Expressivity of Neural Networks via Chaotic Itineraries beyond Sharkovsky's Theorem

Clayton Sanford
Columbia University

Vaggos Chatziafratis
FODSI at MIT/Northeastern &
University of California at Santa Cruz

Abstract

Given a target function f, how large must a neural network be in order to approximate f? Recent works examine this basic question on neural network *expressivity* from the lens of dynamical systems and provide novel "depth-vs-width" tradeoffs for a large family of functions f. They suggest that such tradeoffs are governed by the existence of *periodic* points or *cycles* in f. Our work, by further deploying dynamical systems concepts, illuminates a more subtle connection between periodicity and expressivity: we prove that periodic points alone lead to suboptimal depth-width tradeoffs and we improve upon them by demonstrating that certain "chaotic itineraries" give stronger exponential tradeoffs, even in regimes where previous analyses only imply polynomial gaps. Contrary to prior works, our bounds are nearly-optimal, tighten as the period increases, and handle strong notions of inapproximability (e.g., constant L_1 error). More broadly, we identify a phase transition to the *chaotic regime* that exactly coincides with an abrupt shift in other notions of function complexity, including VC-dimension and topological entropy.

1 Introduction

Whether a neural network (NN) succeeds or fails at a given task crucially depends on whether or not its architecture (depth, width, types of activation units etc.) is suitable for the task at hand. For example, a "size-inflation" phenomenon has occurred in recent years, in which NNs tend to be deeper and/or larger. Recall that in 2012, AlexNet had 8 layers. In 2015, ResNet won the ImageNet competition with 152 layers (Krizhevsky et al., 2012; He et al., 2016), This trend still continues to date, with modern models using billions of parameters (Brown et al., 2020). The empirical success of deep neural networks motivates researchers to ask: What are the theoretical benefits of depth, and what are the depth-vs-width tradeoffs?

This question gives rise to the study of neural network *expressivity*, which characterizes the class of functions that are representable (or approximately representable) by a NN of certain depth, width, and activation. For instance, Eldan and Shamir (2016) propose a family of "radial" functions in $\mathbb{R}^d$ that are easily expressible with 3-layered feedforward neural nets of small width, but require any approximating 2-layer network to have exponentially (in d) many neurons. In other words, they formally show that depth—even if increased by 1—can be exponentially more valuable than width.

Not surprisingly, understanding the expressivity of NNs was an early question asked in 1969, when Minsky and Papert showed that the Perceptron can only learn linearly separable data and fails on simple XOR functions (Minsky and Papert, 1969). The natural question of which functions can multiple such Perceptrons (i.e., multilayer feedforward NN) express was addressed later by Cybenko (1989); Hornik et al. (1989) proving the so-called *universal approximation* theorem. This states, roughly, that just one hidden layer of standard activation units (e.g., sigmoids, ReLUs etc.) suffices to approximate any continuous function arbitrarily well. Taken at face value, any continuous function is a 2-layer (i.e., 1-hidden-layer) network in disguise, and hence, there is no reason to consider deeper networks. However, the width required can grow arbitrarily, and many works in the following decades quantify those depth-vs-width tradeoffs.

Towards this direction, one typically identifies a function together with a "measure of complexity" to

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

demonstrate benefits of depth. For example, the seminal work by Telgarsky (2015, 2016) relies on the number of oscillations of a narrow family of triangle mappings on $[0,1]$ that can be expressed recursively with deep neural networks. Other relevant notions of complexity to the expressivity of NNs include the VC dimension (Warren, 1968; Anthony and Bartlett, 1999; Schmitt, 2000), the number of linear regions (Montufar et al., 2014; Arora et al., 2016) or activation patterns (Hanin and Rolnick, 2019), the dimension of algebraic varieties (Kileel et al., 2019), the Fourier spectrum (Barron, 1993; Eldan and Shamir, 2016; Daniely, 2017; Lee et al., 2017; Bresler and Nagaraj, 2020), fractals (Malach and Shalev-Shwartz, 2019), topological entropy (Bu et al., 2020), Lipschitzness (Safran et al., 2019; Hsu et al., 2021), global curvature and trajectory length (Poole et al., 2016; Raghu et al., 2017) just to name a few.

This work builds upon recent papers (Chatziafratis et al., 2019, 2020), which study expressivity from the lens of discrete-time dynamical systems and extend Telgarsky's results beyond triangle (tent) maps. At a high-level, their idea is the following: if the initial layers of a NN output a real-valued function f, then concatenating the *same* layers k times one after the other outputs $f^k := f \circ f \circ \ldots \circ f$, i.e., the composition of f with itself k times. By associating each discrete timestep k to the output of the corresponding layer in the network, one can study expressivity via the underlying properties of f's trajectories. Indeed, if f contains higher-order fixed points, called *periodic* points, then deeper NNs can efficiently approximate f^k, but shallower nets would require exponential width, governed by f's periodicity.

Inspired by these novel connections to discrete dynamical systems, we pose the following natural question:

Apart from periodicity, are there other properties of f's trajectories that govern the expressivity tradeoffs?

We indeed prove that f's periodicity alone is not the end of the story, and we improve on the known depth-width tradeoffs from several perspectives. We exhibit functions of the same period with very different behaviors (see Sec. 2) that can be distinguished by the concept of "chaotic itineraries." We analyze these here in order to achieve nearly-optimal tradeoffs for NNs. Our work highlights why previous works that examine periodicity alone only obtain loose bounds. More specifically:

- We accurately quantify the oscillatory behavior of a large family of functions f. This leads to sharper and nearly-optimal lower bounds for the width of NNs that approximate f^k.
- Our lower bounds cover a stronger notion of approximation error, i.e., *constant* separations between NNs, instead of bounds that become small depending heavily on f and its periodicity.
- At a conceptual level, we introduce and study certain chaotic itineraries, which supersede Sharkovsky's theorem (see Sec. 1.2).
- We elucidate connections between periodicity and other function complexity measures like the VC-dimension and the topological entropy (Alsedà et al., 2000). We show that all of these measures undergo a phase transition that exactly coincides with the emergence of the chaotic regime based on periods.

To the best of our knowledge, we are the first to incorporate the notion of chaotic itineraries from discrete dynamical systems into the study of NN expressivity. Before stating and interpreting our results, we provide some basic definitions.

1.1 Function Approximation and NNs

This paper employs three notions of approximation to compare functions $f, g : [0,1] \to [0,1]$.

- $L_1(f,g) = \|f-g\|_1 = \int_0^1 |f(x)-g(x)|\,dx$.
- $L_\infty(f,g) = \|f-g\|_\infty = \sup_{x\in[0,1]} |f(x)-g(x)|$.
- Classification error $\mathcal{R}_{S,t}$: For some sample $S = \{x_1,\ldots,x_n\} \subseteq [0,1]$ and threshold $t \in [0,1]$, let $\mathcal{R}_{S,t}(f,g)$ be the fraction of samples that classifiers derived by thresholding f and g disagree on. That is, $\mathcal{R}_{S,t}(f,g) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{[[f(x_i)]]_t \neq [[g(x_i)]]_t\}$ for $[[x]]_t = \mathbb{1}\{x \geq t\}$.

While L_1 and L_∞ directly measure the ability of a hypothesis to approximate a fixed function, $\mathcal{R}_{S,t}$ measures the difference between functions by framing the question as a classification problem.

For what follows, let $\mathcal{N}(u,\ell)$ be the family of feedforward NNs of depth ℓ and width at most u per layer with ReLU activation functions.[1] All our results also hold for the more general family of semialgebraic activations (Telgarsky, 2016).

1.2 Discrete Dynamical Systems

To construct families of functions that yield depth-separation results, we rely on a standard notion of *unimodal* functions from dynamical systems (Metropolis et al., 1973).

[1] Recall $\mathrm{ReLU}(x) = \max(x,0)$.

Definition 1. *Let $f : [0,1] \to [0,1]$ be a continuous and piece-wise differentiable function. We say f is a* unimodal mapping *if:*

1. *$f(0) = f(1) = 0$, and $f(x) > 0$ for all $x \in (0,1)$.*
2. *There exists a unique maximizer $x' \in (0,1)$ of f, i.e., f is strictly increasing on the interval $[0, x')$ and strictly decreasing on $(x', 1]$.*

Our constructions rely on unimodal functions that are concave and also symmetric around $\frac{1}{2}$ (i.e., $f(x) = f(1-x)$ for all $x \in [0,1]$)[2]. We note that the resulting function family is fairly general, already capturing the triangle waves of Telgarsky (2015, 2016) and the logistic map used in previous depth-separation results (Schmitt, 2000). Moreover, the study of one-dimensional discrete dynamical systems by applied mathematicians explicitly identifies unimodal mappings as important objects of study (Metropolis et al., 1973; Alsedà et al., 2000).

Recall that a fixed point x^* of f is a point where $f(x^*) = x^*$. A more general notion of higher-order fixed points is that of *periodicity*.

Definition 2. *For some $p \in \mathbb{N}$, we say that $x_1 \in [0,1]$ is a* point of period p *if $f^p(x_1) = x_1$ and $f^k(x_1) \neq x_1$[3] for all $k \in [p-1]$.*[4] *The sequence $x_1, f(x_1), \ldots, f^{p-1}(x_1)$ is called a* p-cycle*, and f has* periodicity p *if such a cycle exists.*

For example, the identify map $f(x) = x$ has a fixed point (or a point of period 1) at any $x \in [0,1]$. Likewise, $f(x) = 1 - x$ has a fixed point at $x = \frac{1}{2}$ and a point of period 2 at any other choice of x. The triangle map $f(x) = \min(2x, 2(1-x))$ has a fixed point at $x = \frac{2}{3}$; a 2-cycle with $x_1 = \frac{2}{5}$ and $x_2 = \frac{4}{5}$; and a 3-cycle with $x_1 = \frac{2}{9}$, $x_2 = \frac{4}{9}$ and $x_3 = \frac{8}{9}$ (among other cycles of higher periodicity).

Does the existence of some p-cycle in f have any implications about the existence of other cycles? These relations between the periods of f are of fundamental importance to the study of dynamical systems. In particular, Li and Yorke (1975) proved in 1975 that "period 3 implies chaos" in their celebrated work, which also introduced the term "chaos" to mathematics and later spurred the development of chaos theory. Interestingly, an even more general result was already obtained a decade earlier in Eastern Europe, by Sharkovsky (1964, 1965):

[2]Throughout, *symmetric* f refers to such functions that are symmetric around $\frac{1}{2}$.

[3]Throughout the paper, f^k means composition of f with itself k times, or $f^k = \underbrace{f \circ f \circ \cdots \circ f}_{k}$.

[4]As is common, $[m] = \{1, 2, \ldots, m\}$.

Theorem 1 (Sharkovsky's Theorem)**.** *Let $f : [0,1] \to [0,1]$ be continuous. If f contains period p and $p \triangleright p'$, then f also contains period p', where the symbol "$\triangleright$" is defined based on the following (decreasing) ordering:*

$$3 \triangleright 5 \triangleright 7 \triangleright \ldots \triangleright 2 \cdot 3 \triangleright 2 \cdot 5 \triangleright 2 \cdot 7 \triangleright \ldots$$

$$\ldots \triangleright 2^2 \cdot 3 \triangleright 2^2 \cdot 5 \triangleright 2^2 \cdot 7 \triangleright \ldots \triangleright 2^3 \triangleright 2^2 \triangleright 2 \triangleright 1.$$

This ordering, called *Sharkovsky's ordering*, is a total ordering on the natural numbers, where $l \triangleright r$ whenever l is to the left of r. The maximum number in this ordering is 3; if f contains period 3, then it also has all other periods, which is also known as *Li-Yorke chaos*. Chatziafratis et al. (2019, 2020) apply this theorem to obtain depth-width tradeoffs based on periods and obtain their most powerful results when $p = 3$. We go beyond Sharkovsky's theorem and prove that tradeoffs are determined by the "itineraries" of periods.[5]

Definition 3 (Itineraries)**.** *For a p-cycle $x_1, \ldots, x_p$, suppose that $x_{a_1} < \cdots < x_{a_p}$ for $a_j \in [p]$. The* itinerary *of the cycle is the cyclic permutation of $x_{a_1}, \ldots, x_{a_p}$ induced by f, which we represent by the string $\mathbf{a} = a_1 \ldots a_p$. Because cyclic permutations are invariant to rotation, we assume (without loss of generality) that $a_1 = 1$.*

Definition 4 (Chaotic Itineraries)**.** *A p-cycle is a* chaotic itinerary *or an* increasing cycle *if its itinerary is $12 \ldots p$. That is, $x_1 < \cdots < x_p$.*

Examining chaotic itineraries circumvents the limitations of prior works based on periods and yields sharper exponential depth-width tradeoffs. For example, there are two distinct itineraries of 4-cycles on unimodal maps: $\mathbf{a} = 1234$ and $\mathbf{a} = 1324$. The former is chaotic, and repeatedly applying the function yields a complex function that is hard to approximate; the latter does not guarantee hardness of approximation, and there exist easily approximable functions f^k derived recursively from mappings f that have the 1324 itinerary. We discuss this case more thoroughly in Section 2 and explore other examples of chaotic itineraries in App. 7.1. Unlike other function complexity properties, the existence of a chaotic itinerary is easily verifiable (see App. 7.3).

1.3 Our Main Contributions

Our principal goal is to use knowledge about f's itineraries to more accurately quantify the number of oscillations—the number of monotone pieces of a sufficient size, formally defined in Definitions 5 and 6—of f^k as a measure of complexity and draw connections to other complexity measures. Section 3 produces

[5]These are called "patterns" in Alsedà et al. (2000).

sharper and more robust NN approximability tradeoffs than prior works by leveraging chaotic itineraries and unimodality. Section 4 shows how a phase transition in VC-dimension and topological entropy of f occurs exactly when the growth rate of oscillations shifts from polynomial to exponential.

While previous works count oscillations too, they either construct too narrow a range of functions[6], obtain loose depth-width tradeoffs[7], or have unsatisfactory approximation error. [8] In Section 3, we improve along these three directions by taking advantage of the unimodality and itineraries of f. The *unimodality* of f allows us to quantify both the number of piecewise monotone pieces of f^k (i.e., oscillations) and the corresponding height between the highest and lowest values of f^k's oscillations. This improvement on the height enables stronger notions of function approximation (e.g., constant error rates with no dependence on f or its period p). *Chaotic itineraries* allow an improved analysis of the number of oscillations in f^k and grant sharper exponential lower bounds on the width of any shallow net g approximating f^k.

We say that our results are *nearly-optimal* because we exhibit a broad family of functions f that are inapproximable by shallow networks of width $O(\rho^k)$ for ρ arbitrarily close to 2. Because no unimodal function f can induce more than 2^k oscillations in f^k, we cannot aspire to tighter exponent bases in this setting.[9] On the other hand, none of the bounds from previous works (except the narrow bounds of Telgarsky) produce width bounds of more than $\Omega(\phi^k)$, where $\phi \approx 1.618$ is the Golden Ratio. To demonstrate our sharper tradeoffs, we state a special case of our results for the L_∞ error.

Theorem 2. *For $p \geq 3$ and $k \in \mathbb{N}$, consider any symmetric, concave unimodal mapping f with an increasing p-cycle and any $g \in \mathcal{N}(u, \ell)$ with width*

$$u \leq \frac{1}{8}\left(\max\left(2 - \frac{4}{2^p}, \phi\right)\right)^{k/\ell}$$

Then, $L_\infty(f^k, g) = \Omega(1)$, independent of f, p, k.

Remark 1. *When g is shallow with depth $\ell = O(k^{1-\epsilon})$ (e.g., $\ell = k^{0.99}$), then its width must be exponentially large in order to well-approximate f^k. This exponential separation in k is sharper than prior works (Chatziafratis et al., 2019, 2020), and quickly becomes even sharper (tending to 2) with larger values of p. This is counterintuitive as Sharkovsky's ordering implies that period 3 is the most chaotic and prior works recover a suboptimal rate of at most $\phi \approx 1.618$ (see Table 1).*

Remark 2. *Our approximation error is constant independent of all other parameters f, k, p. Previous results (Chatziafratis et al., 2019, 2020; Bu et al., 2020) obtain a gap that depends on f, p and may be arbitrarily small. Moreover, we have required nothing of the Lipschitz constant of f, unlike the strict assumptions on the Lipschitz constant L of f by Chatziafratis et al. (2020) (e.g., they require $L = \phi$ for period $p = 3$). Indeed, Propositions 2 and 3 in the Appendix 9.6 illustrate how their lower bounds break down for large L and how their L_∞ bounds can shrink, becoming arbitrarily weak for certain 3-periodic f.*

We also present analogous results for the classification error and the L_1 errors. Please see the full statements in Theorems 4 and 5. Furthermore, Theorems 6 and 7 offer an improvement on the results of Chatziafratis et al. (2020) by giving constant-accuracy L_∞ lower bounds without needing a chaotic itinerary.

In addition, Section 4 relates our chaotic itineraries to standard notions of function complexity like the VC dimension and the topological entropy (for precise definitions, see Sec. 4). The types of periodic itineraries of f give rise to two regimes: the *doubling* regime and the *chaotic* regime. In the former, we have a polynomial number of oscillations, while the latter is characterized by an exponential number of oscillations. Here we show the following correspondence:

Theorem 3 (Informal)**.** *The transition between these two regimes exactly coincides with a sharp transition in the VC-dimension of the iterated mappings f^k for fixed f (from bounded to infinite) and in the topological entropy (from zero to positive).*

Our Techniques To quantify the oscillations of f^k, we use its chaotic itineraries to decompose the $[0, 1]$ interval into several subintervals $\{I_j\}_{j=1}^{j=p-1}$. We count the number of times f^k "visits" each I_j, by identifying a suitable matrix A whose spectral radius is a lower bound on the growth rate of oscillations. The associated characteristic polynomial of A is $\lambda^p - 2\lambda^{p-1} + 1$ and has larger spectral radius that that of prior works for *all* periods. Moreover, the corresponding oscillations of at least one of the subintervals I_j do not shrink in size, giving a bound on the total number of oscillations of a *sufficient* size. This provides a lower bound on the height between the peak and the bottom of these oscillations that later provides *constant* approx-

[6] e.g., Telgarsky (2015, 2016) analyzes only a restricted family of surjective triangle mappings constructed from neural networks with semi-algebraic gates.

[7] e.g., Chatziafratis et al. (2019, 2020) have a suboptimal dependence on p under stringent Lipschitz assumptions.

[8] e.g., Chatziafratis et al. (2019, 2020); Bu et al. (2020) do not obtain constant error rates.

[9] Our results also transfer to non-unimodal functions via the observation that for bimodal g, there is some unimodal f such that the number of oscillations of g is at most twice those of f.

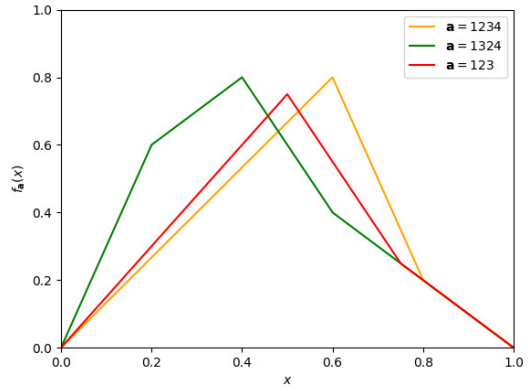

Figure 1: Plots of unimodal mappings with different itineraries f_{1234}, f_{1324}, and f_{123}. Despite their similarities, f_{1234} leads to the most oscillations and sharpest depth-width tradeoffs (see Fig. 2).

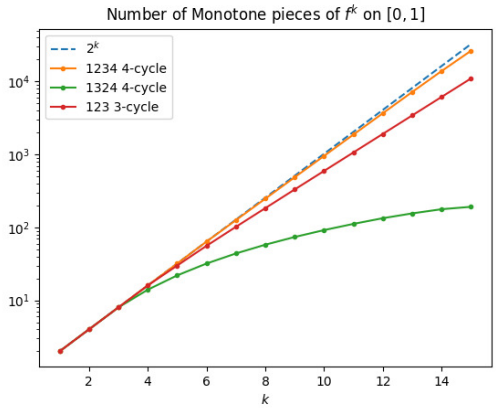

Figure 2: The chaotic itinerary f_{1234} has more oscillations than f_{123} even though $3 \triangleright 4$ by Sharkovsky's Theorem. Itineraries f_{1234} and f_{1324} (both of period 4) differ dramatically in oscillation count, showing why periodicity alone fails to capture the optimal tradeoffs.

imation errors for small shallow NNs.

More broadly, our work builds on the efforts to characterize large families of functions that give depth separations and addresses questions raised by Eldan and Shamir (2016); Telgarsky (2016); Poole et al. (2016); Malach and Shalev-Shwartz (2019) about the properties of hard-to-represent functions. Similar to periods, the concept of chaotic itineraries can serve as a certificate of complexity, which is also easy to verify for unimodal f (see Proposition 1 in Appendix).

2 Warm-up Examples

This section presents illustrative examples and instantiates our results for some simple cases. These highlight the limitations of exclusively considering periodicity of cycles alone—and not itineraries—when developing accurate oscillation/crossing bounds (see also Def. 5, 6) and sharp expressivity tradeoffs.

Consider the three unimodal mappings in Figure 1, $f_{\mathbf{a}}$ with itineraries $\mathbf{a} \in \{1324, 1234, 123\}$. Observe that f_{1234} has the cycle $(\frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5})$, f_{1324} has $(\frac{1}{5}, \frac{3}{5}, \frac{2}{5}, \frac{4}{5})$, and f_{123} has $(\frac{1}{4}, \frac{1}{2}, \frac{3}{4})$. Despite their similarities, they give rise to significantly different behaviours in $f^k_{\mathbf{a}}$.

What do prior works based on NN approximation with respect to periods and Sharkovsky's theorem alone tell us? Chatziafratis et al. (2019, 2020) show that the 3-cycle of f_{123} ensures that f^k has $\Omega(\phi^k)$ oscillations, where $\phi \approx 1.618$ is the golden ratio. However, their theorems do not imply anything for f_{1324} and f_{1234}, since 4 is a power of 2, and they require odd periods.

As it turns out, f_{1234} leads to exponential oscillations and f_{1324} leads only to polynomial oscillations:

- A mapping with a 1324-itinerary is guaranteed no other cycles except the 2-cycle and a fixed point (Metropolis et al., 1973). Sharkovsky's theorem and Chatziafratis et al. (2019) predict this outcome, since 4 is the third-right-most element of the Sharkovsky ordering, and its existence alone promises nothing more. The ordering of itineraries introduced by (Metropolis et al., 1973) (see Table 2 in Appendix) indicates that the particular 1324-itinerary only implies the periods 2 and 1, and confirms this intuition. We classify this itinerary as part of the *doubling regime* and prove in Theorem 8 that any f^k with a *maximal* 1324-itinerary (that is, there is no 8-cycle) cannot exhibit sharp depth-width tradeoffs: for any $\epsilon > 0$, there exists a 2-layer ReLU neural network g of width $O(\frac{k^3}{\epsilon})$ such that $L_\infty(f^k_{1324}, g) \leq \epsilon$.

- Going beyond Sharkovsky's theorem, a mapping with a 1234-itinerary—even though it is of period 4—it is guaranteed to contain a 3-cycle as well (see Table 2 in Appendix). Hence, "itinerary-1234 implies period-3, implies chaos," and f^k_{1234} has at least $\Omega(\phi^k)$ oscillations and is hard to approximate by small shallow NNs. Moreover, Theorem 4 and Table 1 show that f^k_{1234} actually has $\Omega(\rho^k)$ oscillations for $\rho \approx 1.839 > \phi$. A corollary is that any NN g of depth $\sqrt{k}$ and width $O(1.839^{\sqrt{k}})$ has $L_\infty(f^k_{1234}, g) = \Omega(1)$, which is a stronger separation (*constant* error) than the ones given by Chatziafratis et al. (2019, 2020).

The reverse is not true: Sharkovsky's Theorem guarantees that period-3 implies period-4, but the only 4-

cycle guaranteed by the theorem is actually the non-chaotic 1324-itinerary, already shown to lead to minimal function complexity.

Furthermore, as p increases, the existence of a chaotic itinerary $12\dots p$ on f ensures that f^k has $\Omega(\rho^k)$ oscillations for $\rho \to 2$.[10] Figure 2 demonstrates these differences in oscillations (by counting the number of monotone pieces of functions $f^k_{\mathbf{a}}$ with a maximal itinerary-$\mathbf{a}$). As indicated theoretically, the number of oscillations of f_{1324} is polynomially-bounded, while the others grow exponentially fast, with f_{1234} being closer to 2^k. Please see Appendix 6 for more such examples.

Generally, prior constructions where the oscillation count of f^k increase at a rate faster than ϕ^k were too narrow (including only the triangle map). Because f_{1234} breaks the barrier, we abstract away the details and point to chaotic itineraries as the main source of complexity, leading to sharper depth-width tradeoffs.

While periodicity tells a compelling story about why f^k_{123} is difficult to approximate, it fails to explain why f^k_{1234} is even more complex. The exponential-vs-polynomial gap in the function complexity of f_{1234} and f_{1324} depends solely on the order of the elements of the cycle and distinguishes functions that NNs can easily approximate from those they cannot.

The remainder of the paper addresses the question introduced here—when does the itinerary tell us much more than the length of the period—in a general context that explores a "hierarchy" of such chaotic itineraries, strengthens a host of NN inapproximability bounds (Sec. 3), and reveals tight connections with other complexity notions, like the VC-dimension and topological entropy (Sec. 4).

3 Depth-Width Tradeoffs via Chaotic Itineraries

We give our main hardness results on the inapproximability of functions generated by repeated compositions of f to itself when f has certain cyclic behavior. Section 3.2 applies insights about chaotic itineraries to prove constant L_∞ and L_1 lower bounds on the accuracy of approximating f^k when f has an increasing cycle. Section 3.3 strengthens previous bounds on the number of oscillations when f has an odd cycle, which is not necessarily increasing. Appendix 8 presents Table 3 that illustrates the key differences between results.

[10] Similarly to Telgarsky (2016), the optimal achievable rate is $\rho \le 2$ if we start with a unimodal f (e.g., tent map). If one used multimodal functions as a building block (e.g., starting with $f' = f^2$ or $f' = f^3$), we could achieve larger rates (e.g., 4 or 8 respectively).

3.1 Notation

To measure the function complexity of f^k, we count the number of times f^k oscillates. We employ two notions of oscillation counts. The first is relatively weak and counts every interval on which f is either increasing or decreasing, regardless of its size.

Definition 5. *Let $f : [0,1] \to [0,1]$. $M(f)$ represents the* number of monotone pieces *of f. That is, it is the minimum m such that there exists $x_0 = 0 < x_1 < \cdots < x_{m-1} < x_m = 1$ where f is monotone on $[x_{j-1}, x_j]$ for all $j \in [m]$.*

The second instead counts the number of times a fixed interval of size $b-a$ is crossed:

Definition 6. *Let $f : [0,1] \to [0,1]$ and $[a,b] \subseteq [0,1]$. $C_{a,b}(f)$ represents the* number of crossings *of f on the interval $[a,b]$. That is, it is the maximum c such that there exist*

$$0 \le x_1 < x'_1 \le x_2 < x'_2 \le \cdots \le x_c < x'_c \le 1$$

where for all $j \in [c]$, $f([x_j, x'_j]) \subset [a,b]$ and either $f(x_j) = a$ and $f(x'_j) = b$ or vice versa.

Characteristic Polynomials The base of the exponent of our width bounds is shown to equal the largest root of one of two polynomials:

$$P_{\mathrm{inc},p}(\lambda) = \lambda^p - 2\lambda^{p-1} + 1,$$
$$P_{\mathrm{odd},p}(\lambda) = \lambda^p - 2\lambda^{p-2} - 1.$$

Let $\rho_{\mathrm{inc},p}$ and $\rho_{\mathrm{odd},p}$ be the largest roots of $P_{\mathrm{inc},p}$ and $P_{\mathrm{odd},p}$ respectively. Table 1 illustrates that as p grows, $\rho_{\mathrm{inc},p}$ increases to 2, while $\rho_{\mathrm{odd},p}$ drops to $\sqrt{2}$. Note that $\rho_{\mathrm{odd},p} \in (\sqrt{2}, \sqrt{2 + 2/2^{p/2}})$ (Alsedà et al., 2000). We bound the growth rate of $\rho_{\mathrm{inc},p}$ with the following:

Fact 1. *$\rho_{\mathrm{inc},p} \in [\max(2 - \frac{4}{2^p}, \phi), 2)$, where $\phi = \frac{1+\sqrt{5}}{2}$ is the Golden Ratio.*

We prove Fact 1 in Appendix 9.2.

3.2 Inapproximability of Iterated Functions with Increasing Cycles

Our inapproximability results that govern the size of neural network g necessary to adequately approximate f^k when f has an increasing cycle (like Theorem 2) rely on a key lemma that bounds the number of constant-size oscillations of f^k.

Lemma 1 (Oscillation Bound for Increasing Cycles)**.** *Suppose f is a symmetric, concave unimodal mapping with an increasing p-cycle for some $p \ge 3$. Then, there exists $[a,b] \subset [0,1]$ with $b-a \ge \frac{1}{18}$ such that $C_{a,b}(f^k) \ge \frac{1}{2}\rho^k_{\mathrm{inc},p}$ for all $k \in \mathbb{N}$.*

Table 1: Approximate values of $\rho_{\text{inc},p}$, the lower bound on $\rho_{\text{inc},p}$ in Fact 1, and $\rho_{\text{odd},p}$ (for odd p).

p	$\rho_{\text{inc},p}$	Fact 1	$\rho_{\text{odd},p}$
3	1.618	1.618	1.618
4	1.839	1.75	n/a
5	1.928	1.875	1.513
6	1.966	1.938	n/a
7	1.984	1.969	1.466
8	1.992	1.984	n/a
9	1.996	1.992	1.441
10	1.999	1.996	n/a

We prove Lemma 1 in Appendix 9.1. For an increasing p-cycle $x_1, \dots, x_p$, we lower-bound $M(f^k)$ (the total number of monotone pieces, regardless of size) by relating the number of times f^k crosses each interval $[x_j, x_{j+1}]$ to the number of crossings of f^{k-1}. Doing so entails analyzing the largest eigenvalues of a transition matrix, which gives rise to the polynomial $P_{\text{inc},p}$. We prove that the intervals crossed must be sufficiently large due to the symmetry, concavity, and unimodality of f.

Remark 3. *If one does not wish to assume that f is unimodal, symmetric, or concave, then the proof can be modified to show that $C_{a,b}(f^k) = \Omega(\rho^k)$ for the same ρ, but for $b-a$ dependent on f. These results are similar in flavor to those of Chatziafratis et al. (2019, 2020); Bu et al. (2020), and they suffer from the same drawback: potentially vacuous approximation bounds when a and b are close. Appendix 9.6 shows natural functions that are either not symmetric or not concave, whose oscillations shrink in size arbitrarily.*

3.2.1 L_∞ Approximation and Classification

Our first result is a restatement of Theorem 2 that quantifies inapproximability in terms of both L_∞ and classification error, which are comparable to the respective results of Bu et al. (2020) and Chatziafratis et al. (2019).

Theorem 4. *Suppose f is a symmetric concave unimodal mapping with an increasing p-cycle for some $p \geq 3$. Then, any $k \in \mathbb{N}$ and $g \in \mathcal{N}(u, \ell)$ with $u \leq \frac{1}{8}\rho_{\text{inc},p}^{k/\ell}$ have $\|f^k - g\|_\infty = \Omega(1)$.*

Moreover, there exists S with $|S| = \frac{1}{2}\lfloor \rho_{\text{inc},p}^{k/\ell} \rfloor$ and $t \in (0,1)$ such that $\mathcal{R}_{S,t}(f^k, g) \geq \frac{1}{4}$.

The proof follows from our main Lemma 1 above and Theorem 10/Corollary 2 in the Appendix (two previous inapproximability bounds based on oscillations).

Despite relying on unimodality assumptions and the existence of increasing cycles, Theorem 4 obtains much stronger bounds than its previous counterparts:

- The assumption that f has an increasing cycle causes a much larger exponent base for the width bound. Chatziafratis et al. (2019, 2020) only prove that the existence of 3-cycle mandates a width of $\Omega(\phi^{k/\ell})$. We exactly match that bound for $p = 3$, and improve upon it when $p > 3$. As illustrated by Table 1, increasing p pushes the base $\rho_{\text{inc},p}$ rapidly to 2, which is the maximum exponent base for the increase of oscillations of any unimodal map. (And the maximal topological entropy of a unimodal map.) This also approximately matches the bases from Bu et al. (2020), which scale with the topological entropy of f.

- As illustrated in Appendix 9.6, the inaccuracy of neural networks with respect to the L_∞ approximation in Chatziafratis et al. (2019, 2020); Bu et al. (2020) may be arbitrarily small for certain choices of f. Our unimodality assumptions ensure that the oscillations of f^k are large and hence, that the inaccuracy of g is constant.

3.2.2 L_1 Approximation

We also strengthen the bound on L_1-inapproximability given by Chatziafratis et al. (2020) by again introducing a stronger exponent and applying unimodality to yield a constant-accuracy bound.

Theorem 5. *Consider any L-Lipschitz $f : [0,1] \to [0,1]$ with an increasing p-cycle for some $p \geq 3$. If $L = \rho_{\text{inc},p}$, then for any $k \in \mathbb{N}$, any $g \in \mathcal{N}(u, \ell)$ with $u \leq \frac{1}{16}\rho_{\text{inc},p}^{k/\ell}$ has $\|f^k - g\|_1 = \Omega(1)$.*

The proof follows again using our main Lemma 1 and using Theorem 11 in the Appendix.

We make Theorem 5 more explicit by showing that many tent maps meet the Lipschitzness condition. Let $f_{\text{tent},r} = 2r\min(x, 1-x)$ be the tent map, parameterized by $r \in (0,1)$. Our result improves upon Chatziafratis et al. (2020), by obtaining constant approximation error and using the larger $\rho_{\text{inc},p}$ rather than $\rho_{\text{odd},p}$.

Corollary 1. *For any $p \geq 3$ and $k \in \mathbb{N}$, any $g \in \mathcal{N}(u, \ell)$ with $u \leq \frac{1}{16}\rho_{\text{inc},p}^{k/\ell}$ has $\|f_{tent,\rho_{\text{inc},p}}^k - g\|_1 = \Omega(1)$.*

We prove Corollary 1 in Appendix 9.4. The only nontrivial part of the proof involves proving the existence of an increasing p-cycle that causes f^k to have $\Omega(\rho_{\text{inc},p}^k)$ oscillations.

3.3 Improved Bounds for Odd Periods

While Theorems 4 and 5 give stricter bounds on the width of neural networks needed to approximate iterated functions f^k than Chatziafratis et al. (2019, 2020), they also require extra assumptions about the cycles—namely, that the cycles are increasing. However, more powerful inapproximability results with constant error are still possible even without additional assumptions. Specifically, we leverage unimodality to improve the desired inaccuracy to a constant without compromising width.

As before, the results hinge on a key technical lemma that bounds the number of interval crossings.

Lemma 2. *For some odd $p \geq 3$, suppose f is a symmetric concave unimodal mapping with an odd p-cycle. Then, there exists $[a, b] \subset [0, 1]$ with $b - a \geq 0.07$ such that $C_{a,b}(f^k) = \rho_{\text{odd},p}^{k-p}$ for any $k \in \mathbb{N}$.*

We prove Lemma 2 in Appendix 9.5. The challenging part is to find a lower bound on the length of the intervals crossed.

Like before, we provide lower-bounds on approximation up to a constant degree.

Theorem 6. *For some odd $p \geq 3$, suppose f is a symmetric, concave unimodal mapping with any p-cycle. Then, any $k \in \mathbb{N}$ and any $g \in \mathcal{N}(u, \ell)$ with $u \leq \frac{1}{8}\rho_{\text{odd},p}^{(k-p)/\ell}$ have $\|f^k - g\|_\infty = \Omega(1)$.*

Moreover, there exists S with $|S| = \frac{1}{2}\lfloor \rho_{\text{odd},p}^k \rfloor$ and $t \in (0, 1)$ such that $\mathcal{R}_{S,t}(f^k, g) \geq \frac{1}{4}$.

The proof is immediate from Lemma 2, Theorem 10, and Corollary 2 in the Appendix.

We also get the analogous result but for the L_1 error:

Theorem 7. *Consider any L-Lipschitz $f : [0, 1] \to [0, 1]$ with a p-cycle for some odd $p \geq 3$. If $L = \rho_{\text{odd},p}$, then, any $k \in \mathbb{N}$ and $g \in \mathcal{N}(u, \ell)$ with $u \leq \frac{1}{16}\rho_{\text{odd},p}^{(k-p)/\ell}$ have $\|f^k - g\|_1 = \Omega(1)$.*

Remark 4. *We impose strict conditions on the Lipschitz constant because the bounds are vacuous or impossible for functions with other Lipschitz constants. By Lemma 3.1 of Chatziafratis et al. (2020), there are no L-Lipschitz interval mappings f whose iterates f^k have $\Omega(\rho_{\text{odd},p})^k$ oscillations when $L < \rho_{\text{odd},p}$. On the other hand, if $L > \rho_{\text{odd},p}$, then our proofs would yield vacuous lower bounds because they depend on $(\frac{\rho_{\text{odd},p}}{L})^k$, which is arbitrarily small for large k. See Section 3.1 of Chatziafratis et al. (2020) for a more thorough treatment of this issue.*

The proof is immediate from Lemma 1 and Theorem 11 in the Appendix.

4 Periods, Phase Transitions and Function Complexity

We formalize the correspondence between different notions of function complexity in dynamical systems and learning theory: neural network approximation, oscillation count, cycle itinerary, topological entropy, and VC-dimension. We make Theorem 3 rigorous by presenting two regimes into which unimodal mappings can be classified—the *doubling regime* and the *chaotic regime*—and show that all of these measurements of complexity hinge on which regime a function belongs to.[11]

The following two theorems split most of the space of unimodal mappings into one of two regimes and show that the doubling regime (so called because all cycles have power-of-two lengths and their itineraries are not chaotic) is intrinsically simpler from an approximation theoretic and a function complexity standpoint than the chaotic regime (where there exist chaotic itineraries). The pair of theorems combined known facts about approximation and topological entropy with new ideas about VC dimension. They support the claim that the phase transition that separates mappings with chaotic itineraries from those without is meaningful, because it also separates functions f^k that cannot be tractably approximated from those that can and separates highly expressive iterates f^k from those that cannot express complex data patterns.

Some components of the claims regarding the topological entropy are the immediate consequences of other results; however, we include them to give a complete picture of the gap between the two regimes. We believe the upper bound on monotone pieces of f^k in the doubling regime and both VC-dimension bounds below to be novel.

We define VC-dimension and introduce topological entropy in Appendix 10, along with the proofs of both theorems. For the VC-dimension, we consider the hypothesis class $\mathcal{H}_{f,t} := \{[[f^k]]_t : k \in \mathbb{N}\}$, which corresponds to the class of iterated fixed maps.

Theorem 8. *[Doubling Regime] Suppose f is a symmetric unimodal mapping whose maximal cycle is a primary cycle of length $p = 2^q$. That is, there exists a p-cycle but no $2p$-cycles (and thus, no cycles with lengths non-powers-of-two). Then, the following are true:*

[11] These two regimes correspond to different settings of the parameters r in the bifurcation diagram of Figure 7 in the Appendix. The doubling regime is the left-hand-side, where the stable periods routinely split in two before the first chaos is encountered. The chaotic regime is to the right-hand-side, which is characterized by chaos punctuated by intermittent stability.

1. *For any* $k \in \mathbb{N}$, $M(f^k) = O((4k)^{q+1})$.
2. *For any* $k \in \mathbb{N}$, *there exists* $g \in \mathcal{N}(u, 2)$ *with* $u = O((4k)^{q+1}/\epsilon)$ *such that* $\|g - f^k\|_\infty \leq \epsilon$. *Moreover, if* $f = f_{tent,r}$, *then there exists* $g \in \mathcal{N}(u, 2)$ *with* $u = O((4k)^{q+1})$ *and* $g = f^k$.
3. $h_{\text{top}}(f) = 0$.
4. *For any* $t \in (0, 1)$, $VC(\mathcal{H}_{f,t}) \leq 18p^2$.

The proof of Theorem 8 relies on a recursive characterization of f^k whenever f has a maximum cycle length of 2^q. To prove the first claim, we use this recursive structure to bound the number of monotone regions by relating the number of monotone regions of f^k to some g^{2k}, where g has a maximum cycle length no more than 2^{q-1}. The second and third claims are implications of the first. The fourth claim relies on a different recursive argument which shows that the family of iterated maps f^k for fixed f are unable to shatter certain subsets of points.

Theorem 9. *[Chaotic Regime] Suppose* f *is a unimodal mapping that has a* p*-cycle where* p *is not a power-of-two. Then, the following are true:*

1. *There exists some* $\rho \in (1, 2]$ *such that for any* $k \in \mathbb{N}$, $M(f^k) = \Omega(\rho^k)$.
2. *For any* $k \in \mathbb{N}$ *and any* $g \in \mathcal{N}(u, \ell)$ *with* $\ell \leq k$ *and* $u \leq \frac{1}{8}\rho^{k/\ell}$, *there exist samples* S *with* $|S| = \frac{1}{2}\lfloor \rho^k \rfloor$ *such that* $\mathcal{R}_{S,1/2}(f^k, g) \geq \frac{1}{4}$.
3. $h_{\text{top}}(f) \geq \rho > 0$.
4. *There exists a* $t \in (0, 1)$ *such that* $VC(\mathcal{H}_{f,t}) = \infty$.

Remark 5. *As discussed in Appendix 7, any non-primary cycle implies the existence of a cycle whose length is not a power of two. Thus, these results also apply if there exists any non-primary power-of-two cycle, such as the 1234-itinerary 4-cycle.*

The first three claims are implications of the proofs from previous sections of paper and previous works. The fourth claim relies on applying Sharkovsky's theorem to prove the existence of an infinitely large number of cycles with coprime lengths. Then, by considering a set of points each contained in a cycle of different coprime lengths, we show that a large number of iterates k is sufficient to "shatter" the points by realizing every possible labeling.

5 Conclusion

In this work, we build new connections between deep learning theory and dynamical systems by applying results from discrete-time dynamical systems to obtain novel depth-width tradeoffs for the expressivity of neural networks. While prior works relied on Sharkovsky's theorem and periodicity to provide families of functions that are hard-to-approximate with shallow neural networks, we go beyond periodicity. Studying the chaotic itineraries of unimodal mappings, we reveal subtle connections between expressivity and different types of periods, and we use them to shed new light on the benefits of depth in the form of enhanced width lower bounds and stronger approximation errors. More broadly, we believe that it is an exciting direction for future research to exploit similar tools and concepts from the literature of dynamical systems in order to improve our understanding of neural networks, e.g., their dynamics, optimization and robustness properties.

Acknowledgments

CS is grateful for the financial support from an NSF GRFP fellowship, NSF grant CCF-1563155, and Daniel Hsu's Google Faculty Research Award. VC is grateful for the hospitality of Northwestern University, as part of this work was done while VC was a postdoc at Northwestern under Samir Khuller. The authors thank Daniel Hsu, Ioannis Panageas, and five anonymous reviewers for their helpful comments.

References

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940, 2016.

Marvin Minsky and Seymour A Papert. *Perceptrons: An introduction to computational geometry*. MIT press, 1969.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are univer-

sal approximators. *Neural networks*, 2(5):359–366, 1989.

Matus Telgarsky. Representation benefits of deep feedforward networks. *arXiv preprint arXiv:1509.08101*, 2015.

Matus Telgarsky. benefits of depth in neural networks. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1517–1539, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

Hugh E Warren. Lower bounds for approximation by nonlinear manifolds. *Transactions of the American Mathematical Society*, 133(1):167–178, 1968.

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*, volume 9. cambridge university press Cambridge, 1999.

Michael Schmitt. Lower bounds on the complexity of approximating continuous functions by sigmoidal neural networks. In *Advances in neural information processing systems*, pages 328–334, 2000.

Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, pages 2924–2932, 2014.

Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. *arXiv preprint arXiv:1611.01491*, 2016.

Boris Hanin and David Rolnick. Deep relu networks have surprisingly few activation patterns. In *Advances in Neural Information Processing Systems*, pages 359–368, 2019.

Joe Kileel, Matthew Trager, and Joan Bruna. On the expressive power of deep polynomial neural networks. In *Advances in Neural Information Processing Systems*, pages 10310–10319, 2019.

Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

Amit Daniely. Depth separation for neural networks. In *Conference on Learning Theory*, pages 690–696. PMLR, 2017.

Holden Lee, Rong Ge, Tengyu Ma, Andrej Risteski, and Sanjeev Arora. On the ability of neural nets to express distributions. In *Conference on Learning Theory*, pages 1271–1296. PMLR, 2017.

Guy Bresler and Dheeraj Nagaraj. Sharp representation theorems for relu networks with precise dependence on depth. *arXiv preprint arXiv:2006.04048*, 2020.

Eran Malach and Shai Shalev-Shwartz. Is deeper better only when shallow is good? *arXiv preprint arXiv:1903.03488*, 2019.

Kaifeng Bu, Yaobo Zhang, and Qingxian Luo. Depth-width trade-offs for neural networks via topological entropy, 2020.

Itay Safran, Ronen Eldan, and Ohad Shamir. Depth separations in neural networks: what is actually being separated? In *Conference on Learning Theory*, pages 2664–2666. PMLR, 2019.

Daniel Hsu, Clayton Sanford, Rocco A Servedio, and Emmanouil-Vasileios Vlatakis-Gkaragkounis. On the approximation power of two-layer networks of random relus. *Conference on Learning Theory*, 2021.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in neural information processing systems*, pages 3360–3368, 2016.

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl Dickstein. On the expressive power of deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2847–2854. JMLR. org, 2017.

Vaggos Chatziafratis, Sai Ganesh Nagarajan, Ioannis Panageas, and Xiao Wang. Depth-width trade-offs for relu networks via sharkovsky's theorem. *arXiv preprint arXiv:1912.04378*, 2019.

Vaggos Chatziafratis, Sai Ganesh Nagarajan, and Ioannis Panageas. Better depth-width trade-offs for neural networks through the lens of dynamical systems. In *International Conference on Machine Learning*, pages 1469–1478. PMLR, 2020.

Lluís Alsedà, Jaume Llibre, and Michal Misiurewicz. *Combinatorial Dynamics and Entropy in Dimension One*. WORLD SCIENTIFIC, 2nd edition, 2000. doi: 10.1142/4205.

N Metropolis, M.L Stein, and P.R Stein. On finite limit sets for transformations on the unit interval. *Journal of Combinatorial Theory, Series A*, 15(1): 25 – 44, 1973. ISSN 0097-3165. doi: https://doi.org/10.1016/0097-3165(73)90033-2.

Tien-Yien Li and James A Yorke. Period three implies chaos. *The American Mathematical Monthly*, 82(10):985–992, 1975.

OM Sharkovsky. Coexistence of the cycles of a continuous mapping of the line into itself. *Ukrainskij matematicheskij zhurnal*, 16(01):61–71, 1964.

OM Sharkovsky. On cycles and structure of continuous mapping. *Ukrainskij matematicheskij zhurnal*, 17(03):104–111, 1965.

Michal Misiurewicz and Wieslaw Szlenk. Entropy of piecewise monotone mappings. *Studia Mathematica*, 67:45–63, 1980.

Lai-Sang Young. On the prevalence of horseshoes. *Transactions of the American Mathematical Society*, 263:75–88, 1981. ISSN 0002-9947.

Vladimir Naumovich Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of the frequencies of occurrence of events to their probabilities. In *Empirical Inference*, 2013.

Barkley Rosser. Explicit bounds for some functions of prime numbers. *American Journal of Mathematics*, 63(1):211–232, 1941. ISSN 00029327, 10806377.

Expressivity of Neural Networks via Chaotic Itineraries beyond Sharkovsky's Theorem: Supplementary Materials

6 Supplement for Section 2

Figures 3 and 5 demonstrate two emblematic cases where the differences in function complexity of f_{123}, f_{1234}, and f_{1324} are most evident. Both figures provide a function for each $f_{\mathbf{a}}$ that has a maximal itinerary of $\mathbf{a}$. (That is, there is no "higher-ranked" itinerary from Table 2 present in $f_{\mathbf{a}}$; all other cycles are induced by the existence of a cycle with itinerary $\mathbf{a}$.)

Figures 3 and 4 provide a simple case where the elements of the cycles are evenly spaced ($\frac{1}{4}, \frac{1}{2}, \frac{3}{4}$ for f_{123}; $\frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}$ for f_{1234}, f_{1324}). Despite the fact that f_{1234} and f_{1324} have the same maximum value, they exhibit substantially different fractal-like patterns, which produce exponentially more oscillations for f_{1234}.

Figure 5 and 6 instead considers logistic maps of the form $f_{\log,r}(x) = 4rx(1-x)$ for the values of r where itinerary $\mathbf{a}$ is *super-stable*, or when nearby iterates converge to the cycle exponentially fast. These functions are concave, symmetric, and unimodal. Here, complexity strictly increases with the maximum value of $f_{\log,r}$. Indeed, f_{1234}, f_{123} and f_{1324} ordered by height is the order by which they exhibit most to least chaotic behavior.

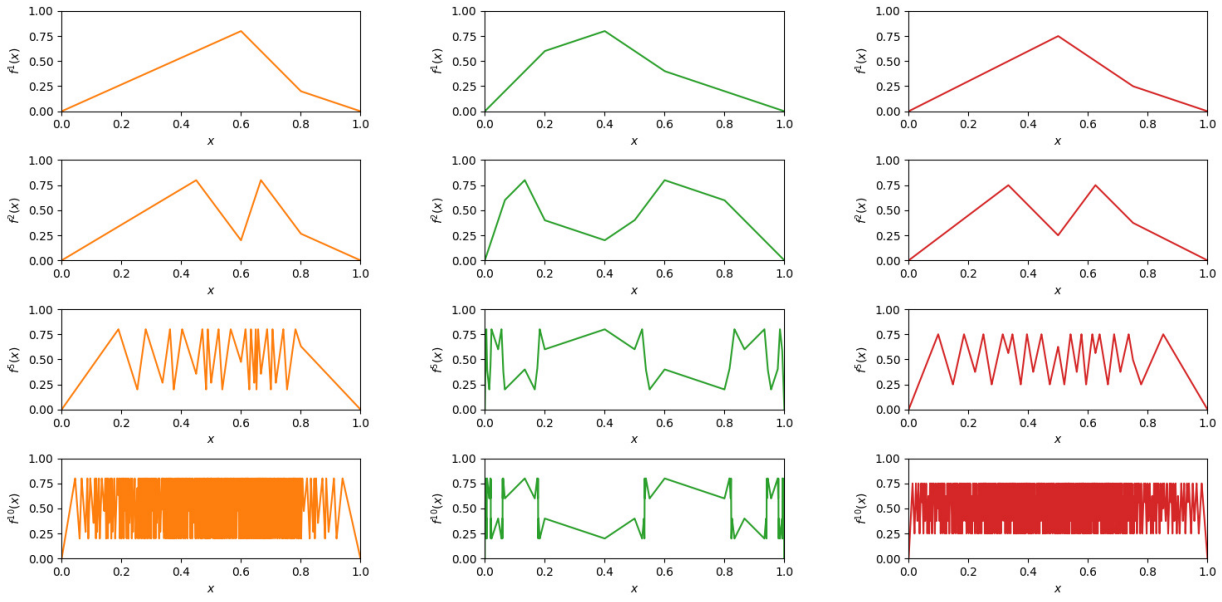

Figure 3: A comparison of the function complexity (as measured by the number of monotone pieces) of f^k for unimodal mappings f having cycles with different itineraries. The left shows f, f^2, f^5, and f^{10} for a function with a 1234 4-cycle. The center has a 1324 4-cycle. The bottom has a 123 3-cycle. Figure 4 shows how the number of monotone pieces of f^k increases with k for each mapping.

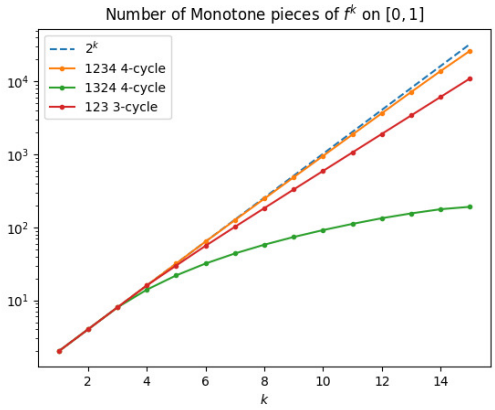

Figure 4: Visualizes the number of monotone pieces of f^k which increases with k for each mapping along with 2^k (the maximum number of monotone pieces of any unimodal f). Note that the 1234 itinerary produces a more "complex" function with more monotone pieces than 123, despite the Sharkovsky analysis from Chatziafratis et al. (2019) arguing that 3-cycles are the most powerful when determining iteration counts. Moreover, the number of monotone pieces of the 1234 and 123 itineraries increases exponentially, while that of the 1324 itineraries does not. (Identical to Figure 2.)

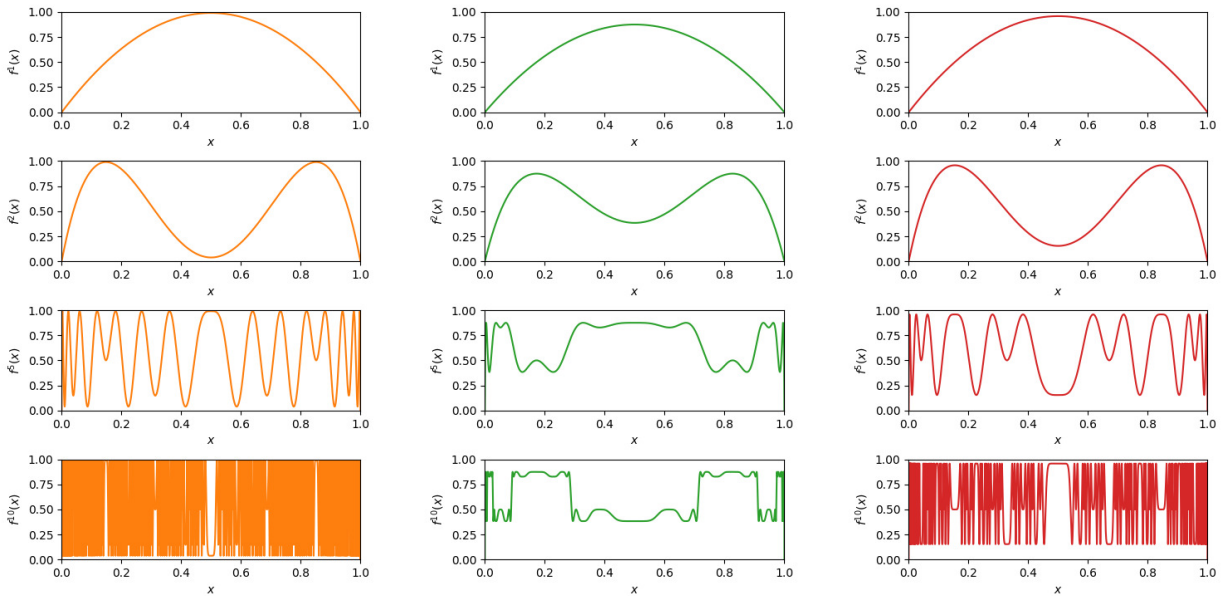

Figure 5: Demonstrates the same ideas as Figure 3, except instead of using asymmetric and non-concave piecewise functions, we use the scaled logistic map, $f_{\log,r}$. Using Table 1 of Metropolis et al. (1973), we set the parameter r to 3.96, 3.50, and 3.83 respectively to ensure that a super-stable 1234, 1324, and 123 cycle exists.

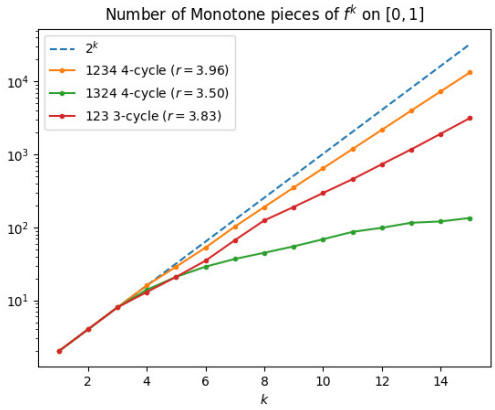

Figure 6: Like Figure 4, vizualizes the differences in number of monotone pieces for the logistic mappings described in Figure 5.

7 More Examples for Itineraries

7.1 Examples of Itineraries

Let the tent map and logistic map be defined by $f_{\text{tent},r}(x) = 2r\max(x, 1-x)$ and $f_{\text{log},r}(x) = 4rx(1-x)$ respectively, for parameter $r \in (0, 1)$.

Example 1. *For all $r \in (\frac{1}{2}, 1]$, there is a two-cycle C of itinerary 12 (which is the only itinerary for a 2-cycle) in $f_{tent,r}$ with*

$$C = \left(\frac{2r}{1+4r^2}, \frac{4r^2}{1+4r^2}\right).$$

Example 2. *When $r = \frac{1+\sqrt{5}}{4}$, there is a two-cycle C of $f_{log,r}$ with*

$$C = \left(\frac{1}{2}, \frac{1+\sqrt{5}}{4}\right).$$

Example 3. *When $r \in [\frac{1+\sqrt{5}}{4}, 1]$, $f_{tent,r}$ has a three-cycle C of itinerary 123 with*

$$C = \left(\frac{2r}{1+8r^3}, \frac{4r^2}{1+8r^3}, \frac{8r^3}{1+8r^3}\right).$$

Note that this and Example 1 are consistent with Sharkovsky's Theorem; whenever there exists a three-cycle, there also exists a two-cycle.

Example 4. *When $r \in [\frac{1}{2}, 1]$, there also exists a four-cycle C of itinerary 1324 for $f_{tent,r}$ with*

$$C = \left(\frac{8r^3-4r^2+2r}{16r^2+1}, \frac{16r^4-8r^3+4r^2}{16r^2+1}, \frac{16r^4-8r^3+2r}{16r^2+1}, \frac{16r^4-4r^2+2r}{16r^2+1}\right).$$

Again, this reaffirms Sharkovsky's Theorem, since this cycle always exists when the above three-cycle exists.

Example 5. *However, when $r \in (0.9196\ldots, 1]$, there also exists a four-cycle C of itinerary 1234 for $f_{tent,r}$ with*

$$C = \left(\frac{2r}{16r^2+1}, \frac{4r^2}{16r^2+1}, \frac{8r^3}{16r^2+1}, \frac{16r^4}{16r^2+1}\right).$$

This demonstrates a relationship beyond Sharkovsky's theorem: whenever a 1234 four-cycle exists, a 123 three-cycle also exists. This will be integral to the bounds we show.

Example 6. *The triangle map from Telgarsky (2016), $f_{tent,1}$ has an increasing p-cycle C_p for every $p \in \mathbb{N}$ with*

$$C_p = \left(\frac{2}{1+2^p}, \frac{2^2}{1+2^p}, \ldots, \frac{2^p}{1+2^p}\right).$$

Thus Theorem 10 and Fact 1 retrieve the fact used by Telgarsky that $M(f_{tent,1}) = \Omega(2^k)$.

Table 2: For any unimodal function f, let $f_r(x) := rf(x)$ for $r > 0$. As r increases, any such family obtains new cycles in the same order, and those cycles are super-stable in the same order. This translates Table 1 of Metropolis et al. (1973) to our notation and shows at what values of r, $f_{\log,r}$ has various super-stable cycles of length at most 6.

Cycle length p	Itinerary	Regime	r s.t. super-stable for $f_{\log,r}$	Cycle Type
2	12	Doubling	0.8090	Primary
4	1324	Doubling	0.8671	Primary
6	143526	Chaotic	0.9069	Primary
5	13425	Chaotic	0.9347	Stefan, Primary
3	123	Chaotic	0.9580	Stefan, Increasing, Primary
6	135246	Chaotic	0.9611	
5	12435	Chaotic	0.9764	
6	124536	Chaotic	0.9844	
4	1234	Chaotic	0.9901	Increasing
6	123546	Chaotic	0.9944	
5	12345	Chaotic	0.9976	Increasing
6	123456	Chaotic	0.9994	Increasing

7.2 Orderings of Itineraries

As has been mentioned before, the existence of some cycles can be shown to imply the existence of other cycles. Sharkovsky's Theorem famously does this by showing that if $p \triangleright p'$, then the existence of a p-cycle implies the existence of a p'-cycle. Proposition 1 can be used to imply that the existence of a chaotic p-cycle implies the existence of a chaotic $(p-1)$-cycle. These pose a broader question: Is there a complete ordering on all cycle itineraries that can appear in unimodal mappings? And does this ordering coincide with the amount of "chaos" induced by a cycle?

Researchers of discrete dynamical systems have thoroughly investigated these questions; we refer interested readers to Metropolis et al. (1973); Alsedà et al. (2000) for a more comprehensive survey. We introduce the basics of this theory as it relates to our results.

Metropolis et al. (1973) present a partial ordering over cyclic itineraries present in unimodal mappings, which serves as a measurement of the complexity of the function. That is, two itineraries $\mathbf{a}$ and $\mathbf{a}'$ may be related analogously to Sharkovsky's Theorem with $\mathbf{a} \triangleright \mathbf{a}'$, if f having itinerary $\mathbf{a}$ implies that f has itinerary $\mathbf{a}'$. This ordering for all cycles of length at most 6 is illustrated in Table 2. For instance, if a unimodal map has a cycle with itinerary 12435, then it also has a cycle with itinerary 135246.

We make several observations about the table and make connections to the itineraries discussed elsewhere in the paper.

- The table does not contradict Sharkovsky's Theorem. Note that $3 \triangleright 5 \triangleright 6 \triangleright 4 \triangleright 2$, and order in which the first itinerary occurs of a period is the same as the Sharkovsky ordering:

$$12 \triangleleft 1324 \triangleleft 143526 \triangleleft 13425 \triangleleft 123.$$

- The last cycle to occur for a given period is its increasing cycle and it occurs as p increases (not with the Sharkovsky ordering of p):

$$12 \triangleleft 123 \triangleleft 1234 \triangleleft 12345 \triangleleft 123456.$$

- The first cycle to appear for every odd period is its *Stefan cycle* (123, 13425). This is proved by Alsedà et al. (2000) and justifies why Theorem 6 relies on the existence of a Stefan cycle whenever there is an odd period.

- There exist cycles of power-of-two length (e.g. 1234) that induce non-power-of-two cycles (e.g. 123).

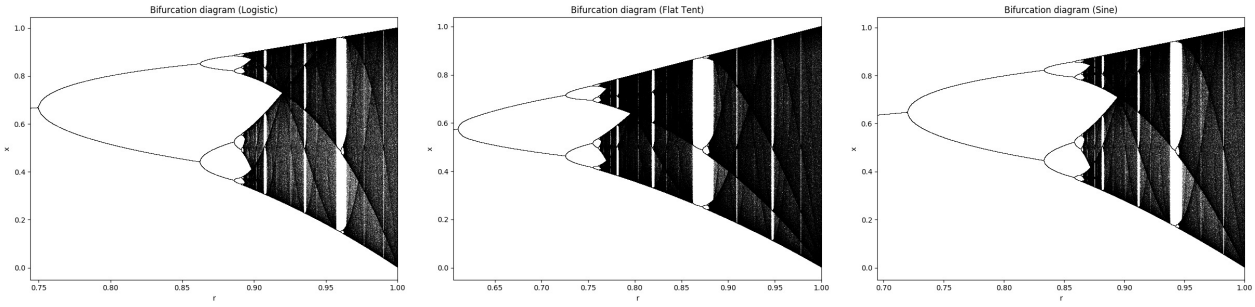

Figure 7: Bifurcation diagrams—which display the qualitative behavior of a family of functions f_r as the parameter $r \in [0,1]$ changes—showing the convergence behavior for iterates $f_r^k(x)$ for large k. For fixed r on the horizontal axis, the points plotted correspond to $f^k(x_0)$ for very large k. Regions of r where a vertical slice contains p discrete points indicates the existence of a *stable* p-cycle, since $f^k(x_0)$ converges exclusively to those points. Regions where the slice has a dispersed mass of points exhibit chaos. As r increases, cycles of different itineraries appear and experience stability in the same order indicated by Table 2. In the first plot, f_r is the logistic map $f_r(x) = f_{\log,r}(x) = 4rx(1-x)$. The second f_r is the"flat tent map," $f_r(x) = \min\{\frac{5rx}{2}, r, \frac{5rx}{2}(1-x)\}$, and the third is the sine map, $f_r(x) = r\sin(\pi x)$. The three are qualitatively identical and exhibit self-similarity.

Following the last bullet point, we distinguish between the 2^q-cycles that only induce cycles of length 2^i for $i < q$ and those that induce non-power-of-two cycles. To do so, we say that the itinerary of a p-cycle is *primary* if it induces no other p-cycle with a different itinerary.

We say that an itinerary $\mathbf{a}' = a'_1 \dots a'_{2p}$ of a $2p$-cycle is a *2-extension* of itinerary $\mathbf{a} = a_1 \dots a_p$ of a p-cycle if

$$a_i = \left\lceil \frac{a'_i}{2} \right\rceil = \left\lceil \frac{a'_{i+p}}{2} \right\rceil$$

for all i. For instance, 12 is a 2-extension of 1, 1324 is of 12, 15472638 is of 1324, and 135246 is of 123.

Theorem 2.11.1 of Alsedà et al. (2000) characterizes which itineraries are primary. It critically shows that a power-of-two cycle is primary if and only if it is composed of iterated 2-extensions of the trivial fixed-point itinerary 1. As a result, 1324 is a primary itinerary and 1234 is not. This sheds further light on the warmup example given in Section 2 and expanded upon in Appendix 6, where f^k_{1324} has a polynomial number of oscillations, while f^k_{1234} has an exponential number.

According to Theorem 2.12.4 of Alsedà et al. (2000), the existence a non-primary itinerary of any period implies the existence of some cycle with period not a power of two. Hence, f can *only* be in the doubling regime (where all periods are powers of two) if all of those power-of-two periods are primary. The existence of any non-primary power-of-two period (such as 1234 or 13726548) implies that the f is in the chaotic regime.

This ordering can also be visualized using the bifurcation diagrams in Figure 7. The diagram plots the convergent behavior of $f_r^k(x)$ for large k, where r is some parameter and reflects the complexity of the unimodal function f_r. (When $r = 0$, $f_r = 0$; when $r = 1$, $x_{\max} = 1$, and $C_{0,1}(f^k) = 2^k$.) As r increases, the number of oscillations of f_r^k increases and with it, new cycles are introduced. Each new cycle has a *stable* region over parameters r where $f_r^k(x)$ converges to the cycle, and the bifurcation diagram visualizes when each of these stable regions occurs. While the three functions families f_r have different underlying unimodal functions, they produce qualitatively identical bifurcation diagrams that feature the same ordering of itineraries.

Our discussions of the *doubling* and *chaotic* regimes in Section 4 are inspired by these bifurcation diagrams. Parameter values r are naturally partitioned into two categories: those on the left side of the diagram where the plot is characterized by a branching of cycles (the doubling regime) and those on the right side where there are extended regions of chaos, interrupted by small stable regions (the chaotic regime).

7.3 Identifying Increasing Cycles in Unimodal Maps

It is straightforward to determine whether a symmetric and unimodal f has an increasing p-cycle. Algorithmically, one can do so by verifying that $f(\frac{1}{2}) > \frac{1}{2}$ and counting how many consecutive values of $k \geq 2$ satisfy $f^k(x_0) < \frac{1}{2}$.

Proposition 1. *Consider some $p \geq 2$ and a symmetric unimodal mapping f. f has an increasing p-cycle if*

$$f^2\left(\frac{1}{2}\right) < \cdots < f^p\left(\frac{1}{2}\right) \leq \frac{1}{2} < f\left(\frac{1}{2}\right),$$

then f has an increasing p-cycle.

Proof. Refer to Figure 8 for a visualization of the variables and inequalities defined.

Let $x' = f(\frac{1}{2})$. By the unimodality of f and the fact that $x' > \frac{1}{2}$, there exists some $x'' > \frac{1}{2}$ such that

$$f(x'') < f^2(x'') < \cdots < f^{p-1}(x'') = \frac{1}{2}.$$

Because f is monotonically increasing on $[0, \frac{1}{2}]$, the following string of inequalities hold.

$$f(x') \leq f(x'') < f^2(x') \leq f^2(x'') < \cdots < f^{p-1}(x') \leq f^{p-1}(x'') = \frac{1}{2} \tag{1}$$

It then must hold that $x' \geq x''$.

Let $g(x) = f^p(x) - x$ and note that g is continuous. Because $\frac{1}{2}$ maximizes f, it must be the case that $f^p(x') \leq x'$ and $g(x') \leq 0$. Because $f^p(x'') = x'$ and $x'' \leq x'$, $g(x'') \geq 0$. Hence, there exists $x^* \in [x'', x']$ such that $g(x^*) = 0$ and $f^p(x^*) = x^*$.

Since $x^* \in [x'', x']$, it must also be the case that $f^j(x^*) \in [f^j(x'), f^j(x'')]$ for $j \in [p-1]$. By Equation (1), it follows that

$$f(x^*) < f^2(x^*) < \cdots < f^{p-1}(x^*) < f^p(x^*) = x^*.$$

Hence, there exists an increasing p-cycle. □

8 Comparison with Prior Works

Given the large number of results presented in this paper and the many axes of comparison one can draw between these results and their predecessors in Telgarsky (2016); Chatziafratis et al. (2019, 2020), we provide Table 3 to illuminate these comparisons. It reinforces our key contributions, namely that (1) the presence of increasing cycles makes a function more difficult to approximate than a 3-cycle alone; (2) requiring that f satisfy unimodality constraints gives lower-bounds to constant accuracy that cannot be made vacuous by adversarial choices of f; and (3) the key distinction between "hard" and "easy" functions is the existence of non-primary power-of-two cycles.

We provide context for each column to clarify what its cells mean and how to compare their values.

- "**Condition**" specifies what must be true of the complexity of f in order for the relevant bounds to occur. All but the latter two conditions describe a very broad array of functions, while the last two focus only on a restricted subset of tent mappings.
  - "Maximal PO2" means that the maximal cycle of f is a primary[12] p-cycle where p is a power of two. This means that f lies in the doubling regime described in Theorem 8.
  - "$h_{\text{top}}(f) \geq \rho$" considers any f with a lower-bound on its topological entropy for some $\rho > 1$. Notably, all conditions other than "Maximal PO2" satisfy this for some ρ.
  - "Non-primary" means that any non-primary cycle exists in f. That is, if f is known to have a non-primary power-of-two cycle, then the results apply.
  - "Non-PO2" refers to any f that has a p-cycle where p is not a power of two.
  - "Odd cycle" includes any f that has a p-cycle where p is odd.

[12] See Appendix 7.2.

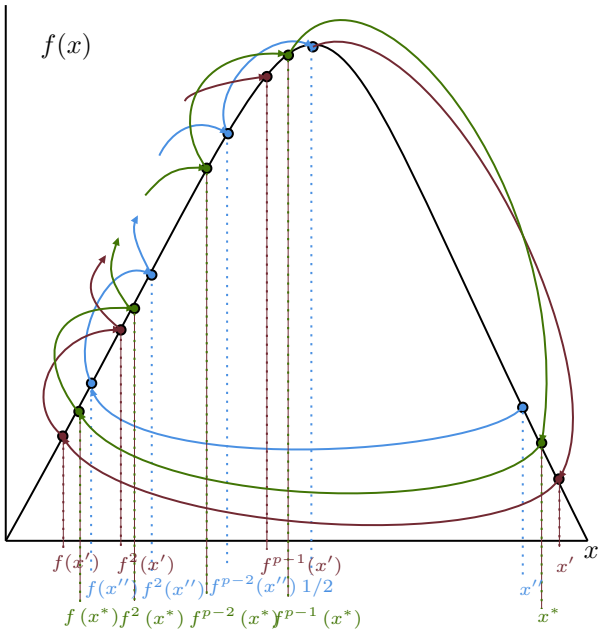

Figure 8: Visualizes the proof of Proposition 1.

	Condition	Approx.	Unimodal?	Concave?	Symmetric?	$L \leq \rho$?	Acc.	Exp.	Hard?	Source
1	**Maximal PO2**	L_∞	**Yes**	**No**	**Yes**	**No**	$\Omega(1)$	**Any**	**No**	**Thm 8**
2	$h_{\text{top}}(f) \geq \rho$	L_∞	No	No	No	No	$\epsilon(f)$	ρ	Yes	BZL Thm 16
3	**Non-primary**	**Cls.**	**No**	**No**	**No**	**No**	$\frac{1}{4}$	$(1, \phi]$	**Yes**	**CNPW Thm 1.6, Remark 5**
4	**Non-primary**	L_∞	**No**	**No**	**No**	**No**	$\epsilon(f)$	$(1, \phi]$	**Yes**	**CNPW Thm 1.6, Remark 5, BZL Thm 16**
5	Non-PO2	Cls.	No	No	No	No	$\frac{1}{4}$	$(1, \phi]$	Yes	CNPW Thm 1.6
6	Non-PO2	L_∞	No	No	No	No	$\epsilon(f)$	$(1, \phi]$	Yes	CNPW Thm 1.6, BZL Thm 16
7	Odd cycle	Cls.	No	No	No	No	$\frac{1}{4}$	$(\sqrt{2}, \phi]$	Yes	CNP Thm 1.1
8	Odd cycle	L_∞	No	No	No	No	$\epsilon(f)$	$(\sqrt{2}, \phi]$	Yes	CNP Thm 1.1, BZL Thm 16
9	**Odd cycle**	L_∞	**Yes**	**Yes**	**Yes**	**No**	$\Omega(1)$	$(\sqrt{1}, \phi]$	**Yes**	**Thm 6**
10	Odd cycle	L_1	No	No	No	Yes	$\epsilon(f)$	$(\sqrt{2}, \phi]$	Yes	CNP Thm 1.2
11	$f_{\text{tent},\rho_p/2}$	L_1	Implied	Implied	Implied	Implied	$\Omega(1)$	$(\sqrt{2}, \phi]$	Yes	CNP Lemma 3.6
12	**Odd cycle**	L_1	**Yes**	**Yes**	**Yes**	**Yes**	$\Omega(1)$	$(\sqrt{2}, \phi]$	**Yes**	**Thm 7**
13	**Inc. Cycle**	**Cls.**	**No**	**No**	**No**	**No**	$\frac{1}{4}$	$[\phi, 2)$	**Yes**	**Thm 4, Remark 3**
14	**Inc. Cycle**	L_∞	**No**	**No**	**No**	**No**	$\epsilon(f)$	$[\phi, 2)$	**Yes**	**Thm 4, Remark 3**
15	**Inc. Cycle**	L_∞	**Yes**	**Yes**	**No**	**No**	$\Omega(1)$	$[\phi, 2)$	**No**	**Prop 2**
16	**Inc. Cycle**	L_∞	**Yes**	**No**	**Yes**	**No**	$\Omega(1)$	$[\phi, 2)$	**No**	**Prop 3**
17	**Inc. Cycle**	L_∞	**Yes**	**Yes**	**Yes**	**No**	$\Omega(1)$	$[\phi, 2)$	**Yes**	**Thm 4**
18	**Inc. Cycle**	L_1	**No**	**No**	**No**	**Yes**	$\epsilon(f)$	$[\phi, 2)$	**Yes**	**Thm 5, CNP Thm 1.2**
19	**Inc. Cycle**	L_1	**Yes**	**Yes**	**Yes**	**Yes**	$\Omega(1)$	$[\phi, 2)$	**Yes**	**Thm 5**
20	$f_{\textbf{tent},\rho_p/2}$	L_1	**Implied**	**Implied**	**Implied**	**Implied**	$\Omega(1)$	$[\phi, 2)$	**Yes**	**Cor 1**
21	$f_{\text{tent},1}$	L_1	Implied	Implied	Implied	Implied	$\Omega(1)$	2	Yes	Telgarsky

Table 3: Compares the conditions and limitations of the theoretical results presented in this paper and its predecessors. New results are bolded.

  - "Inc. cycle" means that f has an increasing p-cycle for some p, i.e. a cycle with itinerary $12\dots p$.
  - $f_{\text{tent},\rho_p/2}$ refers to families of tent maps scaled by ρ_p solving the polynomials from Chatziafratis et al. (2020) Lemma 3.6 (for odd periods) and Corollary 1 (for increasing cycles).
  - The last row refers exclusively to the tent map of height 1 and slope 2.

- "**Approx.**" refers to how difference between neural network g and iterated map f^k is measured. The options are L_1, L_∞, and classification error. It's easier to show that g can L_1-approximate f^k than it is to show that g can L_∞-approximate f; conversely, it's most impressive to show lower bound results with respect to the L_1 error than it is for the L_∞ error.

  Chatziafratis et al. (2019, 2020) consider classification error, Bu et al. (2020) focus on L_∞ approximation, and Chatziafratis et al. (2020) also consider L_1 approximation. We routinely translate classification errors to L_∞ errors using Corollary 2, which draws on Theorem 16 of Bu et al. (2020).

- "**Unimodal?**," "**Concave?**," and "**Symmetric?**," have "Yes" if and only if f must meet the respective property for the proof to hold. They have "Implied" if the value of "Condition" already ensures that the property is satisfied and the requirement need not be enforced.

- "$L \le \rho$?" is "Yes" if the results only hold if f is chosen with a Lipschitz constant less than the rate of growth of its oscillations. This is a very restrictive condition met by very few functions (including no logistic maps with cycles).

- "**Acc.**" specifies the desired accuracy of the hardness result. "$\Omega(1)$" means that there exists some constant ϵ such that for any choice of f in the category, any neural network g will be unable to approximate f up to accuracy ϵ. "$\epsilon(f)$" means that the degree of approximation may depend on the chosen function f (and the period p) that belongs to the category; these bounds may be vacuous by an adversarial choice of f. As a result, hardness results with "$\Omega(1)$" are more impressive.

- "**Exp.**" refers to the base of the exponent of the lower-bound on the width necessary to approximate f^k using a shallow network g. Larger values indicate stronger bounds.

- "**Hard?**" is "Yes" if for every f satisfying the conditions to the left, f cannot be approximated up to the specified accuracy by any neural network g. It is "No" if there exists some f satisfying the conditions that can be approximated to a stronger degree of accuracy.

- "**Source**" denotes where to find the result. Some of the less interesting results are not given their own theorems and rather are immediate implications of several theorems across this body of literature. For the sake of space, we use "CNPW" to refer to (Chatziafratis et al., 2019); "CNP" for (Chatziafratis et al., 2019); "BZL" for (Bu et al., 2020); and "Telgarsky" for (Telgarsky, 2016).

9 Additional Proofs for Section 3

9.1 Proof of Lemma 1

We restate and prove the lemma. This is the main technical lemma that we use to get the sharper depth-width tradeoffs and the improved notion of *constant* approximation.

Lemma 1 (Oscillation Bound for Increasing Cycles)**.** *Suppose f is a symmetric, concave unimodal mapping with an increasing p-cycle for some $p \ge 3$. Then, there exists $[a,b] \subset [0,1]$ with $b-a \ge \frac{1}{18}$ such that $C_{a,b}(f^k) \ge \frac{1}{2}\rho_{\text{inc},p}^k$ for all $k \in \mathbb{N}$.*

Proof. We first lower-bound the total number of oscillations that will appear an increasing p-cycle is present. Later, we show that the size of the oscillations is large as well.

Because we have an increasing cycle of itinerary $12\dots p$, we assume (wlog) that the cycle is $(x_1,\dots,x_p)$ with $x_1 < x_2 < \cdots < x_p$. Define intervals $I_j := [x_j, x_{j+1}]$ for $j \in \{1,\dots,p-1\}$. Because f is continuous, we conclude that $I_{j+1} \subset f(I_j)$ for all $j < p$ and $I_j \subset f(I_{p-1})$ for all j. Figure 9 visualizes these relationships.

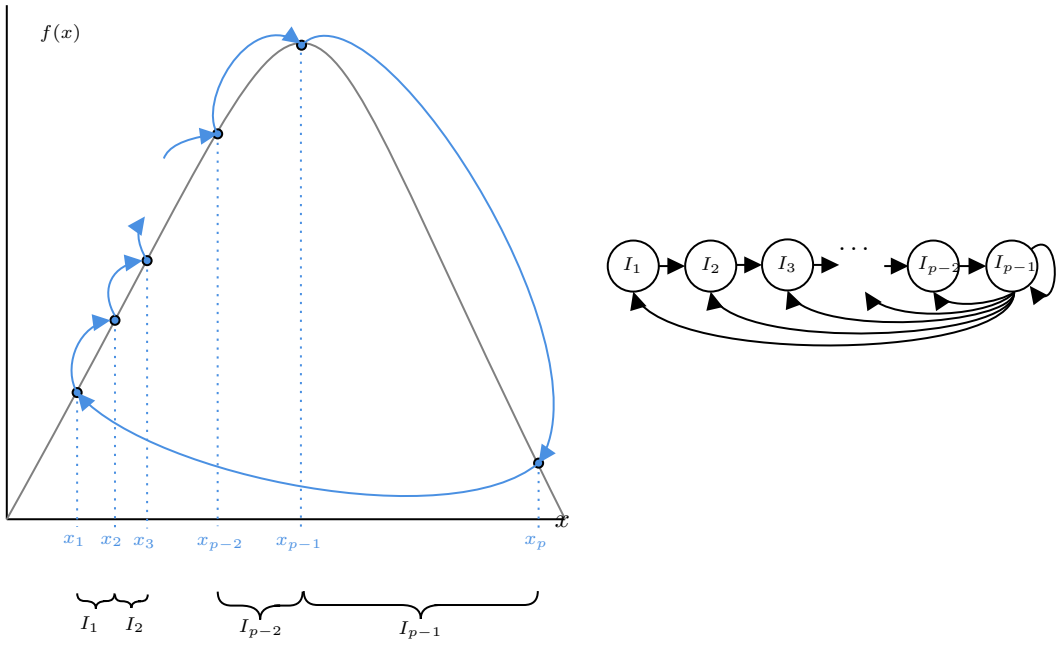

Figure 9: Visualizes the intervals $I_1, \ldots, I_{p-1}$ defined in the proof of Lemma 1 and which intervals f maps to one another when f has an increasing p-cycle.

Using the methods of Chatziafratis et al. (2019), we define $y^{(k)} \in \mathbb{N}^{p-1}$ such that $y_j^{(k)}$ is a lower bound on the number of times f^k passes through interval I_j, or

$$C_{x_j, x_{j+1}}(f^k) \geq y_j^{(k)}.$$

We can then encode the interval relationships above with $y^{(k+1)} = A_p y^{(k)}$ where $y^{(0)}$ is a vector of all ones and and $A_p \in \{0,1\}^{(p-1)\times(p-1)}$ with $(A_p)_{i,j} = \mathbb{1}\{j = p-1 \text{ or } i = j+1\}$. We get the following adjacency matrix for the intervals, capturing the mapping relationships (under f) between them:

$$A_p = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 1 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 1 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 1 \end{bmatrix}.$$

We find the characteristic polynomial of A_p and lower-bound $y^{(k+1)}$ with the spectral radius of A_p. We show by induction on $p \geq 3$ that

$$\det(A_p - \lambda I) = (-1)^{p-1}\left(\lambda^{p-1} - \sum_{i=0}^{p-2} \lambda^i\right).$$

For the base case $p = 3$, we have:

$$\det(A_3 - \lambda I) = \begin{vmatrix} -\lambda & 1 \\ 1 & 1-\lambda \end{vmatrix} = \lambda^2 - \lambda - 1,$$

which satisfies the desired form.

Now, we show the inductive step by expanding the determinant of $A_p - \lambda I$.

$$\det(A_p - \lambda I) = \begin{vmatrix} -\lambda & 0 & 0 & \cdots & 0 & 1 \\ 1 & -\lambda & 0 & \cdots & 0 & 1 \\ 0 & 1 & -\lambda & \cdots & 0 & 1 \\ 0 & 0 & 1 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda & 1 \\ 0 & 0 & 0 & \cdots & 1 & 1-\lambda \end{vmatrix} = -\lambda \begin{vmatrix} -\lambda & 0 & \cdots & 0 & 1 \\ 1 & -\lambda & \cdots & 0 & 1 \\ 0 & 1 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -\lambda & 1 \\ 0 & 0 & \cdots & 1 & 1-\lambda \end{vmatrix} - \begin{vmatrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & -\lambda & \cdots & 0 & 1 \\ 0 & 1 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -\lambda & 1 \\ 0 & 0 & \cdots & 1 & 1-\lambda \end{vmatrix}.$$

The left determinant exactly equals $\det(A_{p-1} - \lambda I)$, which we can expand using the inductive hypothesis. The second equals $(-1)^{p-2}$, because $p-2$ row swaps (which are elementary row operations) can be used to move the first row to the bottom and make the matrix upper-triangular with diagonals of one. We conclude the inductive step below.

$$\begin{aligned} \det(A_p - \lambda I) &= -\lambda \det(A_{p-1} - \lambda I) - (-1)^{p-2} \\ &= -\lambda(-1)^{p-2}\left(\lambda^{p-2} - \sum_{i=0}^{p-3} \lambda^i\right) + (-1)^{p-1} = (-1)^{p-1}\left(\lambda^{p-1} - \sum_{i=0}^{p-2} \lambda^i\right). \end{aligned}$$

We find the eigenvalues of A_p by finding the roots of the polynomial

$$P(x) = \lambda^{p-1} - \sum_{i=0}^{p-2} \lambda^i = 0.$$

Observe that there must be a root greater than 1 because $P(1) = 2 - p < 0$ and $P(2) = 1 > 0$. Equivalently, if $\lambda \neq 1$,

$$P(x) = \lambda^{p-1} - \frac{1-\lambda^{p-1}}{1-\lambda} = \frac{\lambda^p - 2\lambda^{p-1} + 1}{\lambda - 1} = 0.$$

Hence, finding the largest root of P is equivalent to finding the largest root of $\lambda^p - 2\lambda^{p-1} + 1$, which is $\rho_{\text{inc},p}$ by definition.

This implies that the spectral radius of A_p, $\text{sp}(A_p) = \rho_{\text{inc},p} > 1$, and hence, we also have $\text{sp}(A_p^k) = \text{sp}(A_p)^k = \rho_{\text{inc},p}^k$. Since all the elements in A_p and in A_p^k are non-negative, then the infinity norm of A_p^k is by definition the maximum among its row sums. Since the last column of A_p is the all 1's vector, the largest row sum in A_p^k appears at its last row:

$$||A_p^k||_\infty = \sum_{j=1}^{p-1} (A_p^k)_{p-1,j}$$

We can now use the fact that the infinity norm of a matrix is larger than its spectral norm:

$$||A_p^k||_\infty \geq \rho_{\text{inc},p}^k$$

We conclude that there exists at least one interval I_{j^*} (e.g., the interval I_{p-1}) which is crossed at least $\rho_{\text{inc},p}^k$ times by f^k, so $C_{x_{j^*}, x_{j^*+1}}(f^k) \geq \rho_{\text{inc},p}^k$.

Thus, for some a', b' we get $C_{a',b'}(f^k) \geq \rho_{\text{inc},p}^k$. But can we find a', b' with large difference $b' - a'$?

Now, we show that the intervals traversed are sufficiently large, in order to lower-bound $C_{a,b}(f^k)$ with $b - a \geq \frac{1}{18}$. By Lemma 3, there exists some j with $x_{j+1} - x_j \geq \frac{1}{18}$. It suffices to show that f^k traverses the interval I_j sufficiently many times.

From earlier in the proof, there exists some j^* such that f crosses I_{j^*} at least $N := \rho_{\text{inc},p}^k$ times. We conclude by showing that every other interval is traversed at least half as often as this most popular interval, which suggests that $C_{x_j,x_{j+1}}(f) \geq \frac{N}{2}$.

For $A \in \mathbb{R}^{(p-1)\times(p-1)}$ as defined earlier in the section and for $y^{(k)} := A^k\vec{1}$, we argue inductively that the elements of $y^{(k)}$ are non-decreasing and that $y_{p-1}^{(k)} \leq 2y_1^{(k)}$. For the base case, this is trivially true for $k = 0$.

Suppose it holds for k. By construction, we have $y_1^{(k+1)} = y_{p-1}^{(k)}$ and $y_j^{(k+1)} = y_{j-1}^{(k)} + y_{p-1}^{(k)}$ for all $j > 1$. By the inductive hypotheses,

$$y_1^{(k+1)} \leq y_2^{(k+1)} \leq \cdots \leq y_{p-1}^{(k+1)} \leq 2y_1^{(k+1)}.$$

Therefore, f^k crosses interval I_j at least $\frac{N}{2}$ times, and I_j has width at least $\frac{1}{18}$. The claim immediately follows. □

Lemma 3. *For some $p \geq 3$, consider a symmetric concave unimodal function f with an increasing p-cycle of $x_1 < \cdots < x_p$. Then, there exists $j \in [p-1]$ such that $x_{j+1} - x_j \geq \frac{1}{18}$.*

Proof. By the continuity of f, note that $[x_1, x_p] \subset f^3([x_{p-3}, x_{p-2}])$. There then exists some $y_1 \in [x_{p-3}, x_{p-2}]$ such that $f^3(y_1) = y_1$, $y_2 := f(y_1) \in [x_{p-2}, x_{p-1}]$, and $y_3 := f(y_2) \in [x_{p-1}, x_p]$. Thus, if f has a maximal p-cycle, then f also has a 3-cycle corresponding to $x_{p-3} < y_1 < y_2 < y_3 < x_p$.

We now show that $y_3 - y_1$ must be sufficiently large by concavity. For f to be concave, the following inequality must hold:

$$\frac{f(y_1) - f(0)}{y_1 - 0} \geq \frac{f(y_2) - f(y_1)}{y_2 - y_1} > 0 > \frac{f(y_3) - f(y_2)}{y_3 - y_2} \geq \frac{f(1) - f(y_3)}{1 - y_3},$$

or equivalently,

$$\frac{y_2}{y_1} \geq \frac{y_3 - y_2}{y_2 - y_1} > 0 > -\frac{y_3 - x_1}{y_3 - y_2} \geq -\frac{y_1}{1 - y_3}.$$

In addition, note that $y_1 < \frac{1}{2}$ and $y_3 > \frac{1}{2}$. If the former were false, then $f(y_2) \leq f(y_1)$ (by unimodality), which contradicts $y_3 > y_2$. If the latter were false, then $f(y_3) > f(y_2)$, which contradicts $y_1 < y_3$.

We consider two cases and show that either way, the interval must have width at least $\frac{1}{6}$.

- If $y_2 - y_1 \leq \frac{2}{5}(y_3 - y_1)$, then $\frac{y_3 - y_2}{y_2 - y_1} \geq \frac{3}{2}$, which mandates that $y_1 \leq \frac{2y_2}{3}$ to ensure concavity. Thus,

$$y_3 - y_1 \geq y_3 - \frac{2y_2}{3} \geq \frac{y_3}{3} \geq \frac{1}{6}.$$

- If $y_2 - y_1 \geq \frac{2}{5}(y_3 - y_1)$, then $\frac{y_3 - y_1}{y_3 - y_2} \geq \frac{5}{3}$, and thus $y_1 \geq \frac{5}{3}(1 - y_3)$ and $y_3 \geq 1 - \frac{3y_1}{5}$. Then,

$$y_3 - y_1 \geq 1 - \frac{3y_1}{5} - y_1 = 1 - \frac{8y_1}{5} \geq \frac{1}{5}.$$

Thus, we must have

$$\max\{x_{p-2} - x_{p-3}, x_{p-2} - x_{p-1}, x_p - x_{p-1}\} \geq \frac{1}{18}.$$ □

9.2 Proof of Fact 1

Fact 1. *$\rho_{\text{inc},p} \in [\max(2 - \frac{4}{2^p}, \phi), 2)$, where $\phi = \frac{1+\sqrt{5}}{2}$ is the Golden Ratio.*

Proof. Let $P_{\text{inc},p}(\lambda) = \lambda^p - 2\lambda^{p-1} + 1$.

First, observe that $\rho_{\text{inc},p} < 2$, because $P_{\text{inc},p}(\lambda) > 0$ whenever $\lambda \geq 2$. We lower-bound $\rho_{\text{inc},p}$ by finding some λ for each p such that $P_{\text{inc},p}(\lambda) \leq 0$ or equivalently $\lambda^{p-1}(2 - \lambda) \geq 1$ for all $p \geq 3$, which bounds $\rho_{\text{inc},p}$ by the Intermediate Value Theorem.

Consider $\lambda = 2 - \frac{4}{2^p}$. Then,

$$\begin{aligned}\lambda^{p-1}(2-\lambda) &= \left(2 - \frac{4}{2^p}\right)^{p-1} \cdot \frac{4}{2^p} = 2\left(1 - \frac{2}{2^p}\right)^{p-1} \\ &\geq 2\left(1 - \frac{2(p-1)}{2^p}\right) = 2 - 2 \cdot \frac{p-1}{2^{p-1}} \\ &\geq 2 - 2 \cdot \frac{1}{2} = 1.\end{aligned}$$

□

9.3 Previous Results about Hardness of Approximating Oscillatory Functions

We rely on prior results from Chatziafratis et al. (2019, 2020) to show that an iterated function f^k is inapproximable by neural networks. These results hold if f^k has sufficiently many crossings of some interval. We apply these results later with improved bounds on both the number and the size of crossings.

Chatziafratis et al. (2019) show that the classification error of f^k can be bounded if there are enough oscillations.

Theorem 10 ((Chatziafratis et al., 2019), Section 4)**.** *Consider any continuous $f : [0,1] \to [0,1]$ and any $g \in \mathcal{N}(u,\ell)$. Suppose there exists $a < b$ such that $C_{a,b}(f) = \Omega(\rho^t)$ and suppose $u \leq \frac{1}{8}\rho^{k/\ell}$. Then, for $t = \frac{a+b}{2}$, there exists S with $|S| = \frac{1}{2}\lfloor \rho^k \rfloor$ samples such that*

$$\mathcal{R}_{S,t}(f^k, g) \geq \frac{1}{2} - \frac{(2u)^\ell}{n}.$$

We adapt that claim to lower-bound the L_∞ approximation of f^k by g.

Corollary 2. *Consider any continuous $f : [0,1] \to [0,1]$ and any $g \in \mathcal{N}(u,\ell)$. Suppose there exists $a < b$ such that $C_{a,b}(f) = \Omega(\rho^t)$ and suppose $u \leq \frac{1}{8}\rho^{k/\ell}$. Then,*

$$\left\|f^k - g\right\|_\infty \geq \frac{b-a}{2}.$$

Proof. By Theorem 10, there exists some $x \in [0,1]$ such that (wlog) $f^k(x) \leq a$ and $g(x) \geq \frac{a+b}{2}$. The conclusion for the L_∞ error is immediate by definition. □

Chatziafratis et al. (2020) give a lower-bound on the ability of a neural network g to L_1-approximate f^k, provided a correspondence between the Lipschitz constant of f and the rate of oscillations ρ.

Theorem 11 (Chatziafratis et al. (2020) Theorem 3.2)**.** *Consider any L-Lipschitz $f : [0,1] \to [0,1]$ and any $g \in \mathcal{N}(u,\ell)$. Suppose there exists $a < b$ such that $C_{a,b}(f) = \Omega(\rho^t)$. If $L \leq \rho$ and $u \leq \frac{1}{16}\rho^{k/\ell}$, then*

$$\left\|f^k - g\right\|_1 = \Omega((b-a)^2).$$

The Lipschitzness assumption is extremely strict, especially because they show in their Lemma 3.1 that $L \geq \rho$ whenever f has a period of odd length.

9.4 Proof of Corollary 1

Corollary 1. *For any $p \geq 3$ and $k \in \mathbb{N}$, any $g \in \mathcal{N}(u,\ell)$ with $u \leq \frac{1}{16}\rho_{\text{inc},p}^{k/\ell}$ has $\|f^k_{tent,\rho_{\text{inc},p}} - g\|_1 = \Omega(1)$.*

Proof. This theorem follows from Theorem 5 and Lemma 1. Because $f_{\text{tent},\rho_p/2}$ is ρ_p-Lipschitz, it remains only to prove that there exists an increasing p-cycle. We show that

$$\frac{1}{2}, f\left(\frac{1}{2}\right), \ldots, f^{p-1}\left(\frac{1}{2}\right)$$

is such a cycle.

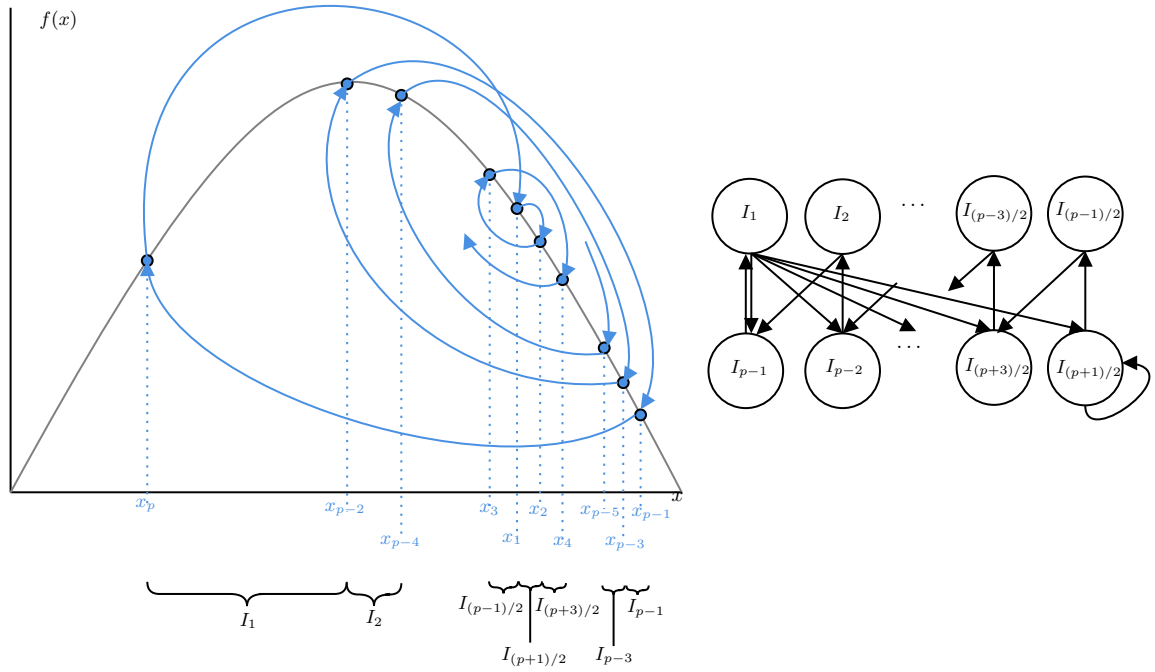

Figure 10: Gives an example of a Stefan p-cycle (which is relied upon in Lemma 2 and demonstrates the interval relationships). Analogous to Figure 9.

By definition of the tent map, $f(\frac{1}{2}) = \frac{\rho_{\text{inc},p}}{2}$ and $f^2(\frac{1}{2}) = \rho_{\text{inc},p}(1 - \frac{\rho_{\text{inc},p}}{2})$. If we assume for now that $f^j(\frac{1}{2}) \leq \frac{1}{2}$ for all $j \in \{2, \ldots, p-1\}$, then

$$f^p\left(\frac{1}{2}\right) = \rho_{\text{inc},p}^{p-1}\left(1 - \frac{\rho_{\text{inc},p}}{2}\right) = -\frac{1}{2}\left(\rho_{\text{inc},p}^{p} - 2\rho_{\text{inc},p}^{p-1} + 1\right) + \frac{1}{2} = 0 + \frac{1}{2}.$$

Because $f^p(\frac{1}{2}) = \frac{1}{2}$ and we assumed that $f^{j+1}(\frac{1}{2}) = \rho_{\text{inc},p} f^j(\frac{1}{2})$ for $j \geq 2$ and $\rho > 1$, it must be the case that $f^j(\frac{1}{2}) \leq \frac{1}{2}$ for all $j \in \{2, \ldots, p-1\}$.

Lemma 1 thus implies that f^k has $\Omega(\rho_{\text{inc},p}^k)$ crossings, which enables us to complete the proof by invoking Theorem 5, since the Lipschitzness condition is met. □

9.5 Proof of Lemma 2

Lemma 2. *For some odd $p \geq 3$, suppose f is a symmetric concave unimodal mapping with an odd p-cycle. Then, there exists $[a, b] \subset [0, 1]$ with $b - a \geq 0.07$ such that $C_{a,b}(f^k) = \rho_{\text{odd},p}^{k-p}$ for any $k \in \mathbb{N}$.*

Proof. By Theorems 2.94 and 3.11.1 of Alsedà et al. (2000), there exists a p-cycle of the form

$$x_p < x_{p-2} < \cdots < x_3 < x_1 < x_2 < x_4 < \cdots < x_{p-1},$$

which is known as a *Stefan cycle*. The analysis of Section 3.2 of Chatziafratis et al. (2020) shows that $C_{[x_1,x_2]}(f^k) \geq \rho_{\text{odd},p}^k$. Their exploitation of the relationships between intervals is visualized in 10. By the continuity of f, applying f an additional $p-1$ times gives $C_{[x_p,x_1]}(f^{k+p-1}) \geq \rho_{\text{odd},p}^k$. Because $[x_{p-2}, x_1] \subset [x_p, x_1]$, applying f one more time gives $C_{[x_2,x_{p-1}]}(f^{k+p}) \geq \rho_{\text{odd},p}^k$.

Hence, by redefining k, we have

$$\max\{C_{[x_1,x_2]}(f^k), C_{[x_2,x_{p-1}]}(f^k), C_{[x_p,x_1]}(f^k)\} \geq \rho_{\text{odd},p}^{k-p}.$$

Since $[x_p, x_{p-1}]$ is the disjoint union of $[x_1, x_2]$, $[x_2, x_{p-1}]$, and $[x_p, x_1]$, there exists $[a, b] \subset [x_p, x_{p-1}]$ with $b - a \geq \frac{1}{3}(x_{p-1} - x_p)$ such that $C_{[a,b]}(f^k) \geq \rho^{k-p}_{\text{odd},p}$.

The problem reduces to placing a lower bound on $x_{p-1} - x_p$. To do so, we derive contradictions on the concavity and symmetry of f. Let $r = f(\frac{1}{2}) \in (x_p, 1)$ be the the largest outcome of f, and let

$$a = \sup_{x,x' \in [1-r,r]} \left| \frac{f(x) - f(x')}{x - x'} \right|$$

be the maximum absolute slope of f on $[1-r, r]$. a must be finite by the concavity and continuity of f, and if f is differentiable, $a = f'(1-r) = -f'(r)$. Thus, f is a-Lipschitz on that interval.

Because $f([x_p, x_{p-1}]) \subseteq [x_p, r] \subset [1-r, r]$, it follows that $\left|f^2(x) - f^2(x')\right| \leq a^2 |x - x'|$. Thus, $x_2 - x_p \leq a^2(x_{p-2} - x_p)$ and $x_2 - x_p \leq x_4 - x_p \leq a^2(x_2 - x_{p-2})$. Averaging the two together, we have $x_2 - x_p \leq \frac{a^2}{2}(x_2 - x_p)$, which means $a \geq \sqrt{2}$.

To satisfy concavity, the following must be true:

$$\frac{f(1-r) - f(0)}{1 - r - 0} = \frac{f(r)}{1 - r} \geq a \geq \sqrt{2}.$$

We rearrange the inequality and apply properties of monotonicity to lower-bound r away from $\frac{1}{2}$:

$$r \geq 1 - \frac{f(r)}{\sqrt{2}} \geq 1 - \frac{f(x_{p-1})}{\sqrt{2}} = 1 - \frac{x_p}{\sqrt{2}} > 1 - \frac{1}{2\sqrt{2}}.$$

It also must be the case for any $x \in [\frac{1}{2}, 1]$, that:

$$\left| \frac{f(x) - f\left(\frac{1}{2}\right)}{x - \frac{1}{2}} \right| \leq 2.$$

Otherwise, the concavity of f would force $f(\frac{1}{2}) > 1$.

We finally assemble the pieces to lower-bound the gap between x_{p-1} and x_p:

$$\begin{aligned} x_{p-1} - x_p \geq x_{p-1} - \frac{1}{2} \geq -\frac{1}{2}\left(f(x_{p-1}) - f\left(\frac{1}{2}\right)\right) = \frac{r}{2} - \frac{x_p}{2} \\ > \frac{1}{2} - \frac{1}{4\sqrt{2}} - \frac{1}{4} = \frac{1}{4} - \frac{1}{4\sqrt{2}} > 0.07. \end{aligned}$$

□

9.6 Necessity of Symmetry and Concavity Assumptions in Theorems 4 and 5

We demonstrate the weakness of the bounds promised by Chatziafratis et al. (2019, 2020); Bu et al. (2020) and argue that our assumptions of symmetry and concavity are necessary in order to avoid such non-vacuous bounds. To do so, we exhibit two families of functions in Propositions 2 and 3 which contain functions with increasing p-cycles for every p that produce large numbers of oscillations, yet are trivial to approximate because their oscillations can be made arbitrarily small. The functions considered in both cases are unimodal and lack symmetry and concavity respectively.

These expose a fundamental shortcoming of other approaches to the hardness of neural network approximation in the aforementioned works because they all rely on showing that for every mapping f meeting some condition (e.g. odd period, positive topological entropy), there exists some $[a, b] \in [0, 1]$ where $C_{a,b}$ is exponentially large, and hence no poly-size shallow neural network g can obtain $L_\infty(f^k, g) \leq P(b - a)$ for some polynomial P. However, because $[a, b]$ depends on f, their difference can potentially be arbitrarily small. The propositions show that this concern is significant and that $[a, b]$ indeed becomes arbitrarily narrow for simple 3-periodic functions. While Chatziafratis et al. (2019) avoid addressing this issue head-on by focusing on classification error over L_∞ error, their classification lower-bounds rely on misclassification of points whose actual distance can be shrinking (see for example Figure 11).

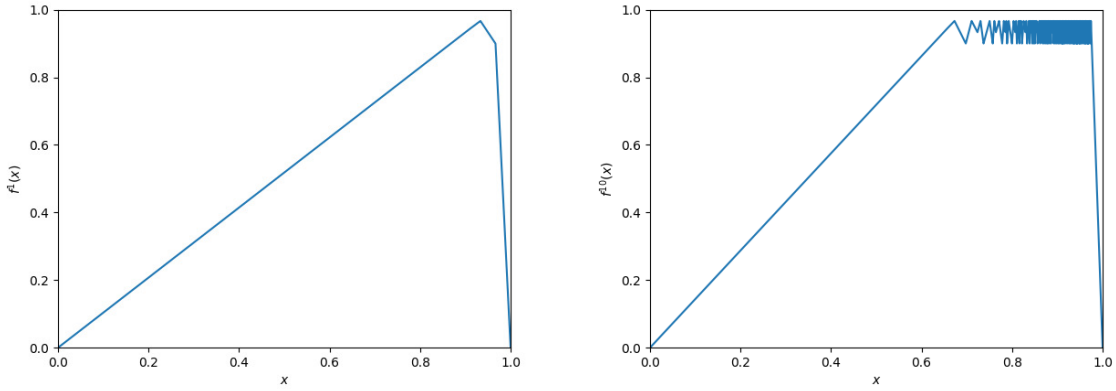

Figure 11: Plots the asymmetric function with a p-cycle referenced in Proposition 2 for $p = 3$ and $\epsilon = 0.1$. While f oscillates frequently, f can be trivially 0.1-approximated by three ReLUs. As $\epsilon \to 0$, the L_∞ approximation hardness guarantees implied by Chatziafratis et al. (2019) become vacuous because the oscillations, even though they are exponentially many, they shrink in size.

The implications of these propositions contrast with the more robust hardness results we present in Theorems 4, 5, 6, and 7, which leverage unimodality, symmetry, and concavity to ensure that the accuracy of approximation can be no better than some constant (independent on f, p) when the neural network g is too small. We show here that those assumptions are necessary by exhibiting functions that satisfy all but one, and become easy to L_∞-approximate with small depth-2 ReLU networks.

Proposition 2. *For $p \geq 3$ and for sufficiently small $\epsilon > 0$, there exists a concave unimodal mapping f with a chaotic p-cycle such that for any k, there exists $g \in \mathcal{N}(3, 2)$ with*

$$L_\infty(f^k, g) \leq \epsilon.$$

Proof. For all $j \in [p]$, let $x_j = 1 - \frac{p-j+1}{p}\epsilon$. Define f to be a piecewise-linear function with $p+1$ pieces chosen with boundaries that satisfy

$$f(0) = 0, f(x_1) = x_2, f(x_2) = x_3, \ldots, f(x_{p-1}) = x_p, f(x_p) = x_1, f(1) = 0.$$

We visualize f for $p = 3$ in Figure 11. f is unimodal because it increases on $[0, x_{p-1}]$ and decreases on $[x_{p-1}, 1]$. It is concave because $f'(x)$ does not increase as x grows, since

$$f'(x) = \begin{cases} \frac{1-\frac{p-1}{p}\epsilon}{1-\epsilon} > 1 & x \in [0, x_1) \\ 1 & x \in (x_1, x_{p-1}) \\ -p+1 & x \in (x_{p-1}, x_p) \\ -\frac{1-\epsilon}{\epsilon} & x \in (x_p, 1], \end{cases}$$

as long as $\frac{1-\epsilon}{\epsilon} > p - 1$.

We show inductively that for all k, there exists $a_k < b_k$ such that $f^k(a_k) = f^k(b_k) = 1-\epsilon$, $f^k([a_k, b_k]) \in [1-\epsilon, 1]$, and f^k has exactly one linear piece for each of the intervals $[0, a_k]$ and $[b_k, 1]$.

These are true for the base case $k = 1$ for $a_1 \in (0, x_1)$ and $b_1 = x_p$.

If the claim holds for k, then there is some $a_{k+1} \in (0, a_k)$ and $b_{k+1} \in (b_k, 1)$ such that $f(a_{k+1}) = f(b_{k+1}) = a_k$. Then, $f^{k+1}(a_{k+1}) = f^{k+1}(b_{k+1}) = 1 - \epsilon$ and $f^{k+1}([0, a_{k+1}]) = f^{k+1}([b_{k+1}, 1]) = [0, 1 - \epsilon]$. For all $x \in [0, a_{k+1}]$, $f^j(x) \leq 1-\epsilon$ for all $j \leq k+1$. Hence, f^{k+1} is linear on $[0, a_{k+1}]$ (and also $[b_{k+1}, 1]$. Because $f([x_1, x_p]) = [x_1, x_p]$, $f^{k+1}([a_{k+1}, b_{k+1}]) \subseteq [x_1, x_p] \subseteq [1 - \epsilon, 1]$. The claim then holds for $k + 1$.

Thus, the piecewise linear mapping g with boundaries $g(0) = 0$, $g(a_k) = 1 - \epsilon$, $g(b_k) = 1 - \epsilon$, and $g(1) = 0$ is an ϵ-approximation of f. Because g has three pieces and contains the origin, it can be exactly represented by a linear combination of four ReLUs, and hence as a depth-2 neural network of width 3. □

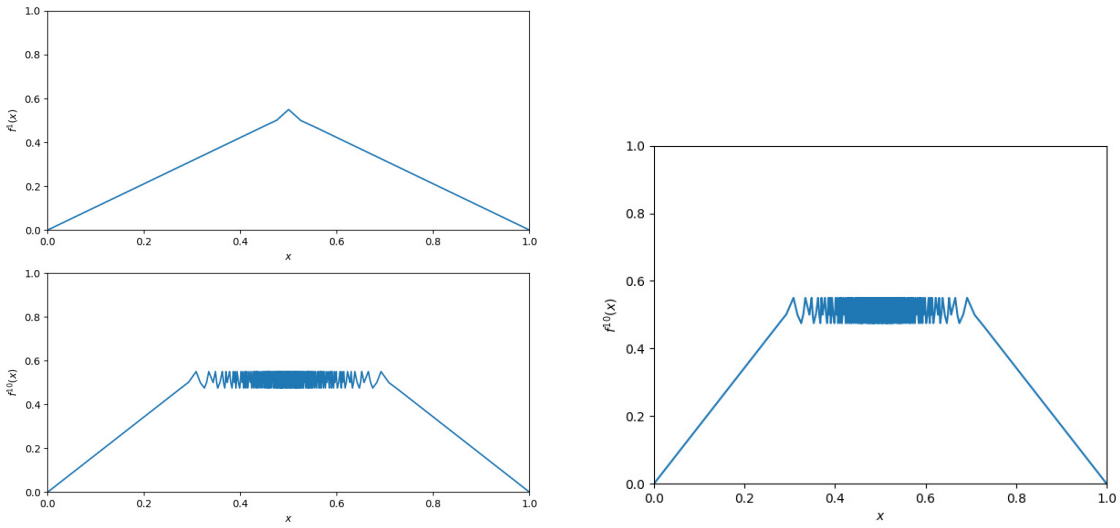

Figure 12: Another example of a function with a 3-cycle that can be ϵ-approximated for arbitrarily small ϵ. (Here, $\epsilon = 0.1$.) This function corresponds to the one in Proposition 3 and the Chatziafratis et al. (2019) bounds are again vacuous for small ϵ. Unlike Figure 11, this function is symmetric, but not concave.

Proposition 3. *For $p \geq 3$ and for sufficiently small $\epsilon > 0$, there exists a symmetric unimodal mapping f with a chaotic p-cycle such that for any k, there exists $g \in \mathcal{N}(3, 2)$ with*

$$L_\infty(f^k, g) \leq \epsilon.$$

Proof. Let $x_j = \frac{1}{2} - \frac{p-1-j}{2(p-1)}\epsilon$ for all $j \in [p-1]$ and $x_p = \frac{1}{2} + \frac{\epsilon}{2}$. Let f be a piecewise-linear function with boundaries

$$f(0) = 0,\ f\left(\frac{1}{2} - \frac{\epsilon}{2}\right) = \frac{1}{2} - \frac{p-2}{p-1} \cdot \frac{\epsilon}{2},\ f\left(\frac{1}{2} - \frac{\epsilon}{2(p-1)}\right) = \frac{1}{2},\ f\left(\frac{1}{2}\right) = \frac{1}{2} + \frac{\epsilon}{2},$$
$$f\left(\frac{1}{2} + \frac{\epsilon}{2(p-1)}\right) = \frac{1}{2},\ f\left(\frac{1}{2} + \frac{\epsilon}{2}\right) = \frac{1}{2} - \frac{p-2}{p-1} \cdot \frac{\epsilon}{2},\ f(1) = 0.$$

We visualize f for $p = 3$ in Figure 12. Note that f is symmetric and unimodal and has an increasing p-cycle $x_1 < \cdots < x_p$. It is *not* concave because $f'(x) = 1$ for $x \in [x_1, x_{p-2}]$ and $f'(x) = 2(p-1)$ for $x \in [x_{p-2}, x_{p-1}]$.

Using a very similar argument to argument from the proof of Proposition 2, for all k, there exists $a_k < b_k$ such that f^k is linear on $[0, a_k]$ and $[b_k, 1]$ and $f^k([a_k, b_k]) \in [\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon]$. As before, there exists a piecewise linear function with three pieces (which can be thought of as a depth-2 neural network of width 3) that ϵ-approximates f. □

10 Additional Proofs for Section 4

10.1 Preliminaries

Before reintroducing and proving the theorems about the doubling and chaotic regime, we introduce topological entropy and define VC-dimension.

10.1.1 Topological Entropy

Topological entropy is a well-known measure of function complexity in dynamical systems that measures the "bumpiness" of a mapping. Like we do with chaotic itineraries, Bu et al. (2020) draw analogies between the neural network approximability of f^k and the topological entropy of f. We do not give a rigorous definition

of topological entropy, but we include a well known result connecting topological entropy to the number of monotone pieces (not constant-sized crossings), which is stated as Lemma 3 of the aforementioned work.

Lemma 4. *[Misiurewicz and Szlenk (1980); Young (1981)] If $f : [0,1] \to [0,1]$ is continuous and piece-wise monotone, then the topological entropy of f satisfies the following:*

$$h_{\text{top}}(f) = \lim_{k\to\infty} \frac{1}{k} \log M(f^k).$$

10.1.2 VC-Dimension

We capture the complexity of the mappings produced by repeated application of f, by measuring the capability of a family of iterates to fit arbitrarily-labeled samples with the VC-dimension. For some threshold parameter $t \in (0,1)$, we first define a hypothesis class that we use to cast this family of iterated functions as Boolean-valued.

Definition 7. *For some unimodal $f : [0,1] \to [0,1]$ and threshold $t \in (0,1)$, let*

$$\mathcal{H}_{f,t} := \{[[f^k]]_t : k \in \mathbb{N}\}$$

be the Boolean-valued hypothesis class of classifiers of composed functions.

The following is the standard definition of the VC-dimension:

Definition 8 (Vapnik and Chervonenkis (2013))**.** *For some hypothesis class $\mathcal{H}$ containing functions $[0,1] \to \{0,1\}$, we say that $\mathcal{H}$* shatters *samples $x_1, \ldots, x_d \in [0,1]$ if for every labeling of the samples $\sigma_1, \ldots, \sigma_d \in \{0,1\}$, there exists some $h \in \mathcal{H}$ such that $h(x_i) = \sigma_i$ for all $i \in [d]$. The* VC-dimension *of $\mathcal{H}$, VC($\mathcal{H}$) is the maximum d such that there exists $x_1, \ldots, x_d \in [0,1]$ that $\mathcal{H}$ shatters.*

$\text{VC}(\mathcal{H}_{f,t})$ will be a useful measurement of complexity of the mapping f, which as we show is tighly connected with the notion of periodicity and oscillations. Notably, this is a measurement of the complexity of iterated maps and is *not* a typical formulation of VC-dimension for neural networks, since those typically would consider a fixed depth and a fixed width, but variable values for the weights, rather than fixed f and variable k.

10.2 Proof for Theorem 8 and 9

Theorem 8. *[Doubling Regime] Suppose f is a symmetric unimodal mapping whose maximal cycle is a primary cycle of length $p = 2^q$. That is, there exists a p-cycle but no $2p$-cycles (and thus, no cycles with lengths non-powers-of-two). Then, the following are true:*

1. *For any $k \in \mathbb{N}$, $M(f^k) = O((4k)^{q+1})$.*
2. *For any $k \in \mathbb{N}$, there exists $g \in \mathcal{N}(u, 2)$ with $u = O((4k)^{q+1}/\epsilon)$ such that $\|g - f^k\|_\infty \le \epsilon$. Moreover, if $f = f_{tent,r}$, then there exists $g \in \mathcal{N}(u, 2)$ with $u = O((4k)^{q+1})$ and $g = f^k$.*
3. *$h_{\text{top}}(f) = 0$.*
4. *For any $t \in (0,1)$, VC($\mathcal{H}_{f,t}$) $\le 18p^2$.*

Proof. Claim 1 follows from a somewhat involved argument in Appendix 10.3 that uses an inductive argument to compare the behavior of a mapping with a maximal p-cycle to one with a maximal $\frac{p}{2}$-cycle. By categorizing intervals of $[0,1]$ based on how f^k behaves on that interval, we analyze how f^{k+1} in turn behaves, which leads to a bound on the monotone pieces $M(f^k)$.

Claim 2 is a simple consequence of Claim 1, by using the fact that a ReLU network can piecewise approximate each monotone piece of f^k. This argument appears in Appendix 10.4.

Claim 3 follows easily from Claim 1 and Lemma 4. We note that this derivation about the topological entropy and the periodicity of f is a known fact in the dynamical systems community.

Claim 4 relies on another recursive argument that frames VC-dimension in terms of the possible trajectories of $f^k(x)$ for fixed x and changing k. We characterize these trajectories by making use of Regular Expressions and by bounding the corresponding VC dimension in Appendix 10.5. □

Theorem 9. *[Chaotic Regime] Suppose f is a unimodal mapping that has a p-cycle where p is not a power-of-two. Then, the following are true:*

1. *There exists some $\rho \in (1, 2]$ such that for any $k \in \mathbb{N}$, $M(f^k) = \Omega(\rho^k)$.*
2. *For any $k \in \mathbb{N}$ and any $g \in \mathcal{N}(u, \ell)$ with $\ell \leq k$ and $u \leq \frac{1}{8}\rho^{k/\ell}$, there exist samples S with $|S| = \frac{1}{2}\lfloor \rho^k \rfloor$ such that $\mathcal{R}_{S,1/2}(f^k, g) \geq \frac{1}{4}$.*
3. *$h_{\text{top}}(f) \geq \rho > 0$.*
4. *There exists a $t \in (0, 1)$ such that $VC(\mathcal{H}_{f,t}) = \infty$.*

Proof. Claims 1 and 2 are immediate implications Theorems 1.5 and 1.6 of Chatziafratis et al. (2019). Claim 3 follows by applying Lemma 4 to Claim 1 (again this derivation about the topological entropy is basic in the literature on dynamical systems).

The most interesting part of the theorem is the last claim. We prove Claim 4 in Appendix 10.6 by showing that the VC-dimension of the class is at least d for all $d \in \mathbb{N}$. The argument relies on the existence of an infinite number of cycles of other lengths, as guaranteed by Sharkovsky's Theorem. □

10.3 Proof of Theorem 8, Claim 1

We restate Claim 1 of the theorem as the following proposition and prove it.

Proposition 4 (Claim 1 of Theorem 8)**.** *Suppose f is a symmetric unimodal mapping whose maximal cycle is of length $p = 2^q$. Then, for any $k \in \mathbb{N}$, $M(f^k) = O((4k)^{q+1})$.*

In order to bound the number of times f oscillates based on its power-of-two periods, we categorize f by its cyclic behavior and the bound the number of local maxima and minima f has based on its characterization.

Definition 9 (Category)**.** *For $q \geq 0$ and $z \in \{0, 1\}$, let $\mathcal{F}_{q,z}$ contain the set of all symmetric unimodal functions f such that (1) f has a 2^q-cycle, (2) f does not have a 2^{q+1}-cycle, and (3) $[[f^{2^q}(\frac{1}{2})]]_{1/2} = z$.*

We abuse notation to let $M(\mathcal{F}^k_{q,z}) = \max_{f \in \mathcal{F}_{q,z}} M(f^k)$. Thus, for f given in the theorem statement with a 2^q-cycle, but not a 2^{q+1}-cycle, our final bound is obtained by

$$M(f^m) \leq \max\{M(\mathcal{F}^m_{q,0}), M(\mathcal{F}^m_{q,1})\}.$$

We let $M(f, a, b)$ represent the number of monotone pieces of f on the sub-interval $[a, b] \subset [0, 1]$.

We build a large-scale inductive argument by first bounding base cases $M(\mathcal{F}^k_{0,0})$ and $M(\mathcal{F}^k_{0,1})$. Then, we relate $M(\mathcal{F}^k_{q,z})$ to $M(\mathcal{F}^k_{q-1,1-z})$ to get the desired outcome.

Before beginning the proof, we state a slight refinement of the part of the theorem, which takes into account the newly-introduced categories, from which the claim follows.

Proposition 5. *For any $k \in \mathbb{N}$, $q \geq 0$, and $z \in \{0, 1\}$,*

$$M(\mathcal{F}^k_{q,z}) \leq \begin{cases} 2(3q)^k & q \text{ is even, } z = 0, \text{ or } q \text{ is odd, } z = 1 \\ 2(3q)^{k+1} & q \text{ is even, } z = 1, \text{ or } q \text{ is odd, } z = 0. \end{cases}$$

Thus, proving Proposition 5 is sufficient to prove Proposition 4. The remainder of the section proves Proposition 5.

10.3.1 Special Case Proof for $q = 1$

We show that $M(\mathcal{F}^k_{0,0}) = 2$ and $M(\mathcal{F}^k_{0,1}) = 2k$.

For f_r as defined above, we characterize the number of oscillations that are added by increasing r past $\frac{1}{2}$, where super-stability of a fixed point exists. Figure 13 illustrates those results.

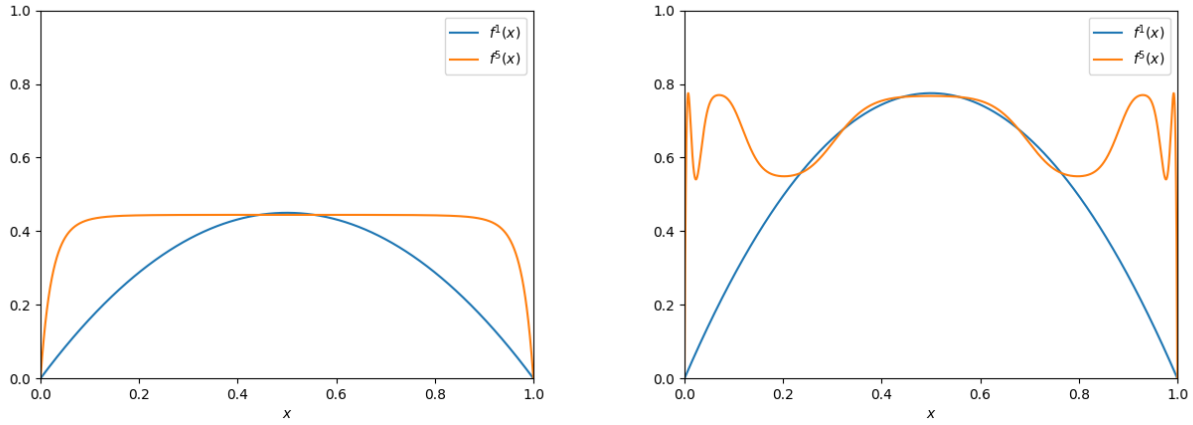

Figure 13: The base case results of Proposition 5 demonstrate the number of oscillations of f^k increases when f moves from $\mathcal{F}_{0,0}$ to $\mathcal{F}_{0,1}$. The plots show f and f^5 for $f \in \mathcal{F}_{0,0}$ ($f = f_{\log,0.45}$) on the left and $f \in \mathcal{F}_{0,1}$ ($f = f_{\log,0.775}$) on the right.

To analyze the oscillation patterns of f^k, we define several "building blocks," which represent disjoint pieces of f^k. That is, the interval $[0,1]$ can be partitioned into several sub-intervals, each of which has f^k follow certain simple behavior that we categorize. We argue that any iterate can be decomposed into those pieces and then show how applying f to f^k modifies the pieces in order to analyze f^{k+1}. Here are the function pieces that we analyze, which map interval $[a,b] \subseteq [0,1]$ to $[0,1]$:

Definition 10. *For any $f : [0,1] \to [0,1]$ and for any $[a,b] \subseteq [0,1]$, f is referred to on interval $[a,b]$ as:*

- *a **increasing crossing piece** Ic if f is strictly increasing on $[a,b]$ and has $f(a) = 0$, $f(b) > \frac{1}{2}$, and $f'(b) > 0$;*
- *a **decreasing crossing piece** Dc if f is strictly decreasing on $[a,b]$ and has $f(a) > \frac{1}{2}$, $f(b) = 0$, and $f(a) < 0$;*
- *a **up peak** Up if there exists some $c \in (a,b)$ that maximizes f on $[a,b]$, f is strictly increasing on $[a,c)$, f is strictly decreasing on $(c,b]$, and $f(x) > \frac{1}{2}$ for all $x \in [a,b]$;*
- *a **up valley** Uv if there exists some $c \in (a,b)$ that minimizes f on $[a,b]$, f is strictly decreasing on $[a,c)$, f is strictly increasing on $(c,b]$, and $f(x) > \frac{1}{2}$ for all $x \in [a,b]$; and*
- *a **down peak** Dp if there exists some $c \in (a,b)$ that maximizes f on $[a,b]$, f is strictly increasing on $[a,c)$, f is strictly decreasing on $(c,b]$, and $f(x) \leq \frac{1}{2}$ for all $x \in [a,b]$.*

If there exists a sequence of intervals $J_1, \ldots, J_m$ such that f is piece η_i on J_i, then we represented f with the string $\eta_1 \ldots \eta_m$.

We specify an invariant for each part of the theorem, such that proving the invariant is sufficient to prove the proposition:

1. If $f \in \mathcal{F}_{0,0}$, then f^k is a down peak on $[0,1]$ for all k, and f^k has two monotone pieces.
2. If $f \in \mathcal{F}_{0,1}$, f is represented by $\mathsf{Ic}(\mathsf{UpUv})^{k-1}\mathsf{UpDc}$. That is, $[0,1]$ can be partitioned into $2k+1$ subsequent intervals $J_1, \ldots J_{2k+1}$ such that f^k is an increasing crossing piece on J_1, a decreasing crossing piece on J_{2k+1} (if $k \neq 0$), an up peak on J_{2j} for $j \in \{1, \ldots, k\}$, and a up valley on J_{2j+1} for $j \in \{1, \ldots, k-1\}$. Hence, f^k has k distinct maxima and $2k$ monotone pieces. Figure 14 illustrates this invariant.

Base Case:

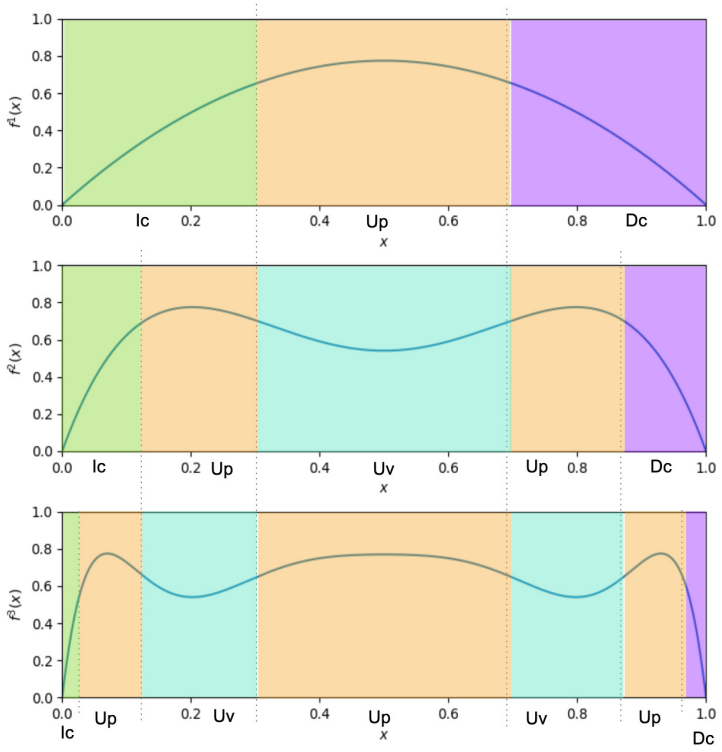

Figure 14: For $f \in \mathcal{F}_{0,1}$ ($f = f_{\log,0.775}$), visualizes the decomposition of f, f^2, and f^3 into IcUpDc, IcUpUvUpDc, and Ic(UpUv)^{2}UpDc respectively.

1. For $f \in \mathcal{F}_{0,0}$, $f^1 = f$ is trivially a down peak on $[0,1]$ by the definition of $\mathcal{F}_{0,0}$, since $\frac{1}{2}$ maximizes f.
2. For $f \in \mathcal{F}_{0,1}$, f can be represented by IcUpDc. That is, $[0,1]$ can be decomposed into intervals I_1, I_2, and I_3, on which f_r is an increasing crossing piece, an up peak, and a decreasing crossing piece respectively.

Inductive Step:

We examine what happens to each function piece when f is applied to it. We can use the following analysis, along with the inductive hypothesis to show that f^{k+1} can be decomposed as we expect it to be.

1. Examining the **down peak** proves first invariant for the case when $f \in \mathcal{F}_{0,0}$. Because f strictly increases on $[0, \frac{1}{2}]$ and because $f([0,1]) \subseteq [0, \frac{1}{2}]$ if f^k is a down peak, $f \circ f^k$ also supports a down peak on $[0,1]$.

   Because we inductively assume that f^k is a low peak on $[0,1]$, it then follows that f^{k+1} is also a down peak on $[0,1]$.

2. We first prove a claim, which implies that f has no down peaks for $f \in \mathcal{F}_{0,1}$. Let $x_{\max} = f(\frac{1}{2})$,

   Claim 1. *If $f \in \mathcal{F}_{0,1}$, then $f((\frac{1}{2}, x_{\max}]) \subseteq (\frac{1}{2}, x_{\max}]$.*

   Proof. Because $\frac{1}{2}$ maximizes f, $f(x) \leq x_{\max}$ for all $x \in [\frac{1}{2}, x_{\max}]$. Since f monotonically decreases, on $[\frac{1}{2}, x_{\max}]$, the claim can only be false if $f(x_{\max}) < \frac{1}{2}$. We show by contradiction that this is impossible.

   Because f is continuous and monotonically increases on $[0, \frac{1}{2}]$ and ranges from 0 to $x_{\max} \geq \frac{1}{2}$, there exists some $x' \leq \frac{1}{2}$ such that $f(x') = \frac{1}{2}$ and $f^2(x') = x_{\max}$.

   Let $g(x) = f^2(x) - x$. By assumption, $g(\frac{1}{2}) = f(x_{\max}) - \frac{1}{2} < 0$. By definition of x', $g(x') = \frac{1}{2} - x' \geq 0$. Because g is continuous, the Intermediate Value Theorem implies the existence of $x'' \in [x', \frac{1}{2})$ such that $g(x'') = 0$ and $f^2(x'') = x''$. Since f has no two-cycles, it must be the cause that $f(x'') = x''$ and $x'' = \frac{1}{2}$. However, this contradicts our finding that $x'' < \frac{1}{2}$, which means that $f(x_{\max}) \geq \frac{1}{2}$ and the claim holds. □

Now, we proceed with analyzing each of the function pieces on some interval $[a, b] \subseteq [0, 1]$ when $f \in \mathcal{F}_{0,1}$. The transformations are visualized in Figure 14.

- **Increasing crossing piece**: If f^k has an Ic on $[a, b]$, then f^{k+1} can be represented by IcUp on $[a, b]$. There exist c and d such that $a < d < c < \frac{1}{2} < b$, $f^k(c) = \frac{1}{2}$, and $f^k(d) = c$. Then, $[a, \frac{1}{2}(c + d)]$ supports an increasing crossing piece on $f \circ f^k$—because $f(f^k(a)) = 0$, $f(f^k(\frac{1}{2}(c+d))) > \frac{1}{2}$, and $f \circ f^k$ is strictly increasing on that interval since f is increasing before reaching $\frac{1}{2}$. $[0.5(c + d), b]$ supports a high peak—because c is a local maxima on $f \circ f^k$, and $f \circ f^k$ is strictly increasing before c and strictly decreasing after c.
- **Decreasing crossing piece**: For the same arguments, f^{k+1} can represented by UpDc on $[a, b]$ if f^k is represented by Dc on $[a, b]$.
- **Up peak**: Because f strictly decreases for $x > \frac{1}{2}$ and because $f^k([a, b]) \subseteq (\frac{1}{2}, x_{\max}]$ if Up represents f^k on $[a, b]$, c becomes a local minimum for $f \circ f^k$, and f^{k+1} is a high valley Uv on $[a, b]$.
- **Up valley**: Because f strictly decreases for $x > \frac{1}{2}$ and because $f^k([a, b]) \subseteq (\frac{1}{2}, x_{\max}]$ if Uv represents f^k on $[a, b]$, c becomes a local maximum for $f \circ f^k$, and f^{k+1} is a high peak Up on $[a, b]$.

Now, consider the inductive hypothesis. Because f^k can be represented by $\mathsf{Ic}(\mathsf{UpUv})^{k-1}\mathsf{UpDc}$, applying the above transformations to each piece implies that f^{k+1} can be represented by $\mathsf{Ic}(\mathsf{UpUv})^{k}\mathsf{UpDc}$. Hence, the inductive argument goes through.

10.3.2 General Case Proof

The argument proceeds inductively. We show that if we have some $f \in \mathcal{F}_{q,k}$, then we can find some other function $h \in \mathcal{F}_{q-1,1-k}$ and characterize the behavior of f in terms of the behavior of h.

Since we assume that $q \geq 1$, there will always exist some $x^* > \frac{1}{2}$ that is a fixed point of f.[13] By symmetry, $f(1 - x^*) = x^*$. Let $\phi : [0, 1] \to [1 - x^*, x^*]$ be a decreasing isomorphism with $\phi(x) = x^* - x(2x^* - 1)$, and let

$$h = \phi^{-1} \circ f^2 \circ \phi.$$

h is a useful construct, because its behavior resembles simpler versions of f, with fewer cycles and oscillations. We use properties of h to relate pieces of f^k to those of $h^{k/2}$. We illustrate this recursive and fractal-like behavior in Figure 15.

Note that $h^k = \phi^{-1} \circ f^{2k} \circ \phi$.

Lemma 5. *h is a symmetric unimodal mapping with $h \in \mathcal{F}_{q-1,1-z}$.*

Proof. We verify the conditions for f to be unimodal mapping.

1. h is continuous and piece-wise differentiable on $[0, 1]$ because f^2 is, and h is merely a linear transformation of f^2.
2. $h(0) = h(1) = 0$. $h((0, 1))$ is strictly positive because $f((1 - x^*, x^*)) = (x^*, x_{\max})$, $f^2((1 - x^*, x^*)) = (f(x_{\max}), x^*)$, and $f(x_{\max}) < f(x^*) = x^*$ by f being decreasing on $[\frac{1}{2}, 1]$.
3. h is uniquely maximized by $\frac{1}{2}$ because $\frac{1}{2}$ minimizes f^2 on the interval $[1 - x^*, x^*]$. f maps both $[1 - x^*, \frac{1}{2}]$ and $[\frac{1}{2}, x^*]$ onto $[x^*, x_{\max}]$ and is increasing and decreasing on the respective intervals. Because f maps $[x^*, x_{\max}]$ onto $[f(x_{\max}), x^*]$ and $f(x_{\max}) < x^*$ and is decreasing on $[x^*, x_{\max}]$, f^2 is increasing on $[1-x^*, \frac{1}{2}]$ and decreasing on $[\frac{1}{2}, x^*]$.

   Thus, h is maximized by $\frac{1}{2}$, increases before $\frac{1}{2}$, and decreases after $\frac{1}{2}$.
4. We must also show that h is well-defined, which entails proving that $h(x) \leq 1$ for all $x \in [0, 1]$. Suppose that were not the case. Then, $h(\frac{1}{2}) > 1$, and there exists some $x' \leq \frac{1}{2}$ with $h(x') = 1$. There also exists some $x^{**} \in [1 - x', 1]$ with $h(x^{**}) = x^{**}$ by the Intermediate Value Theorem.

[13] Sharkovsky's Theorem yields this by showing that the existence of a 2^q-cycle implies the existence of any 2^j-cycle, for all $j \in \{0, \ldots, q - 1\}$. $x^* > \frac{1}{2}$ by our assumption that a 2-cycle $x_1 < x_2$ exists. It must be true that $x_2 > \frac{1}{2}$; otherwise, $f(x_2) > x_2 > x_1$, which breaks the cycle. Because $f(\frac{1}{2}) > \frac{1}{2}$ and $f(x_2) < x_2$, there exists $x^* \in (\frac{1}{2}, x_2)$ such that $f(x^*) = x^*$ by the Intermediate Value Theorem.

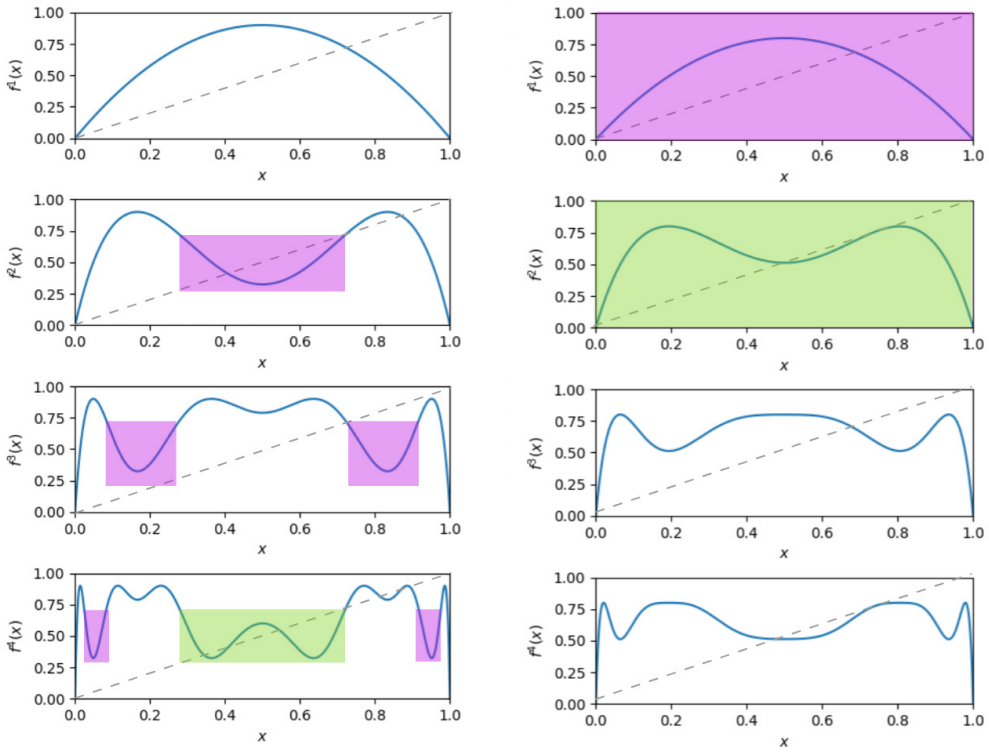

Figure 15: Visualizes the analogy between mappings in $\mathcal{F}_{q,z}$ and $\mathcal{F}_{q-1,1-z}$. The left plots the first 4 iterates of $f = f_{\log,0.9} \in \mathcal{F}_{4,1}$ (has a maximal 4-cycle with $f^4(\frac{1}{2}) > \frac{1}{2}$), while the right plots those of $f = f_{\log,0.85} \in \mathcal{F}_{2,0}$ (has a maximal 2-cycle with $f^2(\frac{1}{2}) < \frac{1}{2}$. The purple highlighted regions on the left behave qualitatively similar to $f_{\log,0.85}$, while the green regions are similar to $f^2_{\log,0.85}$.

Let $g(x) = h^3(x) - x$ and note that g is continuous on $[0, x']$. Observe that $g(1 - x^{**}) = 2x^{**} - 1 > 0$ and $g(x') = -x' < 0$. Thus, there exists $x'' \in [1 - x^{**}, x']$ with $g(x'') = 0$. Because h is increasing on $[0, x']$ and $x' > 1 - x^{**}$, it must be the case that $h(x') > x^{**} > x'$. Thus, x^{**} is not a fixed point and must be on a 3-cycle in h.

However, if x^{**} is on a 3-cycle in h, then $\phi(x^{**})$ must be part of a 6-cycle in f. This contradicts the assumption that f cannot have a 2^{q+1}-cycle, because Sharkovsky's Theorem states that a 6-cycle implies a 2^{q+1}-cycle.

We show that h is symmetric.

$$\begin{aligned} h(x) &= \phi^{-1}(f^2(\phi(x))) = \phi^{-1}(f^2(1 - \phi(x))) = \phi^{-1}(f^2(1 - x^* + x(2x^* - 1))) \\ &= \phi^{-1}(f^2(x^* - (1 - x)(2x^* - 1))) = \phi^{-1}(f^2(\phi(1 - x))) = h(1 - x). \end{aligned}$$

If $f^{2^q}(\frac{1}{2}) \geq \frac{1}{2}$, then $h^{2^{q-1}}(\frac{1}{2}) \leq \frac{1}{2}$, and if $f^{2^q}(\frac{1}{2}) \leq \frac{1}{2}$, then $h^{2^{q-1}}(\frac{1}{2}) \geq \frac{1}{2}$. Thus, $[[h^{2^{q-1}}(\frac{1}{2})]]_{1/2} = [[f^{2^q}(\frac{1}{2})]]_{1/2}$. By Lemma 6, h has a 2^{q-1}-cycle and does not have a 2^q-cycle. Thus, $h \in \mathcal{F}_{q-1,1-z}$. □

Lemma 6. *For $p \in \mathbb{Z}_+$, h has a p-cycle if and only if f has a $2p$-cycle.*

Proof. Suppose $x_1, \ldots, x_p$ is a p-cycle for h. Then, $\phi(x_1), \ldots, \phi(x_p)$ is a p-cycle for f^2. If $x_1, \ldots, x_p$ are distinct, then so must be $\phi(x_1), \ldots, \phi(x_p)$, since ϕ is an isomorphism. Thus,

$$\phi(x_1), f(\phi(x_1)), \ldots, \phi(x_p), f(\phi(x_p))$$

is a $2p$-cycle for f.

Conversely, if $x_1, \ldots, x_{2p}$ is a $2p$-cycle for f, then $x_1, x_3, \ldots, x_{2p-1}$ is a p-cycle for f^2 and

$$\phi^{-1}(x_1), \ldots, \phi^{-1}(x_{2p})$$

is a p-cycle for h. □

We proceed with a proof similar in structure to the one in the last section, where we divide each f^k into intervals and monitor the evolution of each as k increases. We define the classes of the pieces of some 1-dimensional map f^k on interval $[a, b]$ below. We visualize these classes in Figure 16.

- f^k is an **approach** A on $[a, b]$ if f is strictly increasing, $f^k(a) = 0$, and $f^k(b) = 1 - x^*$.
- Similarly, f^k is a **departure** D on $[a, b]$ if f^k is strictly decreasing, $f^k(a) = 1 - x^*$, and $f^k(b) = 0$.
- f^k is an **i-Left Valley** Lv_i on $[a, b]$ if $f^k : [a, b] \to [f(x_{\max}), x^*]$ and if there exists some strictly increasing and bijective $\sigma : [a, b] \to [1 - x^*, x^*]$ such that $f^k = \phi \circ h^i \circ \phi^{-1} \circ \sigma$ on $[a, b]$. Note that $f^k(a) = f^k(b) = x^*$—unless $i = 0$, in which case $f^k(a) = 1 - x^*$ and $f_r^k(b) = x^*$.
- f^k is analogously a **i-Right Valley** Rv_i if the same condition holds, except that σ is strictly decreasing.
- f^k is an **i-Left Peak** Lp_i on $[a, b]$ if f^{k-1} is Lv_{i-1} on $[a, b]$. It follows that $f^k : [a, b] \to [x^*, x_{\max}]$, that there exists some $c \in [a, b]$ such that $f^k(c) = x_{\max}$ (because $\frac{1}{2} \in [f(x_{\max}), x^*]$), and that $f^k(a) = f^k(b) = x^*$.
- f^k is an **i-Right Peak** Rp_i on $[a, b]$ if f^{k-1} is Rv_{i-1} on $[a, b]$. The same claims hold as Lp_i.

Now, the proof of the number of oscillations proceeds in two steps. (1) We analyze how each of the above pieces evolves with each application of f. (2) We show how many maxima and minima each translates to.

Lemma 7. *When $f \in \mathcal{F}_{q,z}$ for $q \geq 1$ and for all $k \in \mathbb{Z}_+$, f^k can be decomposed into $2k + 3$ pieces $\eta_1, \ldots, \eta_{2k+3}$ such that*

$$\eta_i \text{ is } \begin{cases} \mathsf{A} \text{ if } i = 1 \\ \mathsf{Lv}_j \text{ if } i = 2j + 2 \text{ for } j \in \{0, 1, \ldots, \lfloor k/2 \rfloor\} \\ \mathsf{Lp}_j \text{ if } i = 2j + 1 \text{ for } j \in \{1, \ldots, \lfloor (k+1)/2 \rfloor\} \\ \mathsf{Rv}_j \text{ if } i = 2k - 2j + 2 \text{ for } j \in \{0, 1, \ldots, \lfloor (k-1)/2 \rfloor\} \\ \mathsf{Rp}_j \text{ if } i = 2k - 2j + 3 \text{ for } j \in \{1, \ldots, \lfloor k/2 \rfloor\} \\ \mathsf{D} \text{ if } i = 2k + 3 \end{cases}$$

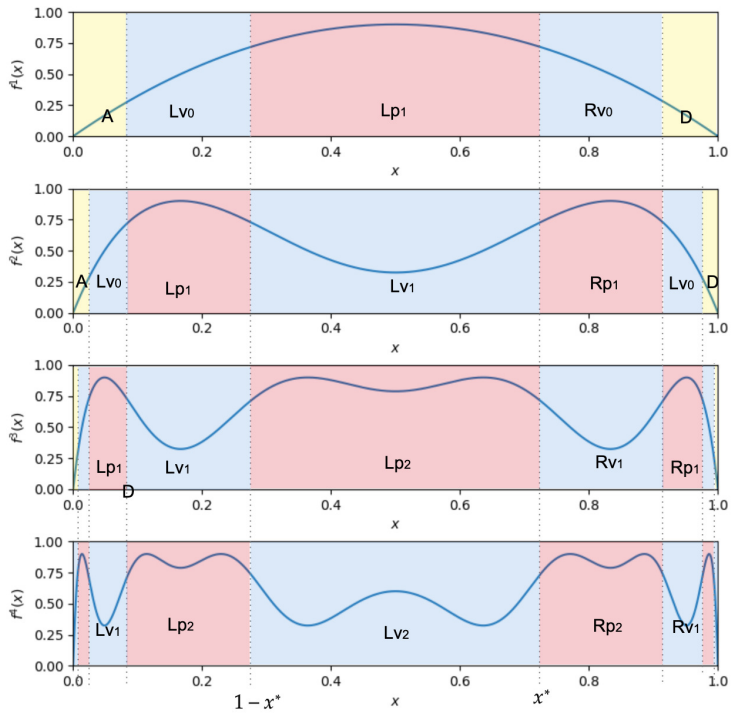

Figure 16: Similar to Figure 14, visualizes the classifications of f, f^2, f^3, f^4 for $f = f_{\log,0.9} \in \mathcal{F}_{4,1}$, and demonstrates that the decompositions are $\mathsf{ALv}_0\mathsf{Lp}_0\mathsf{Rv}_0\mathsf{D}$, $\mathsf{ALv}_0\mathsf{Lp}_1\mathsf{Lv}_1\mathsf{Rp}_1\mathsf{Rv}_0\mathsf{D}$, $\mathsf{ALv}_0\mathsf{Lp}_1\mathsf{Lv}_1\mathsf{Lp}_2\mathsf{Rv}_1\mathsf{Rp}_1\mathsf{Rv}_0\mathsf{D}$, and $\mathsf{ALv}_0\mathsf{Lp}_1\mathsf{Lv}_1\mathsf{Lp}_2\mathsf{Lv}_2\mathsf{Rp}_2\mathsf{Rv}_1\mathsf{Rp}_1\mathsf{Rv}_0\mathsf{D}$ respectively.

That is, if k is even, then f can be represented by

$$\mathsf{ALv}_0\mathsf{Lp}_1\mathsf{Lv}_1 \dots \mathsf{Lv}_{k/2-1}\mathsf{Lp}_{k/2}\mathsf{Lv}_{k/2}\mathsf{Rp}_{k/2}\mathsf{Rv}_{k/2-1} \dots \mathsf{Rv}_1\mathsf{Rp}_2\mathsf{Rv}_0\mathsf{D}.$$

If k is odd, then f is represented by

$$\mathsf{ALv}_0\mathsf{Lp}_1\mathsf{Lv}_1 \dots \mathsf{Lp}_{(k-1)/2}\mathsf{Lv}_{(k-1)/2}\mathsf{Lp}_{(k+1)/2}\mathsf{Rv}_{(k-1)/2}\mathsf{Rp}_{(k-1)/2} \dots \mathsf{Rv}_1\mathsf{Rp}_2\mathsf{Rv}_0\mathsf{D}.$$

Proof. This lemma is proved inductively. f can be decomposed into the pieces $\mathsf{ALv}_0\mathsf{Lp}_1\mathsf{Rv}_0\mathsf{D}$.

- By unimodality and symmetry, f is strictly increasing on $[0, \frac{1}{2})$ and strictly decreasing on $(\frac{1}{2}, 1]$. There exists some x_1 such that $[0, x_1]$ is strictly increasing and $f(x_1) = 1 - x^*$ (because $1 - x^* < x^* < x_{\max}$). Thus, f is A on $[0, x_1]$. Similarly, $[1 - x_1, 1]$ is strictly decreasing and $f(1 - x_1) = 1 - x^*$, which implies that f is D on $[1 - x_1, 1]$.
- Note that $x_1 < 1 - x^* < x^* < 1 - x_1$, and f is increasing on $[x_1, 1 - x^*]$ and decreasing on $[x^*, 1 - x_1]$.

  Because $[x_1, 1 - x^*]$ is monotone, there exists continuous and increasing $\sigma : [x_1, 1 - x^*] \to [1 - x^*, x^*]$ such that $f(x) = \sigma(x)$. Since h^0 is the identity map, it trivially also holds that $f(x) = \phi(h^0(\phi^{-1}(\sigma(x))))$. Because $f(x_1) = 1 - x^*$ and $f(x^*) = x^*$, it follows that f is Lv_0 on $[x_1, 1 - x^*]$.

  By a similar argument, f is Rv_0 on $[x^*, 1 - x_1]$, with the only difference being that σ needs to be strictly decreasing for it to hold.
- $[1 - x^*, x^*]$ is Lp_1 because $[1 - x^*, x^*]$ is Lv_0 on the identity map f^0. This trivially holds using the identity σ map.

Now, we prove the inductive step, which can be summed up by the following line:

$$\mathsf{A} \to \mathsf{ALv}_0; \mathsf{D} \to \mathsf{Rv}_0\mathsf{D}; \mathsf{Lv}_j \to \mathsf{Lp}_{j+1}; \mathsf{Lp}_j \to \mathsf{Lv}_j; \mathsf{Rv}_j \to \mathsf{Rp}_{j+1}; \mathsf{Rp}_j \to \mathsf{Rv}_j.$$

We show each part of the relationship as follows:

- If f^k is A on $[0, b]$, then there exists some $c \in (0, b)$ such that $f^{k+1}(c) = 1 - x^*$ because f^k is an isomorphism between $[0, b]$ and $[0, x^*]$.

  It follows that f^{k+1} is A on $[0, c]$ because f^{k+1} is strictly increasing on the interval from 0 to $1 - x^*$.

  $[c, b]$ is Lv_0 because there must exist some increasing σ such that $f^{k+1}(x) = \sigma(x)$ on that interval. Thus, it follows that $f^{k+1} = \phi \circ h^0 \circ \phi^{-1} \circ \sigma$ on $[0, b]$.

- The same argument holds for D. If f^k is D on $[a, 1]$, then there exists $c \in (a, 1)$ such that $[a, c]$ is Rv_0 and $[c, 1]$ is D.

- If f^k is Lv_j on $[a, b]$, then f^{k+1} is Lp_{j+1} on the same interval by the definition of Lp_{j+1}.

- Similarly, if f^k is Rv_j on $[a, b]$, then f^{k+1} is Rp_{j+1} on the same interval by the definition of Rp_{j+1}.

- If f^k is Lp_j on $[a, b]$, then f^{k-1} is Lv_{j-1} and hence f^{k-1} maps to $[f(x_{\max}), x^*]$ on the interval. Therefore, there exists σ such that $f^{k-1} = \phi \circ h^{j-1} \circ \phi^{-1} \circ \sigma$ on the interval. We use the properties of h to show that f^{k+1} is Lv_j on $[a, b]$. Note that $f^2 = \phi \circ h \circ \phi^{-1}$ on $[f(x_{\max}), x^*]$.

  $$f^{k+1} = f^2 \circ f^{k-1} = \phi \circ h \circ \phi^{-1} \circ \phi \circ h^{j-1} \circ \phi^{-1} \circ \sigma = \phi \circ h^j \circ \phi^{-1} \circ \sigma$$

  Thus, f^{k+1} satisfies the condition to be Lv_j.

- By a identical argument, if f^k is Rp_j on $[a, b]$, then f^{k+1} is Rv_j.

The remainder of this argument follows by applying the above transition rules for each piece to the inductive hypothesis about the ordering of pieces in f^k to obtain the ordering for f^{k+1}. □

Now, we determine how many local maxima and minima are contained in each type of piece. Let maxima(f) and minima(f) represent the number of local maxima and minima respectively on mapping f on interval $[0, 1]$. We bound the total number of monotone pieces with these bounds by using $M(f) = 2\text{maxima}(f)$. We similarly abuse notation to bound the number of maxima and minima in a category with maxima$(\mathcal{F}^k_{q,z})$ and minima$(\mathcal{F}^k_{q,z})$, and in the interval $[a, b]$ with maxima(f, a, b) and minima(f, a, b).

By the base case in the previous section maxima$(\mathcal{F}^k_{0,0}) = 1$, minima$(\mathcal{F}^k_{0,0}) = 2$, maxima$(\mathcal{F}^k_{0,1}) = k$, and minima$(\mathcal{F}^k_{0,1}) = k + 1$. We obtain recurrences to represent maxima$(\mathcal{F}^k_{q,z})$ and minima$(\mathcal{F}^k_{q,z})$.

For each part, we rely on the following facts: If σ is a strictly increasing bijection, then maxima$(f \circ \sigma, a, b) =$ maxima(f, a, b). If σ is strictly decreasing, then minima$(f \circ \sigma, a, b) =$ maxima(f, a, b). (The reverse are true for minima of f.)

We analyze each type of piece individually, considering what happens when f has some kind of piece on interval $[a, b]$.

- Because A and D segments are strictly increasing or decreasing, maxima$(f, a, b) = 0$ when f has either piece on $[a, b]$. minima$(f, a, b) = 1$ because segments that support A contain 0 and segments with D have 1, each of which f maps to 0.

- Because each Lv_i segment of f^k on $[a, b]$ can be represented as $\phi \circ h^i \circ \phi^{-1} \circ \sigma$, and because ϕ is strictly decreasing, maxima$(f^k_r, a, b) =$ minima(h^i) and minima$(f^k_r, a, b) =$ maxima(h^i). By Lemma 5, $h \in \mathcal{F}_{q-1,1-z}$, maxima$(f^k_r, a, b) \le$ minima$(\mathcal{F}^i_{q-1,1-z})$ and minima$(f^k_r, a, b) \le$ maxima$(\mathcal{F}^i_{q-1,1-z})$.

  The same analysis holds for each Rv_i segment.

- Consider an Lp_i segment of f^k on $[a, b]$, which has output spanning the interval $[x^*, x_{\max}]$. Because $x^* > \frac{1}{2}$, f is strictly decreasing on the domain $[x^*, x_{\max}]$. Thus, f^{k+1} must satisfy maxima$(f^{k+1}_r, a, b) =$ minima(f^k_r, a, b) and minima$(f^{k+1}_r, a, b) =$ maxima(f^k_r, a, b).

  Note by the definition of Lp_i that $[a, b]$ must also support an Lv_{i-1} segment on f^{k-1} and an Lv_i segment on f^{k+1}. From the previous bullet, the Lv_i segment must have at most minima$(\mathcal{F}^i_{q-1,1-z})$ maxima and maxima$(\mathcal{F}^i_{q-1,1-z})$ minima. Because there must be a one-to-one correspondence between minima of f^{k+1}

and maxima of f^k on the interval and vice versa, the Lp_i segment has $\text{maxima}(f_r^k, a, b) \le \text{maxima}(\mathcal{F}^i_{q-1,1-z})$ and $\text{minima}(f_r^k, a, b) \le \text{minima}(\mathcal{F}^i_{q-1,1-z})$.

The same analysis hold for each Rp_i segment.

Therefore, we can construct a recurrence relationship for the number of maxima and minima for f_r^k based on the sequences found in Lemma 7.

$$\begin{aligned}\text{maxima}(\mathcal{F}^k_{q,z}) \le &\underbrace{\sum_{i=0}^{\lfloor k/2\rfloor} \text{minima}(\mathcal{F}^k_{q-1,1-z})}_{\mathsf{Lv}_i} + \underbrace{\sum_{i=0}^{\lfloor (k-1)/2\rfloor} \text{minima}(\mathcal{F}^k_{q-1,1-z})}_{\mathsf{Rv}_i} \\ &+ \underbrace{\sum_{i=1}^{\lfloor (k+1)/2\rfloor} \text{maxima}(\mathcal{F}^k_{q-1,1-z})}_{\mathsf{Lp}_i} + \underbrace{\sum_{i=1}^{\lfloor k/2\rfloor} \text{maxima}(\mathcal{F}^k_{q-1,1-z})}_{\mathsf{Rp}_i} \\ \text{minima}(\mathcal{F}^k_{q,z}) = &\underbrace{2}_{\mathsf{A\&D}} + \underbrace{\sum_{i=0}^{\lfloor k/2\rfloor} \text{maxima}(\mathcal{F}^k_{q-1,1-z})}_{\mathsf{Lv}_i} + \underbrace{\sum_{i=0}^{\lfloor (k-1)/2\rfloor} \text{maxima}(\mathcal{F}^k_{q-1,1-z})}_{\mathsf{Rv}_i} \\ &+ \underbrace{\sum_{i=1}^{\lfloor (k+1)/2\rfloor} \text{minima}(\mathcal{F}^k_{q-1,1-z})}_{\mathsf{Lp}_i} + \underbrace{\sum_{i=1}^{\lfloor k/2\rfloor} \text{minima}(\mathcal{F}^k_{q-1,1-z})}_{\mathsf{Rp}_i}\end{aligned}$$

We bound $\text{maxima}(\mathcal{F}^k_{q,z})$ and $\text{minima}(\mathcal{F}^k_{q,z})$ by induction to prove Proposition 5. We use the following inductive assumption over all k, q, and z, which suffices to prove the claim:

$$\text{maxima}(\mathcal{F}^k_{q,z}), \text{minima}(\mathcal{F}^k_{q,z}) \le \begin{cases} (4q)^k & q \text{ is even, } z = 0, \text{ or } q \text{ is odd, } z = 1 \\ (4q)^{k+1} & q \text{ is even, } z = 1, \text{ or } q \text{ is odd, } z = 0. \end{cases}$$

By the previous section, the claim holds for $q = 0$ and all k and z, which gives the base case.

Moving forward, we assume that the claim holds for all values of q' with $q' \le q$ and any k and z. We prove that it holds for $q+1$ with any choices of k and z.

We show that the bound holds for $\text{minima}(\mathcal{F}^k_{q+1,z})$ when $q+1$ is even and $z = 1$, or $q+1$ is odd and $z = 0$. The other cases are nearly identical. Since the bounds are trivial for $k = 1$, we prove them below for $k \ge 2$.

$$\begin{aligned}\text{minima}(\mathcal{F}^k_{q+1,z}) &\le 2 + \sum_{i=0}^{\lfloor k/2\rfloor} (4i)^{q+1} + \sum_{i=0}^{\lfloor (k-1)/2\rfloor} (4i)^{q+1} + \sum_{i=0}^{\lfloor (k+1)/2\rfloor} (4i)^{q+1} + \sum_{i=0}^{\lfloor k/2\rfloor} (4i)^{q+1} \\ &\le 4 \cdot \frac{k}{2} \cdot (2k)^{q+1} + (2(k+1))^{q+1} \le (2k)^{q+2} + (3k)^{q+1} \le (4k)^{q+2}.\end{aligned}$$

10.4 Proof of Theorem 8, Claim 2

We restate the claim:

Proposition 6 (Claim 2 of Theorem 8)**.** *Suppose f is a symmetric unimodal mapping whose maximal cycle is of length $p = 2^q$. For any $k \in \mathbb{N}$, there exists $g \in \mathcal{N}(u, 2)$ with width $u = O((4k)^{q+1}/\epsilon)$ such that $L_\infty f^k, g \le \epsilon$. Moreover, if $f = f_{tent,r}$, then there exists g of width $O((4k)^{q+1})$ with $g = f^k$.*

Proof. This part follows the bound on monotone pieces of f^k given in Proposition 4 and a simple neural network approximation bound.

Lemma 8. *Consider some continuous $f : [0,1] \to [0,1]$ with $M(f) \leq m$. For any $\epsilon \in (0,1)$, there exists $g \in \mathcal{N}(u,2)$ of width $u = O(\frac{m}{\epsilon})$ such that $L_\infty(f,g) \leq \epsilon$.*

Proof. A monotone function mapping to $[0,1]$ can be ϵ-approximated by a piecewise-linear function with $O(\frac{1}{\epsilon})$ pieces, and hence, a 2-layer ReLU network of width $O(\frac{1}{\epsilon})$.

Every monotone piece can be approximated as such, which means that g has width $O(\frac{m}{\epsilon})$. □

For the case where $f = f_{\text{tent},r}$ for some r, it is always true that $\left|\frac{d}{dx} f^k_{\text{tent},r}(x)\right| = (2r)^k$, except when x is a local maximum or minimum. Thus, every monotone piece of f^k is linear, and f can be exactly expressed with a piecewise linear function with $O((4q)^{k+1})$ pieces, and also a ReLU neural network of width $((4q)^{k+1})$. □

10.5 Proof of Theorem 8, Claim 4

Recall that for unimodal $f : [0,1] \to [0,1]$ and threshold $t \in (0,1)$,

$$\mathcal{H}_{f,t} := \{[[f^k]]_t : k \in \mathbb{N}\}$$

is the hypothesis class under consideration.

Proposition 7 (Claim 4 of Theorem 8)**.** *Suppose f is a symmetric unimodal mapping whose maximal cycle is of length $p = 2^q$. For any $t \in (0,1)$, $VC(\mathcal{H}_{f,t}) \leq 18p^2$.*

This proof is involved and requires some setup and new definitions.

10.5.1 Notation

Let $\{0,1\}^{\mathbb{N}}$ represent all countable infinite sequences of Boolean values, and let $\{0,1\}^*$ represent all finite sequences (including the empty sequence).

For $y \in \{0,1\}^{\mathbb{N}}$, let $y_{i:j} = (y_i, \dots, y_j) \in \{0,1\}^{j-i+1}$ and $y_{i:} = (y_i, y_{i+1}, \dots) \in \{0,1\}^{\mathbb{N}}$. For $w \in \{0,1\}^n, w' \in \{0,1\}^{n'}$, let $ww' = w \circ w' \in \{0,1\}^{n+n'}$ be their concatenation. Let $w^j = w \circ w \circ \cdots \circ w \in \{0,1\}^{jn}$.

10.5.2 Iterated Boolean-Valued Functions, Regular Expressions, and VC-Dimension

Before we give the main result, we give a way to upper-bound the VC-dimension of countably infinite hypothesis classes $\mathcal{H} = \{h_1, h_2, \dots, \} \subseteq ([0,1] \to \{0,1\})$. For some $x \in \mathcal{X}$, define $s_{\mathcal{H}} : [0,1] \to \{0,1\}^{\mathbb{N}}$ as $s_{\mathcal{H}}(x) = (h_i(x))_{i \in \mathbb{N}}$. We by $\mathcal{H}$ over all choices of $x \in [0,1]$:

$$\mathcal{S}_{\mathcal{H}} = \{s_{\mathcal{H}}(x) : \ x \in [0,1]\} \subset \{0,1\}^{\mathbb{N}}.$$

With this notation, $\mathcal{H}$ shatters d points if and only if there exist $y^{(1)}, \dots, y^{(d)} \in \mathcal{S}_{\mathcal{H}}$ such that $|\{(y^{(1)}_j, \dots, y^{(n)}_j) : j \in \mathbb{N}\}| = 2^d$. We equivalently say that $y^{(1)}, \dots, y^{(d)}$ are shattered.

Here's where the idea of Regular Expressions (Regexes) comes in. If we can show all elements in $\mathcal{S}_{\mathcal{H}}$ are represented by some infinite-length Regex, then we can upper-bound the number of points $\mathcal{H}$ can shatter, which is necessary to bound the expressive capacity of unimodal functions with recursive properties.

To that end, we first introduce a different notion of shattering. Then, we'll give an upper-bound for the VC-dimension of $\mathcal{H}$ when we have a Regex for $\mathcal{S}_{\mathcal{H}}$.

Definition 11. *We say that $\mathcal{H}$ (or $\mathcal{S}_{\mathcal{H}}$) **weakly shatters** d points if there exist $w^{(1)} \circ y^{(1)}, \dots, w^{(d)} \circ y^{(d)} \in \mathcal{S}_{\mathcal{H}}$ for $w^{(1)}, \dots, w^{(d)} \in \{0,1\}^*$ such that $y^{(1)}, \dots, y^{(d)}$ are shattered. Let the **weak VC-dimension** of $\mathcal{H}$ represent the maximum number of points $\mathcal{H}$ can weakly shatter and denote it $VC_{weak}(\mathcal{H}) = VC_{weak}(\mathcal{S}_{\mathcal{H}})$.*

Using this notation, we can extend our notion of weak VC-dimension to any subset of $\{0,1\}^{\mathbb{N}}$, whether or not it corresponds to a hypothesis class. If $\mathcal{H} \subset S \subset \{0,1\}^{\mathbb{N}}$, then $\text{VC}_{\text{weak}}(\mathcal{H}) \leq \text{VC}_{\text{weak}}(S)$.

Note that if $\mathcal{H}$ shatters d points, then it also trivially weakly shatters d points. We can get this by taking $w_1 = \dots, w_d$ to be the empty strings. Thus, the $\text{VC}(\mathcal{H}) \leq \text{VC}_{\text{weak}}(\mathcal{H})$.

A Regex is a recursively defined subset of $\{0,1\}^{\mathbb{N}}$ that can be represented by a string. We describe how a Regex $R \subseteq \{0,1\}^{\mathbb{N}}$ can be defined below.

- One way to define a Regex is with a repeating sequence w^∞ for $w \in \{0,1\}^n$. That is,
$$w^\infty = \{y \in \{0,1\}^{\mathbb{N}} : y_{in+1:(i+1)n} = w, \forall i \in \mathbb{N}\}.$$
For instance, $(011)^\infty = \{(0,1,1,0,1,1,0,1,1,\dots)\}$.
- For $w \in \{0,1\}^n$, if R is a Regex, then wR is also a Regex. This means satisfying sequences must start with w and then the remainder of the bits must satisfy R.
$$wR = \{y \in \{0,1\}^{\mathbb{N}} : y_{1:n} = w, y_{n+1:} \in R\}.$$
- w^*R is also a Regex, where w^* represents any number of recurrences of the finite sequence s. That is,
$$w^*R = \cup_{j=0}^{\infty} w^j R.$$
- If R' is also a Regex, then so is $R \cup R'$.
- If R' is also a Regex, then so is $R \oplus R'$, where the odd entries of sequences in $R \oplus R'$ concatenated together must be in R and the even entries must be in R'.
$$R \oplus R' = \{y \in \{0,1\}^{\mathbb{N}} : y_{1,3,5,\dots} \in R, y_{2,4,6,\dots} \in R'\}.$$

Now, we can create a recursive upper-bound on the number of points $\mathcal{H}$ can weakly shatter. To do so, we assume that $\mathcal{H} \subseteq R$ for some Regex R and bound the weak VC dimension of R.

Lemma 9. *Consider infinite-length Regexes R, R', R'' and $w \in \{0,1\}^n$.*

1. *If $R = w^\infty$, then $VC_{weak}(R) \le \log_2 n$.*
2. *If $R = wR'$, then $VC_{weak}(R) \le VC_{weak}(R') + \log_2 n + 1$.*
3. *If $R = w^*R'$, then $VC_{weak}(R) \le VC_{weak}(R') + \log_2 n + 1$.*
4. *If $R = R' \cup R''$, then $VC_{weak}(R) \le VC_{weak}(R') + VC_{weak}(R'')$.*
5. *If $R = R' \oplus R''$, then $VC_{weak}(R) \le 4\max(VC_{weak}(R'), VC_{weak}(R'')) + 2$.*

Proof. 1. If $R = w^\infty$, then the set $Y = \{y : w \circ y \in w^\infty, w \in \{0,1\}^*\}$ contains at most n elements. Hence,
$$\left|\{(y_j^{(1)}, \dots, y_j^{(d)}) : j \in \mathbb{N}\}\right| \le n$$
for any fixed $y^{(1)}, \dots, y^{(d)} \in Y$, and no more than $d = \log_2 n$ points can be weakly shattered.

2. Suppose R weakly shatters d points, so $y^{(1)}, \dots, y^{(d)}$ are shattered for some $w^{(1)} \circ y^{(1)}, \dots, w^{(d)} \circ y^{(d)} \in R$. If $Y = \{(y_j^{(1)}, \dots, y_j^{(d)}) : j \in \mathbb{N}\}$ and $Y_n = \{(y_j^{(1)}, \dots, y_j^{(d)}) : j \le n\}$, then $|Y| = 2^d$ and $|Y_n| \le n$. There exists some $v \in \{0,1\}^{1+\log_2 n}$ such that $v \circ \sigma \in Y \setminus Y_n$ for all $\sigma \in \{0,1\}^{d-1-\log_2 n}$. Therefore,
$$|\{(y_j^{(2+\log_2 n)}, \dots, y_j^{(d)}) : j > n\}| = 2^{d-1-\log_2 n},$$
and there exist $d - 1 - \log_2 n$ points that can be weakly shattered by R', since none of the labelings with w are necessary.

3. Once again, suppose R weakly shatters d points, $y^{(1)}, \dots, y^{(d)}$ for $w^{(1)} \circ y^{(1)}, \dots, w^{(d)} \circ y^{(d)} \in R$. Because each $w^{(i)} \circ y^{(i)} \in w^*R'$, there exists an index ℓ_i such that $(w^{(i)} \circ y^{(i)})_{1:\ell_i} = w^{\ell_i/n}$ and $(w^{(i)} \circ y^{(i)})_{\ell_i+1:} \in R'$. Without loss of generality, assume $y^{(1)}, \dots, y^{(d)}$ are ordered such that $\ell_i - \left|w^{(i)}\right|$ decreases. That is, the first $1 + \log_2 n$ sequences are the ones that "leave w^* last." Let $\ell^* := \ell_{1+\log_2 n} - \left|w^{(1+\log_2 n)}\right|$. Define Y and Y_{ℓ^*} analogously to the previous part and note that $|Y| = 2^d$. Because Y_{ℓ^*} corresponds only to labelings where the first $1 + \log_2 n$ elements come from subsets of w^∞, there exists some $v \in \{0,1\}^{1+\log_2 n}$ such that $v \circ \sigma \in Y \setminus Y_{\ell^*}$ for all $\sigma \in \{0,1\}^{d-1-\log_2 n}$. As before, there exist $d - 1 - \log_2 n$ points that can be weakly shattered by R'

4. There is no set of $\mathrm{VC}_{\mathrm{weak}}(R')+1$ and $\mathrm{VC}_{\mathrm{weak}}(R'')+1$ points that can be weakly shattered by R' and R'' respectively. Any $\mathrm{VC}_{\mathrm{weak}}(R')+\mathrm{VC}_{\mathrm{weak}}(R'')+1$ points in R must have at either $\mathrm{VC}_{\mathrm{weak}}(R')+1$ points in R' or $\mathrm{VC}_{\mathrm{weak}}(R'')+1$ points in R''. Thus, at least one subset cannot be shattered.

5. Suppose without loss of generality that $d := \mathrm{VC}_{\mathrm{weak}}(R') \geq \mathrm{VC}_{\mathrm{weak}}(R'')$. Consider any $w^{(1)} \circ y^{(1)}, \ldots, w^{(d)} \circ y^{(4d+3)} \in R$. WLOG, assume that $\left|w^{(1)}\right|, \ldots, \left|w^{(2d+2)}\right|$ are even, which implies that $w^{(1)}_{\mathrm{odd}} \circ y^{(1)}_{\mathrm{odd}}, \ldots w^{(2d+2)}_{\mathrm{odd}} \circ y^{(2d+2)}_{\mathrm{odd}} \in R'$ and $w^{(1)}_{\mathrm{even}} \circ y^{(1)}_{\mathrm{even}}, \ldots w^{(2d+2)}_{\mathrm{even}} \circ y^{(2d+2)}_{\mathrm{even}} \in R''$. Therefore,

$$\begin{aligned}
\left|\{(y_j^{(1)}, \ldots, y_j^{(4d+3)}) : j \in \mathbb{N}\}\right| &\leq 2^{2d+1} \left|\{(y_j^{(1)}, \ldots, y_j^{(2d+2)}) : j \in \mathbb{N}\}\right| \\
&\leq 2^{2d+1} \left(\left|\{(y_j^{(1)}, \ldots, y_j^{(2d+2)}) : j \in \mathbb{N}_{\mathrm{odd}}\}\right| + \left|\{(y_j^{(1)}, \ldots, y_j^{(2d+1)}) : j \in \mathbb{N}_{\mathrm{even}}\}\right|\right) \\
&\leq 2^{2d+1} \cdot 2 \sum_{i=0}^{d} \binom{2d+2}{i} < 2^{2d+2} \cdot 2^{2d+1} = 2^{4d+3}.
\end{aligned}$$

The last line follows by the Sauer Lemma. Thus, R cannot shatter $4d+3$ points if R' and R'' cannot shatter d points.

□

Here's an example of how to apply our regex rules:

$$\begin{aligned}
\mathrm{VC}_{\mathrm{weak}}(1^*0(01)^\infty \cup 10^\infty) &\leq \mathrm{VC}_{\mathrm{weak}}(1^*0(01)^\infty) + \mathrm{VC}_{\mathrm{weak}}(10^\infty) \\
&\leq 1 + \mathrm{VC}_{\mathrm{weak}}(0(01)^\infty) + 1 + \mathrm{VC}_{\mathrm{weak}}(0^\infty) \\
&\leq 2 + 1 + \mathrm{VC}_{\mathrm{weak}}((01)^\infty) \\
&\leq 3 + 1 = 4.
\end{aligned}$$

10.5.3 Proof of the Proposition 7

Recall that we consider the hypothesis class

$$\mathcal{H}_{f,t} := \{[[f^k]]_t : k \in \mathbb{N}\}$$

for symmetric unimodal f and $t \in (0,1)$.

To build up the argument, we first bound the VC-dimension for two simple cases.

- First, we consider the case when f has no fixed point. Thus, for all $x \in (0,1]$, $f(x) < x$, which means that the sequence $f(x), f^2(x), \ldots$ is decreasing.

  If the threshold t is 0 or is greater than $f(\frac{1}{2})$, then the sequence will be all 0's or 1's, which will imply that $\mathrm{VC}(\mathcal{H}_{f,t}) = 0$. Thus, the only interesting thresholds are $t \in (0, f(\frac{1}{2})]$. Because the sequence is decreasing, $\mathcal{S}_{\mathcal{H}_{f,t}} = 1^*0^\infty$. From Lemma 9, $\mathrm{VC}(\mathcal{H}_{f,t}) \leq \mathrm{VC}_{\mathrm{weak}}(\mathcal{H}_{f,t}) \leq 1$.

- Let $x_1 < \cdots < x_m$ be all the fixed points of f. Suppose $x_m \leq \frac{1}{2}$. By symmetry, for all $j \in [m]$, $f(1-x_j) = x_j$.

  To analyze this function, we partition $[0,1]$ into $2m+2$ intervals: $I_0 = [0, x_1), I_0' = (1-x_1, 1]$, $I_m = [x_m, \frac{1}{2}]$, $I_m' = (\frac{1}{2}, 1-x_m]$, $I_j = [x_j, x_{j+1})$, and $I_j' = (1-x_{j+1}, 1-x_j]$ for all $j \in \{1, \ldots, m-1\}$ (visualized in Figure 17).

  Because f is unimodal and because the edges of all intervals map to fixed points, for all $j \in \{0, \ldots, m\}$, $f(I_j') = f(I_j) = I_j$. In this case, it must be the case that $q = 0$ because f cannot have a 2-cycle. Such a cycle is impossible because it would have to be contained entirely in some I_j. In those intervals, it must be the case that either $\forall x \in I_j$, $f(x) \geq x$, or $\forall x \in I_j$, $f(x) \leq x$ (if this were not the case, then this would imply the existence of a fixed point other than x_j in I_j). Thus, cyclic behavior within an interval is impossible.

  Thus, we can construct a Regex to represent the itinerary of any $x \in [0,1]$: $\bigcup_{j=0}^{m} I_j^\infty$.[14] Now, we consider all possible locations of threshold t:

[14] This is a massive abuse of notation, but we use the same Regex notation to denote the intervals that are traversed as we use to denote the values of Boolean sequence.

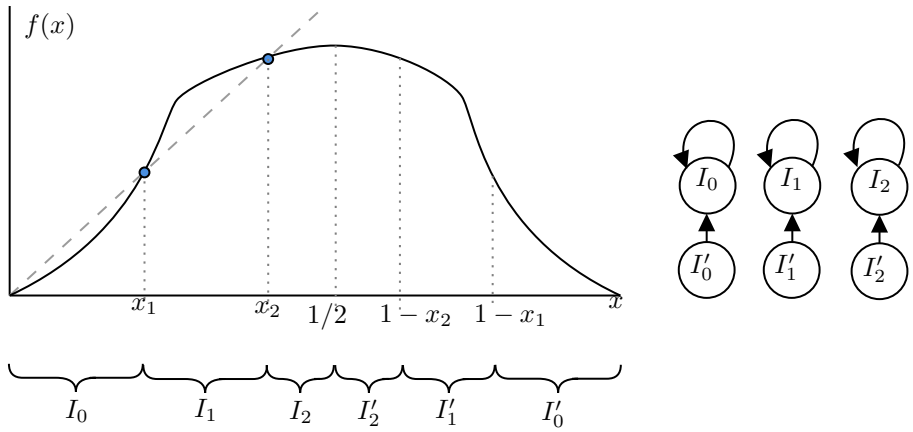

Figure 17: A plot of f with fixed point $m = 2$ fixed points—both less than $\frac{1}{2}$—subdivided into intervals. The relationships of which intervals f maps onto one another are also visualized.

- If $t \in I_j$, such that $f(x) \geq x$ for $x \in I_j$, then $\mathcal{S}_{\mathcal{H}_{f,t}} \subseteq 0^*1^\infty \cup 0^\infty \cup 1^\infty$. By Lemma 9, $\text{VC}_{\text{weak}}(\mathcal{H}_{f,t}) \leq 1$.
- If $t \in I_j$, such that $f(x) \leq x$ for $x \in I_j$, then $\mathcal{S}_{\mathcal{H}_{f,t}} \subseteq 1^*0^\infty \cup 0^\infty \cup 1^\infty$. By Lemma 9, $\text{VC}_{\text{weak}}(\mathcal{H}_{f,t}) \leq 1$.
- If $t \in \bigcup_{j=0}^{m} I_j'$, then $\mathcal{S}_{\mathcal{H}_{f,t}} = 0^\infty$, and $\text{VC}_{\text{weak}}(\mathcal{H}_{f,t}) = 0$.

Now, we give a lemma, which relates the VC-dimension of complex functions to that of simpler ones. Let $\mathcal{F}_q$ refer to the family of symmetric unimodal functions that have a 2^q-cycle but not a 2^{q+1}-cycle.

Lemma 10. *For any $f \in \mathcal{F}_q$ with fixed point $x^* > \frac{1}{2}$ and any $t \in [0,1]$,*

$$VC_{weak}(\mathcal{H}_{f,t}) \leq 4 \max_{g \in \mathcal{F}_{q-1}, t' \in [0,1]} VC_{weak}(\mathcal{H}_{g,t'}) + 10.$$

Proof. Consider some such f. Let $x_1 < \cdots < x_m$ be the fixed points of f where $x_m = x^* > \frac{1}{2}$. Because $\frac{1}{2}$ maximizes f, $f(\frac{1}{2}) \geq x_m > \frac{1}{2}$. This fixed point must the only fixed point no smaller than $\frac{1}{2}$; the existence of another such fixed point would contradict the fact that f is decreasing on $(\frac{1}{2}, 1]$. Thus, $x_1, \ldots, x_{m-1} < \frac{1}{2}$.

We build a recursive relationship by considering f^2 and relating some its output on some segments of $[0,1]$ to other maps with smaller q. For now, we instead attempt to upper-bound the VC-dimension of $\mathcal{H}_{f^2,t}$.

For all $j \in [m]$, unimodality implies that x_j and $1 - x_j$ are the only points that map to x_j and that the following ordering holds.

$$0 < x_1 < \ldots, < x_{m-1} < 1 - x_m < \frac{1}{2} < x_m < 1 - x_{m-1} < \ldots, 1 - x_1 < 1.$$

By the Intermediate Value Theorem, there exists some $x_m' \in (x_m, 1 - x_{m-1})$ such that $f(x_m') = f(1 - x_m') = 1 - x_m$ and $f^2(x_m') = f^2(1 - x_m') = x_m$.

We define intervals as follows:

- $I_0 = [0, x_1)$ and $I_0' = (1 - x_1, 1]$.
- For all $j \in [m-2]$, $I_j = [x_j, x_{j+1})$ and $I_j' = (1 - x_{j+1}, 1 - x_j]$.
- $I_{m-1} = [x_{m-1}, 1 - x_m')$ and $I_{m-1}' = (x_m', 1 - x_{m-1}]$.
- $I_m = [1 - x_m', 1 - x_m)$ and $I_m' = (x_m, x_m']$.
- $I_{m+1} = [1 - x_m, \frac{1}{2})$, and $I_{m+1}' = [\frac{1}{2}, x_m]$.

For any $j \in \{0, \ldots, m+1\}$, f is increasing on all intervals I_j and decreasing on I'_j. By symmetry, $f(I_j) = f(I'_j)$. For all $j \in \{0, \ldots, m-2\}$, $f(I_j) = I_j$. $f(I_{m-1}) = I_{m-1} \cup I_m$, $f(I_m) = I_{m+1} \cup I'_{m+1}$, and $f(I_{m+1}) \subseteq I'_m$, because $f(\frac{1}{2}) \in [x_m, x'_m)$.[15]

From there, we obtain additional properties for f^2: $f^2(I_{m-1}) = I_{m-1} \cup I_m \cup I_{m+1} \cup I'_{m+1}$, $f^2(I_m) \subseteq I'_m$, and $f^2(I_{m+1}) \subset I_{m+1} \cup I'_{m+1}$. This suggests that there is recurrent structure that we can take advantage of to count all of the patterns.

Let $J_{m+1} := I_{m+1} \cup I'_{m+1}$. We create a Regex to track the behavior of iterates f^2, which we visualize in Figure 18:

$$\bigcup_{j=0}^{m-2} I_j^\infty \;\cup\; I_{m-1}^* I_m I_m'^\infty \;\cup\; I_m'^\infty \;\cup\; I_{m-1}^* J_{m+1}^\infty.$$

When an iterate of f^2 gets "stuck" in one of $I_0, I_1, \ldots, I_{m-1}$, it must either be at a fixed point, be strictly increasing, or be strictly decreasing. To suggest otherwise would imply the existence of another fixed point in those intervals, because f^2 is monotonically increasing or decreasing in all of those and either all x yield $f^2(x) \geq x$ or $f^2(x) \leq x$.

For the remaining intervals, one might notice in Figure 18 that zooming in on the intervals I_m, J_{m+1}, and I'_m for f^2 gives what looks like unimodal maps.[16] We take advantage of that structure to bound the complexity of the 0/1 Regexes for those intervals. We can formalize this by defining symmetric unimodal mappings h_m and h_{m+1} and bijective monotonic mappings $\phi_m : I'_m \to (0, 1]$ (increasing) and $\phi_{m+1} : J_{m+1} \to [0, 1]$ (decreasing) such that:

- For $x \in I_m$, $f^2(x) = \phi_m^{-1} \circ h_m \circ \phi_m(1 - x)$.
- For $x \in I'_m$, $f^2(x) = \phi_m^{-1} \circ h_m \circ \phi_m(x)$.
- For $x \in J_{m+1}$, $f^2(x) = \phi_{m+1}^{-1} \circ h_{m+1} \circ \phi_{m+1}(x)$.

Because f cannot have a cycle of length 2^{q+1}, h_m and h_{m+1} may not have cycles of length 2^q. Thus, we can reason inductively about how iterates behave when they're trapped in those intervals.

We do another case analysis of the 0/1 Regexes induced by different choices of t.

- If $t \in I_j$ for $j \in \{0, \ldots, m-1\}$, then $\mathcal{S}_{\mathcal{H}_{f^2,t}} \subseteq 0^\infty \cup 1^\infty \cup 0^*1^\infty \cup 1^*0^\infty$ because a sequence of iterates only crosses t if it enters the correct interval I_j, where the iterate then will be stuck and must monotonically increase or decrease. By Lemma 9, $\mathrm{VC}_{\mathrm{weak}}(\mathcal{H}_{f^2,t}) \leq 2$.
- If $t \in I'_j$ for $j \in \{0, \ldots, m-1\}$, then $\mathcal{S}_{\mathcal{H}_{f^2,t}} = 0^\infty$, and $\mathrm{VC}_{\mathrm{weak}}(\mathcal{H}_{f^2,t}) = 0$.
- If $t \in I_m$, then $\mathcal{S}_{\mathcal{H}_{f^2,t}} = 0^\infty \cup 1^\infty \cup 0^*1^\infty$. Then, $\mathrm{VC}_{\mathrm{weak}}(\mathcal{H}_{f^2,t}) \leq 1$.
- If $t \in J_{m+1}$, then $\mathcal{S}_{\mathcal{H}_{f^2,t}} = 0^\infty \cup 0^*1^\infty \cup 0^*J_{m+1}^\infty$. Because h_{m+1} has at most a cycle of length 2^{q-1}, we have that
$$\mathrm{VC}_{\mathrm{weak}}(\mathcal{H}_{f^2,t}) \leq 2 + \max_{t'} \mathrm{VC}_{\mathrm{weak}}(\mathcal{H}_{h_{m+1},t'}).$$
- If $t \in I'_m$, then $\mathcal{S}_{\mathcal{H}_{f^2,t}} = 0^\infty \cup 0^*I_m'^\infty$. This gives us that
$$\mathrm{VC}_{\mathrm{weak}}(\mathcal{H}_{f^2,t}) \leq 1 + \max_{t'} \mathrm{VC}_{\mathrm{weak}}(\mathcal{H}_{h_m,t'}).$$

[15] This must be the case for the assumptions to be met. If $f(\frac{1}{2}) < x_m$, then x_m cannot be a fixed point because $\frac{1}{2}$ maximizes f. If $f(\frac{1}{2}) > x'_m$, then there exists a 3-cycle with points in I_{m+1}, I'_{m-1}, I_m, which contradicts the assumption that we only have power-of-two cycles.

[16] We use similar techniques here to those used in Section 10.3.2.

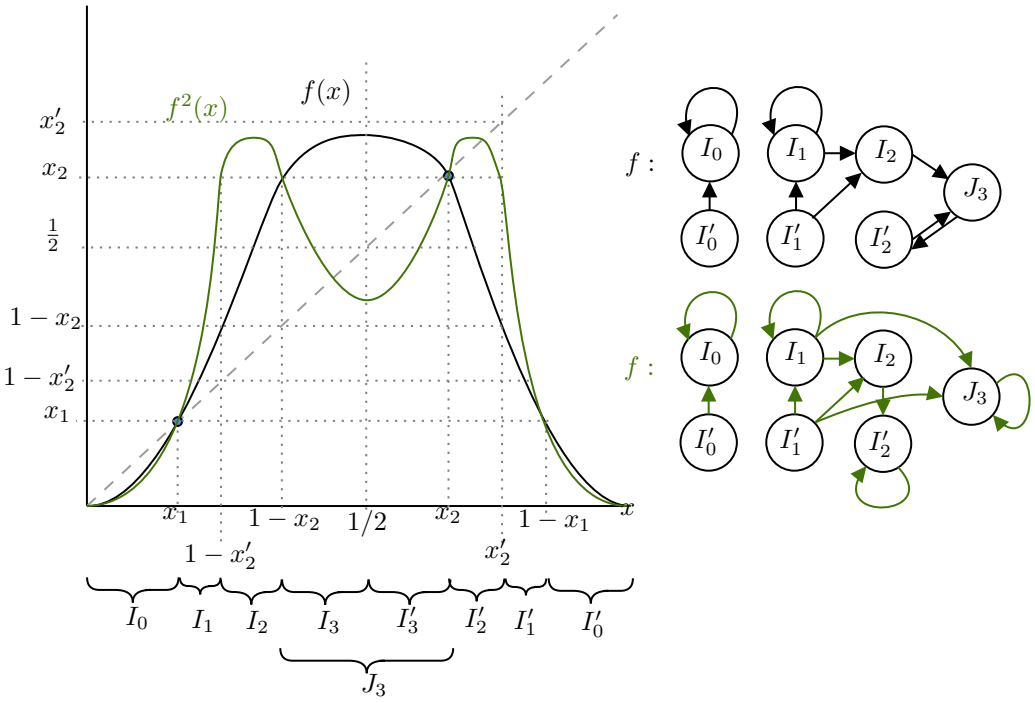

Figure 18: Like Figure 17, plot of f and f^2 with $m = 3$ fixed points with $x_m > \frac{1}{2}$ and visualizes the mappings between intervals.

To get $\mathrm{VC}_{\mathrm{weak}}(\mathcal{H}_{f,t})$, notice that $\mathcal{S}_{\mathcal{H}_{f,t}} = \mathcal{S}_{\mathcal{H}_{f^2,t}} \oplus \mathcal{S}_{\mathcal{H}'_{f^2,t}}$, where $\mathcal{H}'_{f^2,t}$ refers to the outcome of all odd iterates of f. We show that $\mathcal{S}_{\mathcal{H}'_{f^2,t}} \subseteq \mathcal{S}_{\mathcal{H}_{f^2,t}}$ because the latter could induce all sequences produced by the former by starting with some x' such that $f^2(x') = f(x)$. Thus, by Lemma 9,

$$\begin{aligned}
\mathrm{VC}_{\mathrm{weak}}(\mathcal{H}_{f,t}) &\le 4 \max_{t'} \mathrm{VC}_{\mathrm{weak}}(\mathcal{H}_{f^2,t'}) + 2 \\
&\le 4 \max(2 + \max_{t'} \mathrm{VC}_{\mathrm{weak}}(\mathcal{H}_{h_{m+1},t'}), 1 + \max_{t'} \mathrm{VC}_{\mathrm{weak}}(\mathcal{H}_{h_m,t'})) + 2 \\
&\le 4 \max_{g \in \mathcal{F}_{q-1},t'} \mathrm{VC}_{\mathrm{weak}}(\mathcal{H}_{g,t'}) + 10.
\end{aligned}$$

□

Now, we prove a bound on the VC-dimension for arbitrary q by induction with Lemma 10 to show that for $\mathrm{VC}(\mathcal{H}_{f,t}) \le 18 \cdot 4^q$.

This holds when $q = 0$. There are two possible cases for the fixed point of such an f. If the the largest fixed point is smaller than $\frac{1}{2}$, then, by the simple cases explored at the beginning, $\mathrm{VC}_{\mathrm{weak}}(\mathcal{H}_{f,t}) \le 1$. Otherwise, we apply Lemma 10 along with the the other simple case—which tells us what happens when there are no fixed point—to get that $\mathrm{VC}_{\mathrm{weak}}(\mathcal{H}_{f,t}) \le 4(1) + 10 = 14$. This trivially satisfies the proposition.

For the inductive step for arbitrary q, we iteratively apply Lemma 10 to obtain the final bound.

$$\begin{aligned}
\mathrm{VC}(\mathcal{H}_{f,t}) &\le 4 \max_{g \in \mathcal{F}_{q-1},t'} \mathrm{VC}_{\mathrm{weak}}(\mathcal{H}_{g,t'}) + 10 \\
&\le 4^q \max_{g \in \mathcal{F}_0,t'} \mathrm{VC}_{\mathrm{weak}}(\mathcal{H}_{g,t'}) + 10 \sum_{i=0}^{q-1} 4^i \\
&\le 14 \cdot 4^q + \frac{10}{3} 4^q \le 18 \cdot 4^q.
\end{aligned}$$

10.6 Proof of Theorem 9, Claim 4

Proposition 8. *Suppose f is a symmetric unimodal function with a $2^q m$-cycle for odd m. Then for*

$$K = \exp\left(O\left(q + d\log(d+m)\right)\right),$$

$VC(\mathcal{H}_{f,K}) \geq d$ for $\mathcal{H}_{f,K} = \left\{[[f^k]]_{1/2} : k \in [K]\right\}$.

The claim holds by this proposition, since the VC-dimension of $\mathcal{H}_f$ is larger than every d and hence must be infinite.

Proof. The proof of this claim relies on the existence of a lemma that describes a characteristic of odd-period cycles of unimodal functions.

Lemma 11. *Let f be a symmetric unimodal function with some odd cycle $x_1, x_2, \ldots, x_m$ of length $m > 1$ such that $f(x_i) = x_{i+1}$ and $f(x_m) = x_1$. Then, there exists some i such that $x_i < \frac{1}{2}$ and $f(x_i) \geq \frac{1}{2}$.*

Proof. To prove the claim, it suffices to show that the following two cases are impossible: (1) $x_1, \ldots, x_m < \frac{1}{2}$ and (2) $x_1, \ldots, x_m \geq \frac{1}{2}$.

1. Suppose $x_1, \ldots, x_m < \frac{1}{2}$. By unimodality $x_j < x_{j'}$ implies that $f(x_j) < f(x_{j'})$. If x_1 is the smallest element of the cycle, then $f(x_1) > x_1$. For any other x_j, $f(x_j) > x_1$, which means that x_1 cannot be part of a cycle, which contradicts the odd cycle.

2. Suppose instead that $x_1, \ldots, x_m \geq \frac{1}{2}$.

   For this to be the case, $f(\frac{1}{2}) > \frac{1}{2}$ by unimodality. This fact paired with $f(1) < 1$ implies the existence of some $x^* \in (\frac{1}{2}, 1)$ with $f(x^*) = x^*$. Because f is decreasing on $[1/2, 1]$, $f([1/2, x^*)) \subseteq (x^*, 1]$ and $f((x^*, 1]) = [0, x^*)$.

   If $x_1 \in [\frac{1}{2}, x^*)$, then $x_2 \in (x^*, 1]$, and $x_3 \in [\frac{1}{2}, x^*)$. If apply this fact repeatedly, the oddness of m implies that $x_m \in [\frac{1}{2}, x^*)$ and $x_1 \in (x^*, 1]$, a contradiction. □

We show that $\mathrm{VC}(\mathcal{H}_{f^{2^q}, K/2^q}) > d$. If f has a cycle of length $2^q \cdot m$, then f^{2^q} has a cycle of length m. By Sharkovskii's Theorem, for all odd $m' > m$, f^{2^q} also has a cycle of length m'. Let $p_1 < \cdots < p_d$ be the smallest prime numbers greater than m. According to Lemma 12, $p_d \leq \left(\frac{K}{2^q}\right)^{1/d}$ for

$$K = 2^q \left(O(\max(d\log d, m)\right)^d = \exp\left(O\left(q + d\log(d+m)\right)\right).$$

For $j \in [m]$, let $x^{(j)}$ be the point guaranteed by Lemma 11 with $f^{2^q \cdot p_j}(x^{(j)}) = x^{(j)}$, $x^{(j)} < \frac{1}{2}$, and $f^{2^q}(x^{(j)}) \geq \frac{1}{2}$. Therefore, it follows that $f^{2^q \cdot \ell p_j}(x^{(j)}) < \frac{1}{2}$ and $f^{2^q(\ell p_j + 1)}(x^{(j)}) \geq \frac{1}{2}$ for all $\ell \in \mathbb{Z}_{\geq 0}$.

To show that $\mathcal{H}_{f^{2^q}}$ shatters $x^{(1)}, \ldots, x^{(d)}$, we show that for any labeling $\sigma \in \{0,1\}^d$, there exists $h \in \mathcal{H}_{f^{2^q}, K/2^q}$ such that $h(x^{(j)}) = \sigma_j$.

- If $\sigma = (0, \ldots, 0)$, then consider $f^{2^q \cdot k}$, where $k = \prod_{j=1}^n p_j$. Then, for all j, $f^{2^q \cdot k}(x^{(j)}) < \frac{1}{2}$. Because $k \leq p_d^d \leq \frac{K}{2^q}$, there exists some $h \in \mathcal{H}_{f^{2^q}, K/2^q}$ that assigns zero to every $x^{(j)}$.

- Similarly, if $\sigma = (1, \ldots, 1)$, we instead consider $f^{2^q \cdot k}$ for $k = 1 + \prod_{j=1}^n p_j$. Now, for all j, $f^{2^q \cdot k}(x^{(j)}) \geq \frac{1}{2}$, and $k \leq p_d^d \leq \frac{K}{2^q}$, which means there exists satisfactory $h \in \mathcal{H}_{f^{2^q}, K/2^q}$.

- Otherwise, assume WLOG that $(\sigma_1, \ldots, \sigma_\ell) = (0, \ldots, 0)$ and $(\sigma_{\ell+1}, \ldots, \sigma_d) = (1, \ldots, 1)$ for $\ell \in (1, d)$. We satisfy the claim for $f^{2^q \cdot k}$ if we choose some k with $k = q_1 \prod_{i=1}^{\ell} p_i = 1 + q_2 \prod_{i=\ell+1}^{d} p_i$, for some $q_1, q_2 \in \mathbb{Z}_+$.

We find $q_1 \in [\prod_{i=\ell+1}^{d} p_i]$ and $q_2 \in [\prod_{i=1}^{\ell} p_i]$ by choosing them such that:

$$q_1 \prod_{i=1}^{\ell} p_i \equiv 1 \pmod{\prod_{i=\ell+1}^{d} p_i}$$
$$q_2 \prod_{i=\ell+1}^{d} p_i \equiv -1 \pmod{\prod_{i=1}^{\ell} p_i}.$$

This is possible because $p_1, \ldots, p_d$ are prime, and $\gcd\left(\prod_{i=1}^{\ell} p_i, \prod_{i=\ell+1}^{d} p_i\right) = 1$.

Because $k \leq \prod_{i=1}^{d} p_i \leq p_d^d \leq \frac{K}{2^q}$, there must exist some satisfactory $h \in \mathcal{H}_{f^{2^q}, K/2^q}$. □

Lemma 12. *For $m \geq 3$ and any $d \geq 0$, there exist d primes such that $m \leq p_1 < \cdots < p_d$ for*

$$p_d = O(\max(d \log d, m)).$$

Proof. Let $\pi(x) = |\{y \in [x] : y \text{ is prime}\}|$ be the number of primes no larger than x. By the Prime Number Theorem,

$$\frac{x}{\log(x) + 2} \leq \pi(x) \leq \frac{x}{\log(x) - 4},$$

for all $x \geq 55$ (Rosser, 1941). Thus, for some $m' = O(\max(d \log d, m))$, the number of prime numbers smaller than m' is

$$\Omega\left(\frac{d \log d}{\log(d \log d)} + \frac{m}{\log m}\right) = \Omega\left(d + \frac{m}{\log m}\right),$$

and the number between m and m' is $\Omega(d)$. Thus, $p_d \leq m'$. □