On Scrambling Phenomena for Randomly Initialized Recurrent Networks

Vaggos Chatziafratis
UC Santa Cruz

Ioannis Panageas
UC Irvine

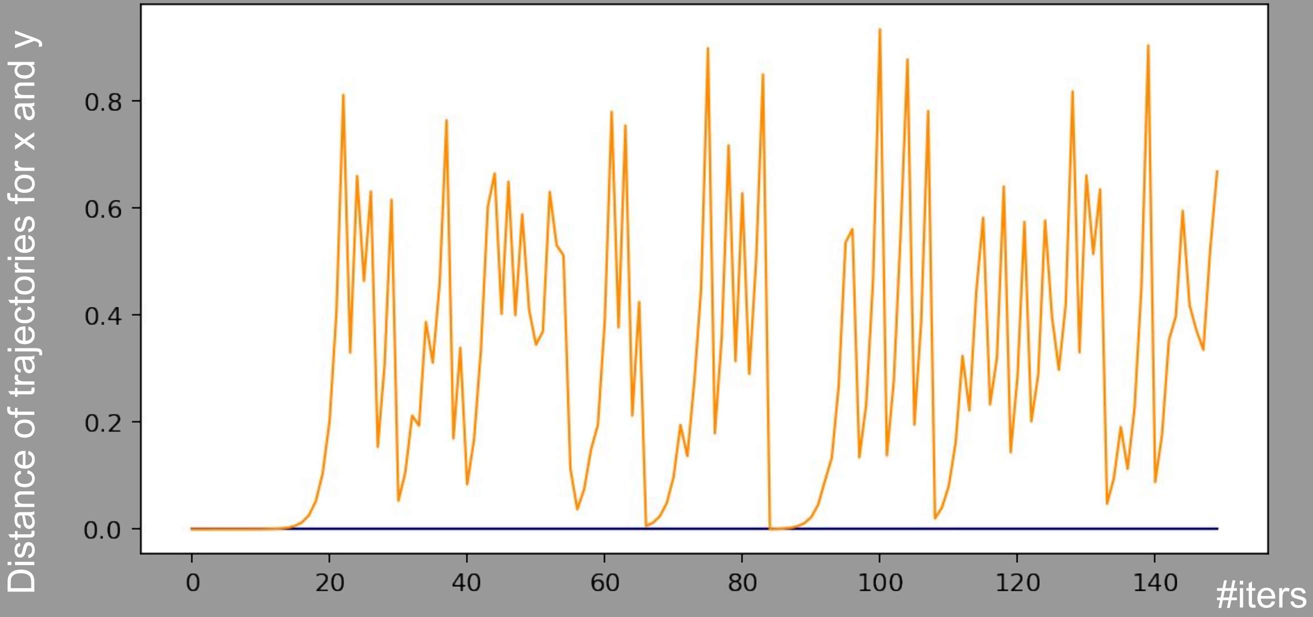
Clayton Sanford
Columbia University

Stelios Stavroulakis
UC Irvine

NeurIPS'22, November, New Orleans

1. Random RNNs exhibit **"scrambling"** phenomena

|Random_RNN(x) - Random_RNN(y)|



⇒ Scrambling ("kiss and separate")

(also see: A recurrent neural network without chaos [Laurent and von Brecht' 16]

Scrambling Phenomena & Li-Yorke Chaos (1975)

Let (X, d) be a compact metric space and let $f : X \to X$ be a continuous map. Two points $x, y \in X$ are called *proximal* if:

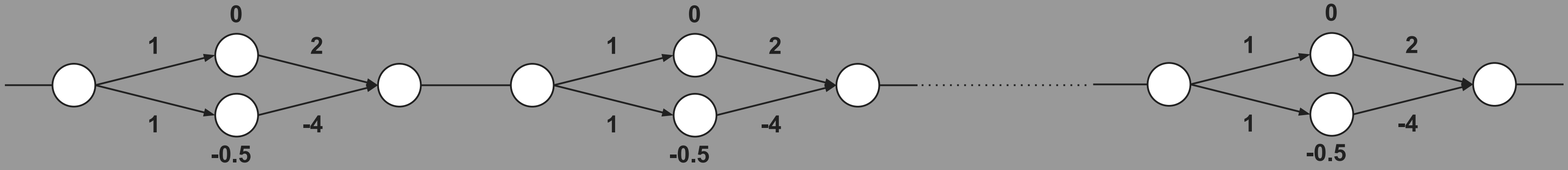
$$\liminf_{n\to\infty} d(f^n(x), f^n(y)) = 0$$

Two points $x, y \in X$ are called *asymptotic* if:

$$\limsup_{n\to\infty} d(f^n(n),\ f^n(y)) = 0$$

A set $Y \subseteq X$ is called scrambled if $\forall x, y \in Y, x \neq y$, the two points are proximal but **not** asymptotic.

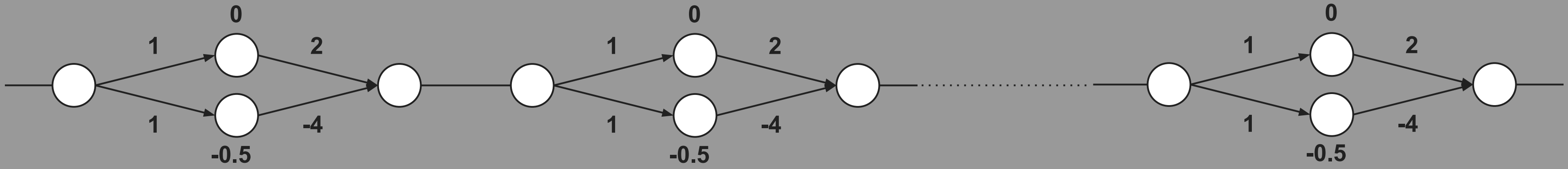
2. RNNs **vs** FNNs, under small noise perturbations

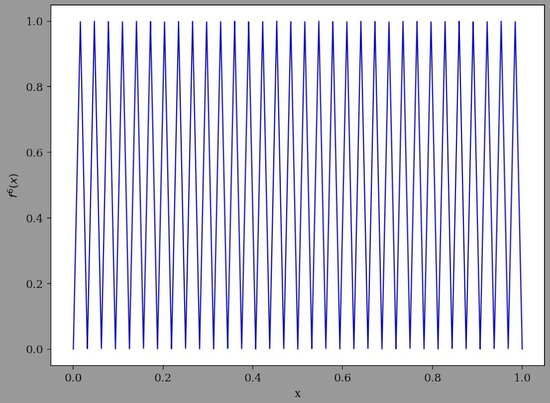


Benefits of Depth in Deep Neural Networks [Telgarsky' 16]

Complexity of Linear Regions in Deep Networks [Hanin, Rolnick '19]

2. RNNs **vs** FNNs, under small noise perturbations

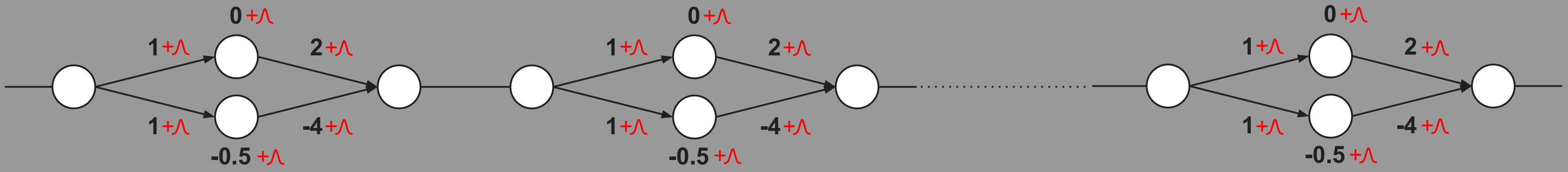


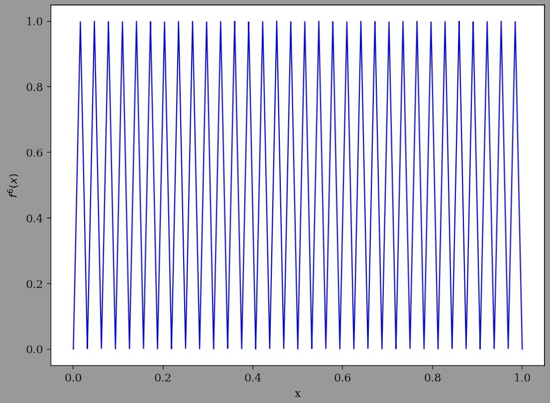


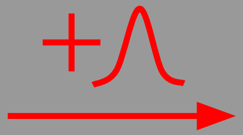
Benefits of Depth in Deep Neural Networks [Telgarsky' 16]

Complexity of Linear Regions in Deep Networks [Hanin, Rolnick '19]

2. RNNs **vs** FNNs, under small noise perturbations



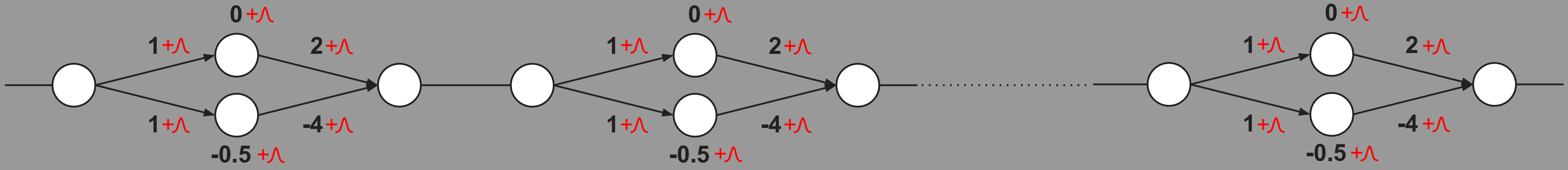


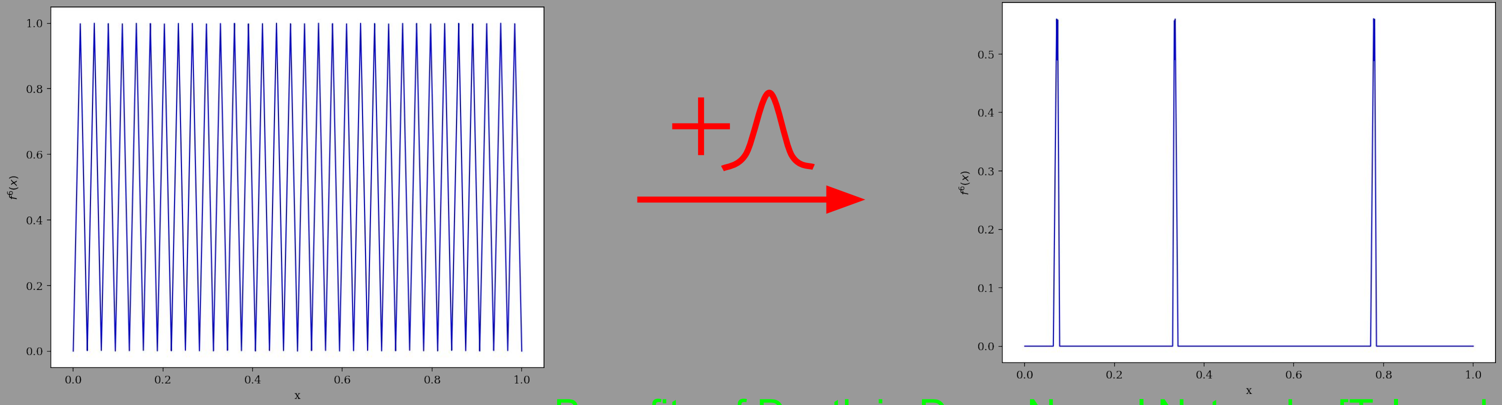


Benefits of Depth in Deep Neural Networks [Telgarsky' 16]

Complexity of Linear Regions in Deep Networks [Hanin, Rolnick '19]

2. RNNs **vs** FNNs, under small noise perturbations

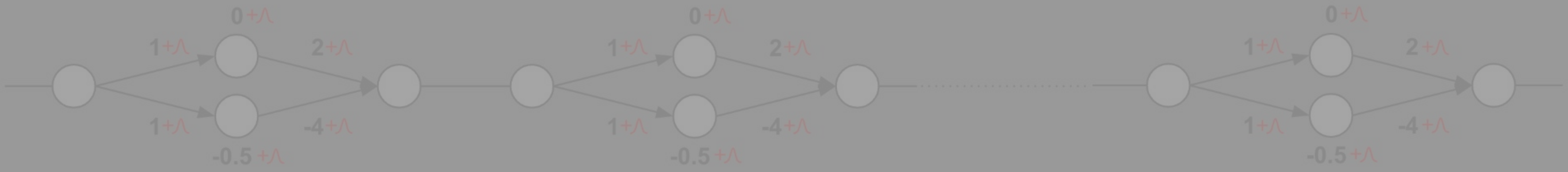


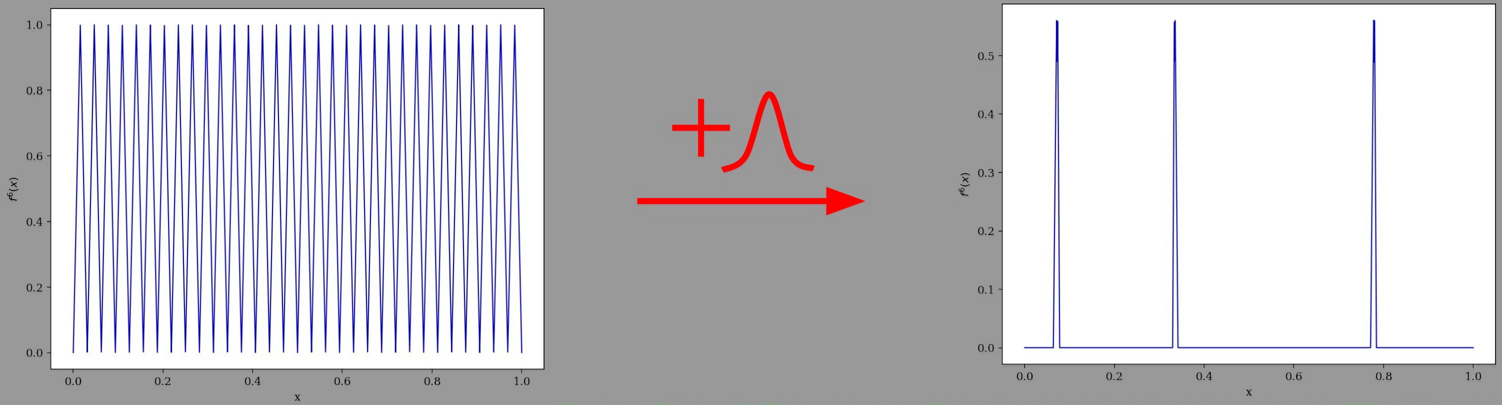


Benefits of Depth in Deep Neural Networks [Telgarsky' 16]

Complexity of Linear Regions in Deep Networks [Hanin, Rolnick '19]

2. RNNs vs FNNs, under small noise perturbations

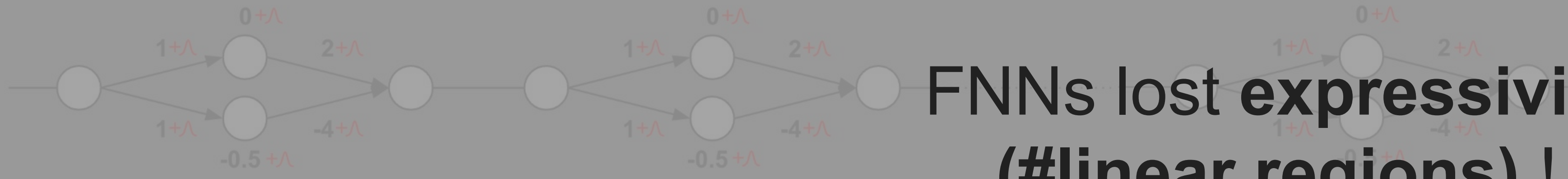




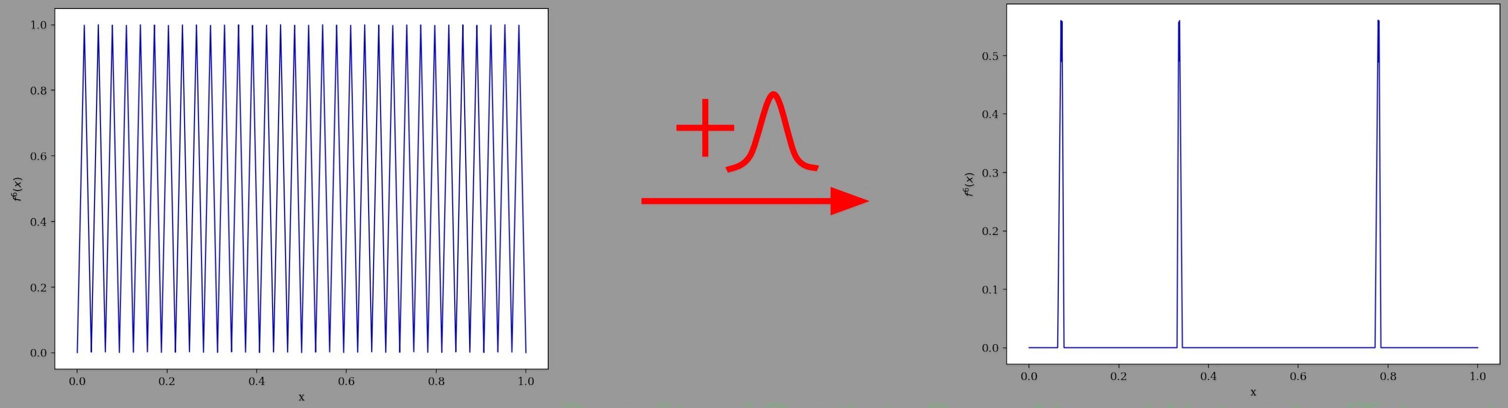
Benefits of Depth in Deep Neural Networks [Telgarsky '16]

Complexity of Linear Regions in Deep Networks [Hanin, Rolnick '19]

2. RNNs **vs** FNNs, under small noise perturbations



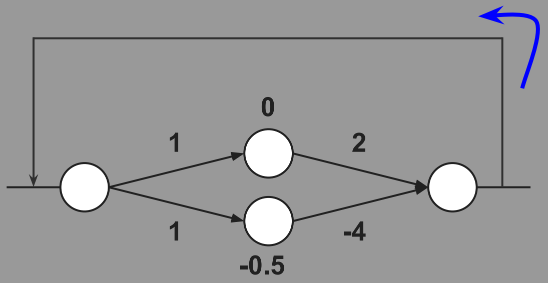
FNNs lost **expressivity (#linear regions)** !



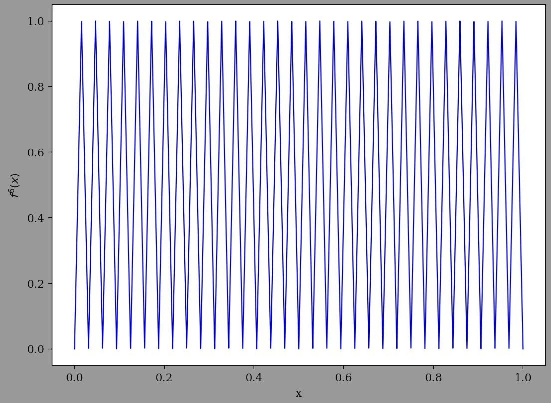
Benefits of Depth in Deep Neural Networks [Telgarsky '16]

Complexity of Linear Regions in Deep Networks [Hanin, Rolnick '19]

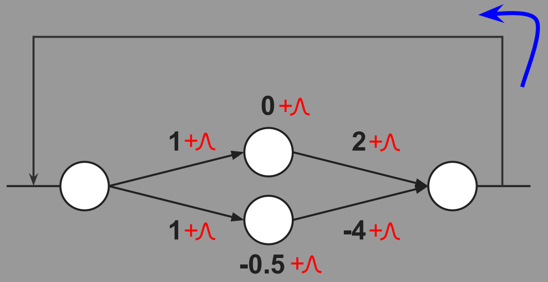
2. But, what happens if we add noise in RNNs ?



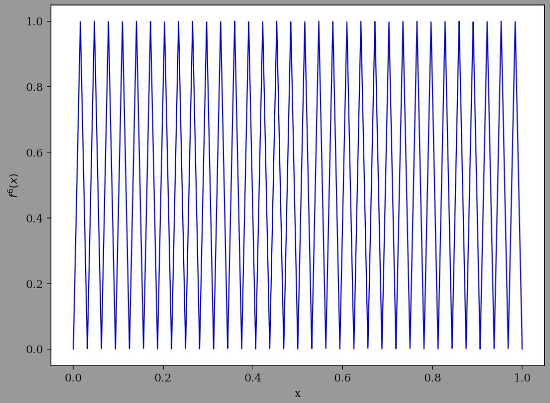
n compositions

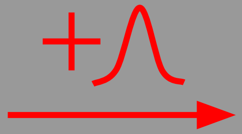


2. But, what happens if we add noise in RNNs ?

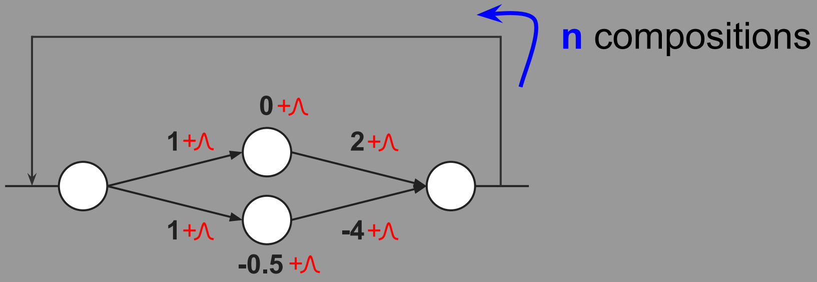


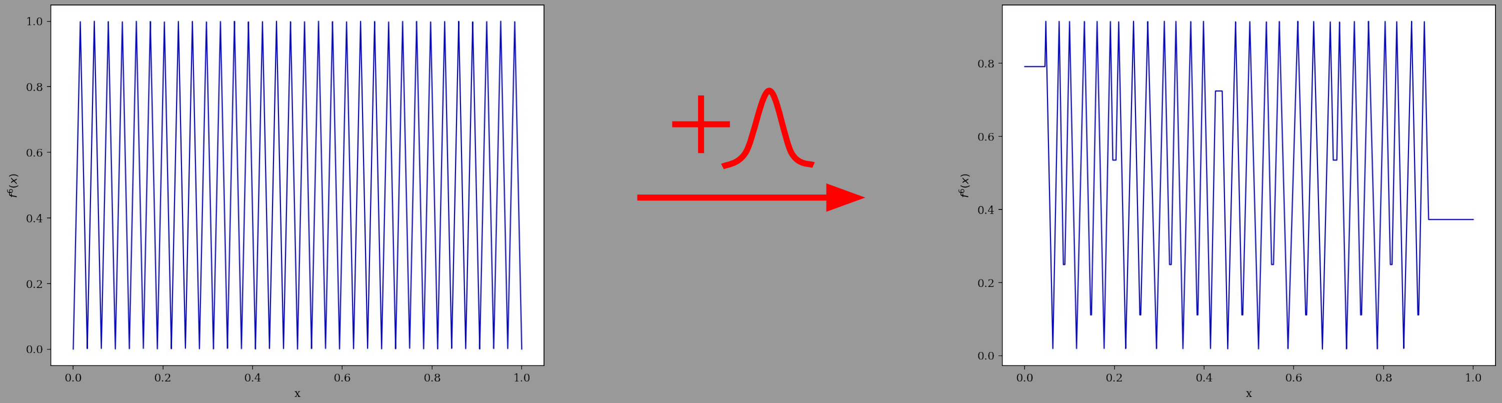
n compositions





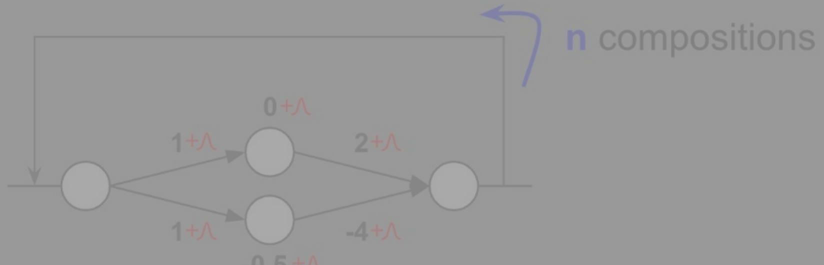
2. But, what happens if we add noise in RNNs ?





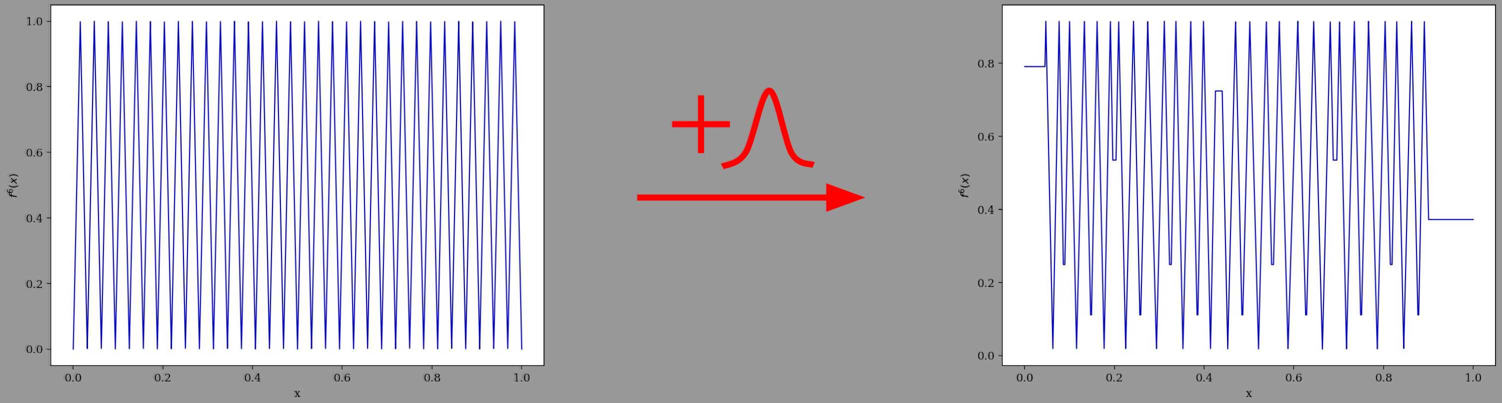
RNNs remain **expressive (#linear regions)** !

2. But, what happens if we add noise in RNNs ?



n compositions

Can we explain this RNNs-vs-FNNs contrast?



RNNs remain **expressive (#linear regions)** !

1. RNNs exhibit “scrambling” phenomena

2. RNNs remain highly expressive
even after small perturbations (**contrary to FNNs**)

RNNs exhibit scrambling phenomena
under standard (e.g., He) random initialization
with probability *p* (small, but constant)
independent of their width

Theorem 1: RNNs may exhibit Chaos at Initialization

Consider $f_k \in RNN\left(k_\sigma, \sigma^2\right)$ initialized with *He* normal initialization $\implies$

$\exists\, \delta > 0,\ k_{He} > 1$: for sufficiently large $k > k_{He},\ f_k$ is Li-Yorke chaotic with probability at least δ, independent of the width.

Theorem 2: Order-to-Chaos Transition for RNNs

For $f_k \in RNN(k, \sigma^2)$ we get 3 regimes as we sweep the init. variance:

1. $a_i \sim \mathcal{N}\left(0, \frac{\Theta(1)}{4k \log k}\right) \Rightarrow \mathbb{P}(f_k \; chaotic) \leq \frac{1}{k}$

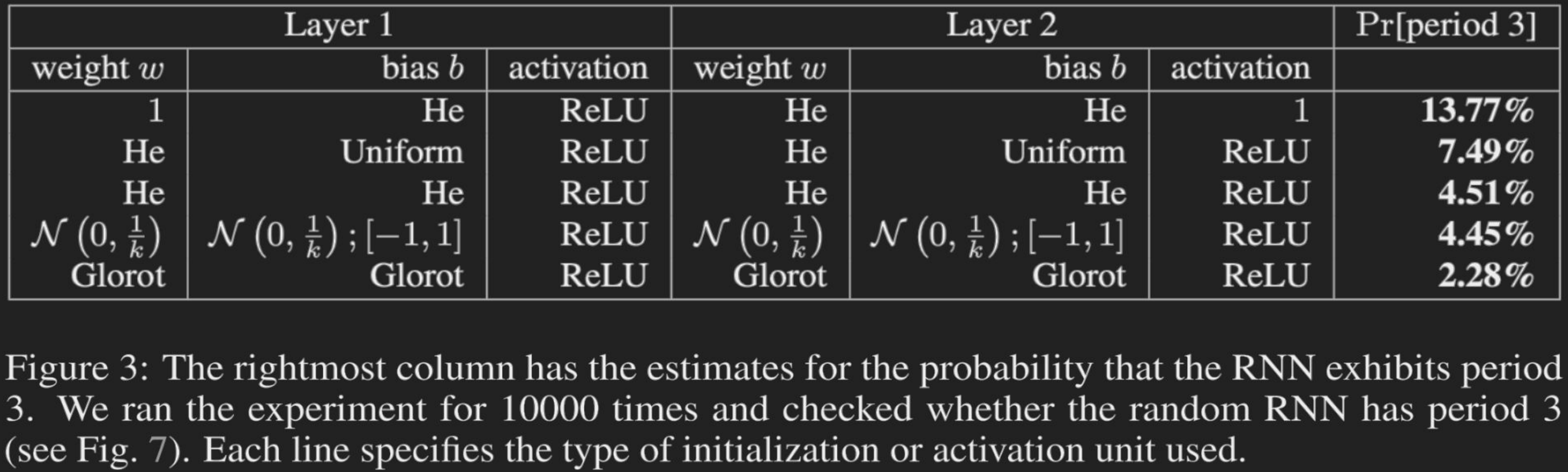
2. $a_i \sim \mathcal{N}\left(0, \frac{2}{k}\right) \Rightarrow \mathbb{P}(f_k \; chaotic) = c > 0$

3. $a_i \sim \mathcal{N}\left(0, \omega\left(\frac{1}{k}\right)\right) \Rightarrow \lim_{k \to \infty} \mathbb{P}(f_k \; chaotic) = 1$

Empirical Verification: Chaos under multiple initializations

We consider the above RNN family with He normal initialization: $Var(w_i) = \frac{2}{fan\!-\!in}$

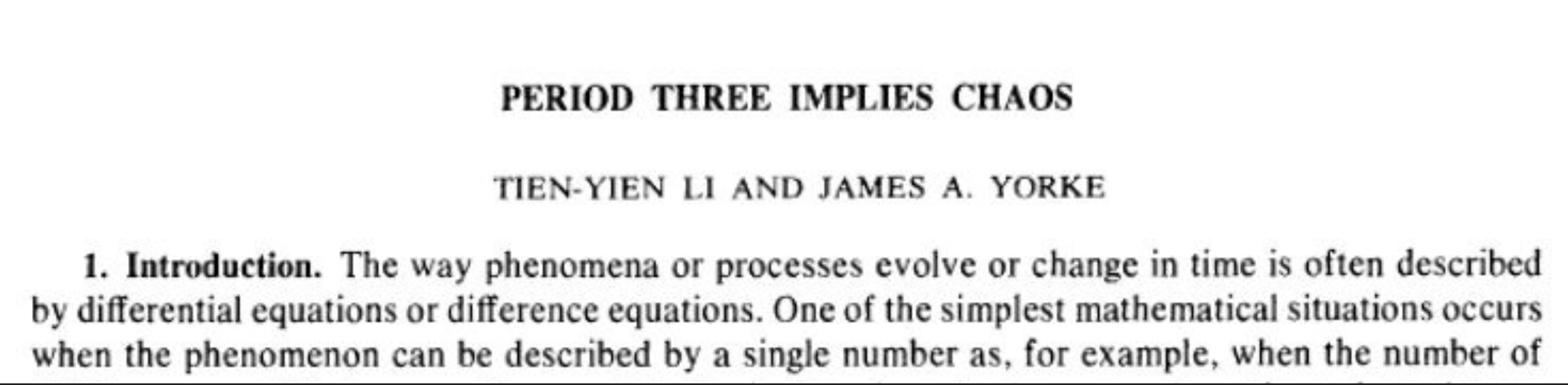
- But also, other initializations, or different architectures, activations.
- Experiments suggest that “Chaos is Robust”.



Layer 1			Layer 2			Pr[period 3]
weight w	bias b	activation	weight w	bias b	activation	
1	He	ReLU	He	He	1	**13.77%**
He	Uniform	ReLU	He	Uniform	ReLU	**7.49%**
He	He	ReLU	He	He	ReLU	**4.51%**
$\mathcal{N}\left(0, \frac{1}{k}\right)$	$\mathcal{N}\left(0, \frac{1}{k}\right); [-1, 1]$	ReLU	$\mathcal{N}\left(0, \frac{1}{k}\right)$	$\mathcal{N}\left(0, \frac{1}{k}\right); [-1, 1]$	ReLU	**4.45%**
Glorot	Glorot	ReLU	Glorot	Glorot	ReLU	**2.28%**

Figure 3: The rightmost column has the estimates for the probability that the RNN exhibits period 3. We ran the experiment for 10000 times and checked whether the random RNN has period 3 (see Fig. 7). Each line specifies the type of initialization or activation unit used.

Why do we care about Period 3? ⇒ Li, Yorke 1975:

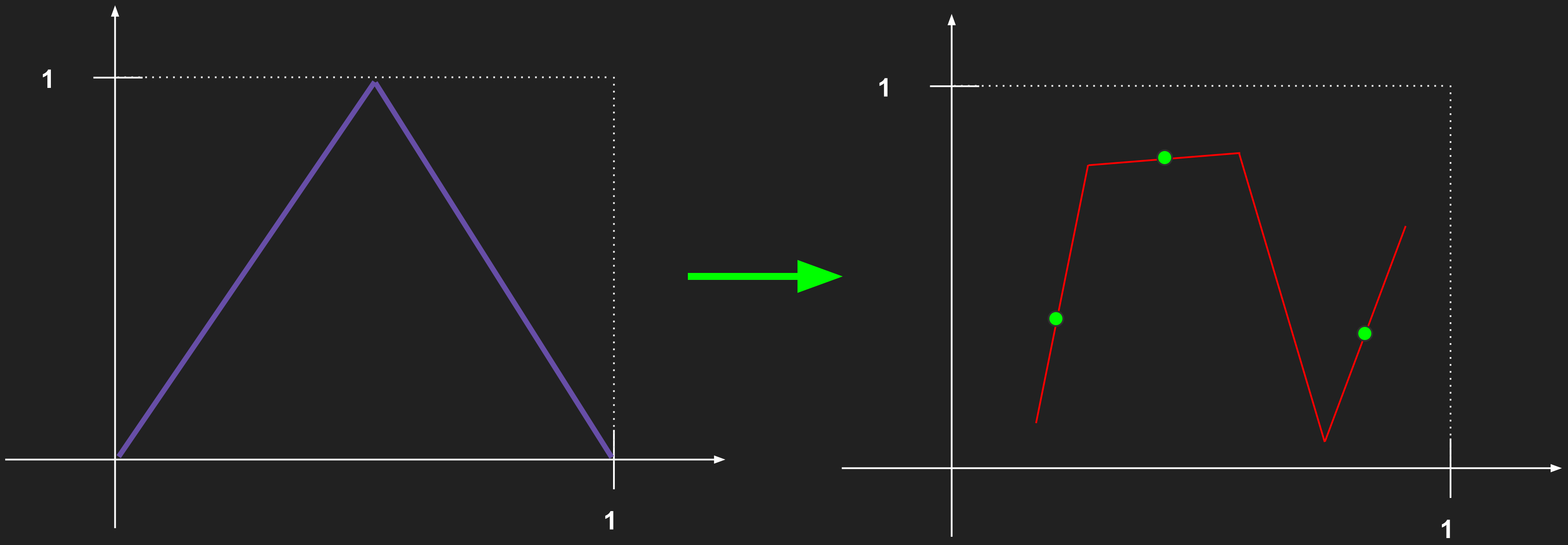


PERIOD THREE IMPLIES CHAOS

TIEN-YIEN LI AND JAMES A. YORKE

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

Construction is now irrelevant. Focus on period-3 points

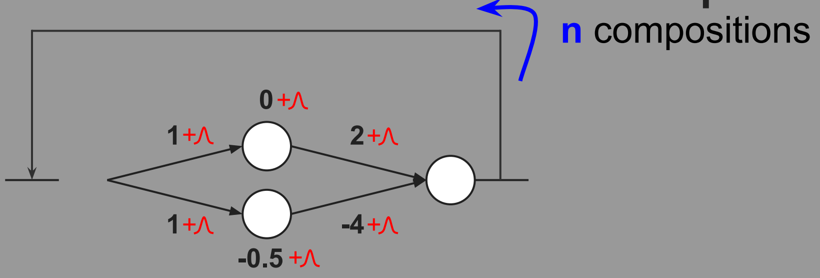


Depth-Width Trade-offs for ReLU Networks via Sharkovsky’s Theorem [Chatziafratis, Nagarajan, Panageas, Wang’20]
Better Depth-Width Trade-offs for Neural Networks through the lens of Dynamical Systems [Chatziafratis, Nagarajan, Panageas’20]
Expressivity of Neural Networks via Chaotic Itineraries beyond Sharkovsky's Theorem [Sanford, Chatziafratis’22]

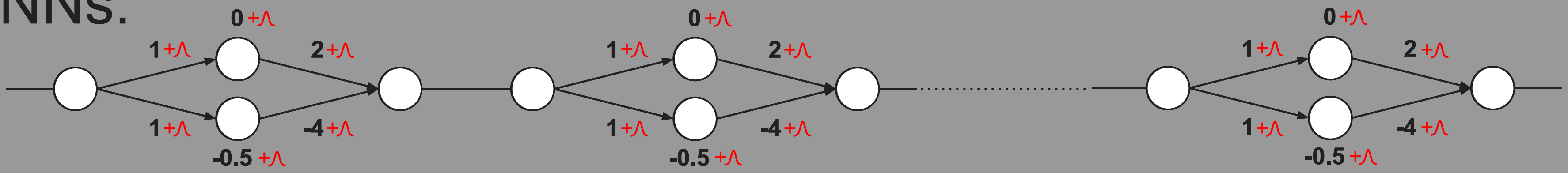
KEY Difference: RNNs vs FNNs

1. Random Perturbation is Done Once in RNNs
2. For FNNs, “fresh” randomness is sampled for parameter.

RNNs:



FNNs:



Complexity of Linear Regions in Deep Networks [Hanin, Rolnick ’19]

A Convergence Theory for Deep Learning via Over-Parameterization [Zeyuan Allen-Zhu ’19]