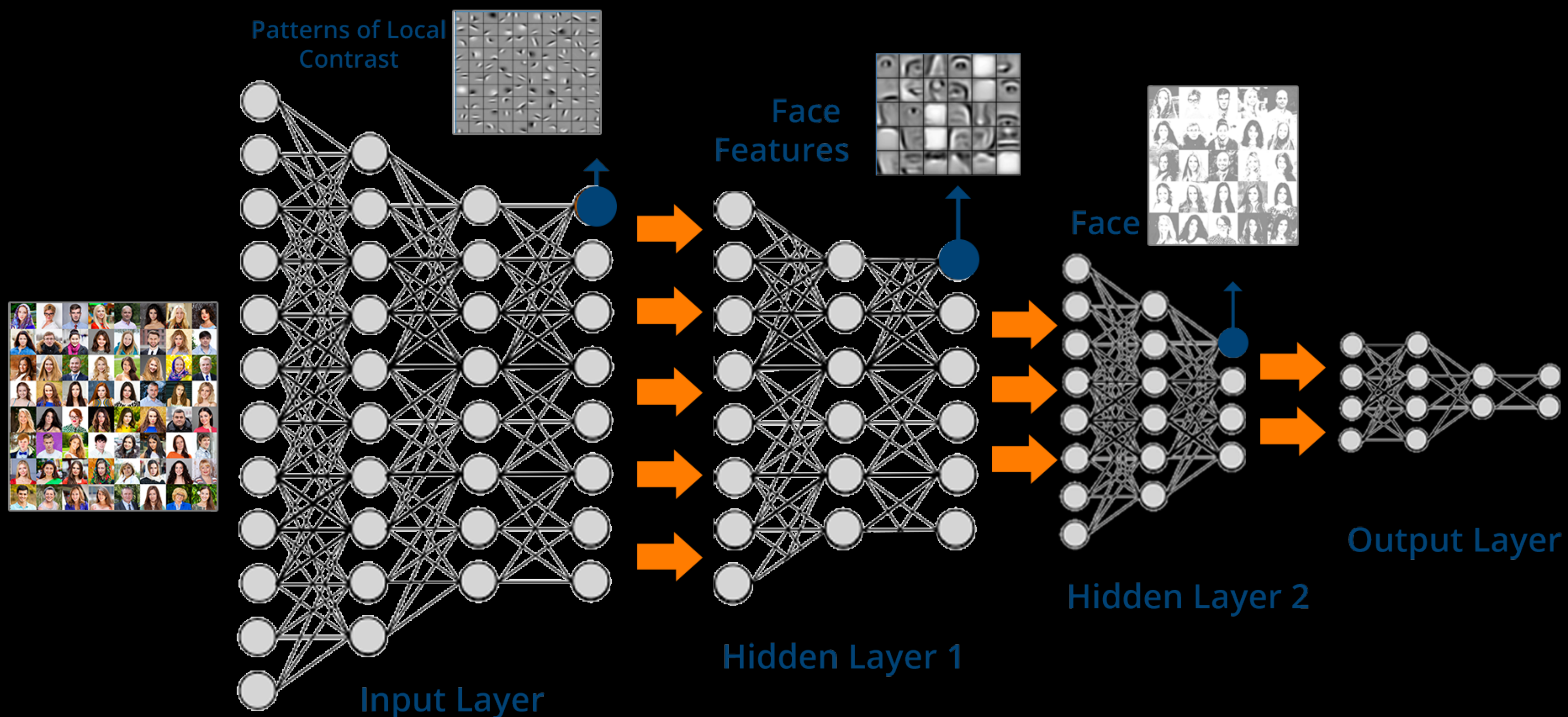
Better Depth-Width Trade-offs for Neural Networks through the lens of Dynamical Systems

Vaggos Chatziafratis
(Stanford & Google NY)

Sai Ganesh Nagarajan
(SUTD)

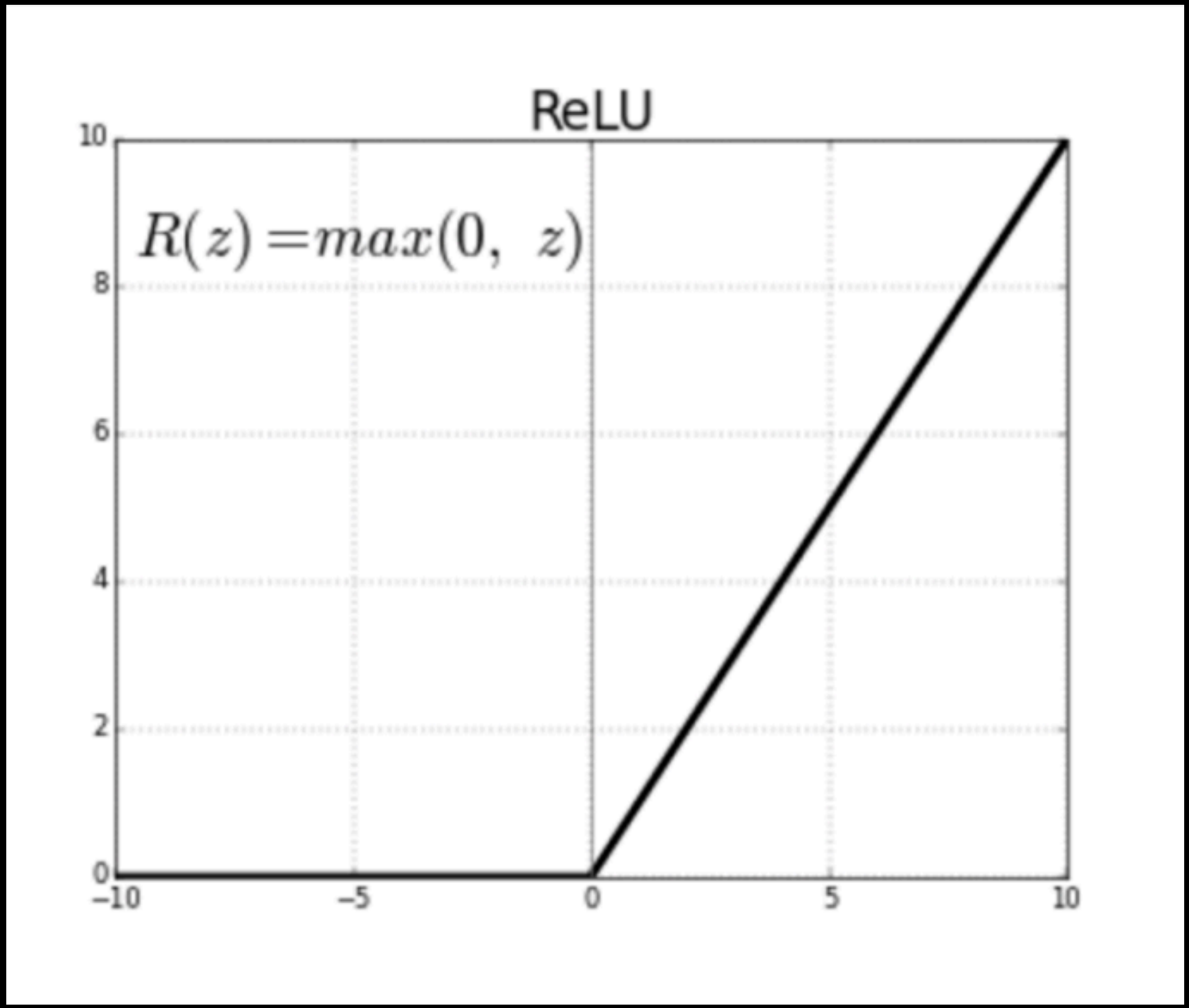
Ioannis Panageas
(SUTD => UC Irvine)

Deep Neural Networks



Are Deeper NNs more powerful?

Approximation Theory (1885-today)



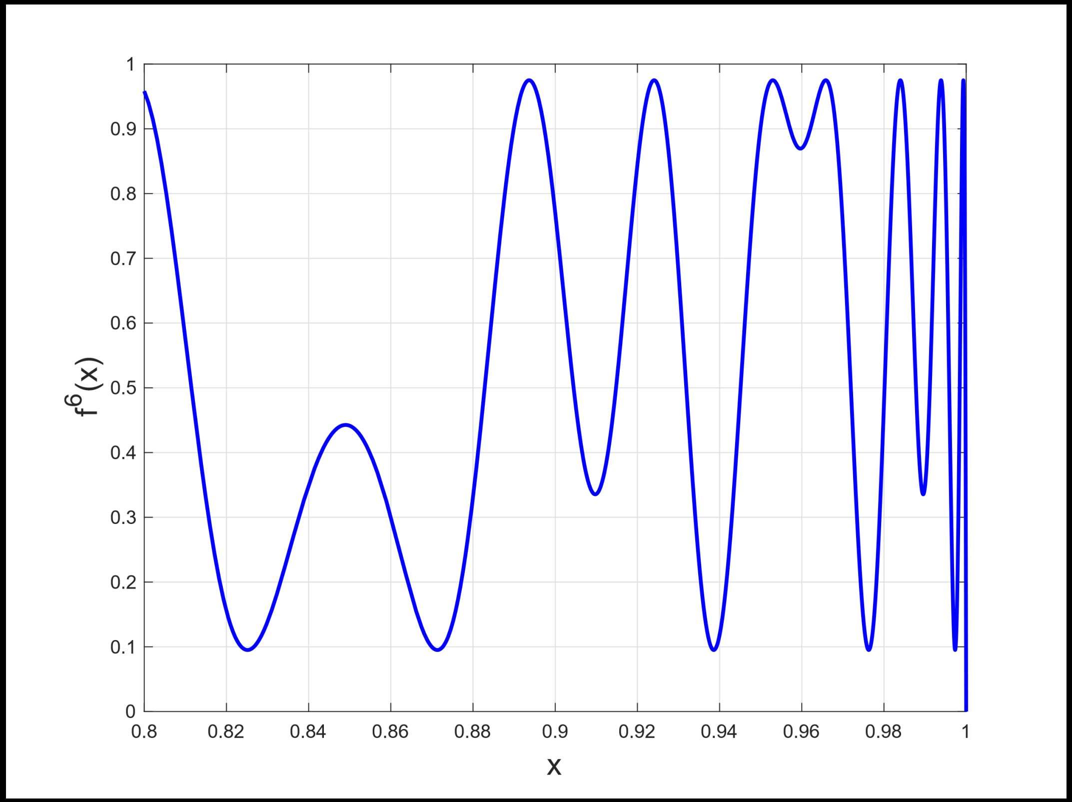
ReLU activation units

Semi-algebraic units [Telgarsky 15’,16’]: piecewise polynomials, max/min gates, and (boosted) decision trees

Expressivity of NNs

Which functions can NNs approximate?

$$\int_{[0,1]} |f(x) - g(x)| dx$$



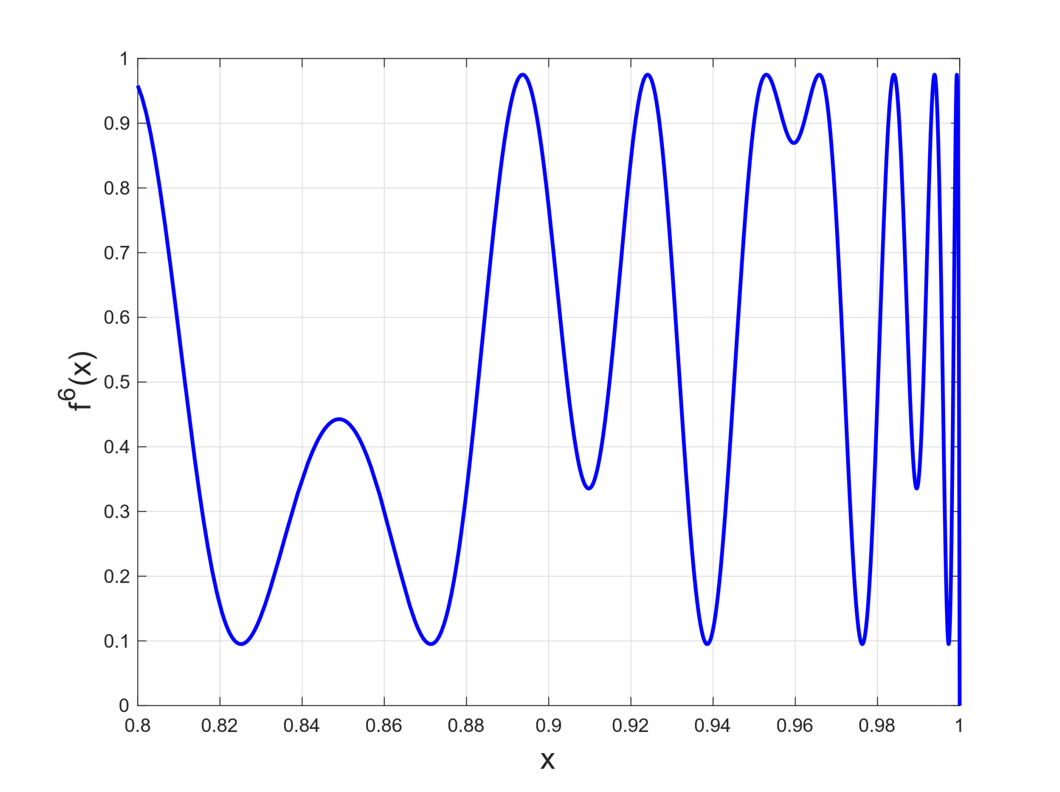
Cybenko [1989]:

Any continuous function can be represented as a (hidden) 1-layer sigmoid net (with “**some**” width).

Expressivity of NNs

Which functions can NNs approximate?

$$\int_{[0,1]} |f(x) - g(x)| dx$$



Cybenko [1989]:

Any continuous function can be represented as a (hidden) 1-layer sigmoid net

in practice: **bounded** resources!

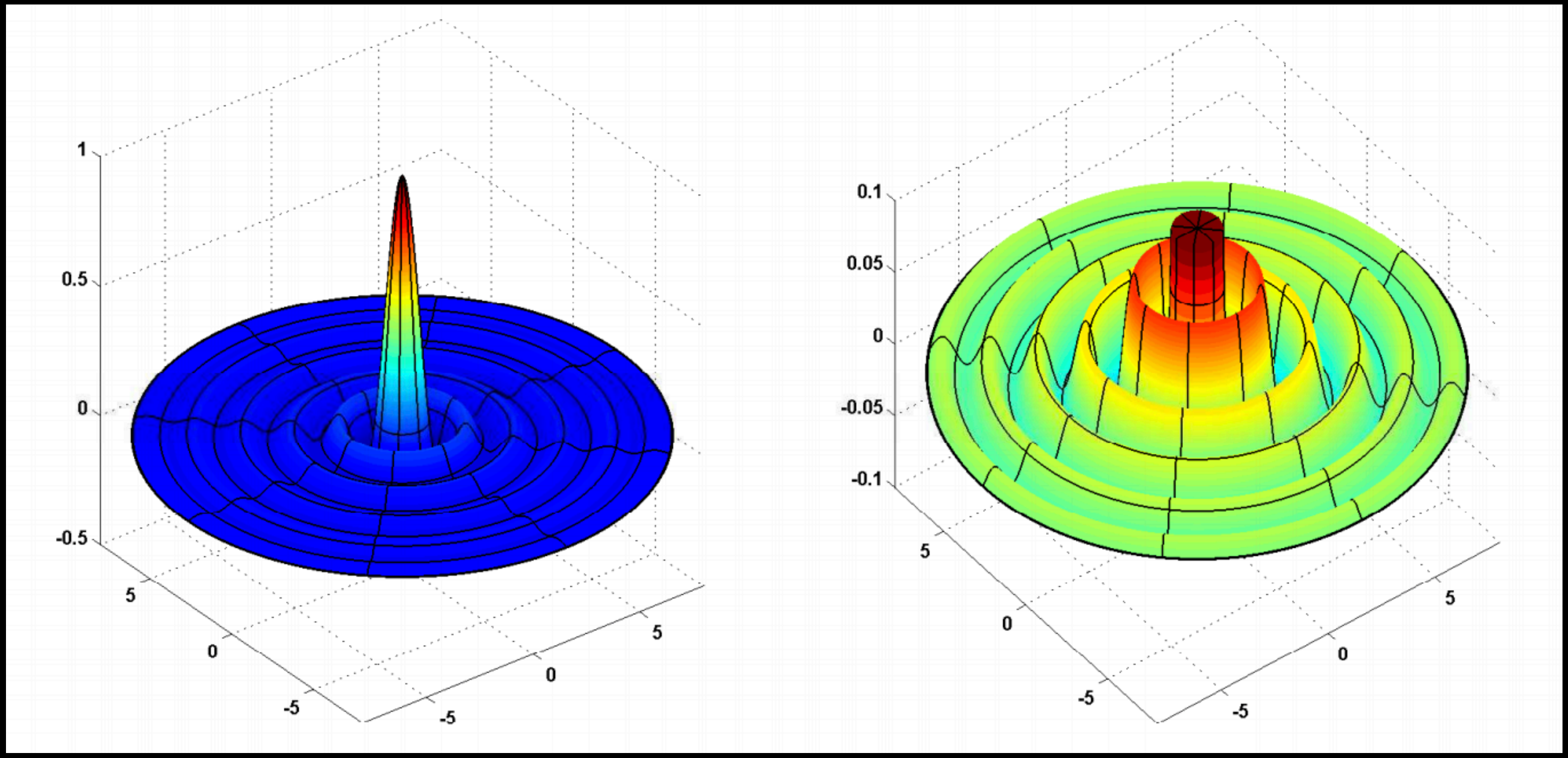
Depth Separation Results

Is there a function expressible by a deep NN
that cannot be approximated with a *much wider* shallow NN?

Yes! Challenging!

[Eldan, Shamir'16] (low depth regime)

Depth:	3	Depth:	2
Width:	poly(d)	Width:	exp(d)



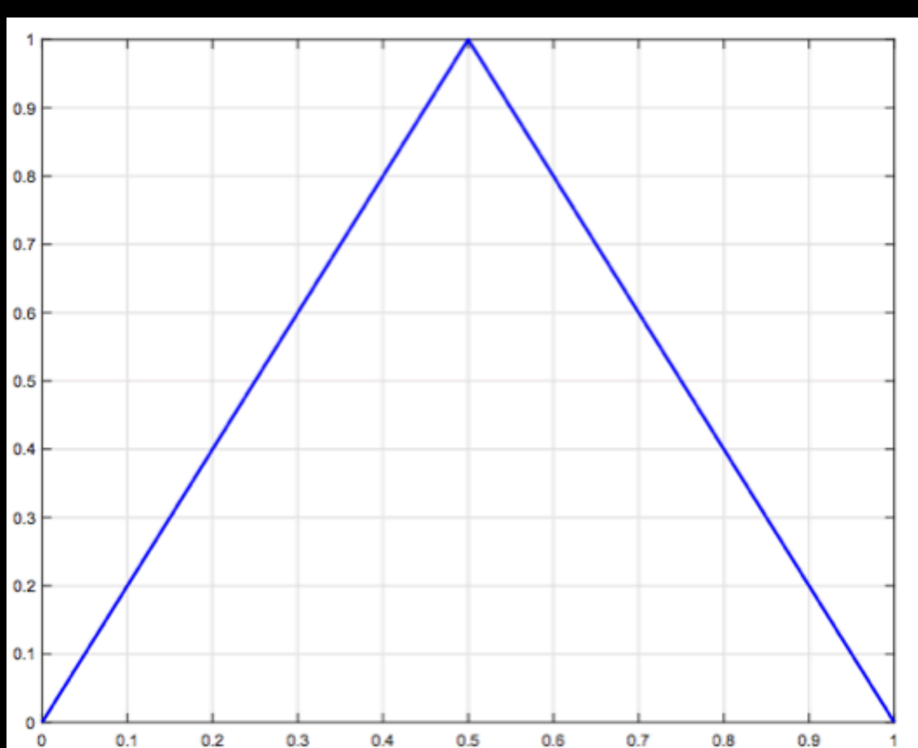
Depth Separation Results

Is there a function expressible by a deep NN
that cannot be approximated with a *much wider* shallow NN?

Yes! Challenging!

[Telgarsky'15,'16] (high depth regime)

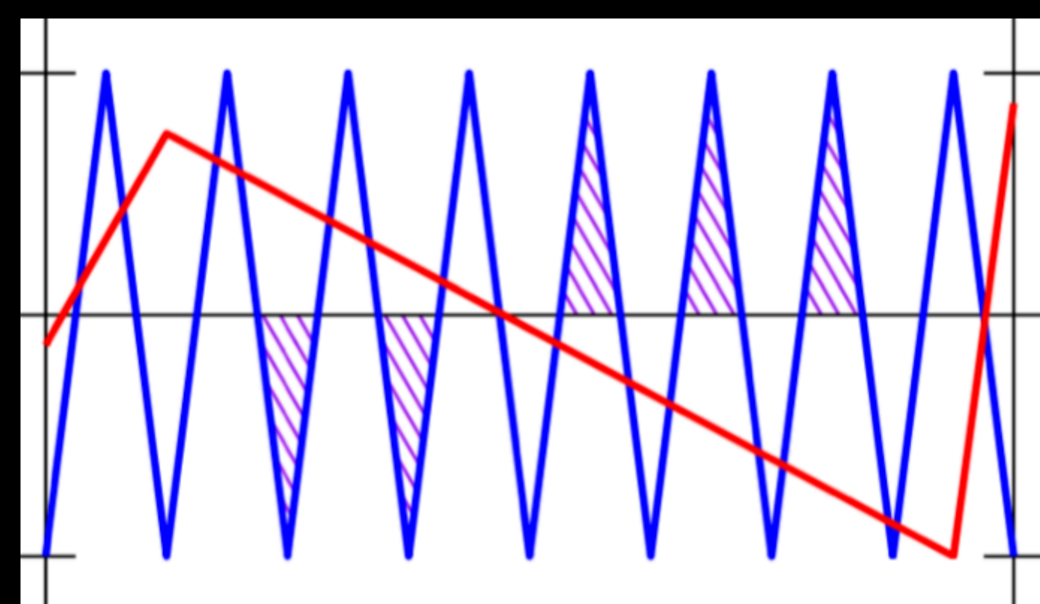
Depth:	$2L$	✓	Depth:	$\sqrt{L}$	✗
Width:	2		Width:	$2^{\sqrt{L}}$	



Tent or Triangle map

L=100

400 vs 10000 ReLUs

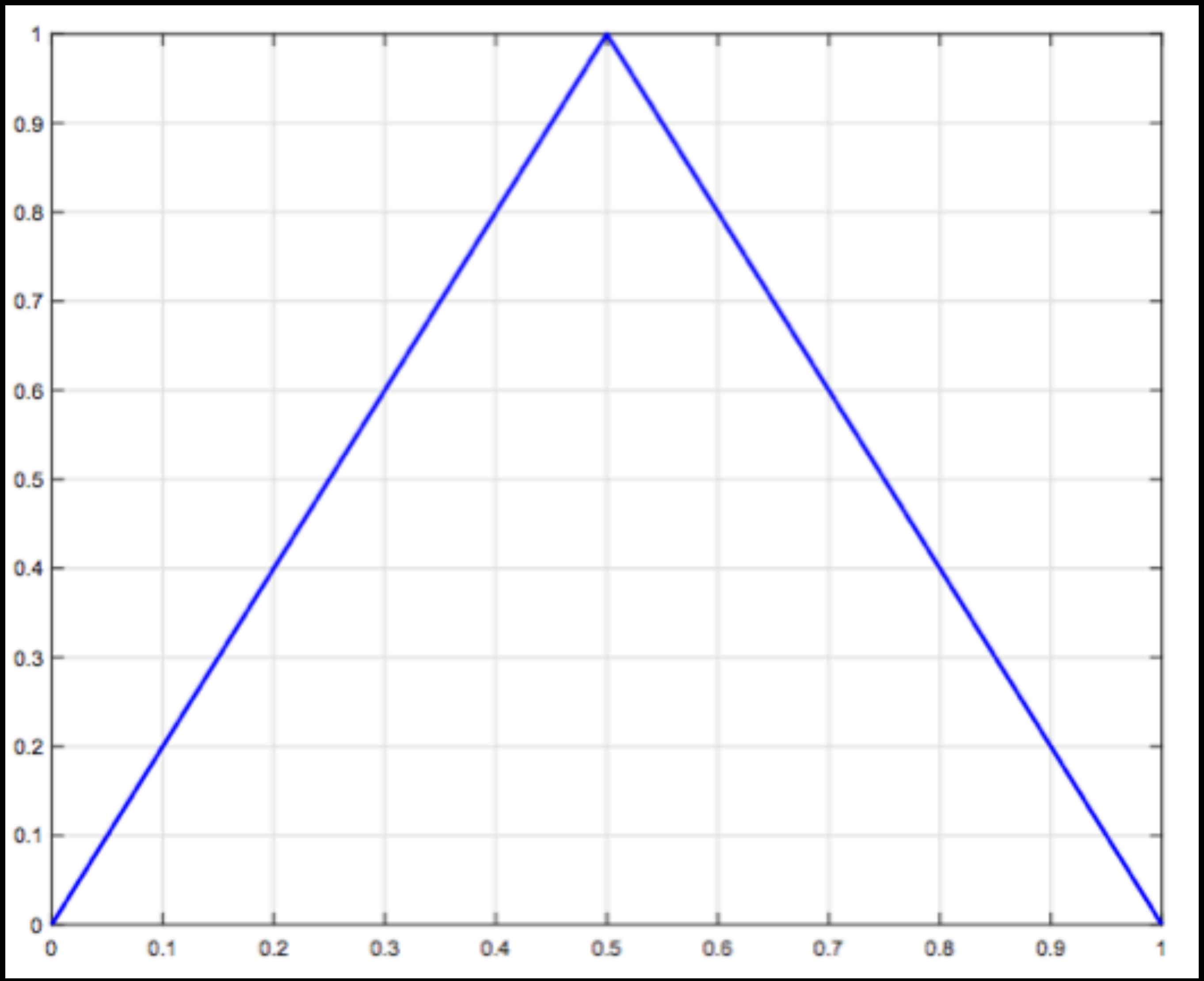


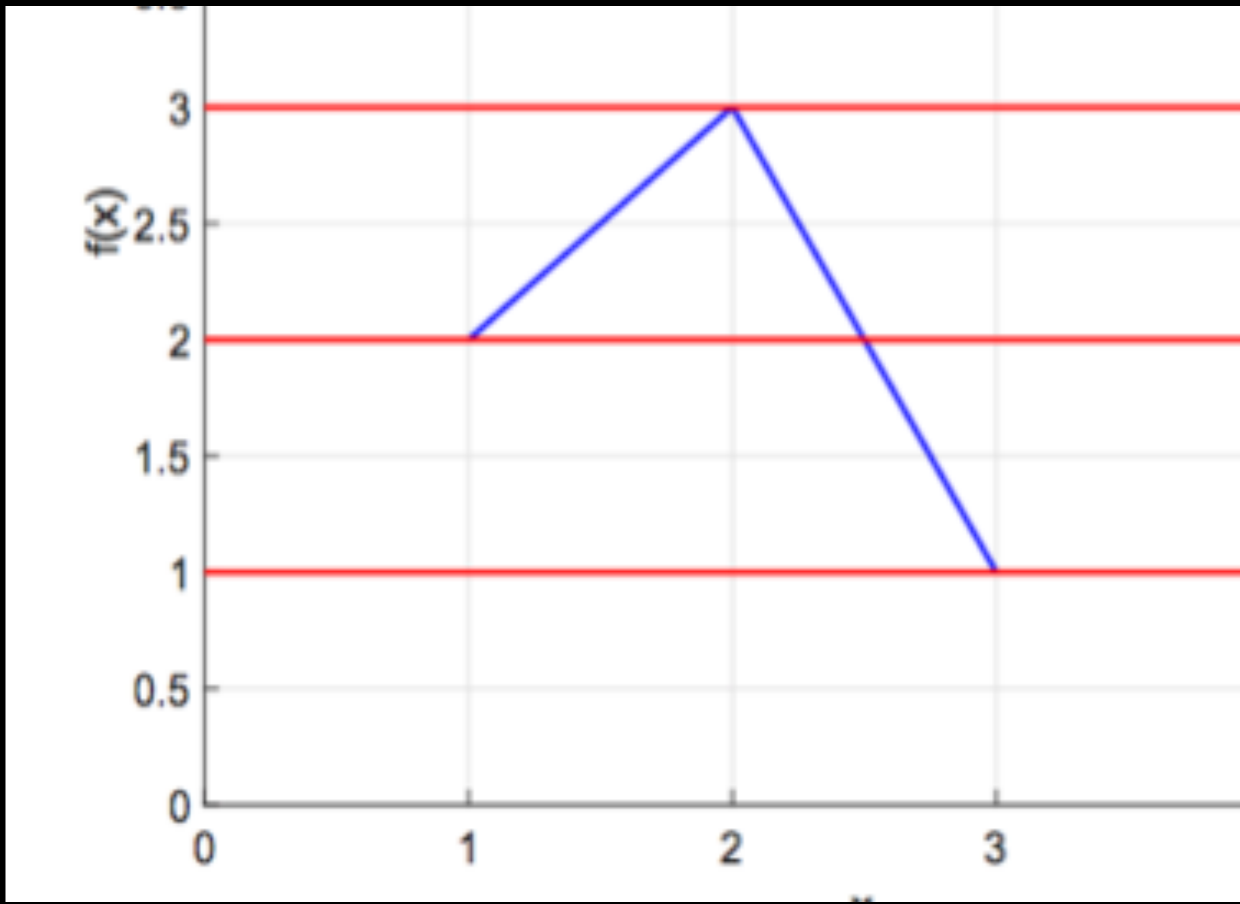
Prior Work

[Telgarsky'15,'16] **Tantalizing open question:**

1. Can we understand **larger** families of functions?
2. Why is the tent map suitable to prove depth separations?

(what if we slightly tweak the tent map?)

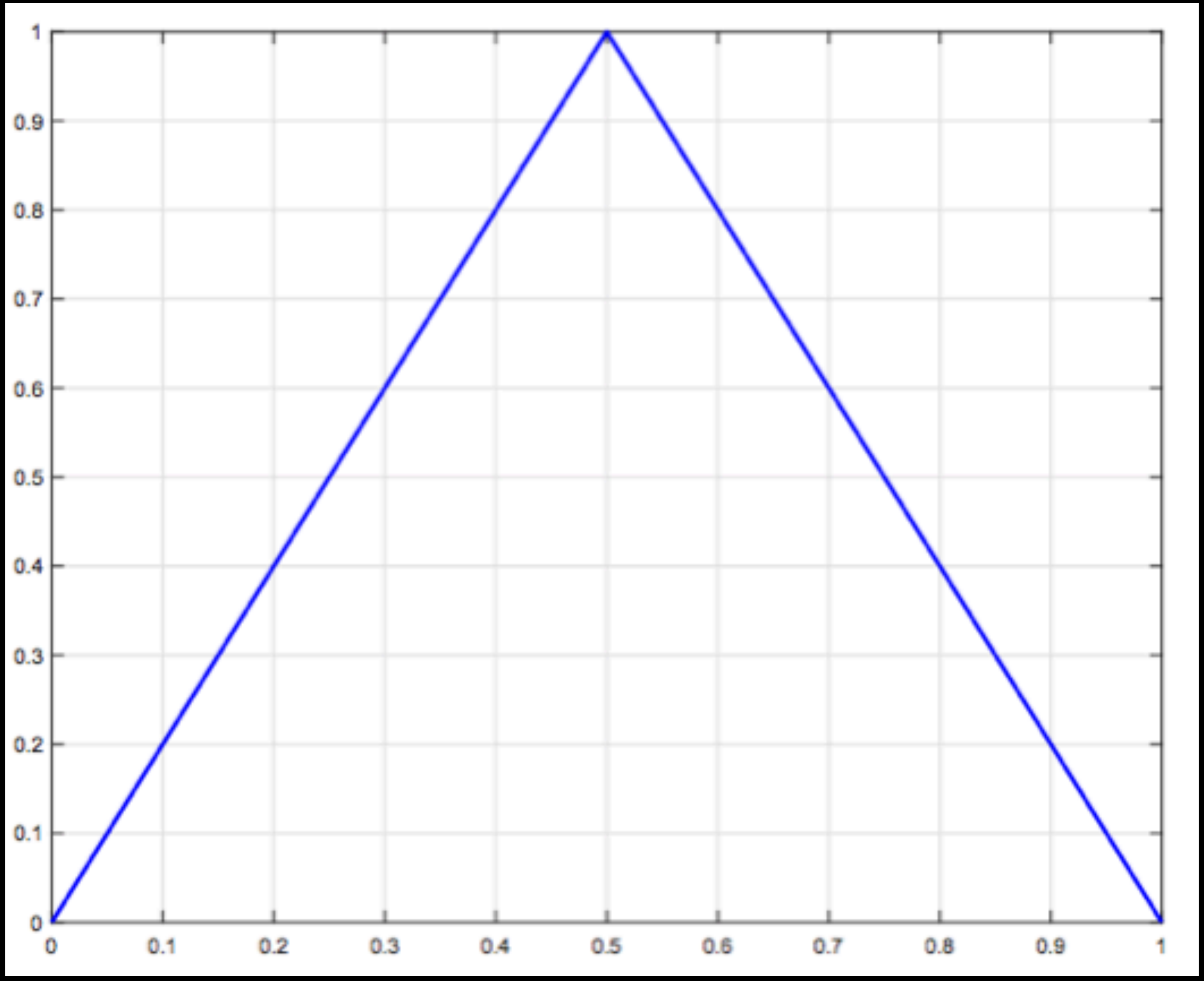




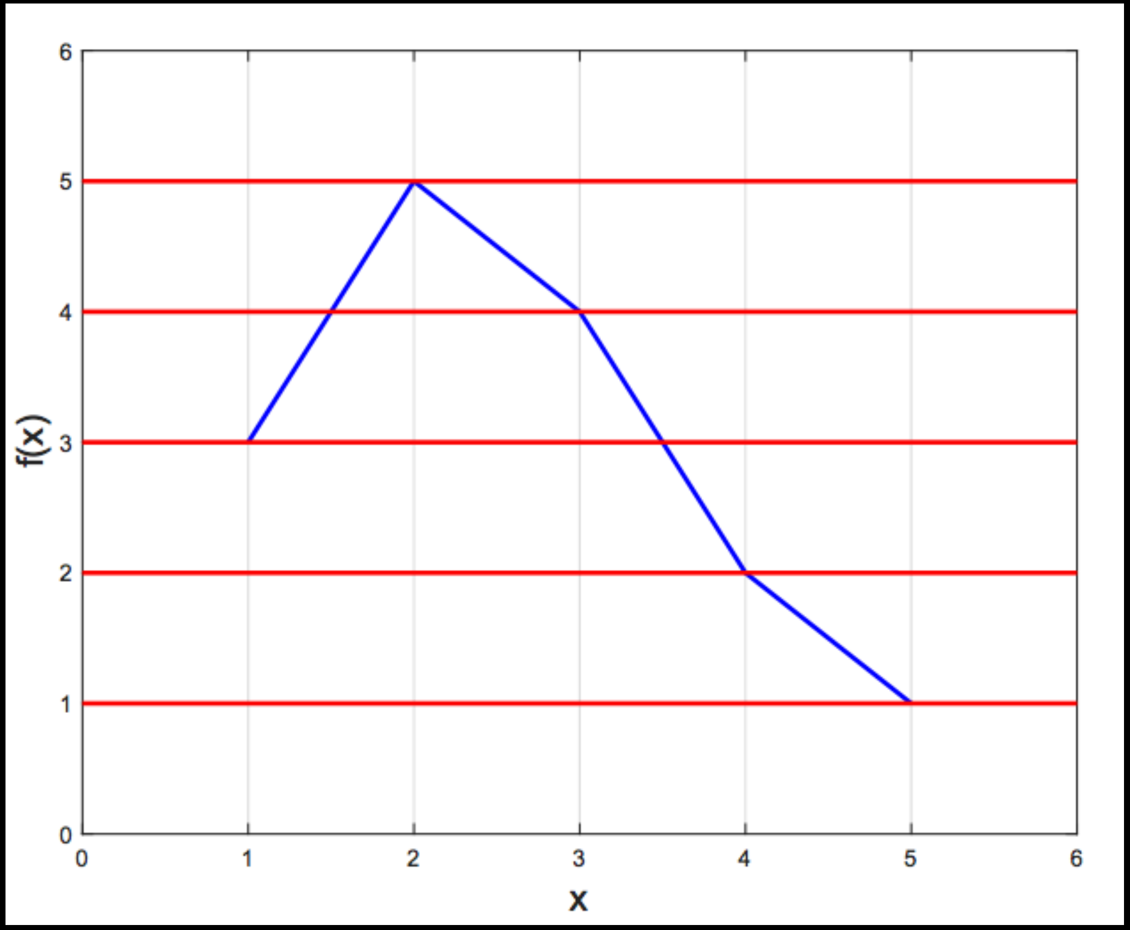
Prior Work

[Telgarsky'15,'16] **Tantalizing open question:**

1. Can we understand **larger** families of functions?
2. Why is the tent map suitable to prove depth separations?



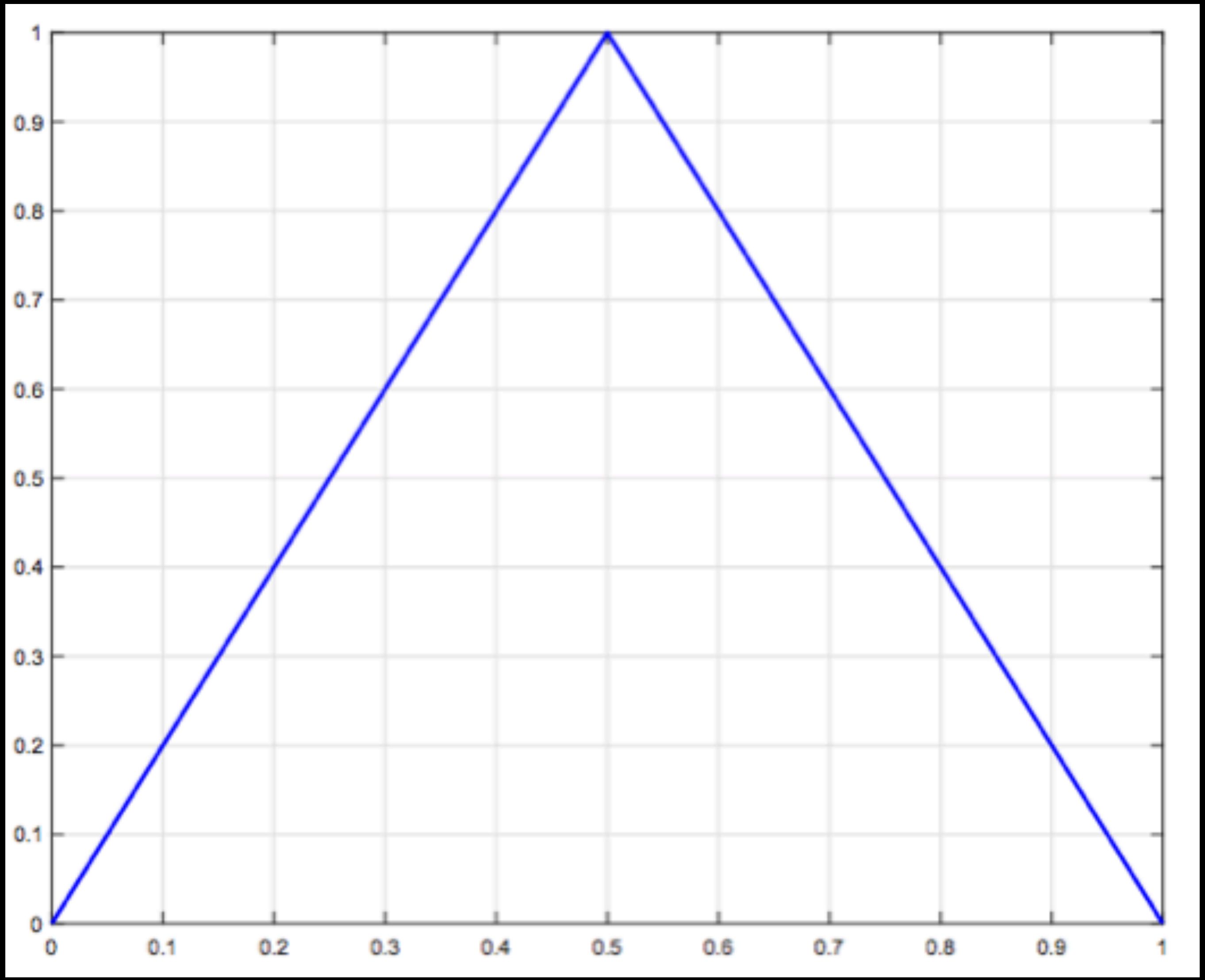
(what if we slightly tweak the tent map?)



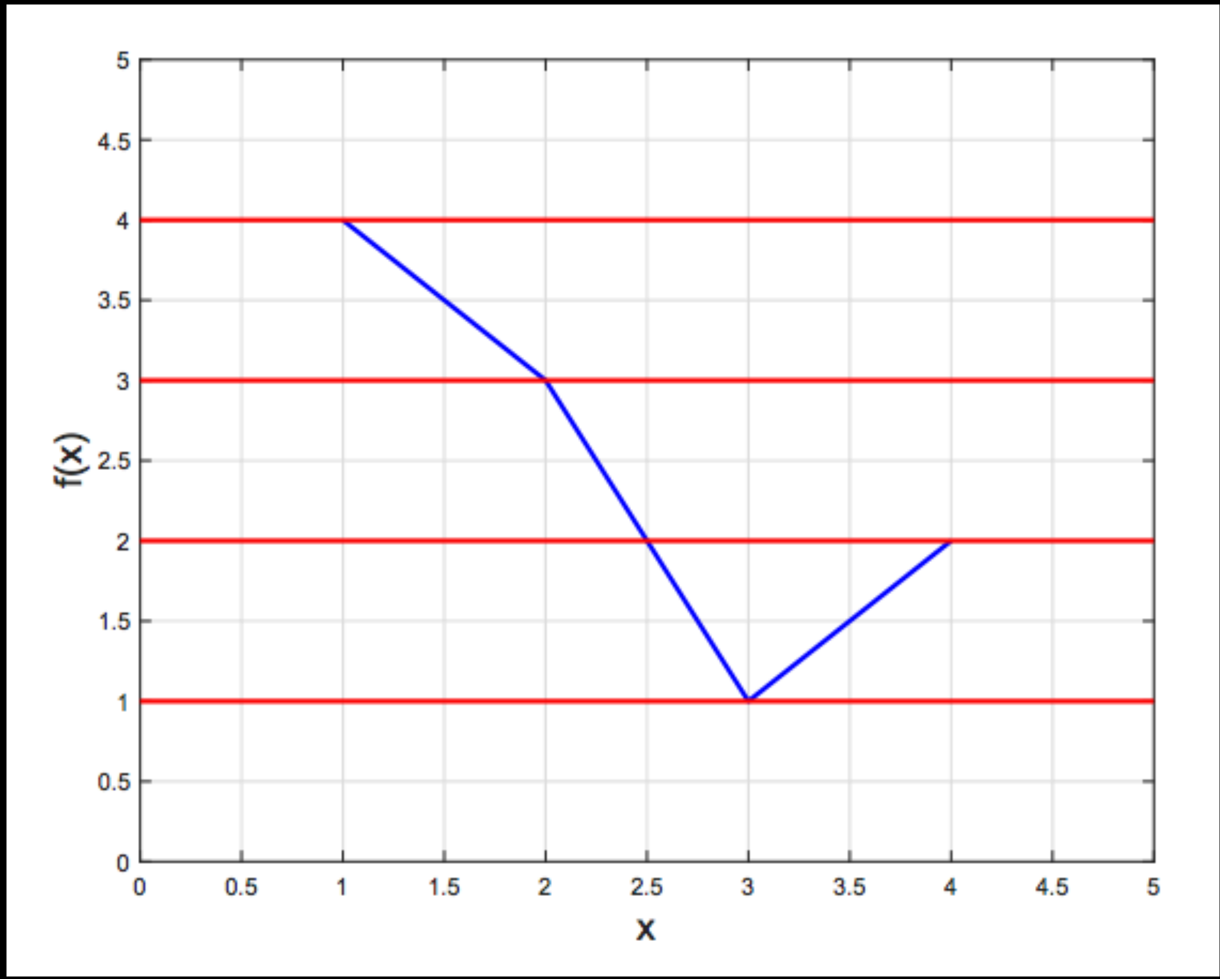
Prior Work

[Telgarsky'15,'16] **Tantalizing open question:**

1. Can we understand **larger** families of functions?
2. Why is the tent map suitable to prove depth separations?



(what if we slightly tweak the tent map?)

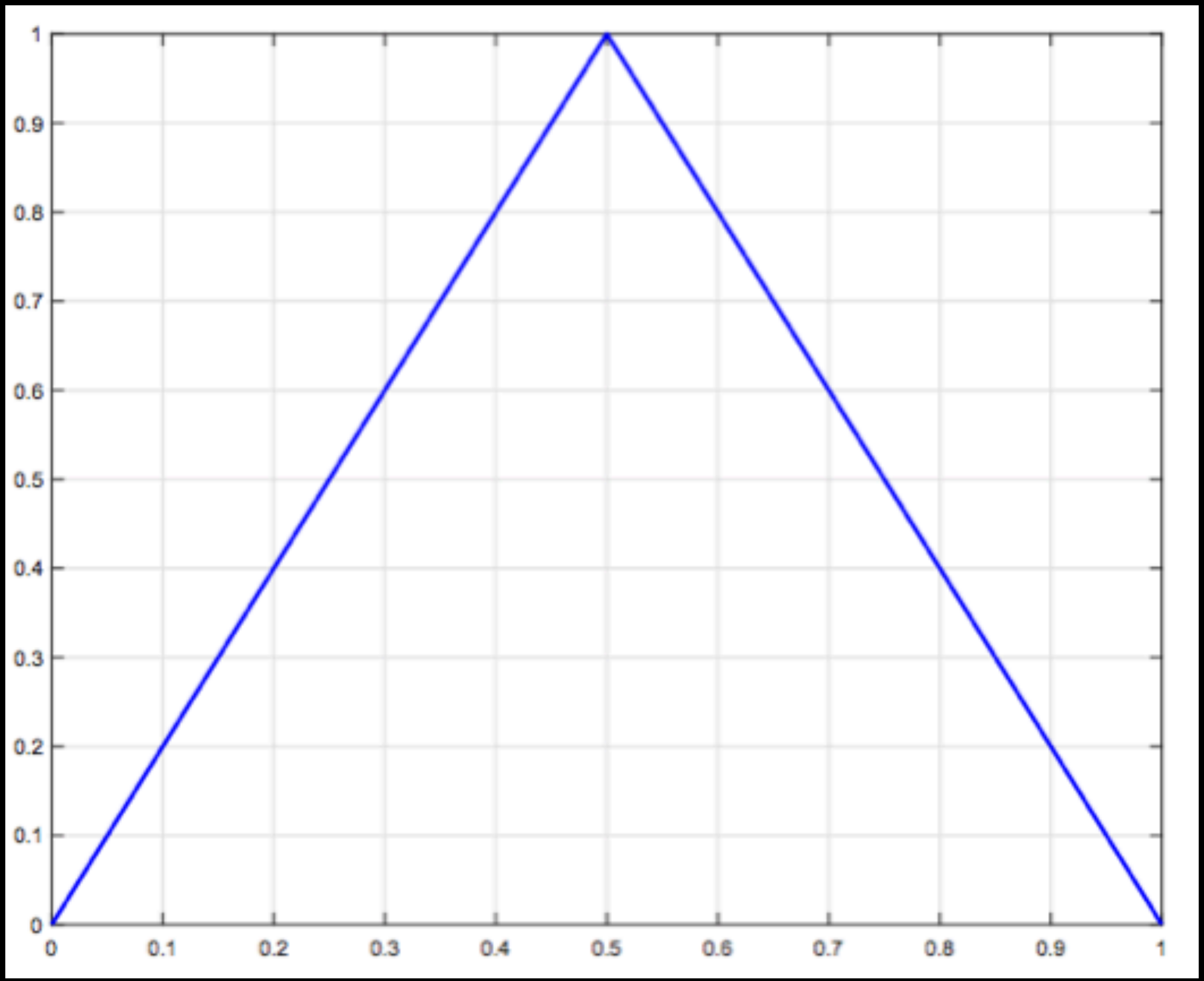


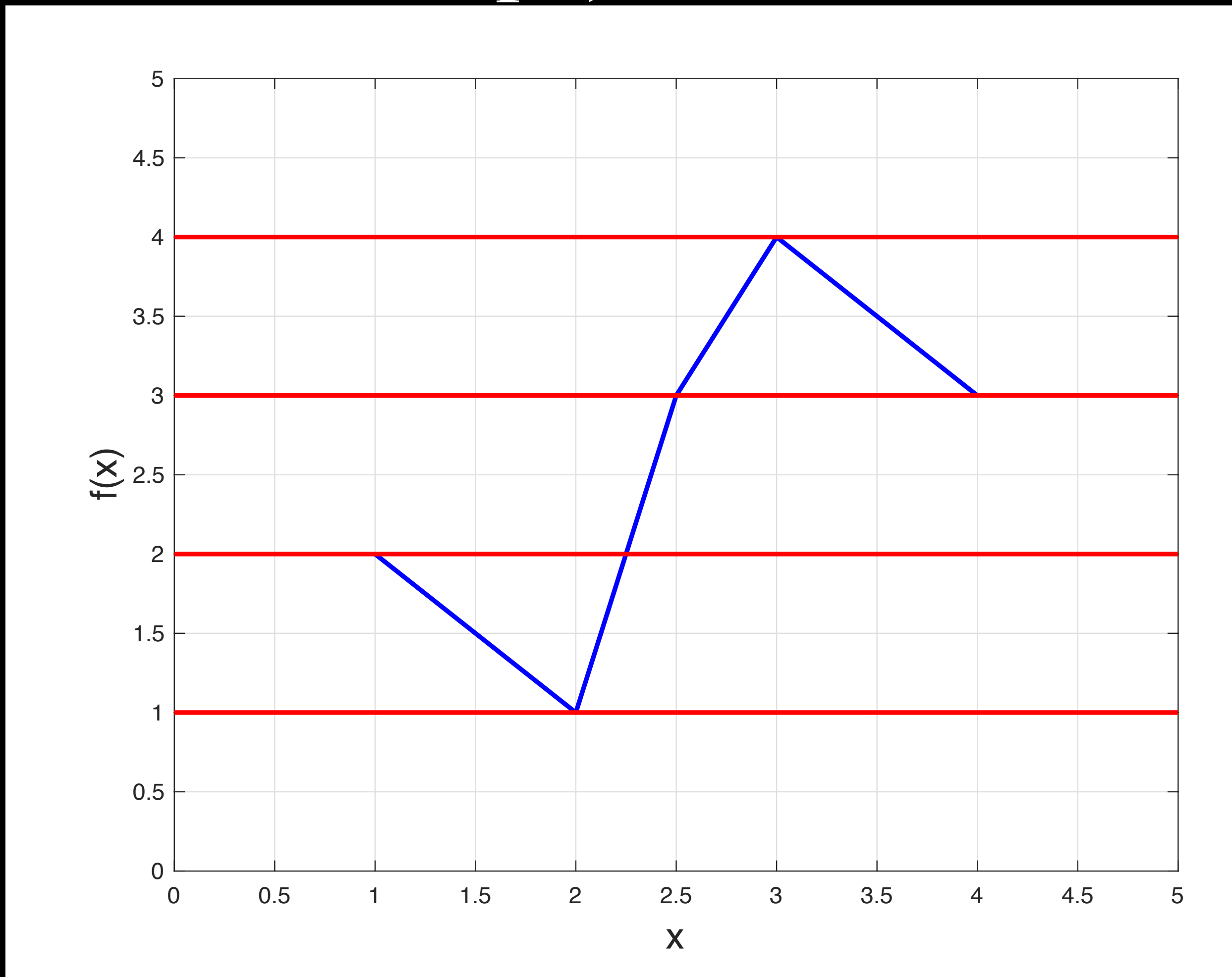
Prior Work

[Telgarsky'15,'16] **Tantalizing open question:**

1. Can we understand **larger** families of functions?
2. Why is the tent map suitable to prove depth separations?

(what if we slightly tweak the tent map?)

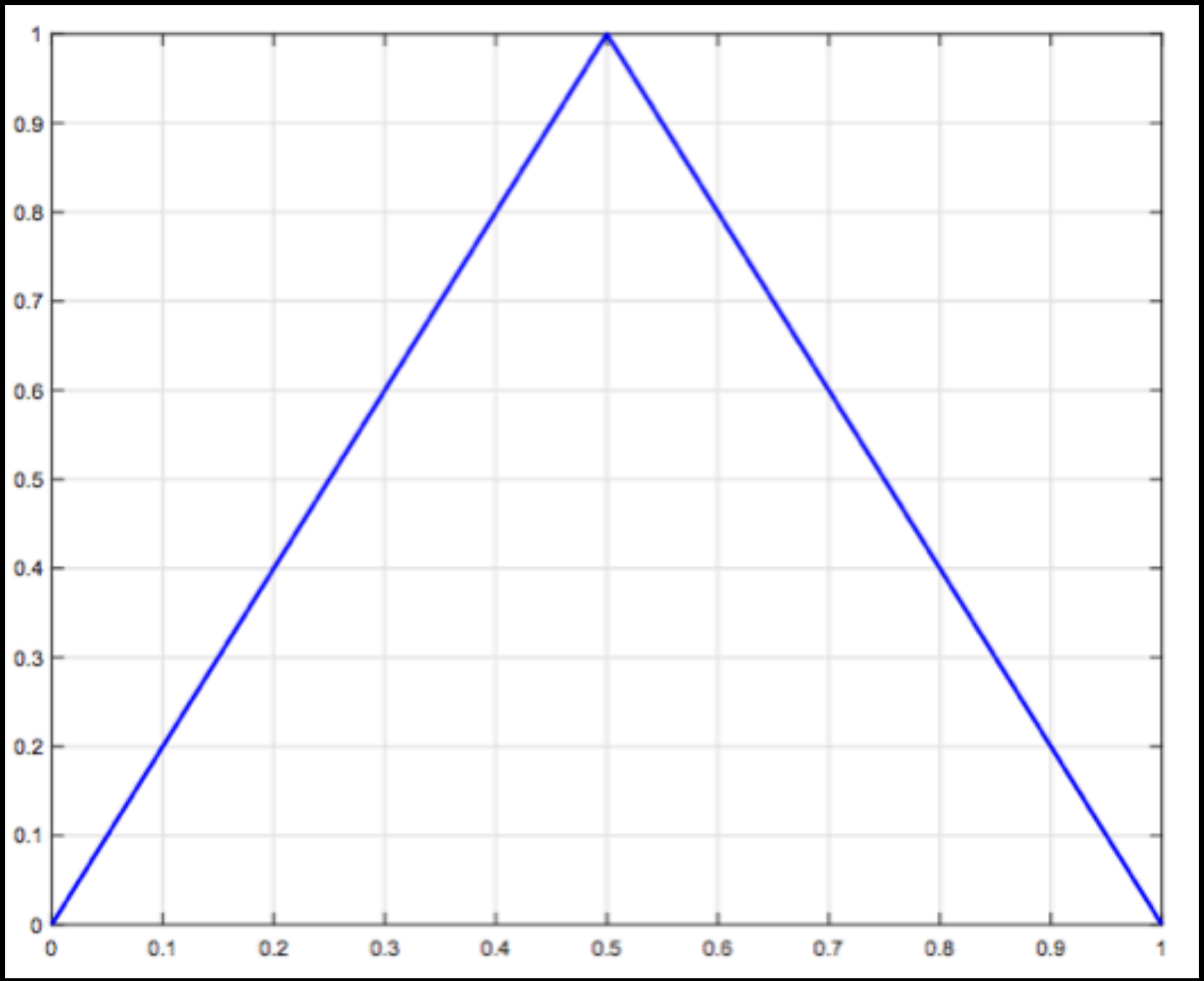




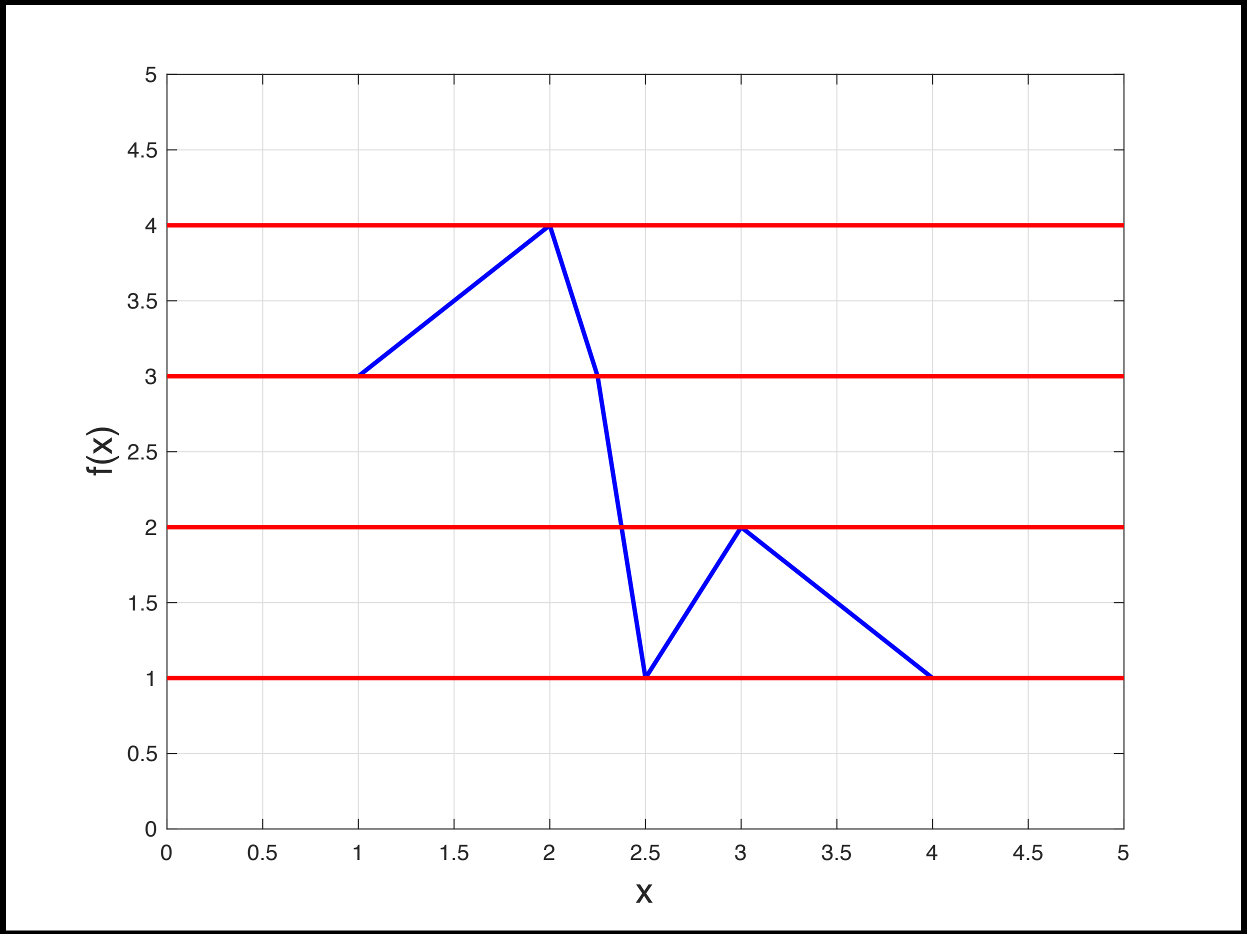
Prior Work

[Telgarsky'15,'16] **Tantalizing open question:**

1. Can we understand **larger** families of functions?
2. Why is the tent map suitable to prove depth separations?



(what if we slightly tweak the tent map?)



Our work in ICML 2020

→ **Connections to Dynamical Systems [ICLR'20]:**

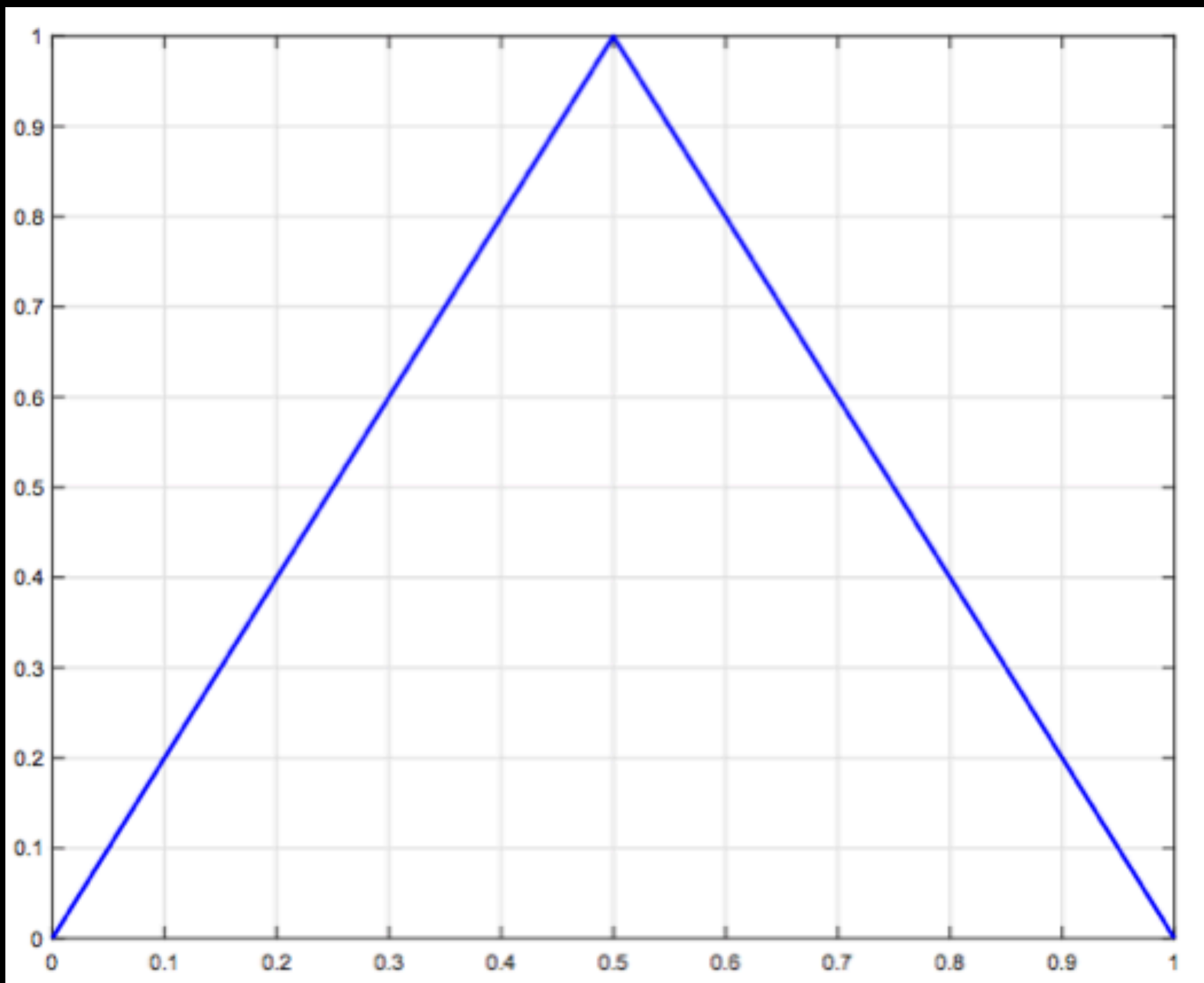
1. We get L1-approximation error and not just classification error.

2. We show tight connections between Lipschitz constant, periods of f, and oscillations.

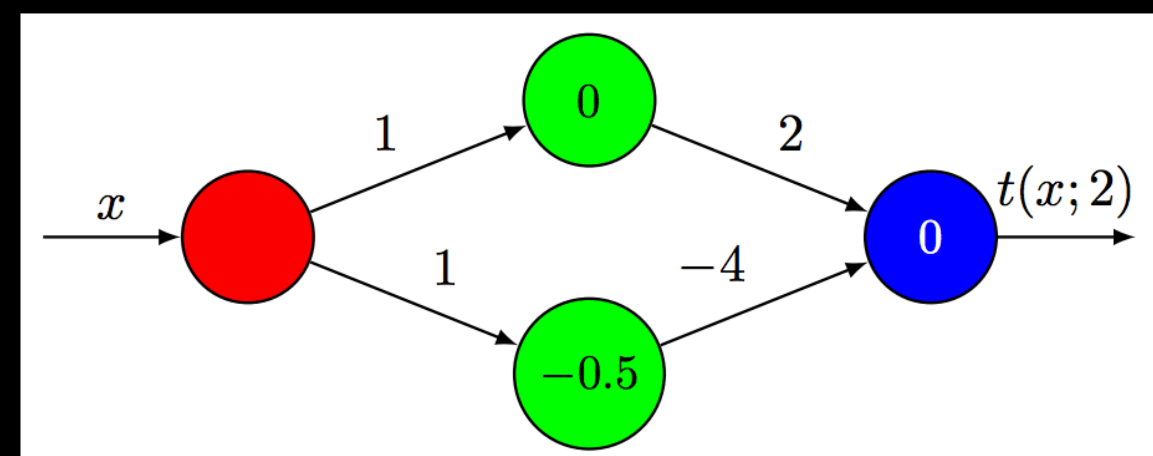
3. Sharper period-dependent depth-width tradeoffs and easy constructions of examples.

4. Experimental validation of our theoretical results.

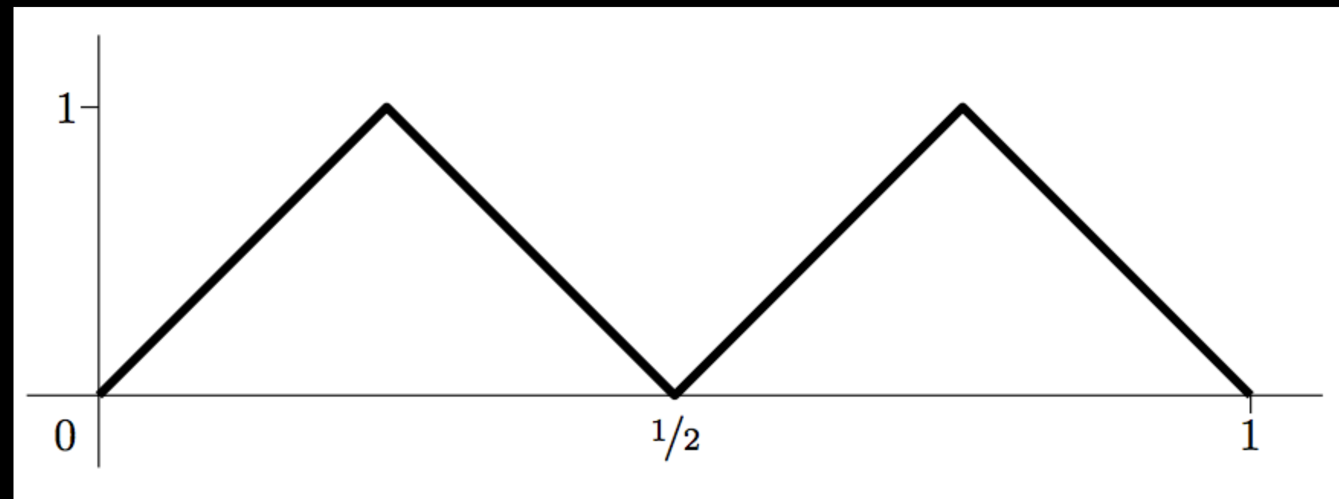
Tent Map (by Telgarsky)

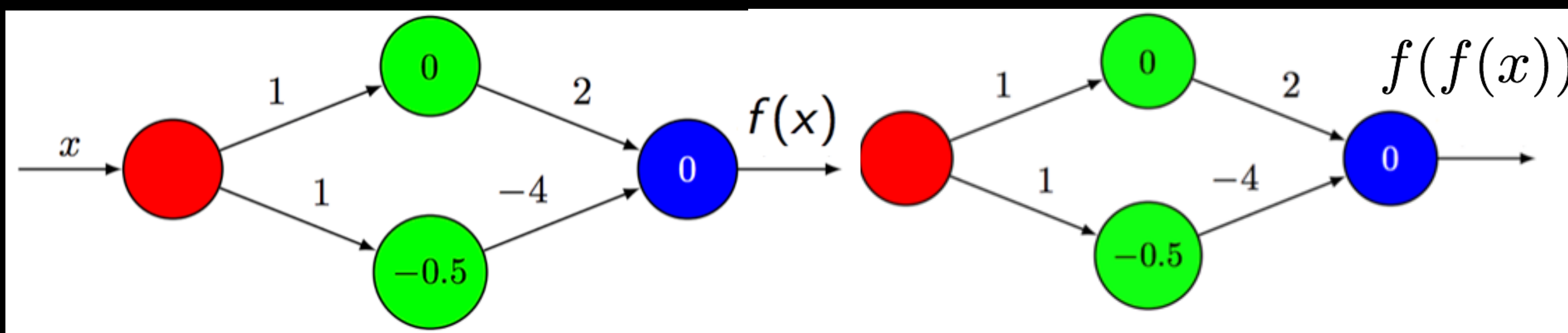


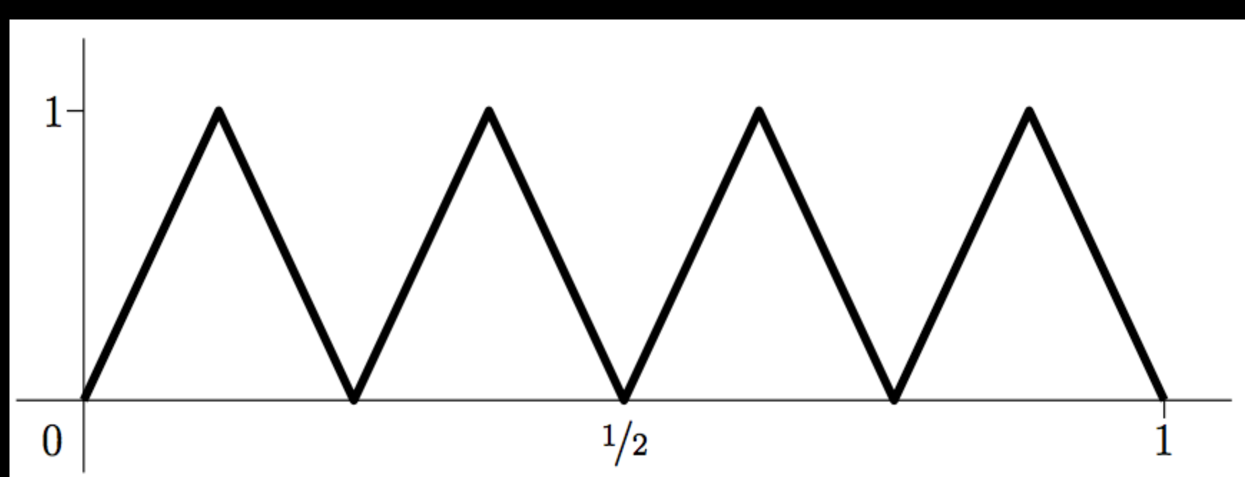
$$f(x) = \begin{cases} 2x, & 0 \leq x \leq 1/2 \\ -2x + 2, & 1/2 \leq x \leq 1 \end{cases}$$



Repeated Compositions

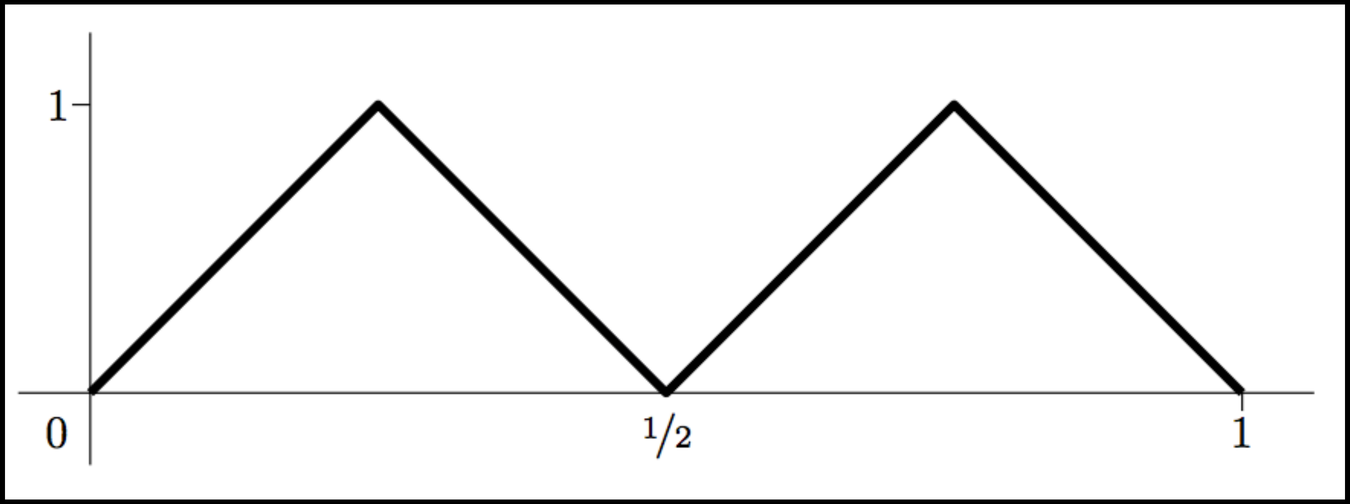


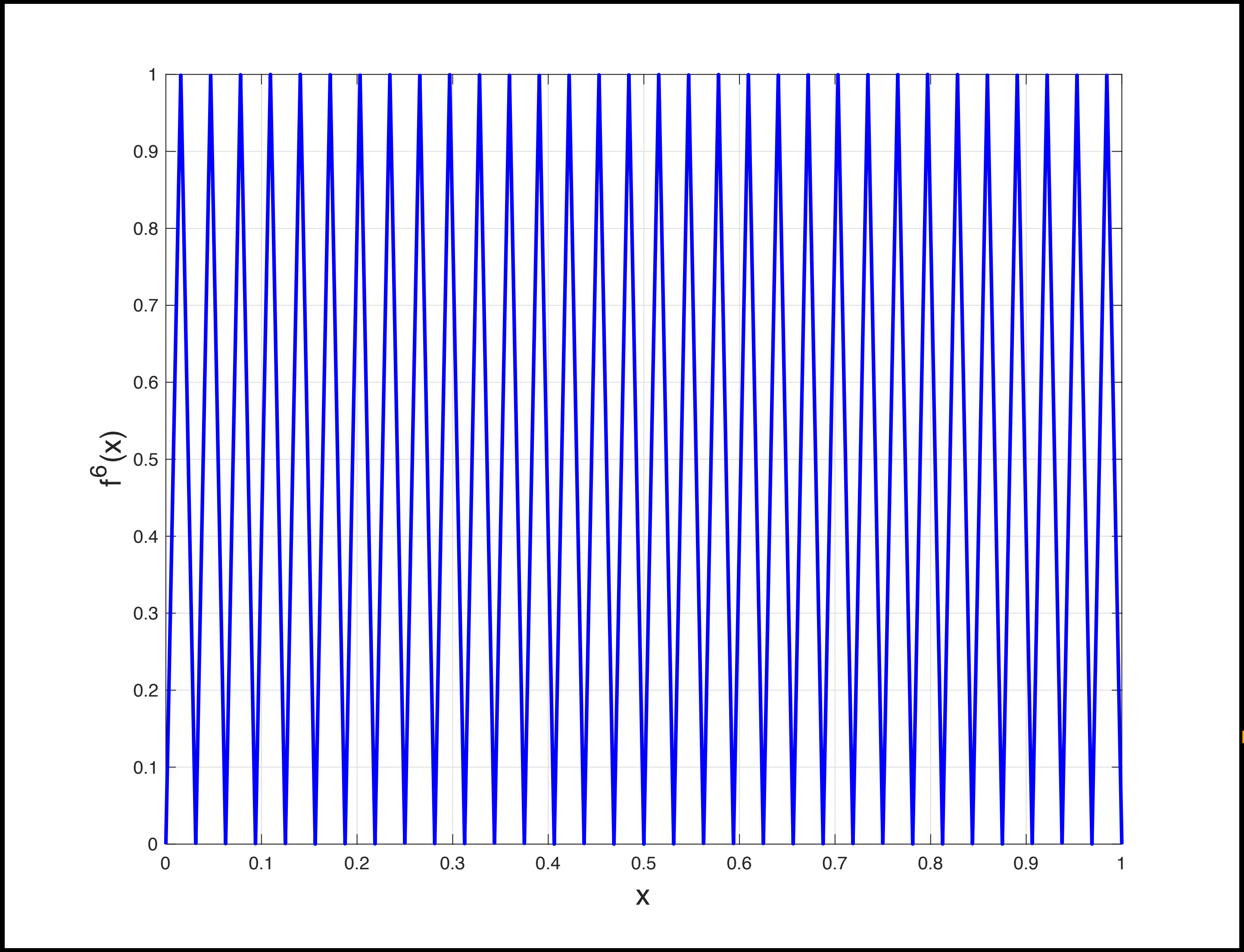




exponentially many bumps
2^t

Repeated Compositions

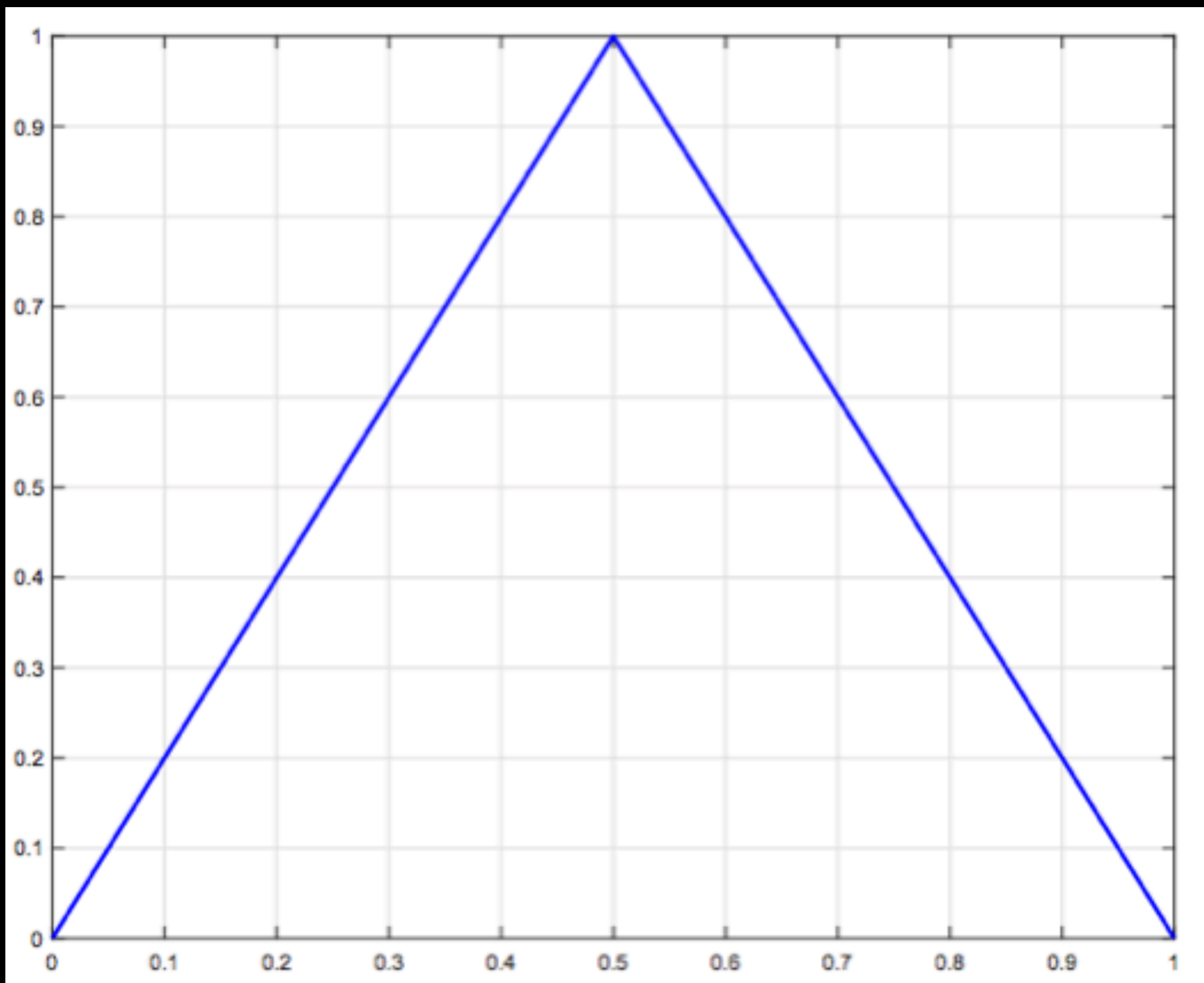




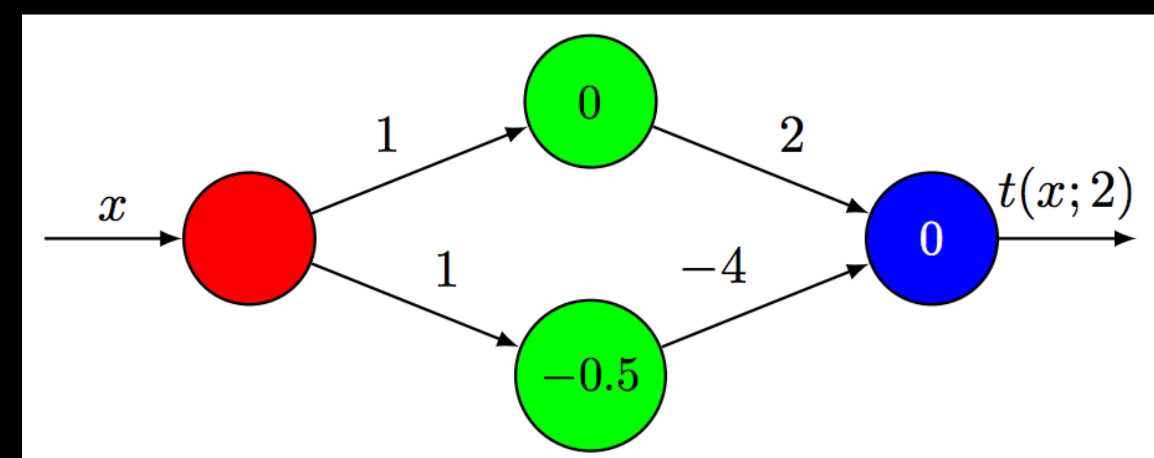
ReLU NN:
#linearRegions:
$(2w)^L \geq 2^t$

exponentially many bumps
2^t

Our starting observation: Period 3



$$f(x) = \begin{cases} 2x, & 0 \leq x \leq 1/2 \\ -2x + 2, & 1/2 \leq x \leq 1 \end{cases}$$



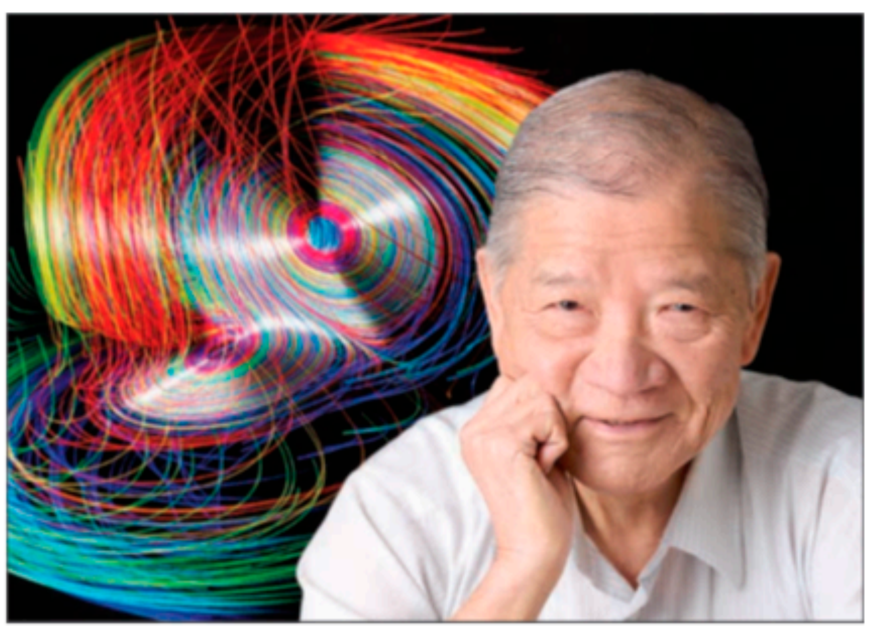
$$\frac{2}{9} \xrightarrow{f} \frac{4}{9} \xrightarrow{f} \frac{8}{9} \xrightarrow{f} \frac{2}{9}$$

Li-Yorke Chaos (1975)

PERIOD THREE IMPLIES CHAOS

TIEN-YIEN LI AND JAMES A. YORKE

1. Introduction. The way phenomena or processes evolve or change in time is often described by differential equations or difference equations. One of the simplest mathematical situations occurs when the phenomenon can be described by a single number as, for example, when the number of



Sharkovsky's Theorem (1964)

$$3 \rhd 5 \rhd 7 \rhd \cdots \rhd$$

$$\rhd 2 \cdot 3 \rhd 2 \cdot 5 \rhd 2 \cdot 7 \rhd \cdots \rhd$$

$$\rhd 2^2 \cdot 3 \rhd 2^2 \cdot 5 \rhd 2^2 \cdot 7 \rhd \cdots \rhd$$

$$\rhd \cdots \rhd 2^4 \rhd 2^3 \rhd 2^2 \rhd 2 \rhd 1$$

Sharkovsky's Theorem (1964)

$$3 \rhd 5 \rhd 7 \rhd \cdots \rhd$$

$$\rhd 2 \cdot 3 \rhd 2 \cdot 5 \rhd 2 \cdot 7 \rhd \cdots \rhd$$

$$\rhd 2^2 \cdot 3 \rhd 2^2 \cdot 5 \rhd 2^2 \cdot 7 \rhd \cdots \rhd$$

$$\rhd \cdots \rhd 2^4 \rhd 2^3 \rhd 2^2 \rhd 2 \rhd 1$$

Period-dependent Trade-offs **[ICLR 2020]**

Main Lemma:

Let f be continuous with odd period $p \geq 3$.

Then f^t oscillates at least c^t times,

where $c > 1$ and is the largest root of $x^{p-1} - x^{p-2} - 1 = 0$.

Period-dependent Trade-offs **[ICLR 2020]**

Main Lemma:

Let f be continuous with odd period $p \geq 3$.

Then f^t oscillates at least c^t times,

where $c > 1$ and is the largest root of $x^{p-1} - x^{p-2} - 1 = 0$.

Informal Main Result:

Using periodic functions f, we construct f^t,

that has c^t oscillations and is the output of a depth t

width 2 neural net, for which any shallow net

(l layers, u width per layer) with $u \leq c^{t/l}/8$

incurs high classification error ($\geq \frac{1}{4}$).

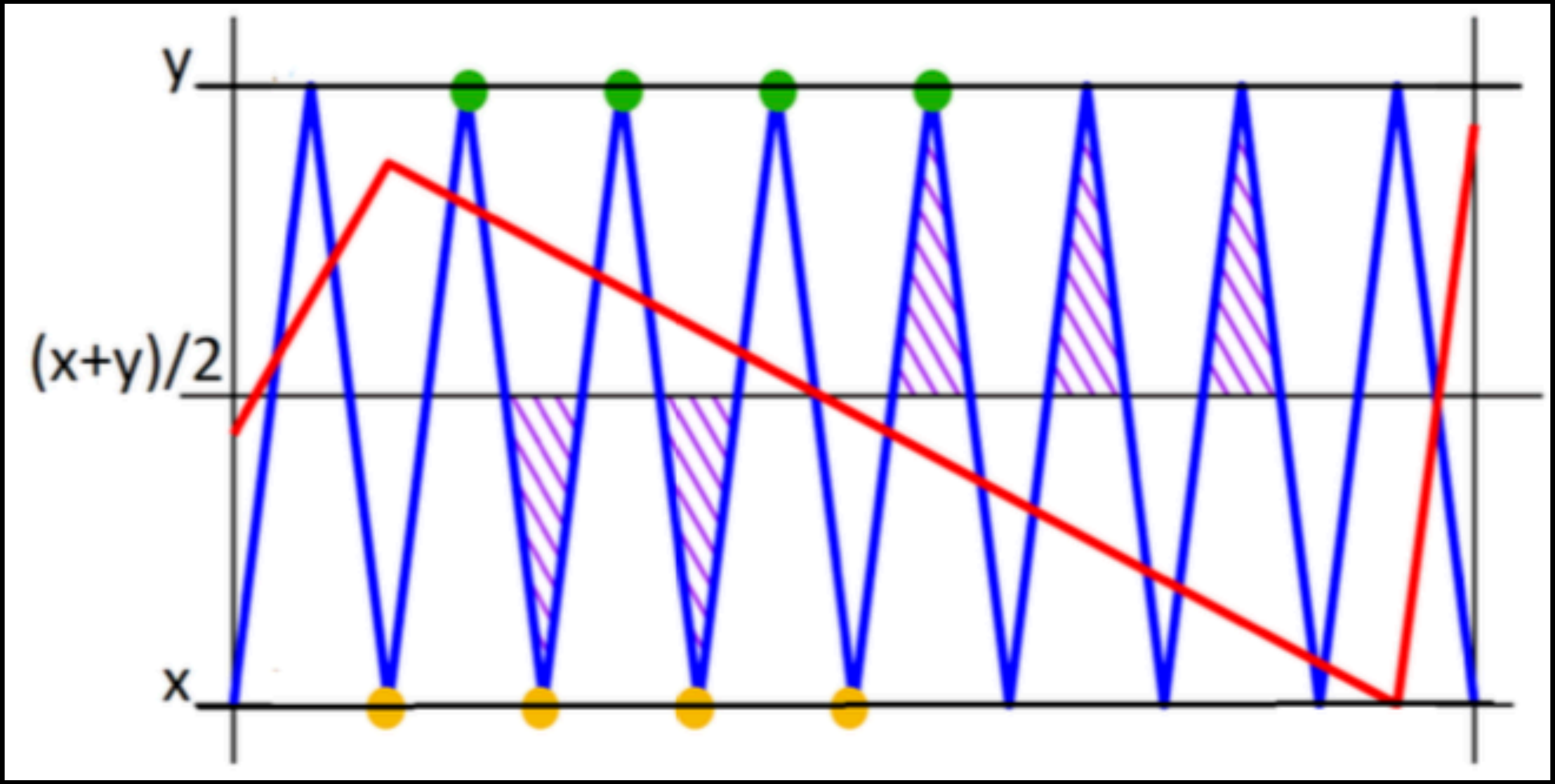
Period-dependent Trade-offs **[ICLR 2020]**

Main Lemma:

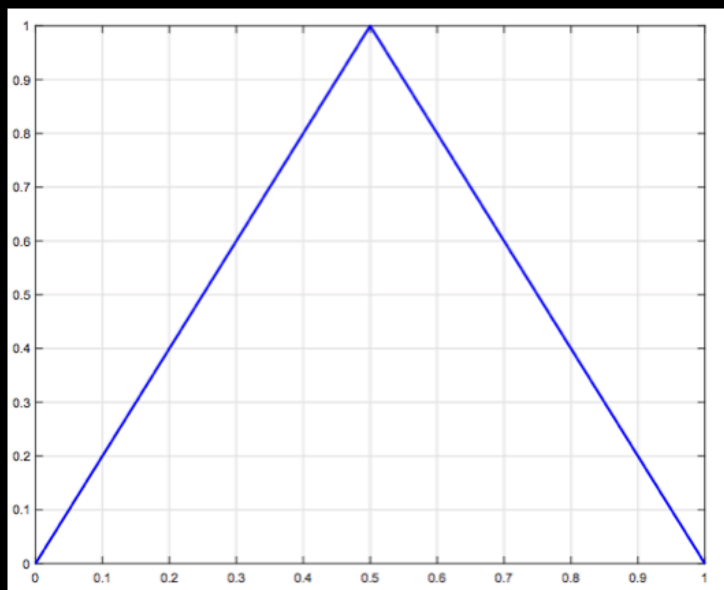
Let f be continuous with odd period $p \geq 3$.

Then f^t oscillates at least c^t times,

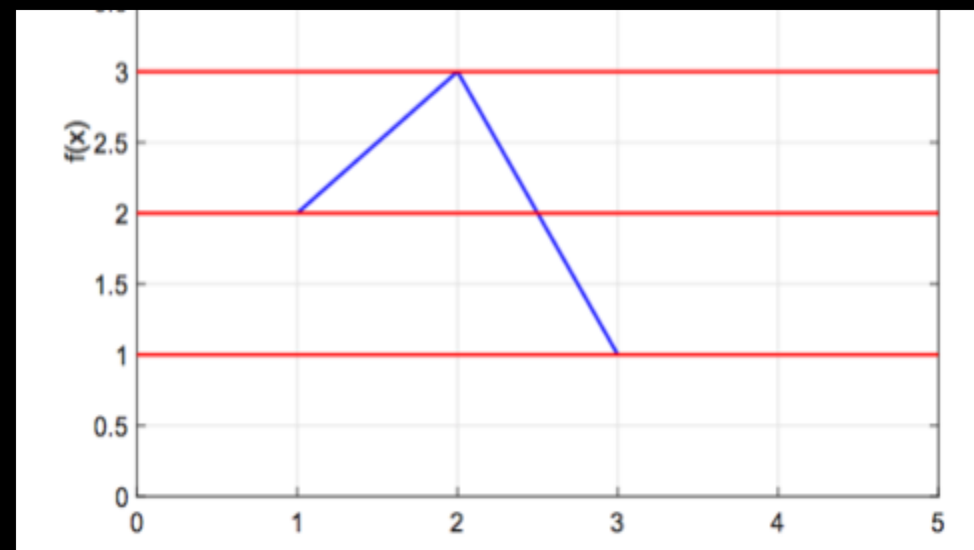
where $c > 1$ and is the largest root of $x^{p-1} - x^{p-2} - 1 = 0$.



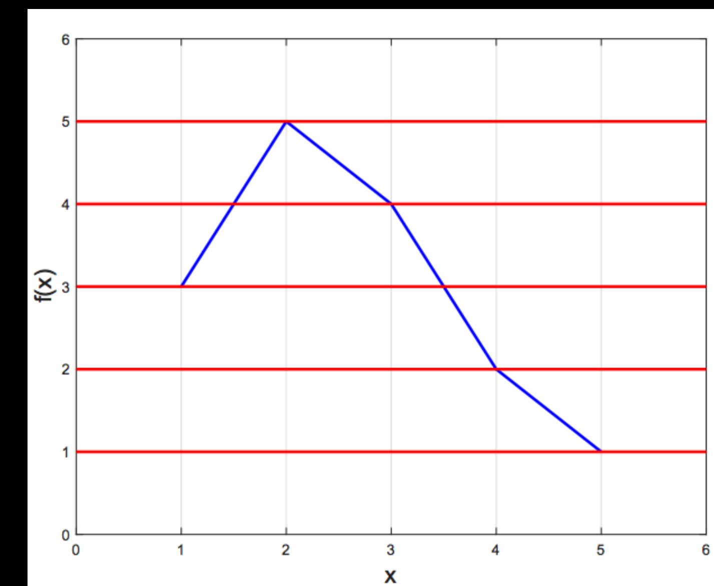
Examples [ICLR 2020]



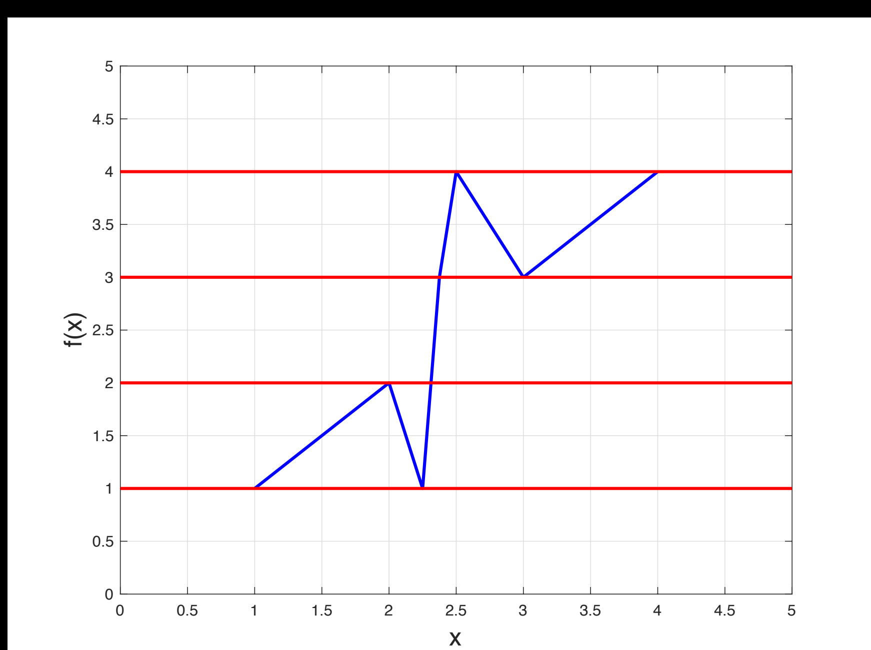
period 3



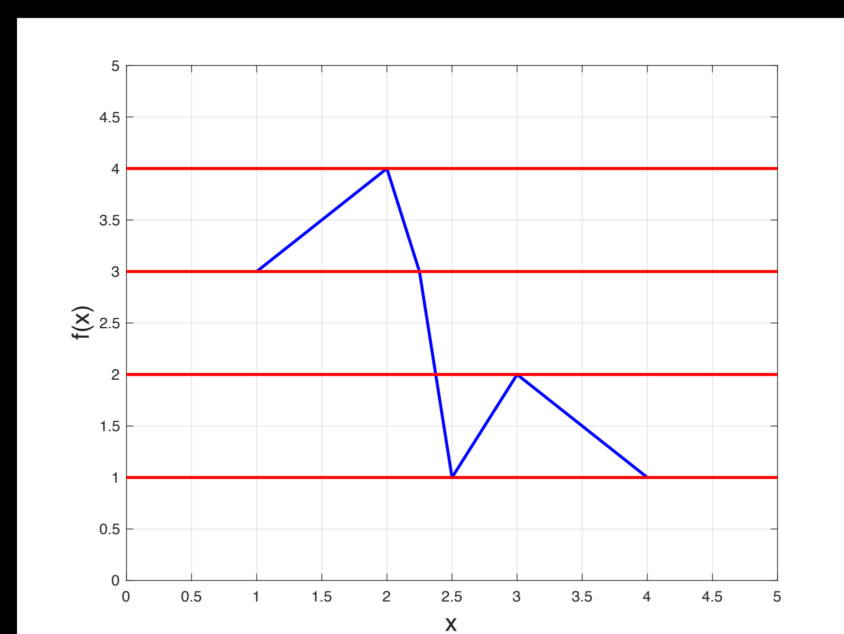
period 3



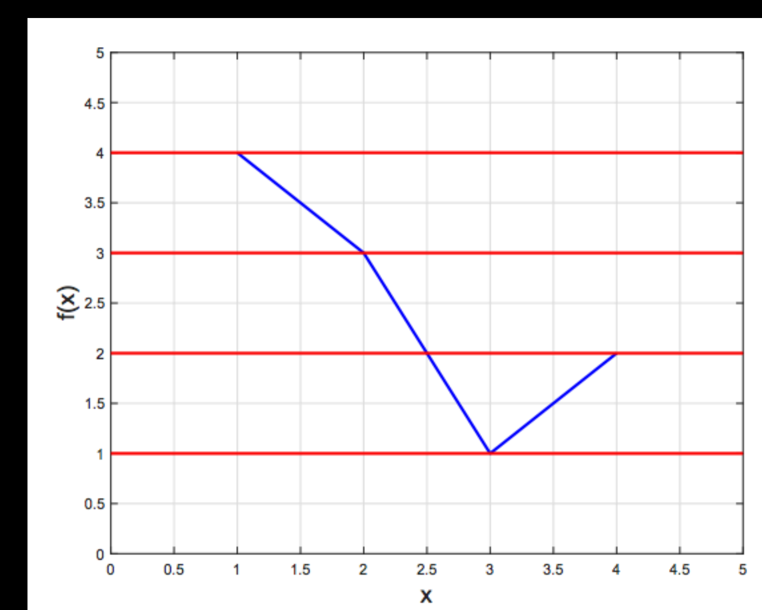
period 5



period 4



period 4



period 4

Period-dependent Trade-offs **[ICLR 2020]**

Main Lemma:

Let f be continuous with odd period $p \geq 3$.
Then f^t oscillates at least c^t times,
where $c > 1$ and is the largest root of $x^{p-1} - x^{p-2} - 1 = 0$.

Informal Main Result:

Using periodic functions f, we construct f^t,
that has c^t oscillations and is the output of a depth t
width 2 neural net, for which any shallow net
(l layers, u width per layer) with $u \leq c^{t/l}/8$
incurs high classification error ($\geq \frac{1}{4}$).

Our work in ICML 2020

Further connections to Dynamical Systems:

1. We get L1-approximation error and not just classification error.

2. We show tight connections between Lipschitz constant, periods of f, and oscillations.

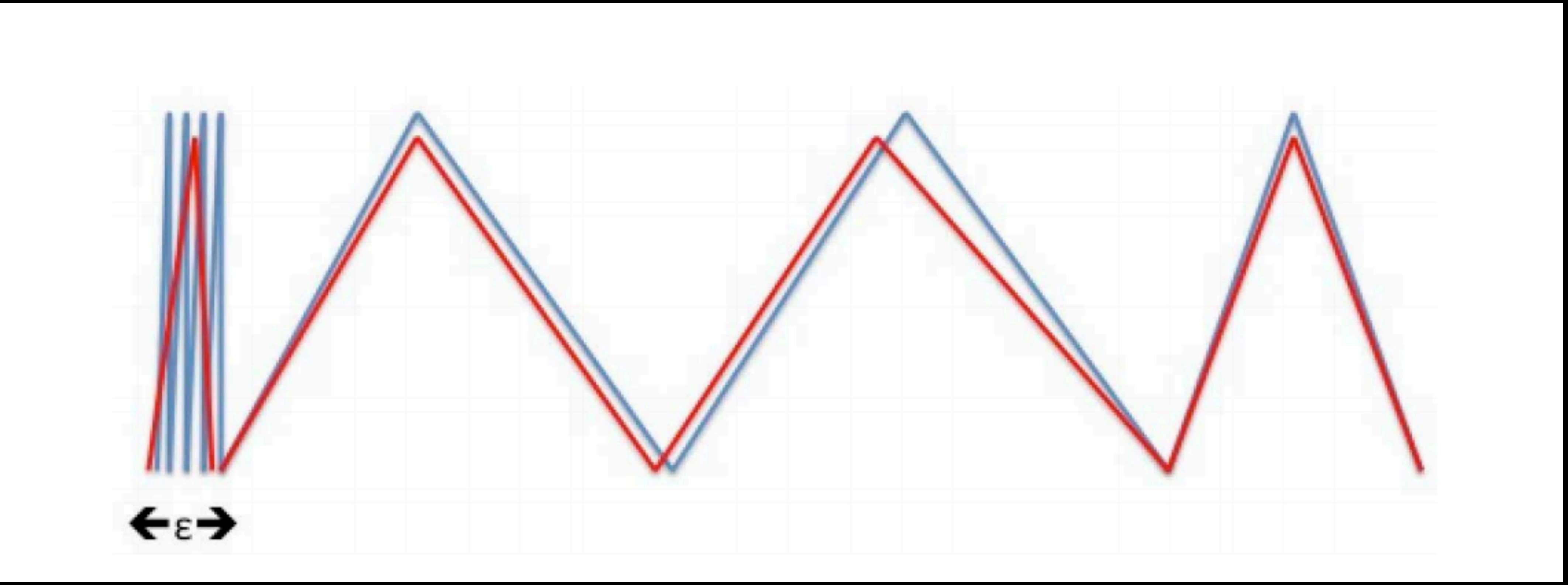
3. Sharper period-dependent depth-width tradeoffs and easy constructions of examples.

4. Experimental validation of our theoretical results.

Our work in ICML 2020

→ **Further connections to Dynamical Systems:**

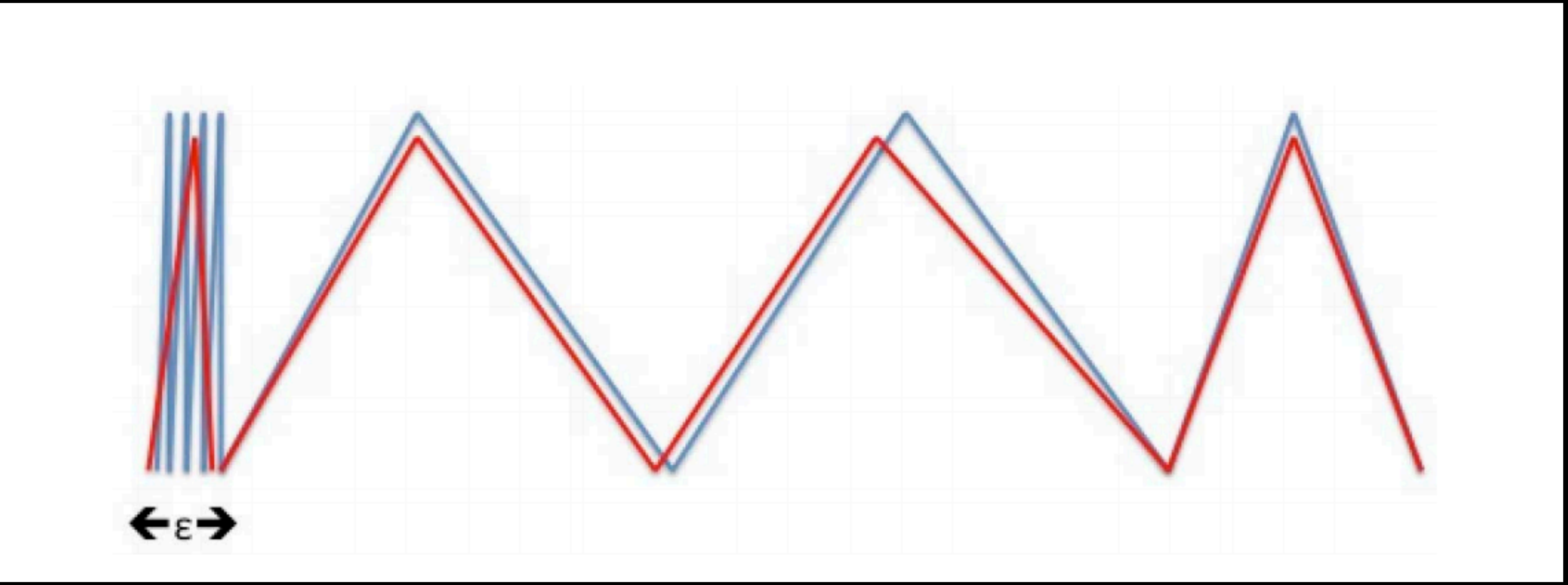
1. We get L1-approximation error and not just classification error.



Our work in ICML 2020

→ **Further connections to Dynamical Systems:**

Is it so hard to obtain L1 guarantees?



Period 3 of f, only informs us on 3 values of f.

Our work in ICML 2020

→ **Further connections to Dynamical Systems:**

1. We get L1-approximation error and not just classification error.

2. We show tight connections between Lipschitz constant, periods of f, and oscillations.

3. Sharper period-dependent depth-width tradeoffs and easy constructions of examples.

4. Experimental validation of our theoretical results.

Periods, Oscillations, Lipschitz

Lemma (Lower Bound on L):

Let $f : [a, b] \to [a, b]$ be L-Lipschitz.

If f^t has at least c^t oscillations, then $L \geq c$.

Informal Main Result (Lipschitz matches oscillations):

Let f as above with c^t oscillations between x, y.
Let g be any ReLU NN with u units per layer and l layers.
As long as $L = c$, and $(2u)^l \leq \frac{c^t}{8}$, we get L^1 separation:

$$\min_g \int_a^b |f^t(z) - g(z)| dz \geq C(x, y) > 0$$

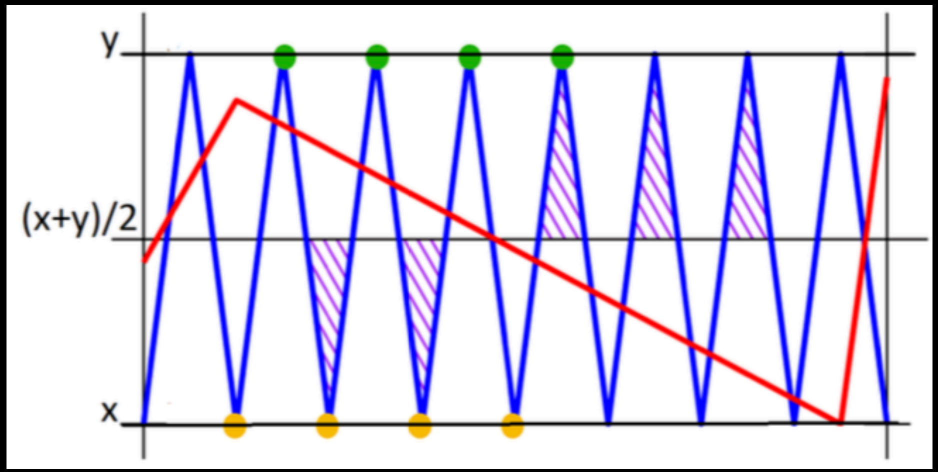
where $C(x, y)$ depends on x, y but not on t.

Proof Sketch

Let f as above with c^t oscillations between x, y.
Let g be any ReLU NN with u units per layer and l layers.
As long as $L = c$, and $(2u)^l \leq \frac{c^t}{8}$, we get L^1 separation:

$$\min_g \int_a^b |f^t(z) - g(z)| dz \geq C(x, y) > 0$$

where $C(x, y)$ depends on x, y but not on t.



Definitions:

let $h = f^t$ for ease of presentation.

$\tilde{h}(z) = \mathbf{1}[\mathbf{h(z)} \geq \frac{\mathbf{x+y}}{\mathbf{2}}]$

$\tilde{g}(z) = \mathbf{1}[\mathbf{g(z)} \geq \frac{\mathbf{x+y}}{\mathbf{2}}]$

Let $\mathcal{I}_{h,x,y}$ be the partition of $[a, b]$, where $\tilde{h}$ is piecewise constant.

Let $\mathcal{J}_{h,x,y} \subseteq \mathcal{I}_{h,x,y}$ be the collection of intervals containing pre-image of y.

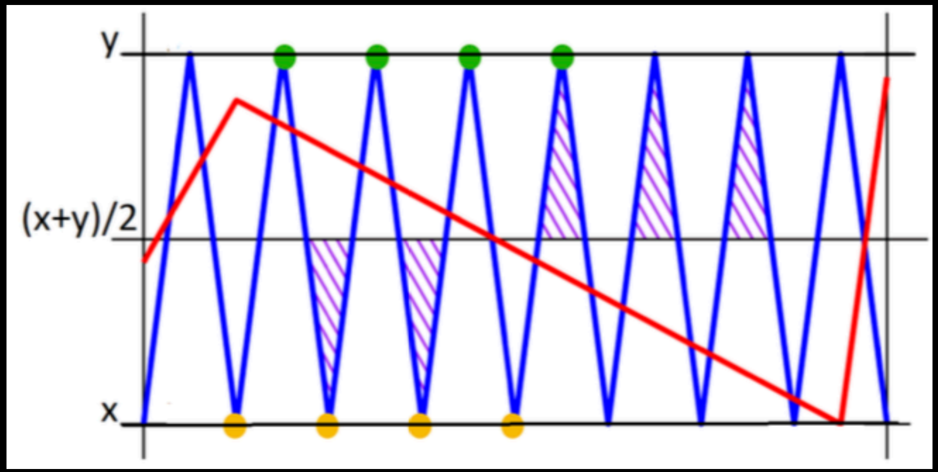
***Fact* [Telgarsky'16]:** $$\frac{1}{|\mathcal{J}_{h,x,y}|} \sum_{U \in \mathcal{J}_{h,x,y}} \mathbf{1}[\forall z \in U. \tilde{h}(z) \neq \tilde{g}(z)] \geq \frac{1}{2}\left(1 - 2\frac{|\mathcal{I}_{g,x,y}|}{|\mathcal{J}_{h,x,y}|}\right)$$

Proof Sketch

Let f as above with c^t oscillations between x, y.
Let g be any ReLU NN with u units per layer and l layers.
As long as $L = c$, and $(2u)^l \leq \frac{c^t}{8}$, we get L^1 separation:

$$\min_g \int_a^b |f^t(z) - g(z)| dz \geq C(x,y) > 0$$

where $C(x,y)$ depends on x, y but not on t.



Definitions:

let $h = f^t$ for ease of presentation.

$\tilde{h}(z) = \mathbf{1}[\mathbf{h(z)} \geq \frac{\mathbf{x+y}}{\mathbf{2}}]$

$\tilde{g}(z) = \mathbf{1}[\mathbf{g(z)} \geq \frac{\mathbf{x+y}}{\mathbf{2}}]$

Let $\mathcal{I}_{h,x,y}$ be the partition of $[a,b]$, where $\tilde{h}$ is piecewise constant.

Let $\mathcal{J}_{h,x,y} \subseteq \mathcal{I}_{h,x,y}$ be the collection of intervals containing pre-image of y.

***Claim*:** Let $U \in \mathcal{J}_{h,x,y}$, then: $\displaystyle\int_U \left| h(z) - \frac{x+y}{2} \right| dz \geq \frac{(y-x)^2}{8L^t}$

Proof Sketch

Let f as above with c^t oscillations between x, y.
Let g be any ReLU NN with u units per layer and l layers.
As long as $L = c$, and $(2u)^l \leq \frac{c^t}{8}$, we get L^1 separation:

$$\min_g \int_a^b |f^t(z) - g(z)| dz \geq C(x, y) > 0$$

where $C(x, y)$ depends on x, y but not on t.

$$\int_a^b |h(z) - g(z)|\, dz = \sum_{U \in \mathcal{I}_{h,x,y}} \int_U |h(z) - g(z)|\, dz$$

$$\geq \sum_{U \in \mathcal{J}_{h,x,y}} \int_U |h(z) - g(z)|\, dz$$

$$\geq \sum_{U \in \mathcal{J}_{h,x,y}} \int_U \left| h(z) - \frac{x+y}{2} \right| \mathbf{1}[\forall z \in U. \tilde{h}(z) \neq \tilde{g}(z)] dz$$

$$\geq \frac{|\mathcal{J}_{h,x,y}|(y-x)^2}{16 L^t} \left(1 - 2 \frac{|\mathcal{I}_{g,x,y}|}{|\mathcal{J}_{h,x,y}|} \right).$$

$$\int_a^b |h(z) - g(z)|\, dz \geq \frac{\left(\frac{c}{L}\right)^t (y-x)^2}{16} \left(1 - 2 \frac{|\mathcal{I}_{g,x,y}|}{c^t} \right)$$

$L = c$

$$\int_a^b |h(z) - g(z)|\, dz \geq \frac{(y-x)^2}{32}.$$

Our work in ICML 2020

→ **Further connections to Dynamical Systems:**

1. We get L1-approximation error and not just classification error.

2. We show tight connections between Lipschitz constant, periods of f, and oscillations.

3. Sharper period-dependent depth-width tradeoffs and easy constructions of examples.

4. Experimental validation of our theoretical results.

Periods, Oscillations

If f has period p, how many oscillations?

Main Lemma:

Let f be continuous with odd period $p \geq 3$.
Then f^t oscillates at least c^t times,
where $c > 1$ and is the largest root of $x^{p-1} - x^{p-2} - 1 = 0$.

$$x^p - 2x^{p-2} - 1 = 0$$

The root $c(p)$ is decreasing, and always $c \geq \sqrt{2}$.

Period-specific threshold phenomenon: shallow g has $(2u)^l \leq \frac{c^t}{8}$

Proof Sketch

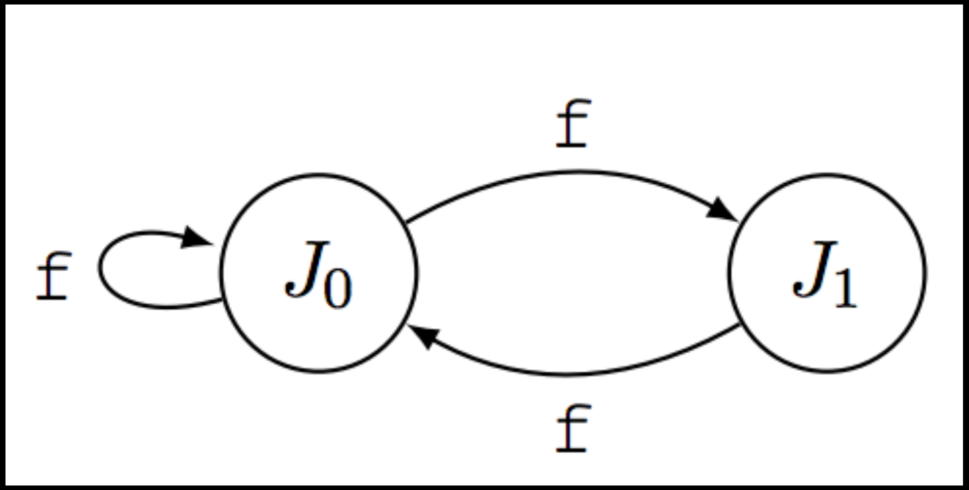
If f has period p, how many oscillations?

Let f be continuous with odd period $p \geq 3$.
Then f^t oscillates at least c^t times,

$$x^p - 2x^{p-2} - 1 = 0$$

$$\frac{2}{9} \xrightarrow{f} \frac{4}{9} \xrightarrow{f} \frac{8}{9} \xrightarrow{f} \frac{2}{9}$$

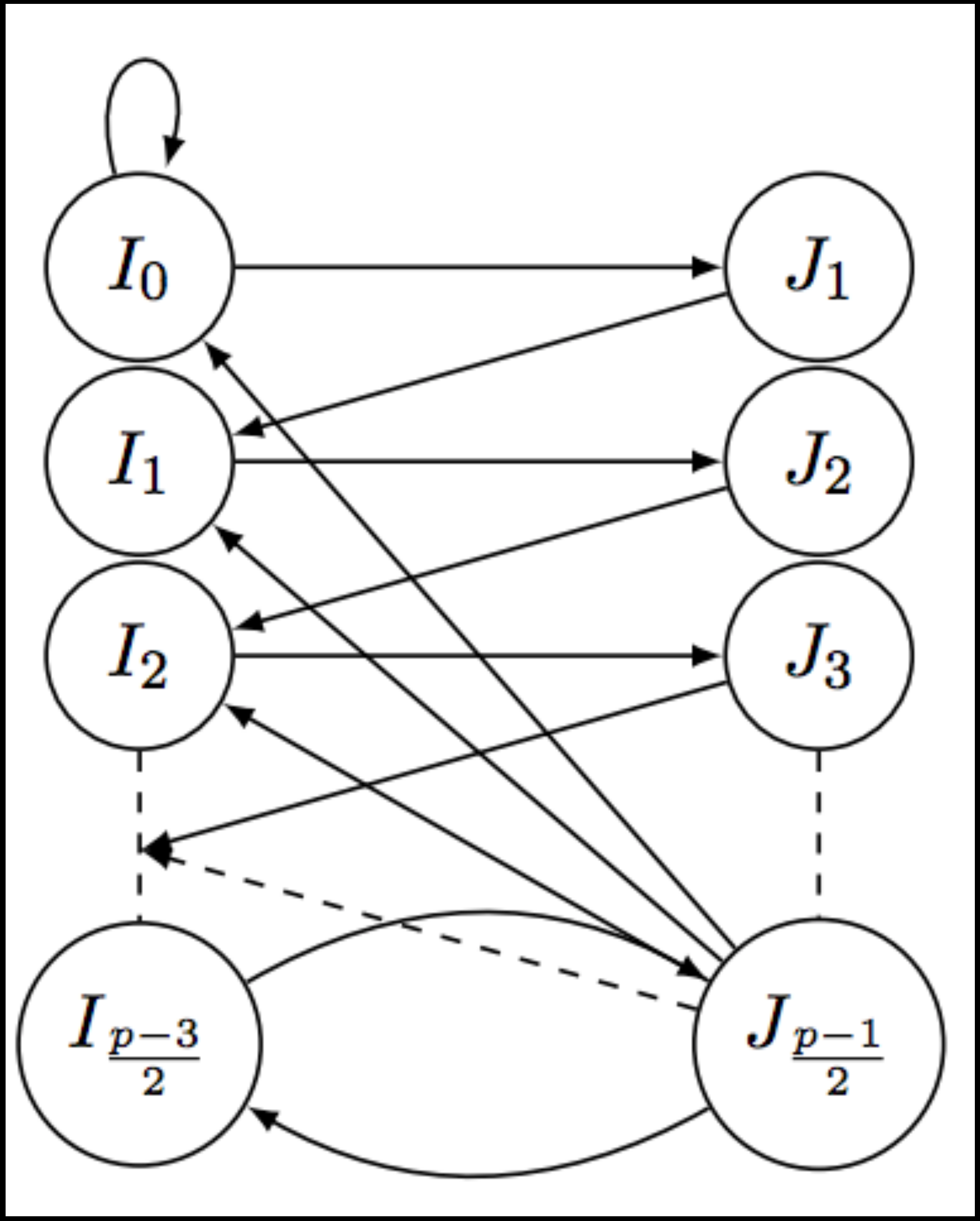
$$J_0 = [\tfrac{2}{9}, \tfrac{4}{9}] \quad J_1 = [\tfrac{4}{9}, \tfrac{8}{9}]$$



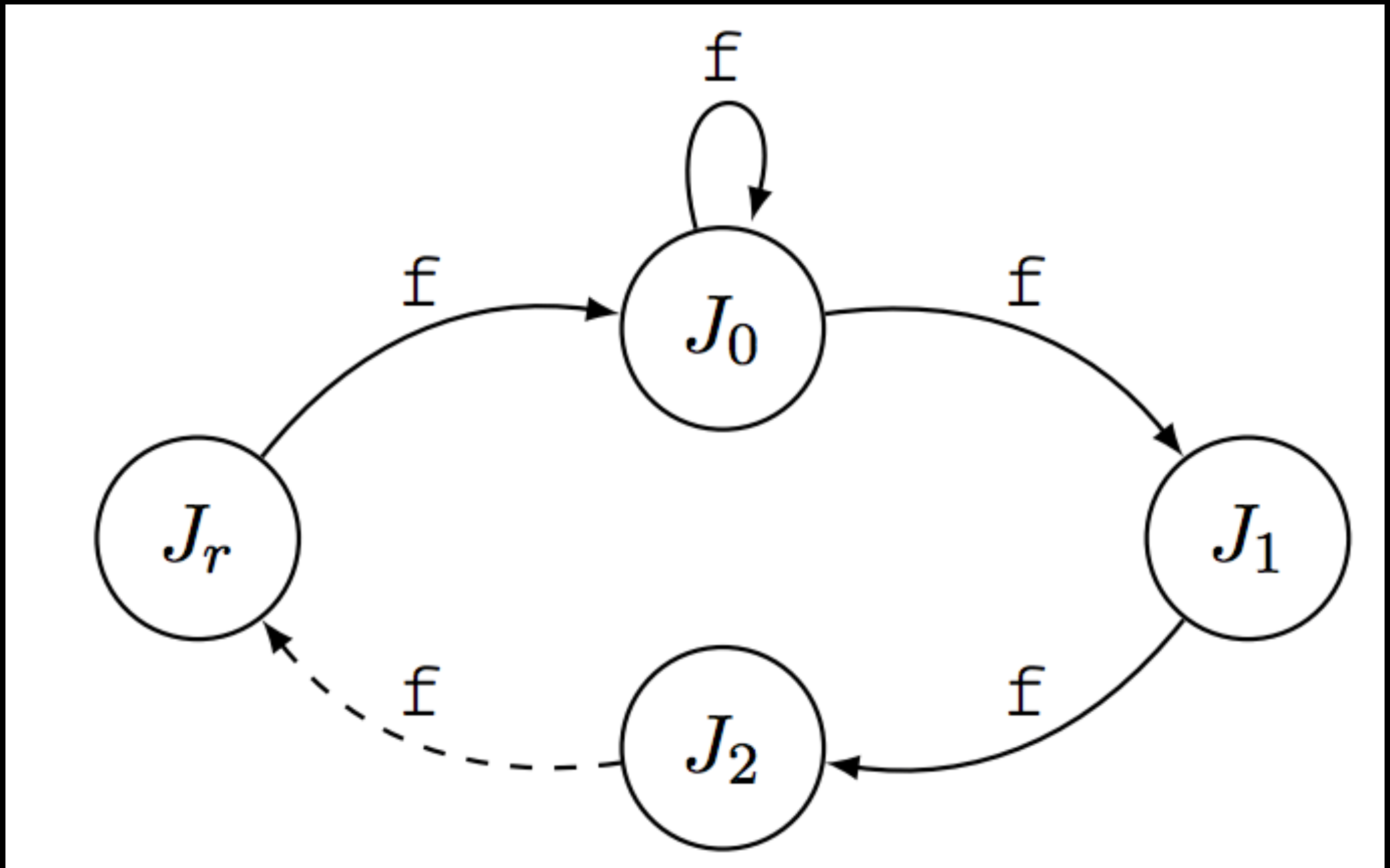
Oscillations → $\|A^t\|_\infty \geq \mathrm{sp}(A^t)$ ← **Root of characteristic**

Proof Sketch

If f has period p, how many oscillations?



$$x^p - 2x^{p-2} - 1 = 0$$

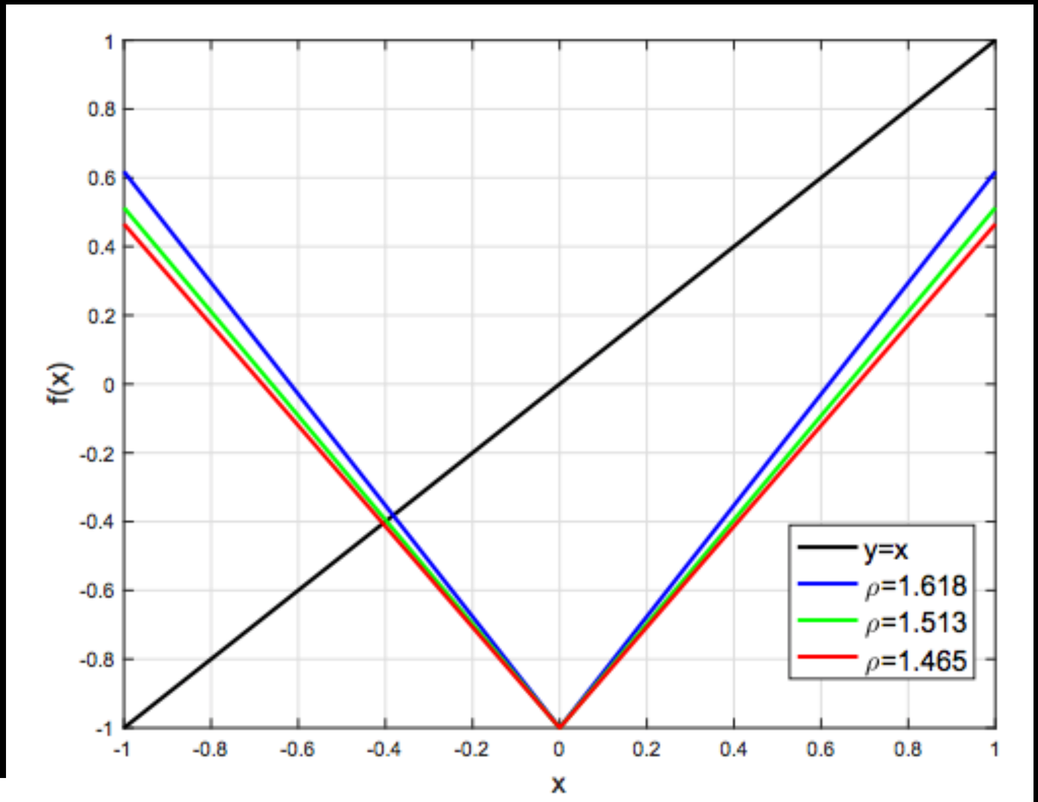


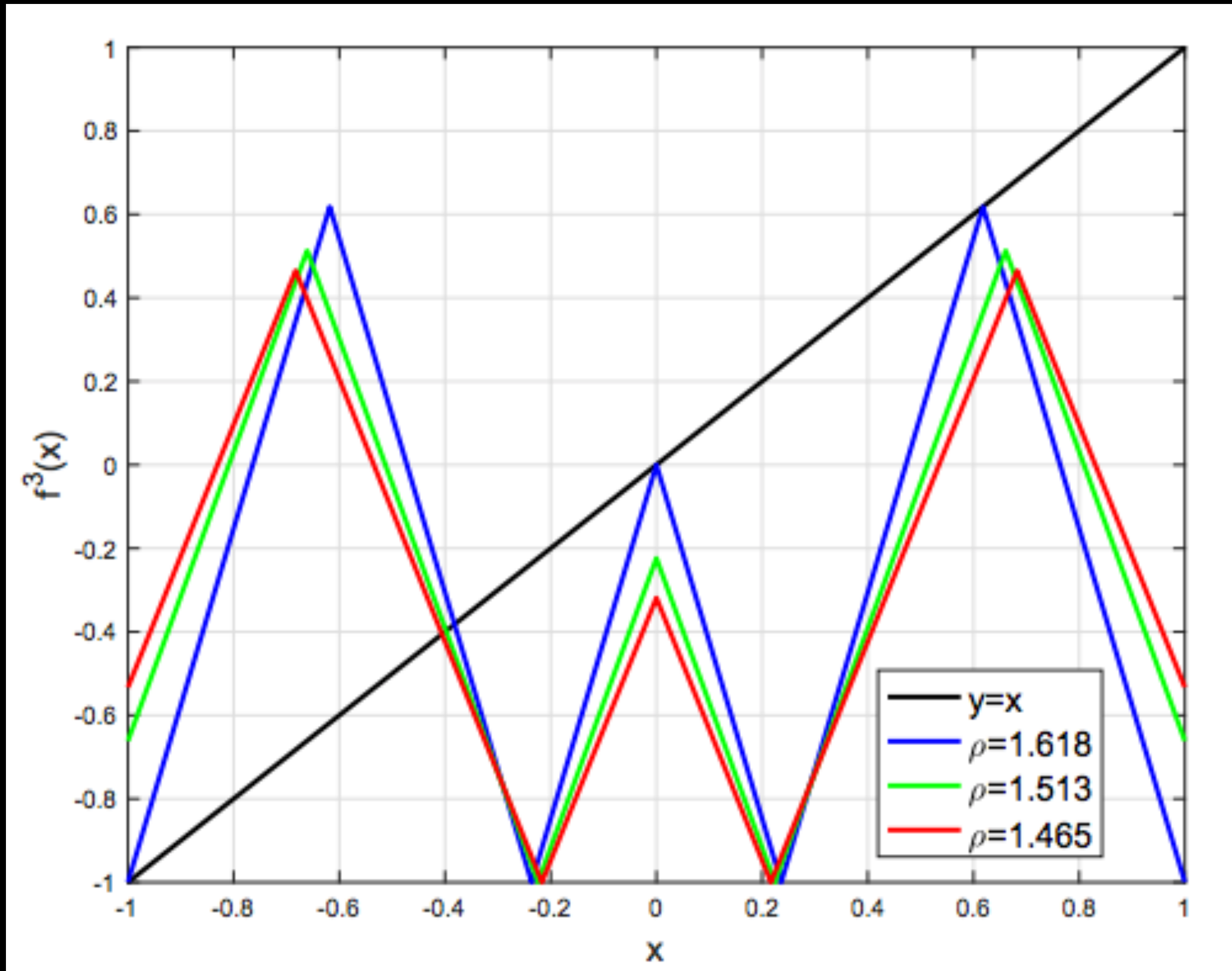
Tight examples - Sensitivity

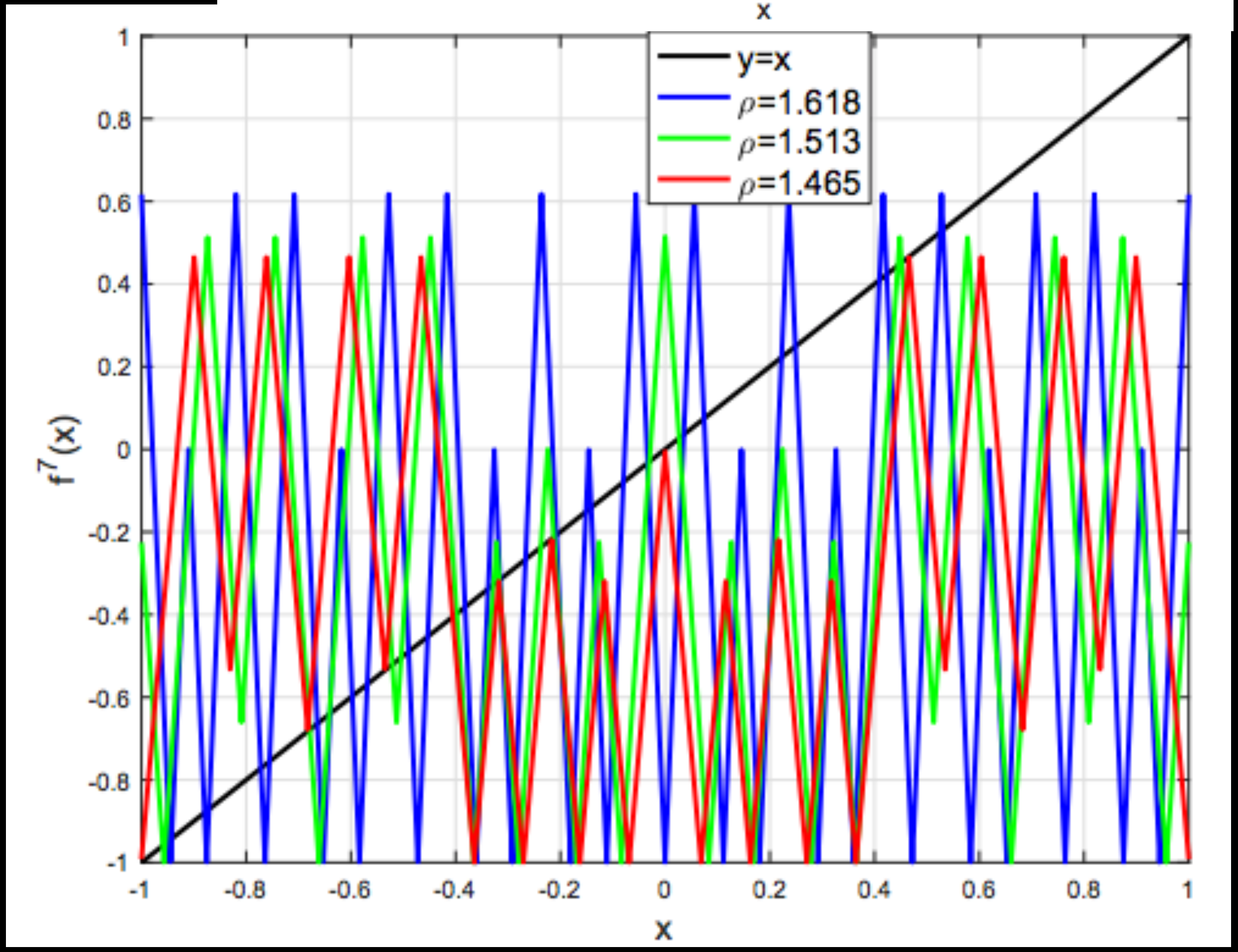
Function of period p & Lipschitz matching oscillation growth:

$$f(x) = c(p)|x| - 1$$

If slope is less than 1.618, then no period 3 appears







Our work in ICML 2020

Further connections to Dynamical Systems:

1. We get L1-approximation error and not just classification error.

2. We show tight connections between Lipschitz constant, periods of f, and oscillations.

3. Sharper period-dependent depth-width tradeoffs and easy constructions of examples.

4. Experimental validation of our theoretical results.

Experimental Section

Goals: 1. Instantiate benefits of depth for a period-specific task.

2. Validate our theoretical threshold for separating shallow NNs from deep.

Setting: $f(x)=1.618|x|-1$ Width: 20, #layers: 1 up to 5

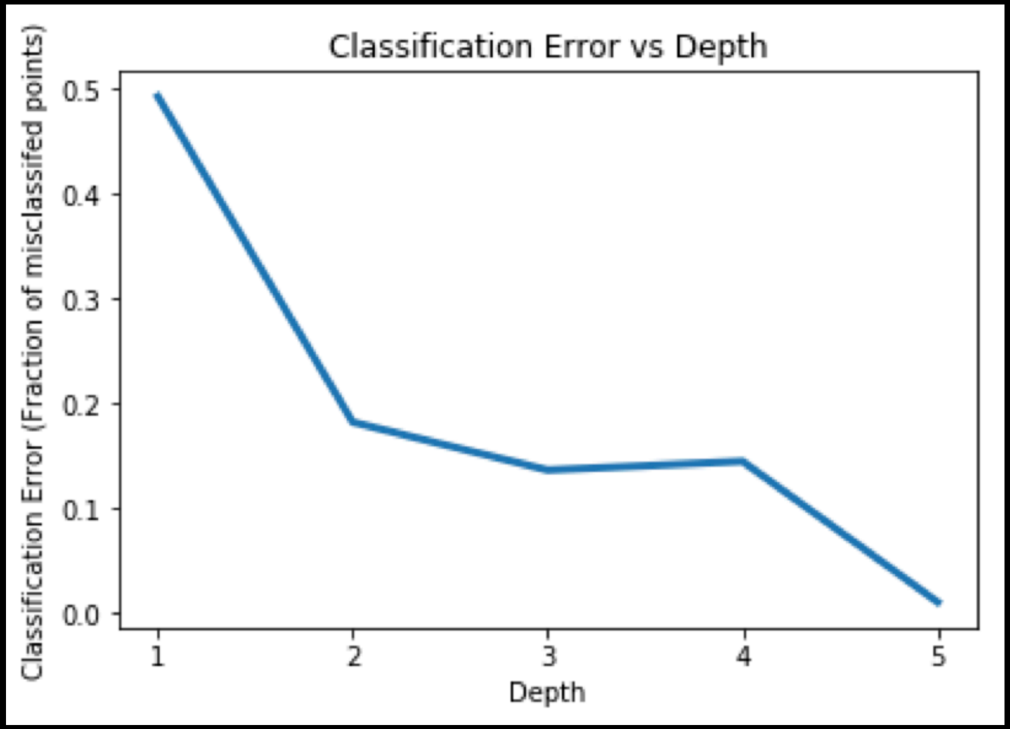
Easy Task: We take only 8 compositions of f.

Hard Task: We take 40 compositions of f. $(2u)^l \leq \frac{c^t}{8}$

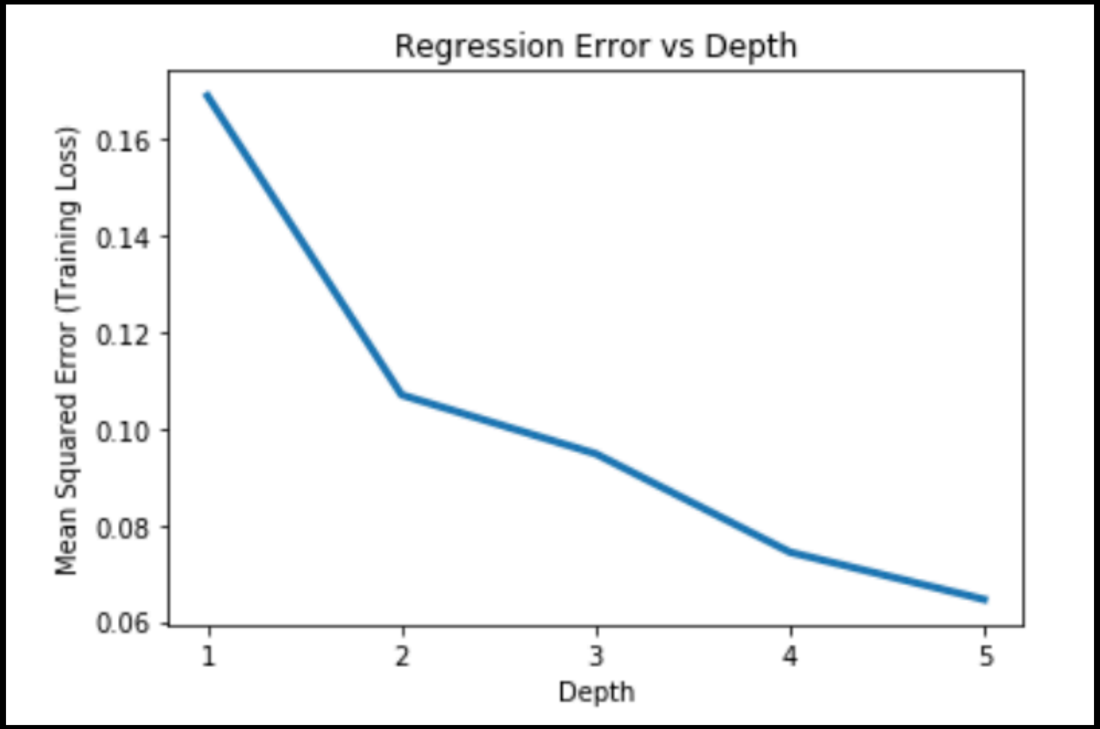
Training: Define a regression task on 10K datapoints chosen uniformly at random by evaluating f. We use Adam as the optimizer and train for 1500 epochs.

Overfitting: We are interested in representation.

Easy Task: We take only 8 compositions of f.



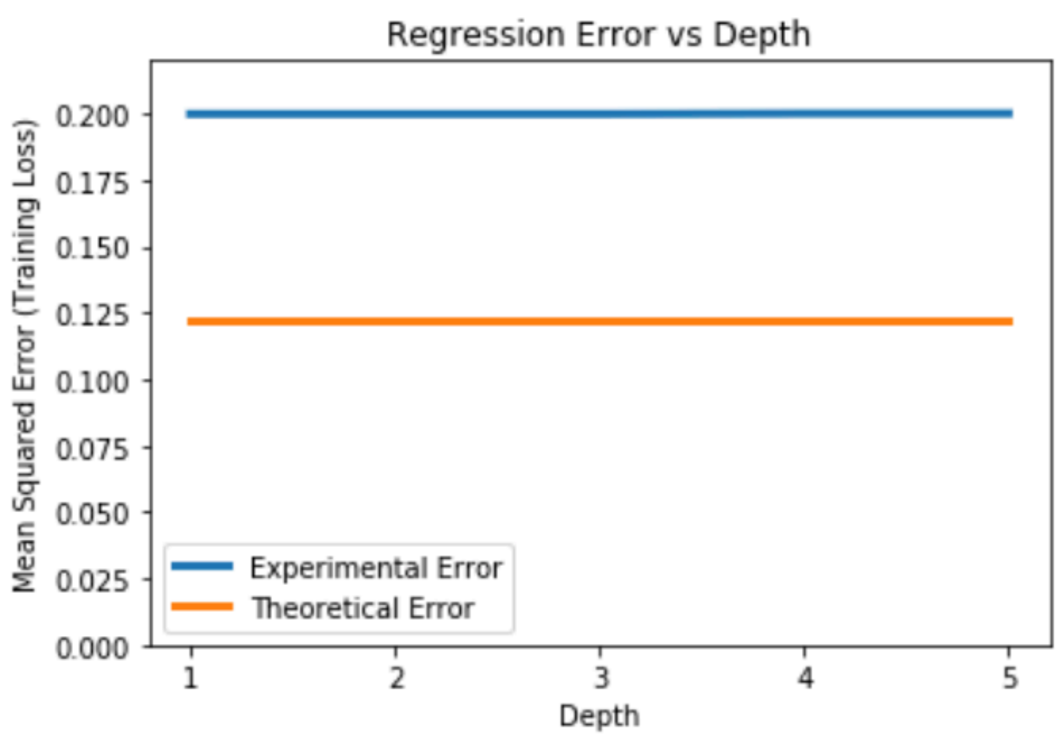
Classification error vs depth for the easy task appearing in our ICLR 2020 paper



Regression error vs depth for easy task

Adding depth does help in reducing error.

Hard Task: We take 40 compositions of f.

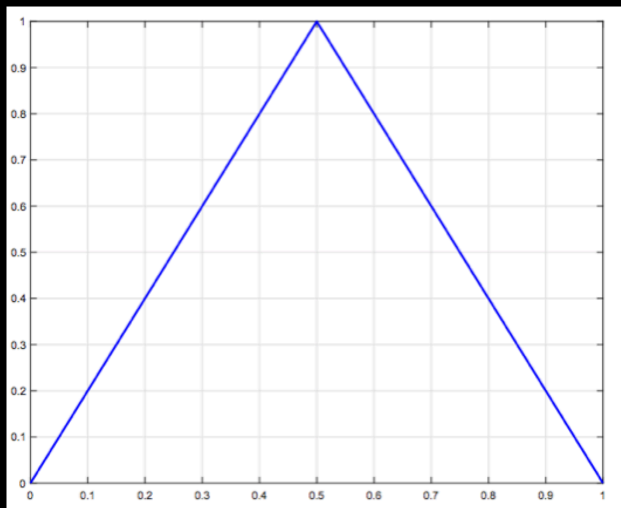


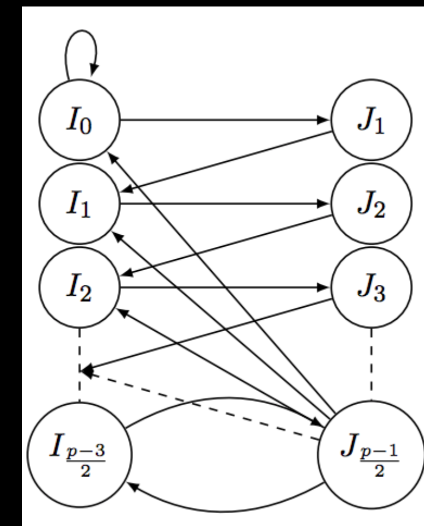
Regression error vs depth

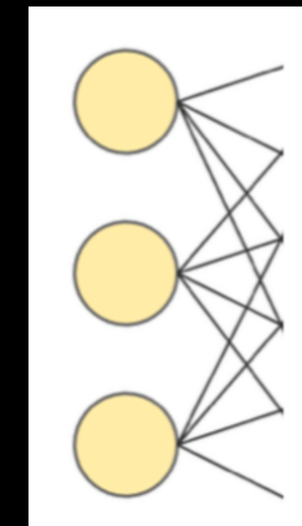
$$(2u)^l \leq \frac{c^t}{8}$$

Error (blue line) is independent of depth and is extremely close to theoretical bound (orange line).

Recap







Natural property of continuous funcitons: Period

1. **Sharp depth-width tradeoffs and L1-separations**
2. **Tight connections between Lipschitz, periods,oscillations.**

Simple constructions useful for proving separations.

Future Work

Understanding optimization (e.g., Malach, Shalev-Shwartz'19)

Unifying notions of complexity used for separations: trajectory length, global curvature, algrebraic varieties

Topological Entropy from Dynamical Systems

Better Depth-Width Trade-offs for Neural Networks through the lens of Dynamical Systems

MIT Mifods Talk by Panageas (2020):

https://www.youtube.com/watch?v=HNQ204BmOQ8

ICLR 2020 spotlight talk:

https://iclr.cc/virtual_2020/poster_BJe55gBtvH.html

Vaggos Chatziafratis
(Stanford & Google NY)

Sai Ganesh Nagarajan
(SUTD)

Ioannis Panageas
(SUTD => UC Irvine)