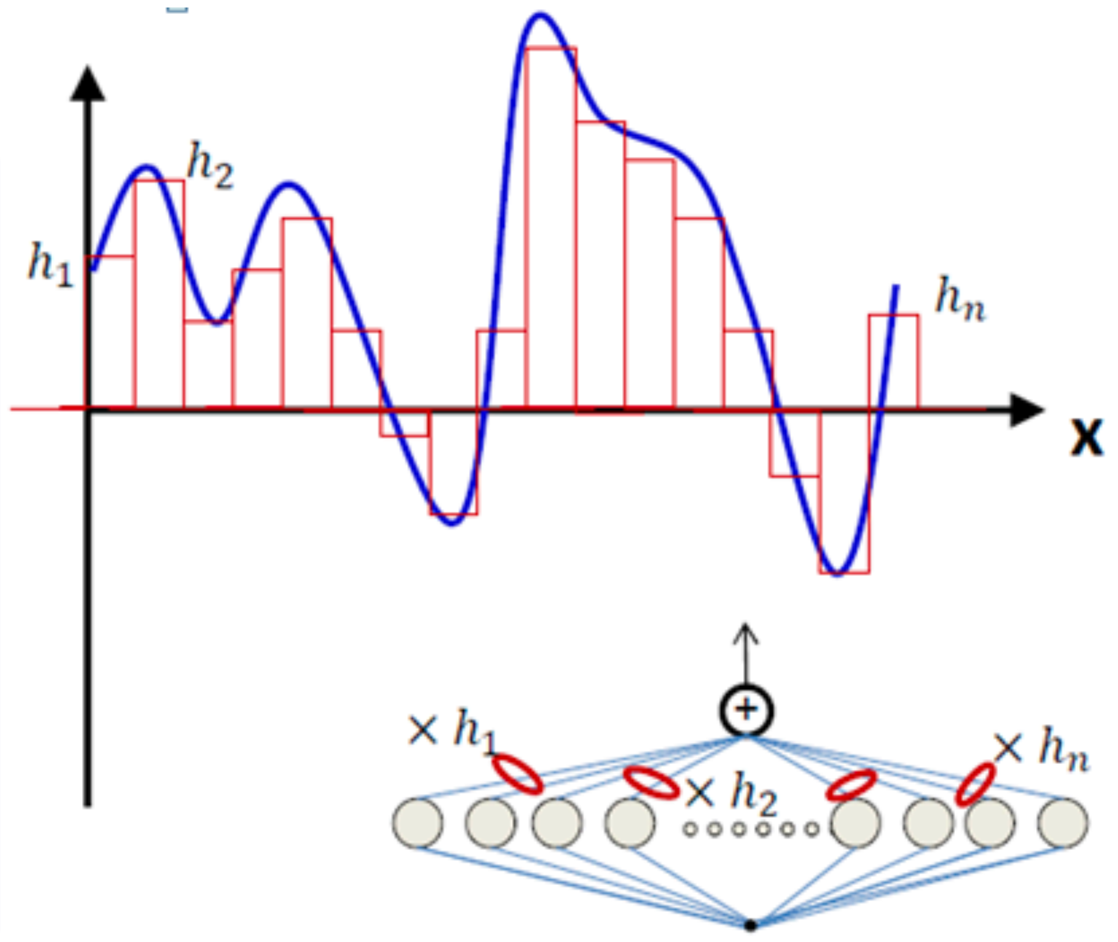
In July, 1957, Project PARA (Perceiving and Recognizing Automaton), an internal research program which had been in progress for over a year at Cornell Aeronautical Laboratory, received the support of the Office of Naval Research. The program had been concerned primarily with the application of probability theory to



Universal Approximation Theorem

- **Informal Theorem [Cybenko; Funahashi; Hornik, Stinchcombe, White ’89]:**
    - For any continuous $f : \mathbb{R}^d \to \mathbb{R}$, $\epsilon > 0$, and compact $S \subset \mathbb{R}^d$, there exists a two-layer neural network g that ϵ-pointwise approximates f on S.
- Problem: no bound on the width of the network needed!





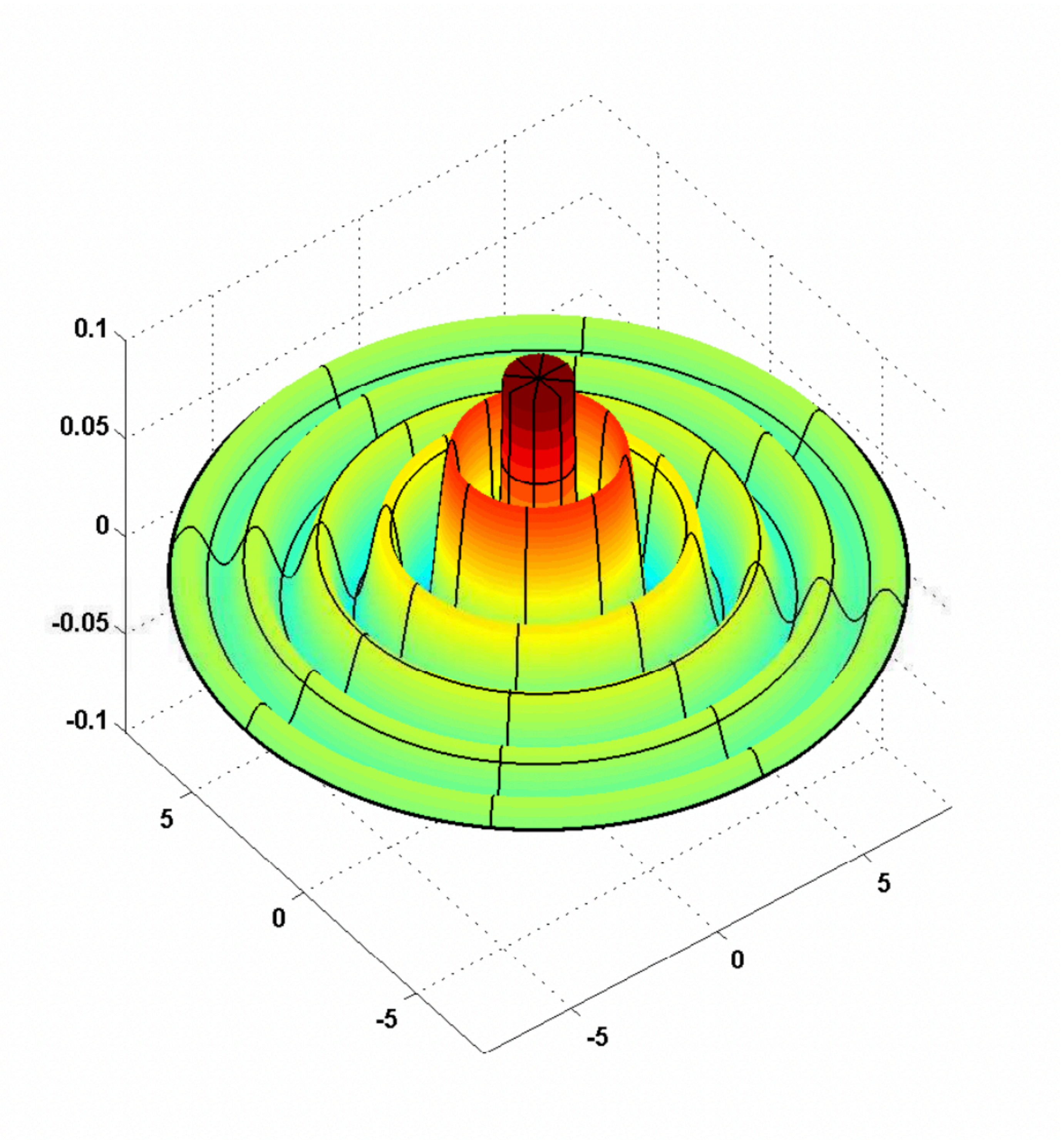
Amended approximation theory question

- **Separation:** What functions can be represented **efficiently** (i.e. poly width in relevant parameters) by one model, but not by another?

- **Depth separation:** What functions $f : \mathbb{R}^d \to \mathbb{R}$ can be ϵ-approximated with $\mathrm{poly}(d)$-width NNs of depth-$(k + 1)$ and require $\exp(d)$-width to 0.1-approximate depth-k NNs?

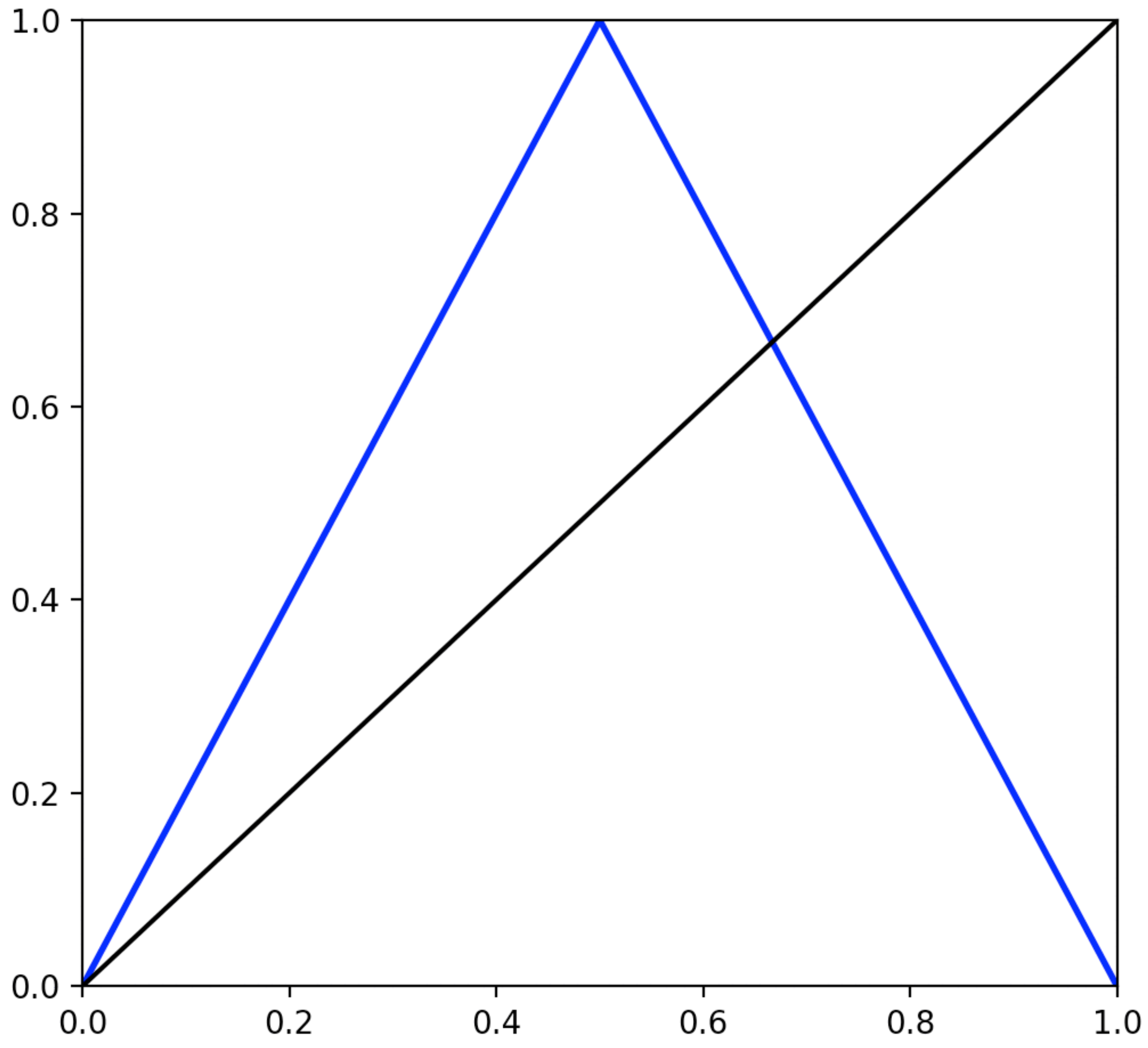
2 vs 3 separations

- **[Daniely '17]** $f(x) = \sin(\pi d^3 \langle x, x' \rangle)$ can be approximated by $\mathrm{poly}(d)$-width 3-layer NN, but requires $\exp(d)$-width (or $\exp(d)$ weights) to approximate with 2-layer NN.
    - Positive result: 1st approximate inner product, 2nd approximate 1-d function
    - Negative result: spherical harmonics, inapproximability of f by low-degree polynomials
- Other 2 vs 3 separation: **[Eldan, Shamir '16], [Safran, Shamir '16]**



$\sqrt{k}$ vs k separations
[Telgarsky '16]

- Triangle map $g : [0,1] \to [0,1]$ with $g(x) = \min(2x, 1 - 2x)$.
- $f(x) = g^k(x)$ can be represented by $\Theta(k)$ -depth NN of constant width, but requires $\exp(k)$-width to approximate with $\Theta(\sqrt{k})$ -depth NN.
- Positive result: directly construct triangle map with 2 ReLUs and iterate
- Negative result: bound maximum number of oscillations of NN with width m and depth ℓ



$\sqrt{k}$ vs k separations + dynamical systems
[Chatziafratis, et. al. ’20, ’21], [Sanford, Chatziafratis, ’22]

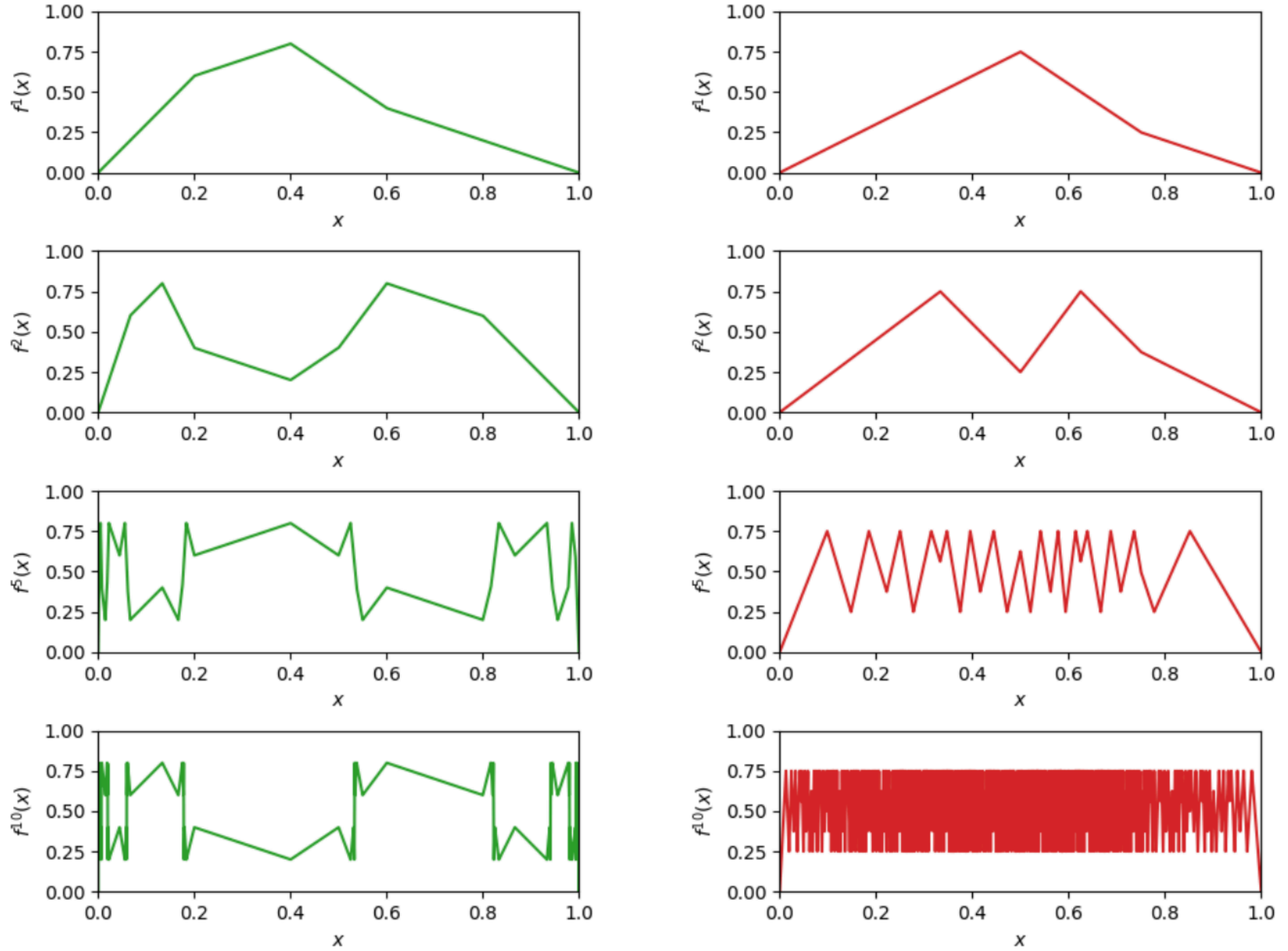
- Question: Do other iterated functions $f(x) = g^k(x)$ provide the highly-oscillatory property needed for depth separation?

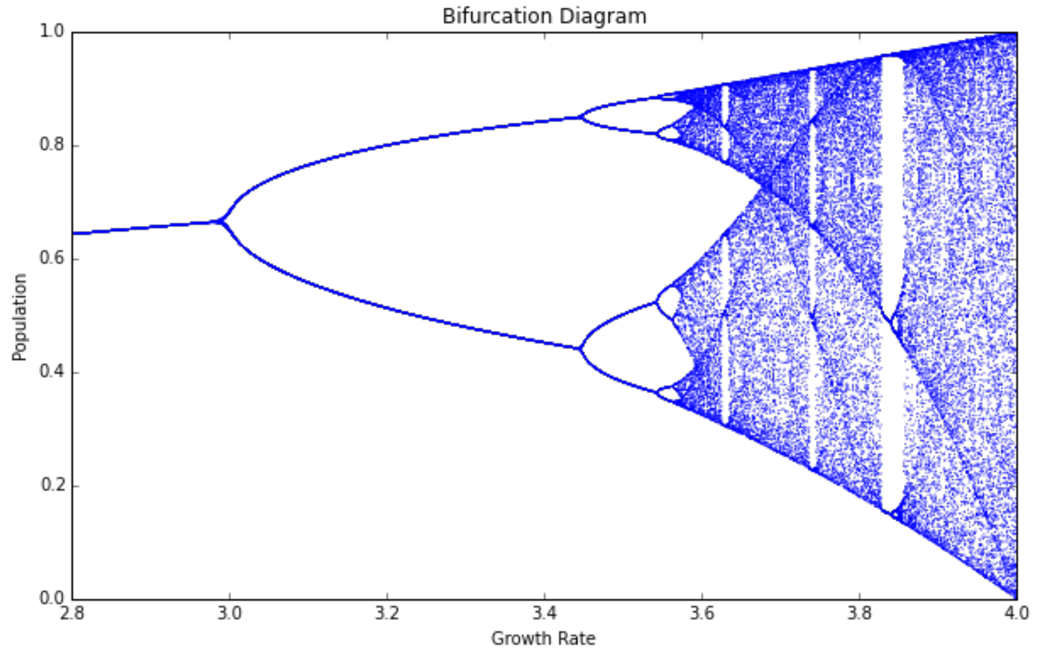
- Yes. If g is a unimodal mapping, then:

  - If g has a cycle of length 3 (or any non-power-of-two), then requires depth $\Omega(k)$ to approximate f with poly width.

  - If g only has power-of-two cycles, then a poly-width two-layer NN can approximate f.

- Relates to Li-Yorke chaos: Period 3 $\Longrightarrow$ Chaos



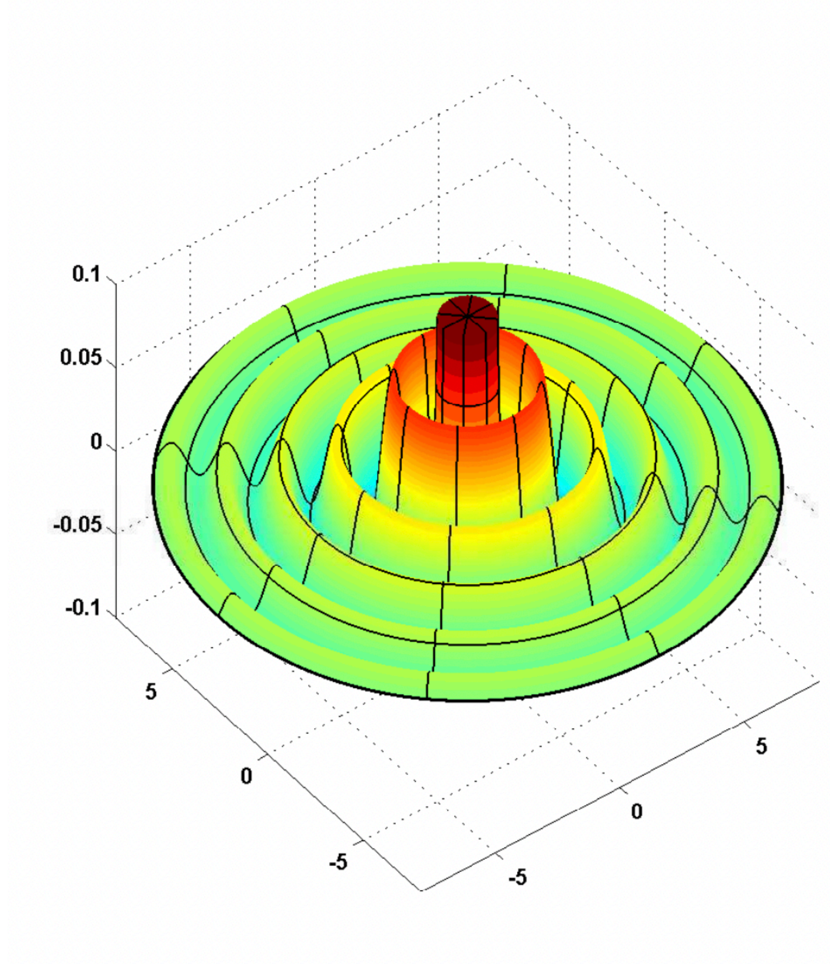


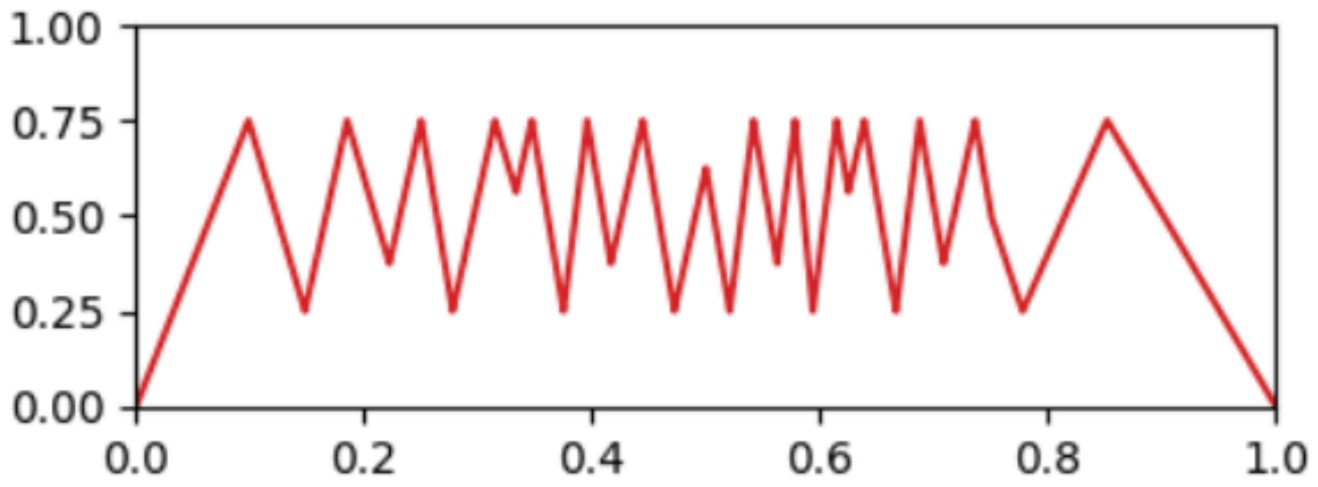
Limitations of depth-separation

Problem #1: All inapproximable functions seem to be adversarial somehow, and “natural” functions are easy to approximate.

- **[Safran, Eldan, Shamir ’19]** All 1-Lipschitz *radial* functions can be 0.1-approximated w.r.t. L_∞ over $\mathbb{B}^d(1)$ with $\mathrm{poly}(d)$-width.

- Question: Does there exist a 1-Lipschitz function with a 2-vs-3 separation?





Limitations of depth-separation

Problem #2: Depth-separation does not imply optimization-separation.

- **[Malach, Yehudai, Shalev-Shwartz, Shamir '21]**
  f cannot be efficiently **weakly**-approximated by depth-3 neural net $\Longrightarrow$
  f cannot be efficiently **weakly**-learned by gradient descent by *any* poly-size neural net.

- Relies on ability to L_2-approximate Lipschitz functions with depth-3 neural nets.

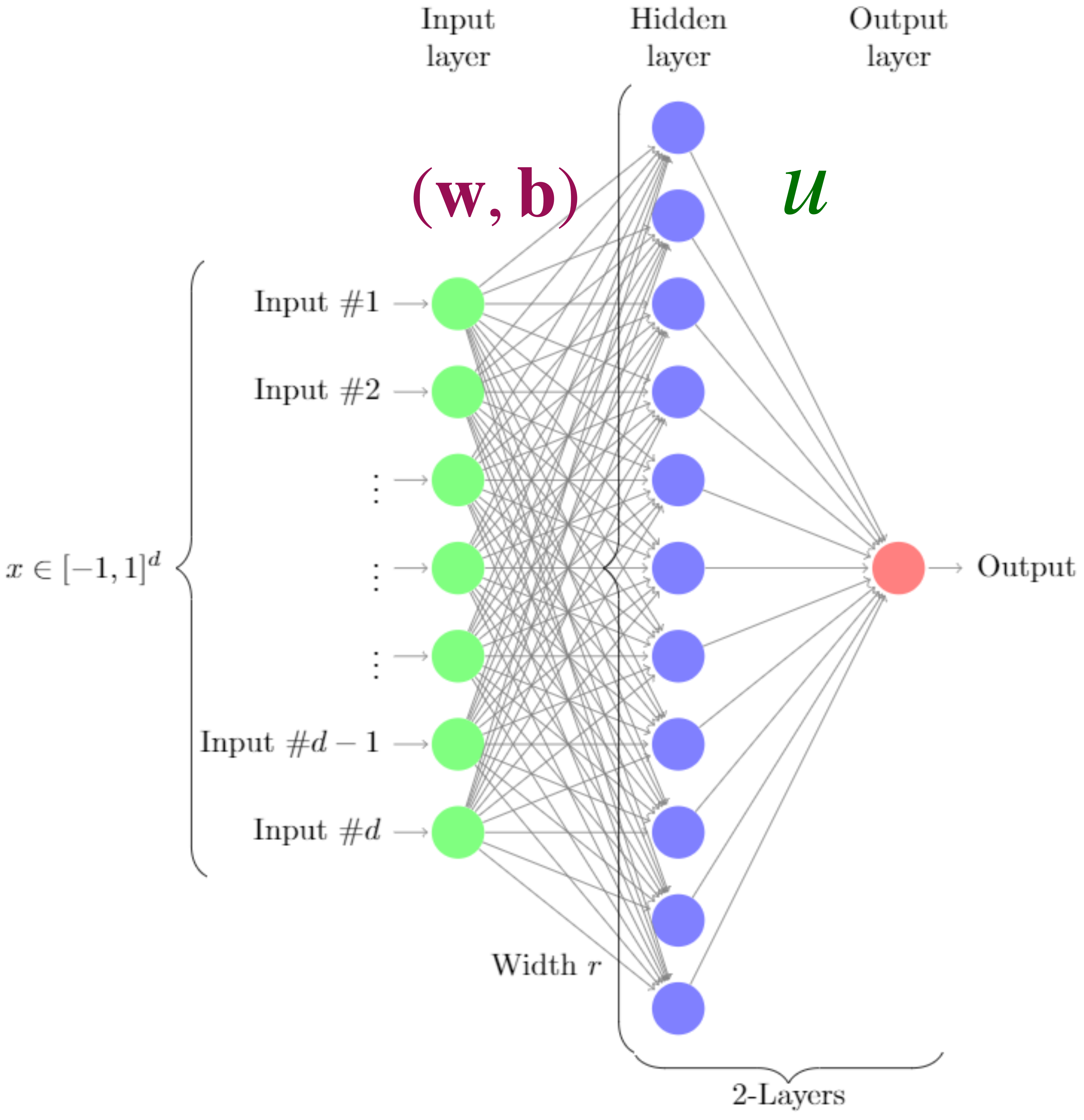
Approximation properties of random feature models

[Hsu, Sanford, Servedio, Vlatakis '21]

- **Question**: What are the approximation powers and limitations of depth-2 neural networks with random bottom-layer weights?

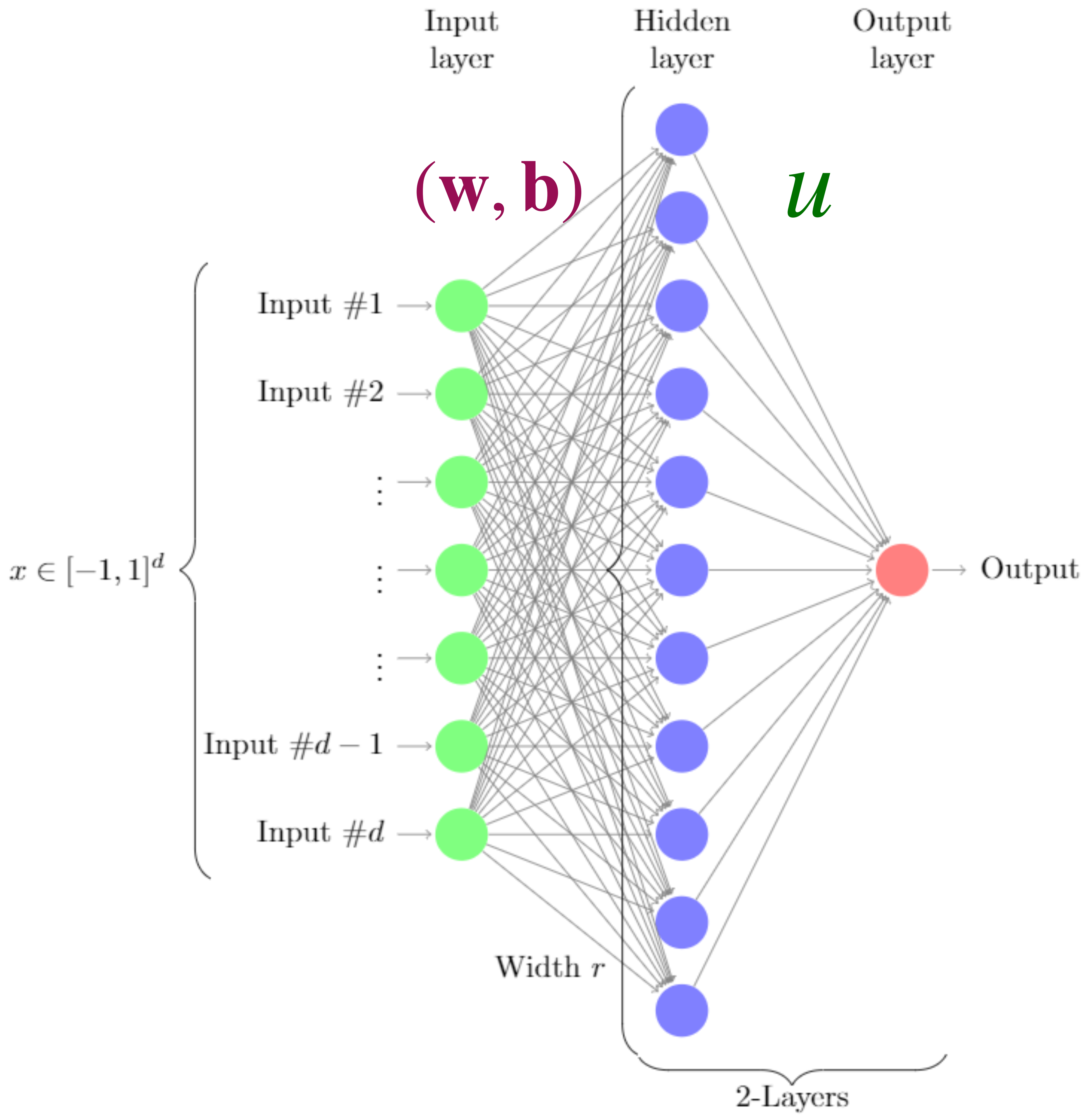
- **Answer**: Width necessary and sufficient to approximate an L-Lipschitz function $f \in L_2([-1,1]^d)$ is:
  - $\mathrm{poly}(d)$ if $L = \Theta(1)$;
  - $\mathrm{poly}(L)$ if $d = \Theta(1)$;
  - and $\exp(\Theta(d))$ if $L = \Theta(\sqrt{d})$.

- Some overlap in methodology and results with [Bresler, Nagaraj '20]



Our setting

- f is L-**Lipschitz** if for all $x, x' \in [-1,1]^d$, $|f(x) - f(x')| \leq L\|x - x'\|_2$.
- Neural net: $g(x) = \sum_{i=1}^{m} u^{(i)}\sigma(\langle \mathbf{w}^{(i)}, x\rangle - \mathbf{b}^{(i)})$ for $(\mathbf{w}^{(i)}, \mathbf{b}^{(i)}) \sim \mathscr{D}$, ReLU $\sigma(z) = \max(0,z)$.
- g **approximates** f if $\|f - g\| = \sqrt{\mathbb{E}_{\mathbf{x}\sim[-1,1]}[(f(\mathbf{x}) - g(\mathbf{x}))^2]} \leq 0.1$.
- $\mathrm{MinWidth}_{f,\mathscr{D}}$ is the smallest m such that with probability 0.9 over $(\mathbf{w}^{(i)}, \mathbf{b}^{(i)})_{i\in[r]}$, there exists a corresponding g with u that approximates f.



Our results

Theorem 1 [Upper-bound]: For any L, d, there exists symmetric $\mathscr{D}$ such that for all L-Lipschitz $f \in L_2([-1,1]^d)$:

$$\text{MinWidth}_{f,\mathscr{D}} = \min(d^{\tilde{O}(L^2)}, L^{\tilde{O}(d)}).$$

Theorem 2 [Lower-bound]: For any L, d and any symmetric $\mathscr{D}$, there exists L-Lipschitz $f(x) = \sin(L\langle u, x\rangle)$ such that:

$$\text{MinWidth}_{f,\mathscr{D}} = \min(d^{\tilde{\Omega}(L^2)}, L^{\tilde{\Omega}(d)}).$$

Proving our upper-bound

Theorem 1 [Upper-bound]: For any L, d, there exists symmetric $\mathscr{D}$ such that for all L-Lipschitz $f \in L_2([-1,1]^d)$, $\text{MinWidth}_{f,\mathscr{D}} = \min(d^{\tilde{O}(L^2)}, L^{\tilde{O}(d)})$.

Lemma 7: Every L-Lipschitz f can be approximated by a trigonometric polynomial of degree $O(L)$.

- Orthonormal basis for $L_2([-1,1]^d)$ with $\sqrt{2}\sin(\pi\langle K, x\rangle), \sqrt{2}\cos(\pi\langle K, x\rangle)$ terms

Lemma 9: Exists symmetric $\mathscr{D}_k$ such that every k-degree trigonometric polynomial P has

$$\text{MinWidth}_{f,\mathscr{D}} = \min(d^{\tilde{O}(k^2)}, k^{\tilde{O}(d)})$$

- Express each basis element as $\sqrt{2}\sin(\pi\langle K, x\rangle) = \mathbb{E}_{\mathbf{w},\mathbf{b}}[h_K(\mathbf{b}, \mathbf{w})\sigma(\langle \mathbf{w}, x\rangle - \mathbf{b})]$
- Concentration bounds for Hilbert spaces

Proving our lower-bound

Theorem 2 [Lower-bound]: For any L, d and any symmetric $\mathcal{D}$, exists L-Lipschitz $f(x) = \sin(L\langle u, x\rangle)$ such that $\mathrm{MinWidth}_{f,\mathcal{D}} = \min(d^{\tilde{\Omega}(L^2)}, L^{\tilde{\Omega}(d)})$.

Lemma 11: For orthonormal $\varphi_1, \ldots, \varphi_N \in L_2([-1,1]^d)$ and $N \gg r$, then at least one φ_i will be inapproximable by the span of r functions.

The family

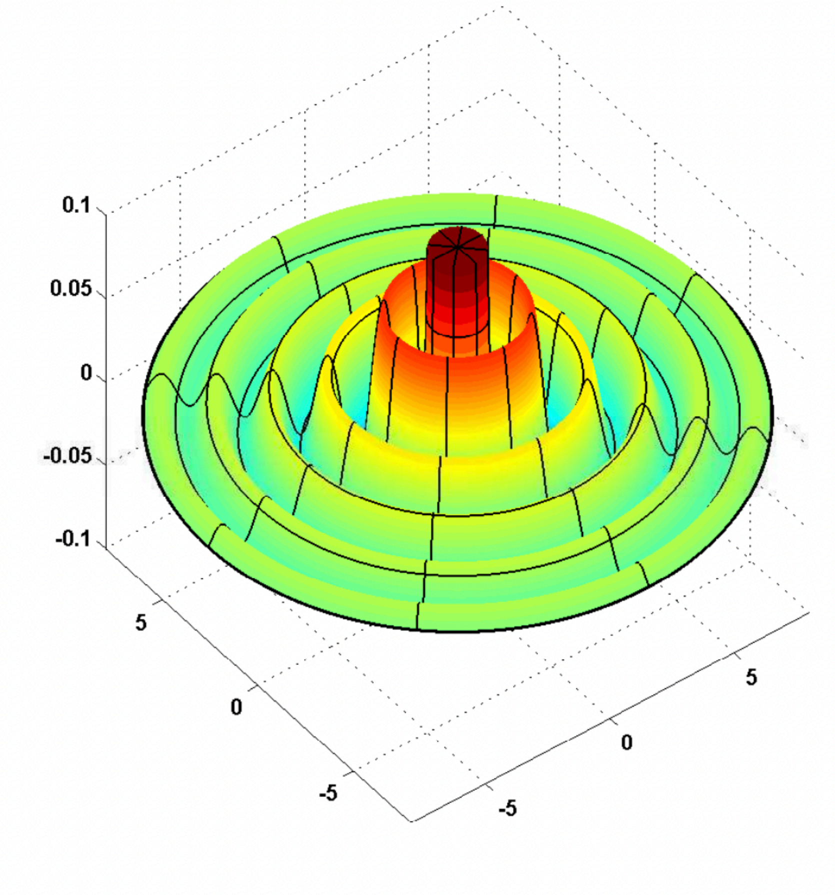
$\mathcal{T}_k = \{x \mapsto \sqrt{2}\sin(\pi\langle K, x\rangle) : \|K\|_2 \leq k\}$

contains $\min(d^{\tilde{\Omega}(L^2)}, L^{\tilde{\Omega}(d)})$ orthonormal $\Theta(k)$ -Lipschitz functions.

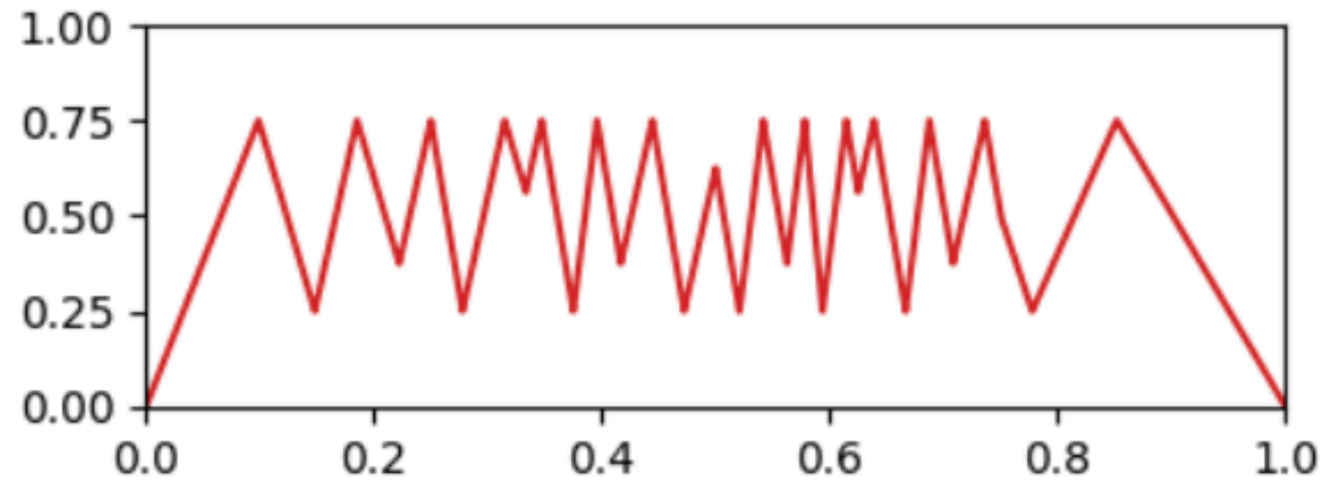
Limitations of depth-separation

Problem #1: All inapproximable functions seem to be adversarial somehow, and "natural" functions are easy to approximate.

- **[Safran, Eldan, Shamir '19]** All 1-Lipschitz *radial* functions can be 0.1-approximated w.r.t. L_∞ over $\mathbb{B}^d(1)$ with $\mathrm{poly}(d)$-width.



- Question: Does there exist a 1-Lipschitz function with a 2-vs-3 separation?



- **Answer: No (for L_2)—every 1-Lipschitz function can be represented with a poly-width 2-layer random bottom-layer NN.**

Limitations of depth-separation

Problem #2: Depth-separation does not imply optimization-separation.

now depth-2!!

- **[Malach, Yehudai, Shalev-Shwartz, Shamir '21]**
  f cannot be efficiently **weakly**-approximated by depth-3 neural net $\implies$ f cannot be efficiently **weakly**-learned by gradient descent by *any* poly-size neural net.

depth-2

- Relies on ability to L_2-approximate Lipschitz functions with depth-3 neural nets.

Interesting current and future work

- **Optimization separation:** What functions can be provably learned with gradient descent by one model, but not even approximated by another?
  - **[Safran, Lee ’22]:** Ball-indicator function can be learned with 2-layer NNs with activations on both layers, but not by 2-layer NNs with activations on only one.
- **Norm-based separation**: What functions can represented with low weight norms in one architecture but not in another?
  - Closer relationship to optimization/implicit biases of gradient descent.
  - [Ongie, Willets, Soudry, Srebro ’19], [Sanford, Ardeshir, Hsu ’22 🤞]
- **Architecture-specific separations:** Can certain functions be efficiently represented with transformer models (or CNNs), but not with other models?

The End