On the approximation power of two-layer networks of random ReLUs

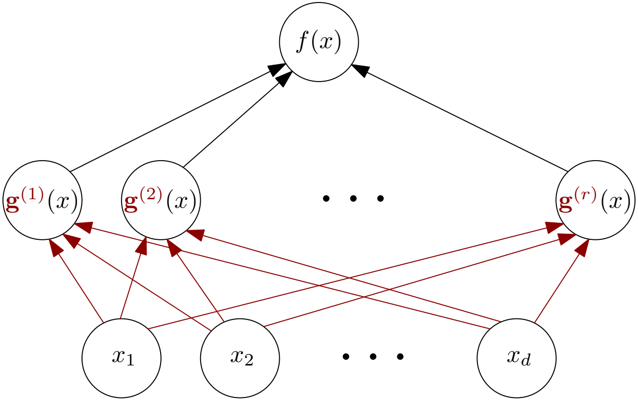
Clayton Sanford

Joint work with Daniel Hsu, Rocco Servedio, Manolis Vlatakis

Columbia University

April 8, 2022
MIT Algorithms & Complexity Seminar

Two-layer networks of random ReLUs ("random ReLU networks")



$$f \in \operatorname{span}\Big\{ \underbrace{x \mapsto \max\{0, \mathbf{w}^{(i)} \cdot x - \mathbf{b}^{(i)}\}}_{\mathbf{g}^{(i)}} : i \in [r] \Big\}, \qquad \big((\mathbf{w}^{(i)}, \mathbf{b}^{(i)})\big)_{i=1}^{r} \sim \mathcal{D}$$

Approximating Lipschitz functions by two-layer networks of random ReLUs

Two-layer networks of random ReLUs:

$$\mathcal{F}_r \;:=\; \operatorname{span}\Big\{\underbrace{x \mapsto \max\{0,\, \mathbf{w}^{(i)} \cdot x - \mathbf{b}^{(i)}\}}_{\mathbf{g}^{(i)}} : i \in [r]\Big\}\;, \qquad \big((\mathbf{w}^{(i)}, \mathbf{b}^{(i)})\big)_{i=1}^{r} \sim \mathcal{D},$$

where $\mathcal{D}$ is probability distribution for bottom-level parameters $(\mathbf{w}^{(i)}, \mathbf{b}^{(i)}) \in \mathbb{S}^{d-1} \times \mathbb{R}$

Approximating Lipschitz functions by two-layer networks of random ReLUs

Two-layer networks of random ReLUs:

$$\mathcal{F}_r \;:=\; \operatorname{span}\Big\{\underbrace{x \mapsto \max\{0, \mathbf{w}^{(i)} \cdot x - \mathbf{b}^{(i)}\}}_{\mathbf{g}^{(i)}} : i \in [r]\Big\}\,, \qquad \big((\mathbf{w}^{(i)}, \mathbf{b}^{(i)})\big)_{i=1}^{r} \sim \mathcal{D},$$

where $\mathcal{D}$ is probability distribution for bottom-level parameters $(\mathbf{w}^{(i)}, \mathbf{b}^{(i)}) \in \mathbb{S}^{d-1} \times \mathbb{R}$

Question:

What is the minimum width r s.t. $\mathcal{F}_r$ can ε-approximate any L-Lipschitz functions in $\mathcal{L}^2([-1,1]^d)$ (with high probability)?

$$\Pr\Big[\inf_{\hat{f} \in \mathcal{F}_r} \|\hat{f} - f^\star\|_{\mathcal{L}^2([-1,1]^d)} \;\le\; \varepsilon\Big] \ge 0.9 \quad \text{for all } L\text{-Lipschitz } f^*\colon [-1,1]^d \to \mathbb{R}$$

$$\|f\|_{\mathcal{L}^2([-1,1]^d)} = \sqrt{\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathrm{Unif}([-1,1]^d)}[f(\mathbf{x})^2]}$$

Approximating Lipschitz functions by two-layer networks of random ReLUs

Two-layer networks of random ReLUs:

$$\mathcal{F}_r := \operatorname{span}\Big\{\underbrace{x \mapsto \max\{0, \mathbf{w}^{(i)} \cdot x - \mathbf{b}^{(i)}\}}_{\mathbf{g}^{(i)}} : i \in [r]\Big\}, \qquad \big((\mathbf{w}^{(i)}, \mathbf{b}^{(i)})\big)_{i=1}^{r} \sim \mathcal{D},$$

where $\mathcal{D}$ is probability distribution for bottom-level parameters $(\mathbf{w}^{(i)}, \mathbf{b}^{(i)}) \in \mathbb{S}^{d-1} \times \mathbb{R}$

Question:

What is the minimum width r s.t. $\mathcal{F}_r$ can ε-approximate any L-Lipschitz functions in $\mathcal{L}^2([-1,1]^d)$ (with high probability)?

$$\Pr\Big[\inf_{\hat{f} \in \mathcal{F}_r} \|\hat{f} - f^\star\|_{\mathcal{L}^2([-1,1]^d)} \le \varepsilon\Big] \ge 0.9 \quad \text{for all } L\text{-Lipschitz } f^*: [-1,1]^d \to \mathbb{R}$$

Our work: upper- and lower-bounds on this minimum width, for all d, ε, and L

$$\|f\|_{\mathcal{L}^2([-1,1]^d)} = \sqrt{\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathrm{Unif}([-1,1]^d)}[f(\mathbf{x})^2]}$$

Motivations

Motivations

1. Approximation capability of neural networks at (or near) random initialization

   [Andoni, Panigrahy, Valiant, & Zhang, '14; Bach, '17; Ji, Telgarsky, & Xian, '19; Yehudai & Shamir, '19; . . .]

   and kernel methods

   [Aizerman, Braverman, Rozonoer, '64; Cho & Saul, '09; . . .]

Motivations

1. Approximation capability of neural networks at (or near) random initialization

   [Andoni, Panigrahy, Valiant, & Zhang, '14; Bach, '17; Ji, Telgarsky, & Xian, '19; Yehudai & Shamir, '19; . . .]

   and kernel methods

   [Aizerman, Braverman, Rozonoer, '64; Cho & Saul, '09; . . .]

2. Interplay between dimension d and relative error ε/L

Our results (informally)

Question: What width is needed to approximate L-Lipschitz functions up to $\mathcal{L}^2([-1,1]^d)$ error ε?

Our results (informally)

Question: What width is needed to approximate L-Lipschitz functions up to $\mathcal{L}^2([-1,1]^d)$ error ε?

Answer: It depends!

Our results (informally)

Question: What width is needed to approximate L-Lipschitz functions up to $\mathcal{L}^2([-1,1]^d)$ error ε?

Answer: It depends!

$$\leq \ \mathrm{poly}(d) \qquad \text{if } L/\varepsilon = O(1)$$

Our results (informally)

Question: What width is needed to approximate L-Lipschitz functions up to $\mathcal{L}^2([-1,1]^d)$ error ε?

Answer: It depends!

$$\begin{aligned} &\leq \mathrm{poly}(d) && \text{if } L/\varepsilon = O(1) \\ &\leq \mathrm{poly}(L/\varepsilon) && \text{if } d = O(1) \end{aligned}$$

Our results (informally)

Question: What width is needed to approximate L-Lipschitz functions up to $\mathcal{L}^2([-1,1]^d)$ error ε?

Answer: It depends!

$$\begin{aligned} &\leq \operatorname{poly}(d) && \text{if } L/\varepsilon = O(1) \\ &\leq \operatorname{poly}(L/\varepsilon) && \text{if } d = O(1) \\ &\geq \exp(\Omega(d)) && \text{if } L/\varepsilon = \Omega(\sqrt{d}) \end{aligned}$$

Some prior work

Question: What width is needed to approximate L-Lipschitz functions up to $\mathcal{L}^2([-1,1]^d)$ error ε?

	Width	Comments
Maiorov, '99	$\geq \exp(\Omega(d))$	$L/\varepsilon \to \infty$
Yehudai & Shamir, '19; Kamath, Montasser, & Srebro, '20	$\geq \exp(\Omega(d))$	$L/\varepsilon \geq \mathrm{poly}(d)$
Andoni, Panigrahy, Valiant, & Zhang, '14	$\leq d^{O(L/\varepsilon)^2}$	exp activation
Bach, '17; Ji, Telgarsky, & Xian, '19	$\leq (L/\varepsilon)^{O(d)}$	$\mathcal{L}^\infty$ approx

Some prior work

Question: What width is needed to approximate L-Lipschitz functions up to $\mathcal{L}^2([-1,1]^d)$ error ε?

	Width	Comments
Maiorov, '99	$\geq \exp(\Omega(d))$	$L/\varepsilon \to \infty$
Yehudai & Shamir, '19; Kamath, Montasser, & Srebro, '20	$\geq \exp(\Omega(d))$	$L/\varepsilon \geq \text{poly}(d)$
Andoni, Panigrahy, Valiant, & Zhang, '14	$\leq d^{O(L/\varepsilon)^2}$	exp activation
Bach, '17; Ji, Telgarsky, & Xian, '19	$\leq (L/\varepsilon)^{O(d)}$	$\mathcal{L}^\infty$ approx

Maiorov's bound (for $H^1([-1,1]^d)$) applies to networks with arbitrary bottom-level weights, but only holds asymptotically as $L/\varepsilon \to \infty$

Some prior work

Question: What width is needed to approximate L-Lipschitz functions up to $\mathcal{L}^2([-1,1]^d)$ error ε?

	Width	**Comments**
Maiorov, '99	$\geq \exp(\Omega(d))$	$L/\varepsilon \to \infty$
Yehudai & Shamir, '19; Kamath, Montasser, & Srebro, '20	$\geq \exp(\Omega(d))$	$L/\varepsilon \geq \mathrm{poly}(d)$
Andoni, Panigrahy, Valiant, & Zhang, '14	$\leq d^{O(L/\varepsilon)^2}$	$\exp$ activation
Bach, '17; Ji, Telgarsky, & Xian, '19	$\leq (L/\varepsilon)^{O(d)}$	$\mathcal{L}^\infty$ approx

Hard function of YS and KMS has $\mathrm{poly}(d)$ Lipschitz constant

Some prior work

Question: What width is needed to approximate L-Lipschitz functions up to $\mathcal{L}^2([-1,1]^d)$ error ε?

	Width	Comments
Maiorov, '99	$\geq \exp(\Omega(d))$	$L/\varepsilon \to \infty$
Yehudai & Shamir, '19; Kamath, Montasser, & Srebro, '20	$\geq \exp(\Omega(d))$	$L/\varepsilon \geq \text{poly}(d)$
Andoni, Panigrahy, Valiant, & Zhang, '14	$\leq d^{O(L/\varepsilon)^2}$	exp activation
Bach, '17; Ji, Telgarsky, & Xian, '19	$\leq (L/\varepsilon)^{O(d)}$	$\mathcal{L}^\infty$ approx

Some prior work

Question: What width is needed to approximate L-Lipschitz functions up to $\mathcal{L}^2([-1,1]^d)$ error ε?

	Width	**Comments**
Maiorov, '99	$\geq \exp(\Omega(d))$	$L/\varepsilon \to \infty$
Yehudai & Shamir, '19; Kamath, Montasser, & Srebro, '20	$\geq \exp(\Omega(d))$	$L/\varepsilon \geq \text{poly}(d)$
Andoni, Panigrahy, Valiant, & Zhang, '14	$\leq d^{O(L/\varepsilon)^2}$	$\exp$ activation
Bach, '17; Ji, Telgarsky, & Xian, '19	$\leq (L/\varepsilon)^{O(d)}$	$\mathcal{L}^\infty$ approx

$\mathcal{L}^\infty$ approximation is stronger than $\mathcal{L}^2$ approximation

Some prior work

Question: What width is needed to approximate L-Lipschitz functions up to $\mathcal{L}^2([-1,1]^d)$ error ε?

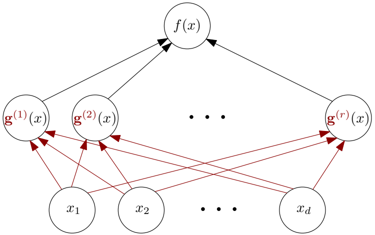
	Width	Comments
Maiorov, '99	$\geq \exp(\Omega(d))$	$L/\varepsilon \to \infty$
Yehudai & Shamir, '19; Kamath, Montasser, & Srebro, '20	$\geq \exp(\Omega(d))$	$L/\varepsilon \geq \mathrm{poly}(d)$
Andoni, Panigrahy, Valiant, & Zhang, '14	$\leq d^{O(L/\varepsilon)^2}$	exp activation
Bach, '17; Ji, Telgarsky, & Xian, '19	$\leq (L/\varepsilon)^{O(d)}$	$\mathcal{L}^\infty$ approx

Upshot: Prior work doesn't reveal the correct minimum width for arbitrary d and L/ε

Outline for rest of talk

1. Upper- and lower-bounds on the minimum width
2. Proof sketches
3. Some consequences

Part 1. Upper- and lower-bounds on the minimum width



Our main results

$$\mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \;:=\; \min\left\{ r \in \mathbb{N} \;:\; \Pr\left[\inf_{\hat{f} \in \mathcal{F}_r} \|\hat{f} - f^\star\|_{\mathcal{L}^2([-1,1]^d)} \;\leq\; \varepsilon \right] \geq 0.9 \right\}$$

smallest width r s.t. $\mathcal{F}_r$ (with bottom-level weights $\sim \mathcal{D}$) ε-approximates $f^\star$ with probability $\geq 90\%$

Our main results

$$\mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \;:=\; \min\left\{ r \in \mathbb{N} \;:\; \Pr\left[\inf_{\hat{f} \in \mathcal{F}_r} \|\hat{f} - f^\star\|_{\mathcal{L}^2([-1,1]^d)} \;\le\; \varepsilon \right] \ge 0.9 \right\}$$

smallest width r s.t. $\mathcal{F}_r$ (with bottom-level weights $\sim \mathcal{D}$) ε-approximates $f^\star$ with probability $\geq 90\%$

$$Q_{k,d} \;:=\; |\{\alpha \in \mathbb{Z}^d : \|\alpha\|_2 \le k\}|$$ number of integer lattice points in radius k ball in $\mathbb{R}^d$

Our main results

$$\mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \;:=\; \min\left\{ r \in \mathbb{N} \;:\; \Pr\left[\inf_{\hat{f} \in \mathcal{F}_r} \|\hat{f} - f^\star\|_{\mathcal{L}^2([-1,1]^d)} \;\le\; \varepsilon \right] \ge 0.9 \right\}$$

smallest width r s.t. $\mathcal{F}_r$ (with bottom-level weights $\sim \mathcal{D}$) ε-approximates $f^\star$ with probability $\ge 90\%$

$Q_{k,d} \;:=\; |\{\alpha \in \mathbb{Z}^d : \|\alpha\|_2 \le k\}|$ number of integer lattice points in radius k ball in $\mathbb{R}^d$

Theorem 1 (upper bound). For any L, ε, d, there exists a parameter distribution $\mathcal{D}$ such that

$$\sup_{L\text{-Lipschitz } f^\star : [-1,1]^d \to \mathbb{R}} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \;\le\; Q_{2L/\varepsilon,d}^{O(1)}$$

Our main results

$$\mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \;:=\; \min\left\{ r \in \mathbb{N} \;:\; \Pr\left[\inf_{\hat{f} \in \mathcal{F}_r} \|\hat{f} - f^\star\|_{\mathcal{L}^2([-1,1]^d)} \;\le\; \varepsilon \right] \ge 0.9 \right\}$$

smallest width r s.t. $\mathcal{F}_r$ (with bottom-level weights $\sim \mathcal{D}$) ε-approximates $f^\star$ with probability $\ge 90\%$

$Q_{k,d} \;:=\; |\{\alpha \in \mathbb{Z}^d : \|\alpha\|_2 \le k\}|$ number of integer lattice points in radius k ball in $\mathbb{R}^d$

Theorem 1 (upper bound). For any L, ε, d, there exists a parameter distribution $\mathcal{D}$ such that

$$\sup_{L\text{-Lipschitz } f^\star : [-1,1]^d \to \mathbb{R}} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \;\le\; Q_{2L/\varepsilon,d}^{O(1)}$$

Theorem 2 (lower bound). For any L, ε, d, and parameter distribution $\mathcal{D}$,

$$\sup_{L\text{-Lipschitz } f^\star : [-1,1]^d \to \mathbb{R}} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \;\ge\; \Omega(Q_{\frac{1}{18}L/\varepsilon,d})$$

Our main results

$$\mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \ := \ \min\left\{ r \in \mathbb{N} \ : \ \Pr\left[\inf_{\hat f \in \mathcal{F}_r} \|\hat f - f^\star\|_{\mathcal{L}^2([-1,1]^d)} \ \leq \ \varepsilon \right] \geq 0.9 \right\}$$

smallest width r s.t. $\mathcal{F}_r$ (with bottom-level weights $\sim \mathcal{D}$) ε-approximates $f^\star$ with probability $\geq 90\%$

$Q_{k,d} \ := \ |\{\alpha \in \mathbb{Z}^d : \|\alpha\|_2 \leq k\}|$ number of integer lattice points in radius k ball in $\mathbb{R}^d$

Theorem 1 (upper bound). For any L, ε, d, there exists a parameter distribution $\mathcal{D}$ such that

$$\sup_{L\text{-Lipschitz } f^\star : [-1,1]^d \to \mathbb{R}} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \ \leq \ Q_{2L/\varepsilon,d}^{O(1)}$$

Theorem 2 (lower bound). For any L, ε, d, and parameter distribution $\mathcal{D}$,

$$\sup_{L\text{-Lipschitz } f^\star : [-1,1]^d \to \mathbb{R}} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \ \geq \ \Omega(Q_{\frac{1}{18}L/\varepsilon,d})$$

Lower-bound, in fact, applies to any target-independent $\mathcal{F}_r$ (not just span of random ReLUs)

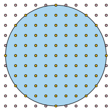
Counting integer lattice points in a ball

$Q_{k,d} := |\{\alpha \in \mathbb{Z}^d : \|\alpha\|_2 \leq k\}|$ number of integer lattice points in radius k Euclidean ball

Counting integer lattice points in a ball

$Q_{k,d} \ := \ |\{\alpha \in \mathbb{Z}^d : \|\alpha\|_2 \leq k\}|$ number of integer lattice points in radius k Euclidean ball

Generalized Gauss Circle Problem: As $k \to \infty$,

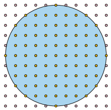


$$Q_{k,d} \ = \ \mathrm{vol}(B_d) \cdot k^d \cdot (1 + o(1)) \ \approx \ \frac{1}{\sqrt{\pi d}} \left(\frac{2\pi e k^2}{d}\right)^{d/2} \cdot (1 + o(1))$$

Counting integer lattice points in a ball

$Q_{k,d} := |\{\alpha \in \mathbb{Z}^d : \|\alpha\|_2 \leq k\}|$ number of integer lattice points in radius k Euclidean ball

Generalized Gauss Circle Problem: As $k \to \infty$,



$$Q_{k,d} = \mathrm{vol}(B_d) \cdot k^d \cdot (1 + o(1)) \approx \frac{1}{\sqrt{\pi d}} \left(\frac{2\pi e k^2}{d}\right)^{d/2} \cdot (1 + o(1))$$

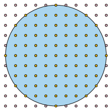
But when $d > k^2$, more favorable bounds via (simple) combinatorics:

$$\binom{d}{\lfloor k^2 \rfloor} \leq Q_{k,d} \leq \binom{k^2 + 2d - 1}{k^2}$$

Counting integer lattice points in a ball

$Q_{k,d} := |\{\alpha \in \mathbb{Z}^d : \|\alpha\|_2 \leq k\}|$ number of integer lattice points in radius k Euclidean ball

Generalized Gauss Circle Problem: As $k \to \infty$,



$$Q_{k,d} = \mathrm{vol}(B_d) \cdot k^d \cdot (1 + o(1)) \approx \frac{1}{\sqrt{\pi d}} \left(\frac{2\pi e k^2}{d}\right)^{d/2} \cdot (1 + o(1))$$

But when $d > k^2$, more favorable bounds via (simple) combinatorics:

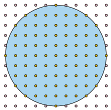
$$\binom{d}{\lfloor k^2 \rfloor} \leq Q_{k,d} \leq \binom{k^2 + 2d - 1}{k^2}$$

$$\implies Q_{k,d} = \exp\left(\Theta\left(\min\left(d \log\left(\frac{k^2}{d} + 2\right), k^2 \log\left(\frac{d}{k^2} + 2\right)\right)\right)\right)$$

Counting integer lattice points in a ball

$Q_{k,d} := |\{\alpha \in \mathbb{Z}^d : \|\alpha\|_2 \leq k\}|$ number of integer lattice points in radius k Euclidean ball

Generalized Gauss Circle Problem: As $k \to \infty$,



$$Q_{k,d} = \mathrm{vol}(B_d) \cdot k^d \cdot (1 + o(1)) \approx \frac{1}{\sqrt{\pi d}} \left(\frac{2\pi e k^2}{d} \right)^{d/2} \cdot (1 + o(1))$$

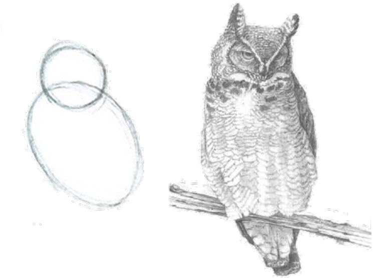
But when $d > k^2$, more favorable bounds via (simple) combinatorics:

$$\binom{d}{\lfloor k^2 \rfloor} \leq Q_{k,d} \leq \binom{k^2 + 2d - 1}{k^2}$$

$$\implies Q_{k,d} = \exp\left(\Theta\left(\min\left(d \log\left(\frac{k^2}{d} + 2 \right), k^2 \log\left(\frac{d}{k^2} + 2 \right) \right) \right) \right)$$

Theorems 1 & 2 $\implies \sup_{L\text{-Lipschitz } f^\star} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) = \begin{cases} \mathrm{poly}(d) & \text{if } L/\varepsilon = \Theta(1) \\ \mathrm{poly}(L/\varepsilon) & \text{if } d = \Theta(1) \\ \exp(\Theta(d)) & \text{if } L/\varepsilon = \Theta(\sqrt{d}) \end{cases}$

Part 2. Proof sketches



Proof of upper-bound (sketch)

Theorem 1 (upper bound). For any L, ε, d, there exists a parameter distribution $\mathcal{D}$ such that

$$\sup_{L\text{-Lipschitz } f^\star : [-1,1]^d \to \mathbb{R}} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \;\le\; Q_{2L/\varepsilon,d}^{O(1)}$$

Proof of upper-bound (sketch)

Theorem 1 (upper bound). For any L, ε, d, there exists a parameter distribution $\mathcal{D}$ such that

$$\sup_{L\text{-Lipschitz } f^\star : [-1,1]^d \to \mathbb{R}} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \;\leq\; Q_{2L/\varepsilon,d}^{O(1)}$$

Follow standard recipe [e.g., Andoni, Panigrahy, Valiant, & Zhang, '14; Bresler & Nagaraj, '20], with some tweaks:

Proof of upper-bound (sketch)

Theorem 1 (upper bound). For any L, ε, d, there exists a parameter distribution $\mathcal{D}$ such that

$$\sup_{L\text{-Lipschitz } f^\star : [-1,1]^d \to \mathbb{R}} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \ \le \ Q_{2L/\varepsilon,d}^{O(1)}$$

Follow standard recipe [e.g., Andoni, Panigrahy, Valiant, & Zhang, '14; Bresler & Nagaraj, '20], with some tweaks:

1. Get $\varepsilon/2$-approximation of L-Lipschitz $f^\star$ using orthonormal basis functions

$$\sqrt{2}\sin(\pi\alpha \cdot x/2) \quad \text{and} \quad \sqrt{2}\cos(\pi\alpha \cdot x/2)$$

   for $\alpha \in \mathbb{Z}^d$ with $\|\alpha\|_2 \le 2L/\varepsilon$

Proof of upper-bound (sketch)

Theorem 1 (upper bound). For any L, ε, d, there exists a parameter distribution $\mathcal{D}$ such that

$$\sup_{L\text{-Lipschitz } f^\star : [-1,1]^d \to \mathbb{R}} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \;\leq\; Q_{2L/\varepsilon,d}^{O(1)}$$

Follow standard recipe [e.g., Andoni, Panigrahy, Valiant, & Zhang, '14; Bresler & Nagaraj, '20], with some tweaks:

1. Get $\varepsilon/2$-approximation of L-Lipschitz $f^\star$ using orthonormal basis functions

   $$\sqrt{2}\sin(\pi\alpha \cdot x/2) \quad \text{and} \quad \sqrt{2}\cos(\pi\alpha \cdot x/2)$$

   for $\alpha \in \mathbb{Z}^d$ with $\|\alpha\|_2 \leq 2L/\varepsilon$
2. Construct suitable parameter distribution $\mathcal{D}$, so every trigonometric polynomial

   $$p^\star \;\in\; \mathrm{span}\Big\{\sin(\pi\alpha \cdot x),\, \cos(\pi\alpha \cdot x) \;:\; \alpha \in \mathbb{Z}^d,\, \|\alpha\|_2 \leq k\Big\}$$

   with bounded coefficients has

   $$\mathrm{MinWidth}_{\varepsilon/2,d,\mathcal{D}}(p^\star) \;\leq\; \mathrm{poly}(d,k,1/\varepsilon) \cdot Q_{k,d}^{O(1)}$$

Proof of upper-bound (sketch)

Theorem 1 (upper bound). For any L, ε, d, there exists a parameter distribution $\mathcal{D}$ such that

$$\sup_{L\text{-Lipschitz } f^\star : [-1,1]^d \to \mathbb{R}} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \leq Q_{2L/\varepsilon,d}^{O(1)}$$

Follow standard recipe [e.g., Andoni, Panigrahy, Valiant, & Zhang, '14; Bresler & Nagaraj, '20], with some tweaks:

1. Get $\varepsilon/2$-approximation of L-Lipschitz $f^\star$ using orthonormal basis functions

   $$\sqrt{2}\sin(\pi\alpha \cdot x/2) \quad \text{and} \quad \sqrt{2}\cos(\pi\alpha \cdot x/2)$$

   for $\alpha \in \mathbb{Z}^d$ with $\|\alpha\|_2 \leq 2L/\varepsilon$
2. Construct suitable parameter distribution $\mathcal{D}$, so every trigonometric polynomial

   $$p^\star \in \mathrm{span}\Big\{\sin(\pi\alpha \cdot x), \cos(\pi\alpha \cdot x) \ : \ \alpha \in \mathbb{Z}^d, \|\alpha\|_2 \leq k\Big\}$$

   with bounded coefficients has

   $$\mathrm{MinWidth}_{\varepsilon/2,d,\mathcal{D}}(p^\star) \leq \mathrm{poly}(d, k, 1/\varepsilon) \cdot Q_{k,d}^{O(1)}$$

   Basis of "sinusoidal ridge functions" are especially convenient for this step

Proof of lower-bound (sketch)

Theorem 2 (lower bound). For any L, ε, d, and parameter distribution $\mathcal{D}$,

$$\sup_{L\text{-Lipschitz } f^\star : [-1,1]^d \to \mathbb{R}} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \;\geq\; \Omega(Q_{\frac{1}{18}L/\varepsilon,d})$$

Proof of lower-bound (sketch)

Theorem 2 (lower bound). For any L, ε, d, and parameter distribution $\mathcal{D}$,

$$\sup_{L\text{-Lipschitz } f^\star : [-1,1]^d \to \mathbb{R}} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \;\geq\; \Omega(Q_{\frac{1}{18}L/\varepsilon,d})$$

We generalize a dimension argument of [Barron, '93]:

Proof of lower-bound (sketch)

Theorem 2 (lower bound). For any L, ε, d, and parameter distribution $\mathcal{D}$,

$$\sup_{L\text{-Lipschitz } f^\star : [-1,1]^d \to \mathbb{R}} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \;\geq\; \Omega(Q_{\frac{1}{18}L/\varepsilon,d})$$

We generalize a dimension argument of [Barron, '93]:

1. If $\varphi_1, \ldots, \varphi_N \in \mathcal{L}^2$ are orthonormal with $N \geq r$, then $\mathcal{F}_r$ is $\sqrt{1 - \frac{r}{N}}$-far from at least one φ_i

Proof of lower-bound (sketch)

Theorem 2 (lower bound). For any L, ε, d, and parameter distribution $\mathcal{D}$,

$$\sup_{L\text{-Lipschitz } f^\star : [-1,1]^d \to \mathbb{R}} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \geq \Omega(Q_{\frac{1}{18}L/\varepsilon,d})$$

We generalize a dimension argument of [Barron, '93]:

1. If $\varphi_1, \ldots, \varphi_N \in \mathcal{L}^2$ are orthonormal with $N \geq r$, then $\mathcal{F}_r$ is $\sqrt{1 - \frac{r}{N}}$-far from at least one φ_i
   - $\mathcal{F}_r$ (or any dimension r subspace of $\mathcal{L}^2$) cannot approximate them all if $r \ll N$

Proof of lower-bound (sketch)

Theorem 2 (lower bound). For any L, ε, d, and parameter distribution $\mathcal{D}$,

$$\sup_{L\text{-Lipschitz } f^\star : [-1,1]^d \to \mathbb{R}} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \;\geq\; \Omega(Q_{\frac{1}{18}L/\varepsilon,d})$$

We generalize a dimension argument of [Barron, '93]:

1. If $\varphi_1, \ldots, \varphi_N \in \mathcal{L}^2$ are orthonormal with $N \geq r$, then $\mathcal{F}_r$ is $\sqrt{1 - \frac{r}{N}}$-far from at least one φ_i
   - $\mathcal{F}_r$ (or any dimension r subspace of $\mathcal{L}^2$) cannot approximate them all if $r \ll N$
2. The $N = Q_{k,d}$ sinusoidal ridge functions (from upper-bound proof) are $O(k)$-Lipschitz

Proof of lower-bound (sketch)

Theorem 2 (lower bound). For any L, ε, d, and parameter distribution $\mathcal{D}$,

$$\sup_{L\text{-Lipschitz } f^\star : [-1,1]^d \to \mathbb{R}} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \;\geq\; \Omega(Q_{\frac{1}{18}L/\varepsilon,d})$$

We generalize a dimension argument of [Barron, '93]:

1. If $\varphi_1, \ldots, \varphi_N \in \mathcal{L}^2$ are orthonormal with $N \geq r$, then $\mathcal{F}_r$ is $\sqrt{1 - \frac{r}{N}}$-far from at least one φ_i
   - $\mathcal{F}_r$ (or any dimension r subspace of $\mathcal{L}^2$) cannot approximate them all if $r \ll N$
2. The $N = Q_{k,d}$ sinusoidal ridge functions (from upper-bound proof) are $O(k)$-Lipschitz
3. Combine these facts + scaling argument, with $k = \Theta(L/\varepsilon)$

Proof of lower-bound (sketch)

Theorem 2 (lower bound). For any L, ε, d, and parameter distribution $\mathcal{D}$,

$$\sup_{L\text{-Lipschitz } f^\star : [-1,1]^d \to \mathbb{R}} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \;\geq\; \Omega(Q_{\frac{1}{18}L/\varepsilon,d})$$

We generalize a dimension argument of [Barron, '93]:

1. If $\varphi_1, \ldots, \varphi_N \in \mathcal{L}^2$ are orthonormal with $N \geq r$, then $\mathcal{F}_r$ is $\sqrt{1 - \frac{r}{N}}$-far from at least one φ_i
   - $\mathcal{F}_r$ (or any dimension r subspace of $\mathcal{L}^2$) cannot approximate them all if $r \ll N$
2. The $N = Q_{k,d}$ sinusoidal ridge functions (from upper-bound proof) are $O(k)$-Lipschitz
3. Combine these facts + scaling argument, with $k = \Theta(L/\varepsilon)$

If $\mathcal{D}_{\mathrm{weights}}$ is invariant to coordinate permutations, then the hard-to-approximate function is *explicit*:

$$x \;\mapsto\; \varepsilon \sin\left(\pi \sum_{i=1}^{\ell} x_i\right), \quad \ell = \min\{\Theta(d),\, \Theta(L^2/\varepsilon^2)\}$$

Key lemma

Lemma. Let H be a Hilbert space, and fix orthonormal $\varphi_1, \ldots, \varphi_N \in H$. Let $\mathbf{W}$ be (possibly random) finite-dimensional subspace of H with $r := \mathbb{E}[\dim(\mathbf{W})] < \infty$. Then there is some $i \in [N]$ such that

$$\mathbb{E}\left[\inf_{g \in \mathbf{W}} \|g - \varphi_i\|_H^2\right] \;\geq\; 1 - \frac{r}{N}.$$

Key lemma

Lemma. Let H be a Hilbert space, and fix orthonormal $\varphi_1, \ldots, \varphi_N \in H$. Let $\mathbf{W}$ be (possibly random) finite-dimensional subspace of H with $r := \mathbb{E}[\dim(\mathbf{W})] < \infty$. Then there is some $i \in [N]$ such that

$$\mathbb{E}\left[\inf_{g \in \mathbf{W}} \|g - \varphi_i\|_H^2\right] \ \geq \ 1 - \frac{r}{N}.$$

Proof. Let $\mathbf{u}_1, \ldots, \mathbf{u}_\mathbf{d}$ be ONB for $\mathbf{W}$, with $\mathbf{d} := \dim(\mathbf{W})$, and let $\Pi_\mathbf{W}$ be orthoprojector for $\mathbf{W}$.

Key lemma

Lemma. Let H be a Hilbert space, and fix orthonormal $\varphi_1, \ldots, \varphi_N \in H$. Let $\mathbf{W}$ be (possibly random) finite-dimensional subspace of H with $r := \mathbb{E}[\dim(\mathbf{W})] < \infty$. Then there is some $i \in [N]$ such that

$$\mathbb{E}\left[\inf_{g \in \mathbf{W}} \|g - \varphi_i\|_H^2\right] \ \geq \ 1 - \frac{r}{N}.$$

Proof. Let $\mathbf{u}_1, \ldots, \mathbf{u}_\mathbf{d}$ be ONB for $\mathbf{W}$, with $\mathbf{d} := \dim(\mathbf{W})$, and let $\Pi_\mathbf{W}$ be orthoprojector for $\mathbf{W}$.

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\inf_{g \in \mathbf{W}} \|g - \varphi_i\|_H^2\right]$$

Key lemma

Lemma. Let H be a Hilbert space, and fix orthonormal $\varphi_1, \dots, \varphi_N \in H$. Let $\mathbf{W}$ be (possibly random) finite-dimensional subspace of H with $r := \mathbb{E}[\dim(\mathbf{W})] < \infty$. Then there is some $i \in [N]$ such that

$$\mathbb{E}\left[\inf_{g \in \mathbf{W}} \|g - \varphi_i\|_H^2\right] \;\geq\; 1 - \frac{r}{N}.$$

Proof. Let $\mathbf{u}_1, \dots, \mathbf{u}_{\mathbf{d}}$ be ONB for $\mathbf{W}$, with $\mathbf{d} := \dim(\mathbf{W})$, and let $\Pi_{\mathbf{W}}$ be orthoprojector for $\mathbf{W}$.

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\inf_{g \in \mathbf{W}} \|g - \varphi_i\|_H^2\right] \quad = \quad \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[1 - \|\Pi_{\mathbf{W}} \varphi_i\|_H^2\right]$$

Key lemma

Lemma. Let H be a Hilbert space, and fix orthonormal $\varphi_1, \ldots, \varphi_N \in H$. Let $\mathbf{W}$ be (possibly random) finite-dimensional subspace of H with $r := \mathbb{E}[\dim(\mathbf{W})] < \infty$. Then there is some $i \in [N]$ such that

$$\mathbb{E}\left[\inf_{g \in \mathbf{W}} \|g - \varphi_i\|_H^2\right] \;\geq\; 1 - \frac{r}{N}.$$

Proof. Let $\mathbf{u}_1, \ldots, \mathbf{u}_{\mathbf{d}}$ be ONB for $\mathbf{W}$, with $\mathbf{d} := \dim(\mathbf{W})$, and let $\Pi_{\mathbf{W}}$ be orthoprojector for $\mathbf{W}$.

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\inf_{g \in \mathbf{W}} \|g - \varphi_i\|_H^2\right] &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[1 - \|\Pi_{\mathbf{W}}\varphi_i\|_H^2\right] \\ &= 1 - \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N}\|\Pi_{\mathbf{W}}\varphi_i\|_H^2\right]\end{aligned}$$

Key lemma

Lemma. Let H be a Hilbert space, and fix orthonormal $\varphi_1, \ldots, \varphi_N \in H$. Let $\mathbf{W}$ be (possibly random) finite-dimensional subspace of H with $r := \mathbb{E}[\dim(\mathbf{W})] < \infty$. Then there is some $i \in [N]$ such that

$$\mathbb{E}\left[\inf_{g \in \mathbf{W}} \|g - \varphi_i\|_H^2\right] \;\geq\; 1 - \frac{r}{N}.$$

Proof. Let $\mathbf{u}_1, \ldots, \mathbf{u}_\mathbf{d}$ be ONB for $\mathbf{W}$, with $\mathbf{d} := \dim(\mathbf{W})$, and let $\Pi_\mathbf{W}$ be orthoprojector for $\mathbf{W}$.

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\inf_{g \in \mathbf{W}} \|g - \varphi_i\|_H^2\right] &= \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[1 - \|\Pi_\mathbf{W}\varphi_i\|_H^2\right] \\
&= 1 - \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^N \|\Pi_\mathbf{W}\varphi_i\|_H^2\right] \\
&= 1 - \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^N \sum_{k=1}^{\mathbf{d}} \langle \mathbf{u}_k, \varphi_i \rangle_H^2\right]
\end{aligned}$$

Key lemma

Lemma. Let H be a Hilbert space, and fix orthonormal $\varphi_1, \dots, \varphi_N \in H$. Let $\mathbf{W}$ be (possibly random) finite-dimensional subspace of H with $r := \mathbb{E}[\dim(\mathbf{W})] < \infty$. Then there is some $i \in [N]$ such that

$$\mathbb{E}\left[\inf_{g \in \mathbf{W}} \|g - \varphi_i\|_H^2\right] \;\geq\; 1 - \frac{r}{N}.$$

Proof. Let $\mathbf{u}_1, \dots, \mathbf{u}_\mathbf{d}$ be ONB for $\mathbf{W}$, with $\mathbf{d} := \dim(\mathbf{W})$, and let $\Pi_\mathbf{W}$ be orthoprojector for $\mathbf{W}$.

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\inf_{g \in \mathbf{W}} \|g - \varphi_i\|_H^2\right] &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[1 - \|\Pi_\mathbf{W}\varphi_i\|_H^2\right] \\
&= 1 - \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N} \|\Pi_\mathbf{W}\varphi_i\|_H^2\right] \\
&= 1 - \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N}\sum_{k=1}^{\mathbf{d}} \langle \mathbf{u}_k, \varphi_i\rangle_H^2\right] \\
&\geq 1 - \frac{1}{N}\mathbb{E}\left[\sum_{k=1}^{\mathbf{d}} 1\right]
\end{aligned}$$

Key lemma

Lemma. Let H be a Hilbert space, and fix orthonormal $\varphi_1, \ldots, \varphi_N \in H$. Let $\mathbf{W}$ be (possibly random) finite-dimensional subspace of H with $r := \mathbb{E}[\dim(\mathbf{W})] < \infty$. Then there is some $i \in [N]$ such that

$$\mathbb{E}\left[\inf_{g \in \mathbf{W}} \|g - \varphi_i\|_H^2\right] \;\geq\; 1 - \frac{r}{N}.$$

Proof. Let $\mathbf{u}_1, \ldots, \mathbf{u}_{\mathbf{d}}$ be ONB for $\mathbf{W}$, with $\mathbf{d} := \dim(\mathbf{W})$, and let $\Pi_{\mathbf{W}}$ be orthoprojector for $\mathbf{W}$.

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\inf_{g \in \mathbf{W}} \|g - \varphi_i\|_H^2\right] &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[1 - \|\Pi_{\mathbf{W}}\varphi_i\|_H^2\right] \\
&= 1 - \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N} \|\Pi_{\mathbf{W}}\varphi_i\|_H^2\right] \\
&= 1 - \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N}\sum_{k=1}^{\mathbf{d}} \langle \mathbf{u}_k, \varphi_i\rangle_H^2\right] \\
&\geq 1 - \frac{1}{N}\mathbb{E}\left[\sum_{k=1}^{\mathbf{d}} 1\right] = 1 - \frac{r}{N}. \qquad \square
\end{aligned}$$

Part 3. Some consequences

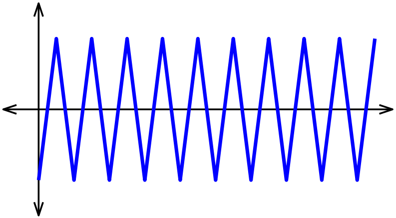
Depth separation

- Recent line-of-inquiry on separations between poly-size “shallow” nets and poly-size “deep” nets
  [Telgarsky, ’16; Eldan & Shamir, ’16; Daniely, ’17; Safran & Shamir, ’17; Safran, Eldan, & Shamir, ’19; ...]

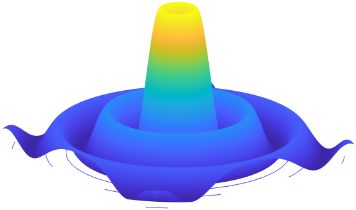
Depth separation

- Recent line-of-inquiry on separations between poly-size "shallow" nets and poly-size "deep" nets
  [Telgarsky, '16; Eldan & Shamir, '16; Daniely, '17; Safran & Shamir, '17; Safran, Eldan, & Shamir, '19; ...]

All known "hard" functions exhibiting the separation have been highly oscillatory or jagged



Telgarsky's iterated tent map

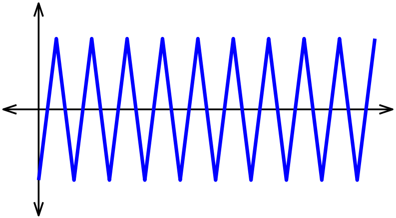


Oscillatory radial function

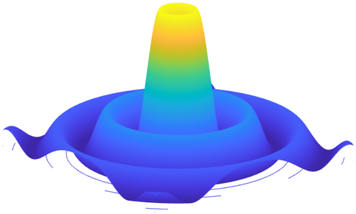
Depth separation

- Recent line-of-inquiry on separations between poly-size “shallow” nets and poly-size “deep” nets [Telgarsky, ’16; Eldan & Shamir, ’16; Daniely, ’17; Safran & Shamir, ’17; Safran, Eldan, & Shamir, ’19; . . .]

All known “hard” functions exhibiting the separation have been highly oscillatory or jagged



Telgarsky’s iterated tent map



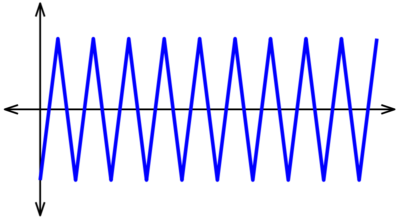
Oscillatory radial function

- [Safran, Eldan, & Shamir, ’19]: Is there a 1-Lipschitz function that separates $\mathrm{poly}(d)$-size depth-2 nets from $\mathrm{poly}(d)$-size depth-3 nets?

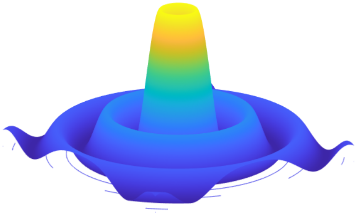
Depth separation

- Recent line-of-inquiry on separations between poly-size "shallow" nets and poly-size "deep" nets [Telgarsky, '16; Eldan & Shamir, '16; Daniely, '17; Safran & Shamir, '17; Safran, Eldan, & Shamir, '19; ...]

All known "hard" functions exhibiting the separation have been highly oscillatory or jagged



Telgarsky's iterated tent map



Oscillatory radial function

- [Safran, Eldan, & Shamir, '19]: Is there a 1-Lipschitz function that separates $\mathrm{poly}(d)$-size depth-2 nets from $\mathrm{poly}(d)$-size depth-3 nets?

Our results $\Rightarrow$ No, for constant $\mathcal{L}^2$ approximation error

Hardness of learning with deeper networks

- "Hardness of approximation implies hardness of learning." [Malach, Yehudai, Shamir, & Shalev-Shwartz, '21]

Hardness of learning with deeper networks

- "Hardness of approximation implies hardness of learning." [Malach, Yehudai, Shamir, & Shalev-Shwartz, '21]
- Suppose no $\mathrm{poly}(d)$-size three-layer neural network can weakly approximate f. Then, no $\mathrm{poly}(d)$-size NN of *any* depth and a standard Xavier initialization will weakly learn f with $\mathrm{poly}(d)$ steps of gradient descent.

Hardness of learning with deeper networks

- "Hardness of approximation implies hardness of learning." [Malach, Yehudai, Shamir, & Shalev-Shwartz, '21]
- Suppose no $\mathrm{poly}(d)$-size three-layer neural network can weakly approximate f. Then, no $\mathrm{poly}(d)$-size NN of *any* depth and a standard Xavier initialization will weakly learn f with $\mathrm{poly}(d)$ steps of gradient descent.
- Proof idea:
  - Initialized gradients are Lipschitz near initialization and can be approximated using three-layer neural networks.

Hardness of learning with deeper networks

- "Hardness of approximation implies hardness of learning." [Malach, Yehudai, Shamir, & Shalev-Shwartz, '21]
- Suppose no $\mathrm{poly}(d)$-size three-layer neural network can weakly approximate f. Then, no $\mathrm{poly}(d)$-size NN of *any* depth and a standard Xavier initialization will weakly learn f with $\mathrm{poly}(d)$ steps of gradient descent.
- Proof idea:
  - Initialized gradients are Lipschitz near initialization and can be approximated using three-layer neural networks.
  - Because f cannot be weakly approximated, gradients cannot correlate strongly with f.

Hardness of learning with deeper networks

- "Hardness of approximation implies hardness of learning." [Malach, Yehudai, Shamir, & Shalev-Shwartz, '21]
- Suppose no $\mathrm{poly}(d)$-size three-layer neural network can weakly approximate f. Then, no $\mathrm{poly}(d)$-size NN of *any* depth and a standard Xavier initialization will weakly learn f with $\mathrm{poly}(d)$ steps of gradient descent.
- Proof idea:
  - Initialized gradients are Lipschitz near initialization and can be approximated using three-layer neural networks.
  - Because f cannot be weakly approximated, gradients cannot correlate strongly with f.
  - $\mathrm{poly}(d)$ gradient steps are extremely small and remain near initialization.

Hardness of learning with deeper networks

- "Hardness of approximation implies hardness of learning." [Malach, Yehudai, Shamir, & Shalev-Shwartz, '21]
- Suppose no $\mathrm{poly}(d)$-size **two**-layer neural network can weakly approximate f. Then, no $\mathrm{poly}(d)$-size NN of *any* depth and a standard Xavier initialization will weakly learn f with $\mathrm{poly}(d)$ steps of gradient descent.
- Proof idea:
  - Initialized gradients are Lipschitz near initialization and can be approximated using **two**-layer neural networks. [Our upper bound]
  - Because f cannot be weakly approximated, gradients cannot correlate strongly with f.
  - $\mathrm{poly}(d)$ gradient steps are extremely small and remain near initialization.

Lower-bounds for kernel methods

- Lower-bound applies to all methods that pick $\hat{f}$ from a target-independent subspace of dimension r — including **kernel methods** based on $r = n$ examples $(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})$:

$$\hat{f} \in \operatorname{span}\left\{ \mathsf{K}(x^{(i)}, \cdot) \; : \; i = 1, \dots, n \right\}$$

Lower-bounds for kernel methods

- Lower-bound applies to all methods that pick $\hat{f}$ from a target-independent subspace of dimension r — including **kernel methods** based on $r = n$ examples $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$:

$$\hat{f} \in \operatorname{span}\Big\{ \mathsf{K}(x^{(i)}, \cdot) \;:\; i = 1, \ldots, n \Big\}$$

  Example: Lower-bound for learning parity functions under uniform distribution on $\{-1, 1\}^d$ with non-adaptive membership queries (MQs) [Bubeck (after Allen-Zhu & Li), '20]

Lower-bounds for kernel methods

- Lower-bound applies to all methods that pick $\hat{f}$ from a target-independent subspace of dimension r — including **kernel methods** based on $r = n$ examples $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$:

$$\hat{f} \in \operatorname{span}\left\{\mathsf{K}(x^{(i)}, \cdot) \ : \ i = 1, \ldots, n\right\}$$

Example: Lower-bound for learning parity functions under uniform distribution on $\{-1, 1\}^d$ with non-adaptive membership queries (MQs) [Bubeck (after Allen-Zhu & Li), '20]

- Why? Learnable — with noise! — using non-adaptive MQs in $\operatorname{poly}(d)$ time [e.g., Feldman, '07]
(Learner allowed to choose $x^{(1)}, \ldots, x^{(n)} \in \{-1, 1\}^d$, which subsequently get labels $y^{(i)}$'s)

Lower-bounds for kernel methods

- Lower-bound applies to all methods that pick $\hat{f}$ from a target-independent subspace of dimension r — including **kernel methods** based on $r = n$ examples $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$:

$$\hat{f} \in \operatorname{span}\left\{\mathsf{K}(x^{(i)}, \cdot) \;:\; i = 1, \ldots, n\right\}$$

Example: Lower-bound for learning parity functions under uniform distribution on $\{-1,1\}^d$ with non-adaptive membership queries (MQs) [Bubeck (after Allen-Zhu & Li), '20]

  - Why? Learnable — with noise! — using non-adaptive MQs in $\operatorname{poly}(d)$ time [e.g., Feldman, '07] (Learner allowed to choose $x^{(1)}, \ldots, x^{(n)} \in \{-1,1\}^d$, which subsequently get labels $y^{(i)}$'s)
  - Let $\varphi_1, \ldots, \varphi_N$ be the $N = 2^d$ parity functions on $\{-1,1\}^d$, which is ONB for $\mathcal{L}^2(\{-1,1\}^d)$

Lower-bounds for kernel methods

- Lower-bound applies to all methods that pick $\hat{f}$ from a target-independent subspace of dimension r — including **kernel methods** based on $r = n$ examples $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$:

$$\hat{f} \in \operatorname{span}\Big\{\mathsf{K}(x^{(i)}, \cdot) \;:\; i = 1, \ldots, n\Big\}$$

Example: Lower-bound for learning parity functions under uniform distribution on $\{-1,1\}^d$ with non-adaptive membership queries (MQs) [Bubeck (after Allen-Zhu & Li), '20]

- Why? Learnable — with noise! — using non-adaptive MQs in $\operatorname{poly}(d)$ time [e.g., Feldman, '07] (Learner allowed to choose $x^{(1)}, \ldots, x^{(n)} \in \{-1,1\}^d$, which subsequently get labels $y^{(i)}$'s)
- Let $\varphi_1, \ldots, \varphi_N$ be the $N = 2^d$ parity functions on $\{-1,1\}^d$, which is ONB for $\mathcal{L}^2(\{-1,1\}^d)$
- **Proposition [B/AZL, '20]:** Every kernel method, even if allowed non-adaptive MQs, needs

$$n \;\geq\; (1 - \varepsilon) \cdot 2^d$$

examples to guarantee mean squared error $\leq \varepsilon$ when any of the φ_i could be the true target

Lower-bounds for kernel methods

- Lower-bound applies to all methods that pick $\hat{f}$ from a target-independent subspace of dimension r — including **kernel methods** based on $r = n$ examples $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$:

$$\hat{f} \in \operatorname{span}\left\{\mathsf{K}(x^{(i)}, \cdot) \ : \ i = 1, \ldots, n\right\}$$

Example: Lower-bound for learning parity functions under uniform distribution on $\{-1,1\}^d$ with non-adaptive membership queries (MQs) [Bubeck (after Allen-Zhu & Li), '20]

- Why? Learnable — with noise! — using non-adaptive MQs in $\operatorname{poly}(d)$ time [e.g., Feldman, '07] (Learner allowed to choose $x^{(1)}, \ldots, x^{(n)} \in \{-1,1\}^d$, which subsequently get labels $y^{(i)}$'s)
- Let $\varphi_1, \ldots, \varphi_N$ be the $N = 2^d$ parity functions on $\{-1,1\}^d$, which is ONB for $\mathcal{L}^2(\{-1,1\}^d)$
- **Proposition [B/AZL, '20]:** Every kernel method, even if allowed non-adaptive MQs, needs

$$n \ \geq \ (1 - \varepsilon) \cdot 2^d$$

examples to guarantee mean squared error $\leq \varepsilon$ when any of the φ_i could be the true target
- Easy consequence of the key lemma!

Recap and closing

1. Width needed to approximate L-Lipschitz functions up to $\mathcal{L}^2([-1,1]^d)$ error ε:

$$\sup_{L\text{-Lipschitz } f^\star} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \;=\; Q^{\Theta(1)}_{\Theta(L/\varepsilon),d} \;=\; \begin{cases} \mathrm{poly}(d) & \text{if } L/\varepsilon = \Theta(1) \\ \mathrm{poly}(L/\varepsilon) & \text{if } d = \Theta(1) \\ \exp(\Theta(d)) & \text{if } L/\varepsilon = \Theta(\sqrt{d}) \end{cases}$$

Recap and closing

1. Width needed to approximate L-Lipschitz functions up to $\mathcal{L}^2([-1,1]^d)$ error ε:

$$\sup_{L\text{-Lipschitz } f^\star} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \; = \; Q^{\Theta(1)}_{\Theta(L/\varepsilon),d} \; = \; \begin{cases} \mathrm{poly}(d) & \text{if } L/\varepsilon = \Theta(1) \\ \mathrm{poly}(L/\varepsilon) & \text{if } d = \Theta(1) \\ \exp(\Theta(d)) & \text{if } L/\varepsilon = \Theta(\sqrt{d}) \end{cases}$$

2. Sheds some light on other questions related to neural nets & kernel methods . . .

Recap and closing

1. Width needed to approximate L-Lipschitz functions up to $\mathcal{L}^2([-1,1]^d)$ error ε:

$$\sup_{L\text{-Lipschitz } f^\star} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \;=\; Q^{\Theta(1)}_{\Theta(L/\varepsilon),d} \;=\; \begin{cases} \mathrm{poly}(d) & \text{if } L/\varepsilon = \Theta(1) \\ \mathrm{poly}(L/\varepsilon) & \text{if } d = \Theta(1) \\ \exp(\Theta(d)) & \text{if } L/\varepsilon = \Theta(\sqrt{d}) \end{cases}$$

2. Sheds some light on other questions related to neural nets & kernel methods . . .
3. Also have results for Sobolev classes H^s for $s \geq 1$ (see paper: `arXiv:2102.02336`)

Recap and closing

1. Width needed to approximate L-Lipschitz functions up to $\mathcal{L}^2([-1,1]^d)$ error ε:

$$\sup_{L\text{-Lipschitz } f^\star} \mathrm{MinWidth}_{\varepsilon,d,\mathcal{D}}(f^\star) \;=\; Q^{\Theta(1)}_{\Theta(L/\varepsilon),d} \;=\; \begin{cases} \mathrm{poly}(d) & \text{if } L/\varepsilon = \Theta(1) \\ \mathrm{poly}(L/\varepsilon) & \text{if } d = \Theta(1) \\ \exp(\Theta(d)) & \text{if } L/\varepsilon = \Theta(\sqrt{d}) \end{cases}$$

2. Sheds some light on other questions related to neural nets & kernel methods ...
3. Also have results for Sobolev classes H^s for $s \geq 1$ (see paper: `arXiv:2102.02336`)

Thank you!

We gratefully acknowledge support from the NSF (CCF-{1563155, 1703925, 1740833, 1763970, 1814873} and IIS-{1563785, 1838154}), a Google Faculty Research Award, an Onassis Foundation Scholarship, a Sloan Research Fellowship, and the Simons Collaboration on Algorithms and Geometry.