On the Expressive Power of Deep Neural Networks

Maithra Raghu [1,2] **Ben Poole** [3] **Jon Kleinberg** [1] **Surya Ganguli** [3] **Jascha Sohl Dickstein** [2]

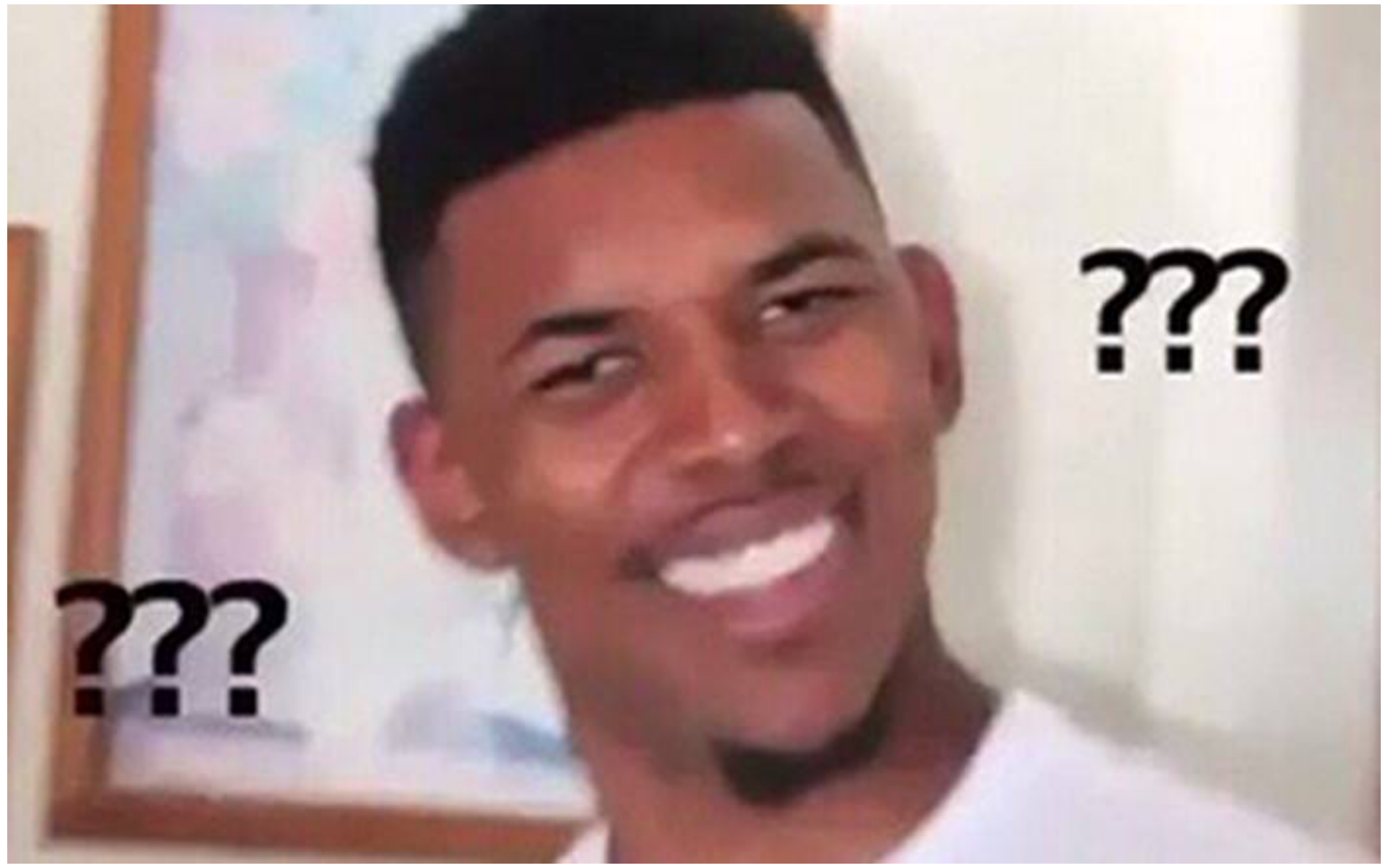
Jiaru Zhang
5.21.2019

Contents

Introduction

On the Expressive Power of Deep Neural Networks

Introduction



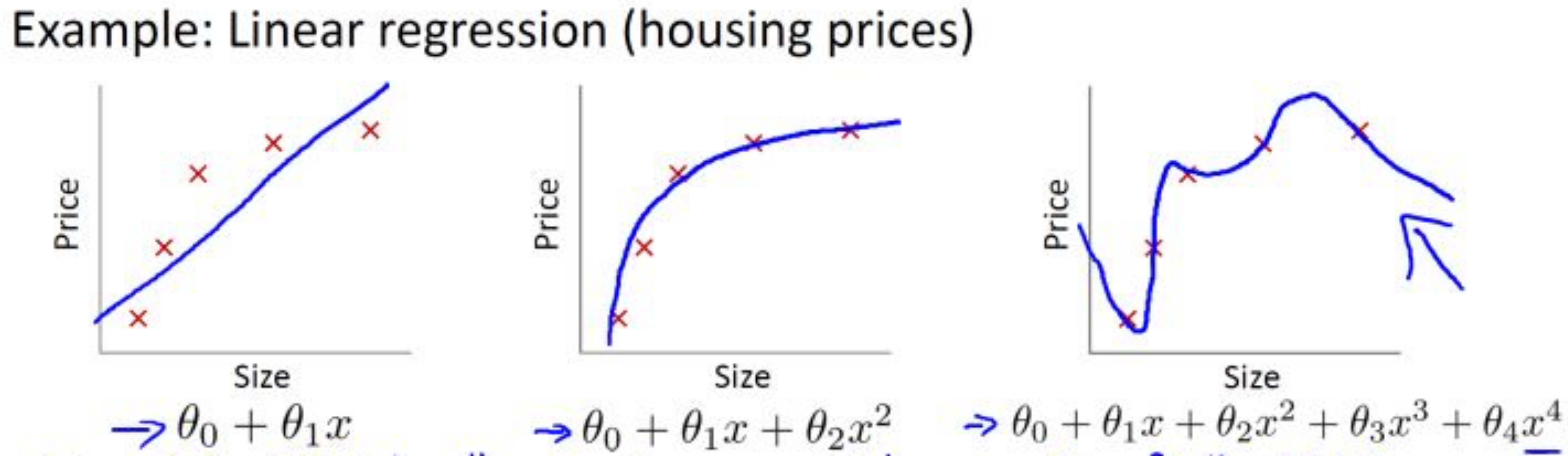
Expressive Power

1. What is it?
2. How to measure?
3. What determines it?
4. Usage?

Contents

Activation Pattern

What is Expressive Power (for a machine learning model)?



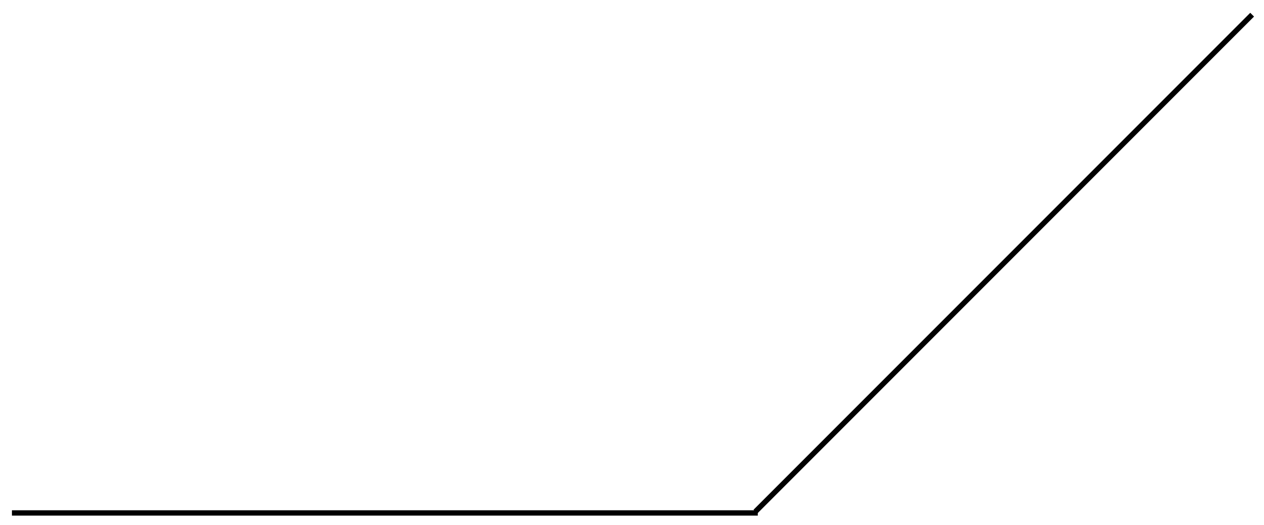
Expressive power: Size of model space.

Figure Source: Andrew Ng, https://www.coursera.org/learn/machine-learning

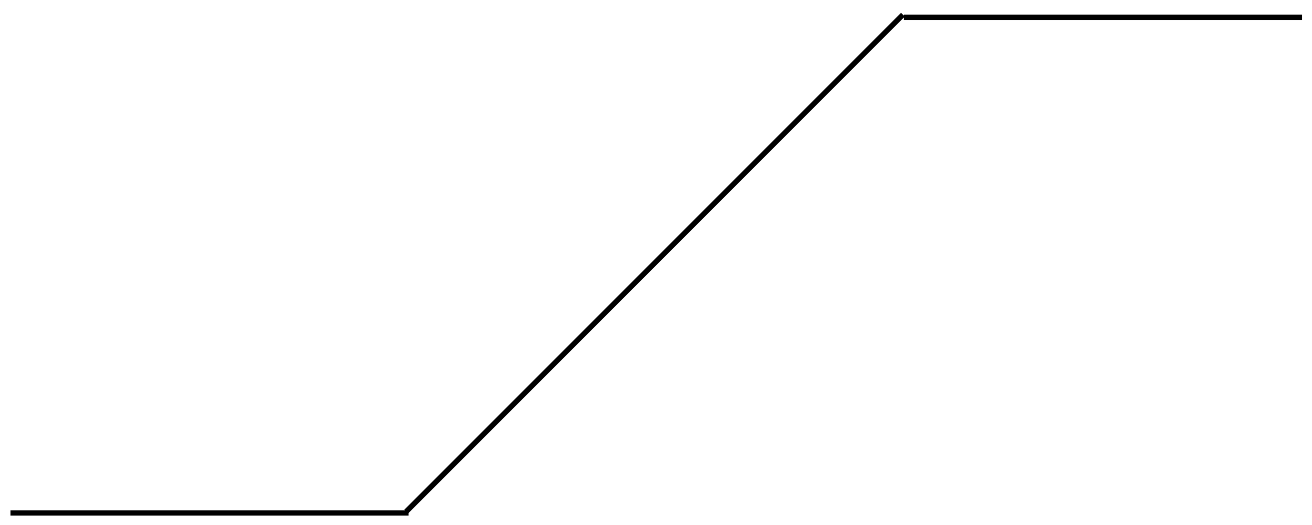
Activation Pattern

Settings for Neural Networks

- Only ReLU and hard tanh



ReLU

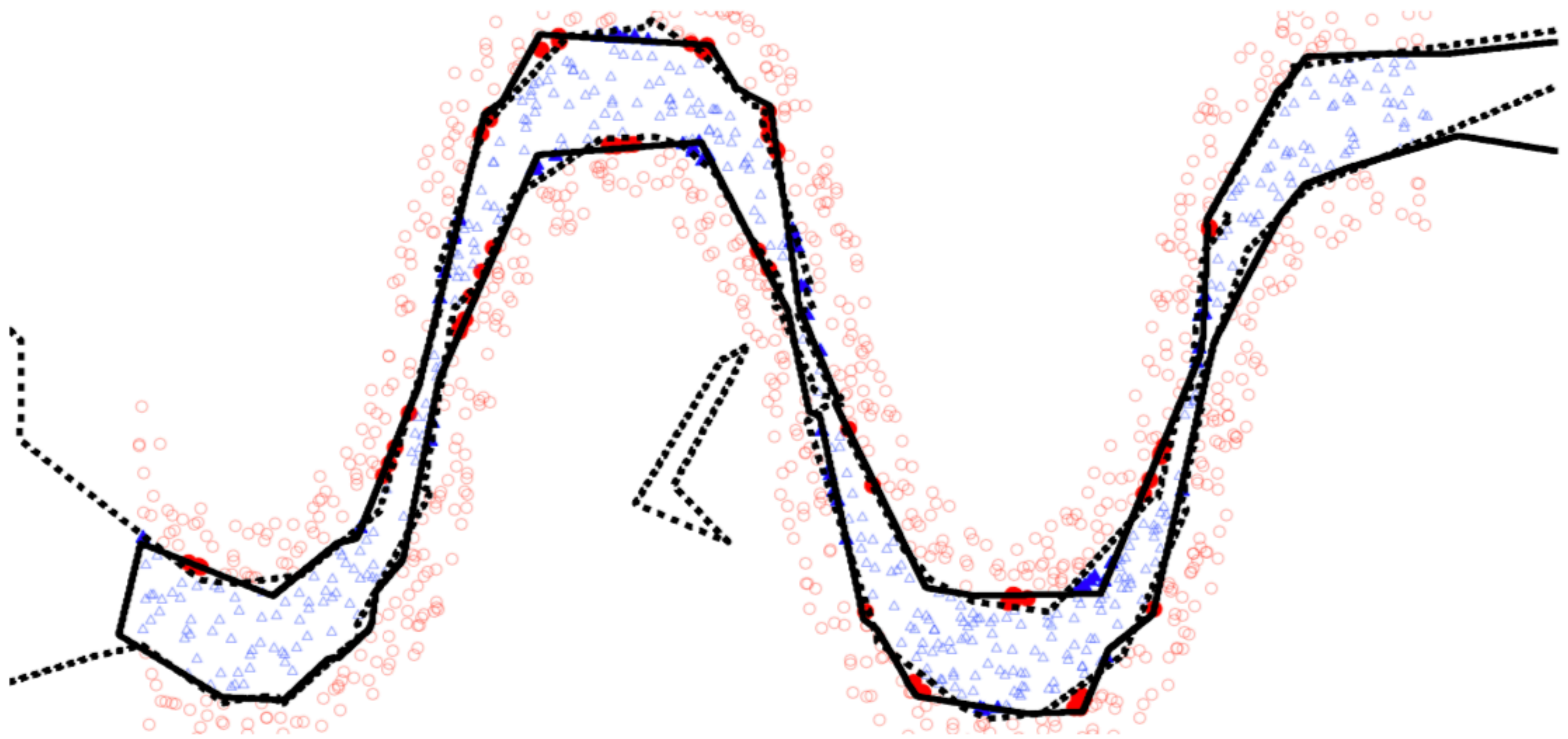


hard tanh

Activation Pattern

Settings for Neural Networks

- Only ReLU and hard tanh



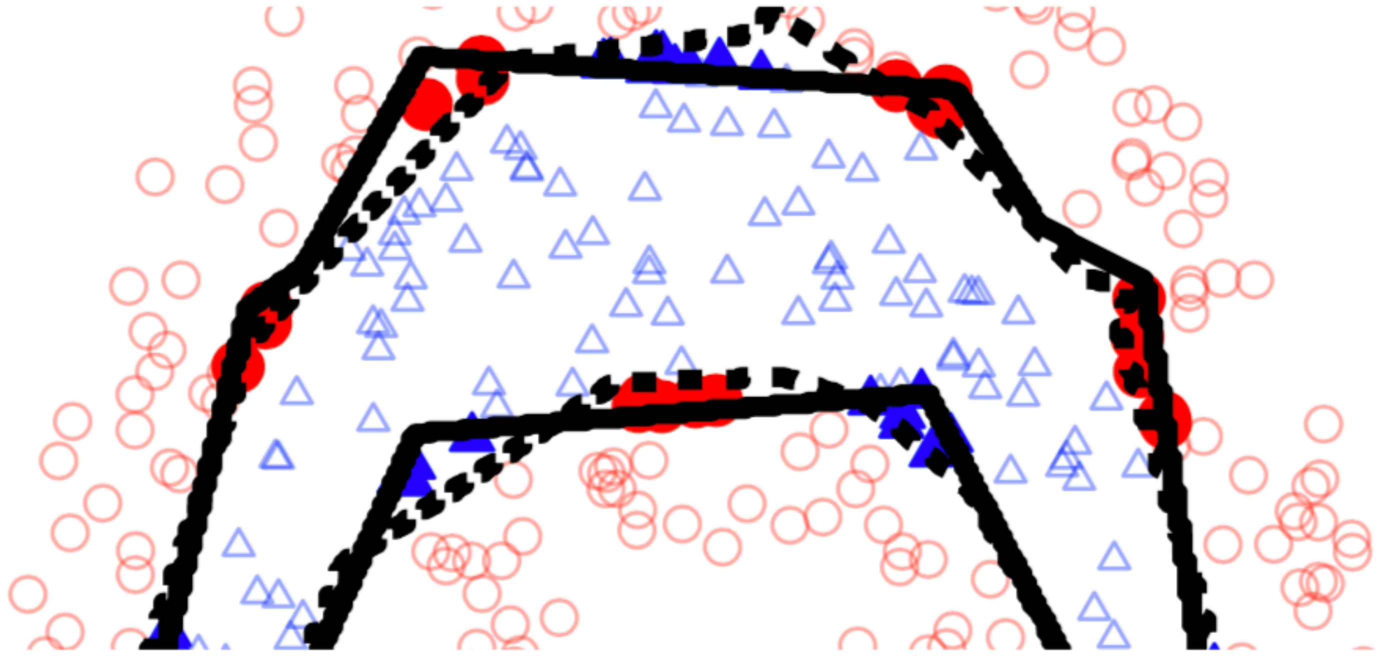
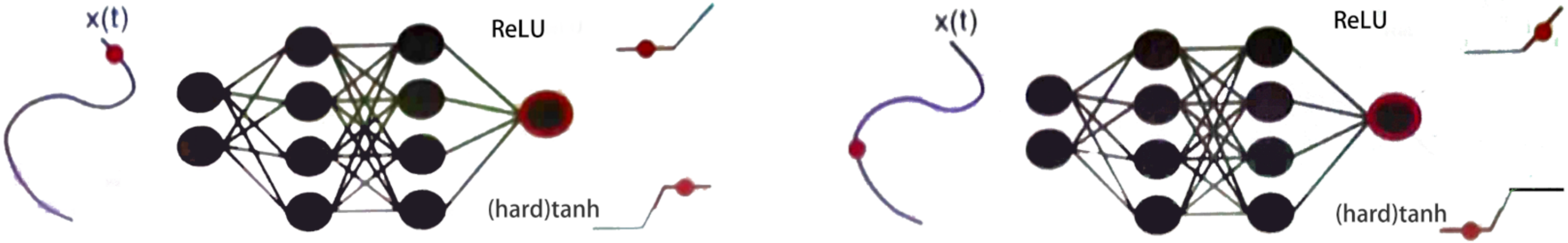


Figure Source: Guido Montufar et al. On the Number of Linear Regions of Deep Neural Networks.

Activation Pattern

How to measure?



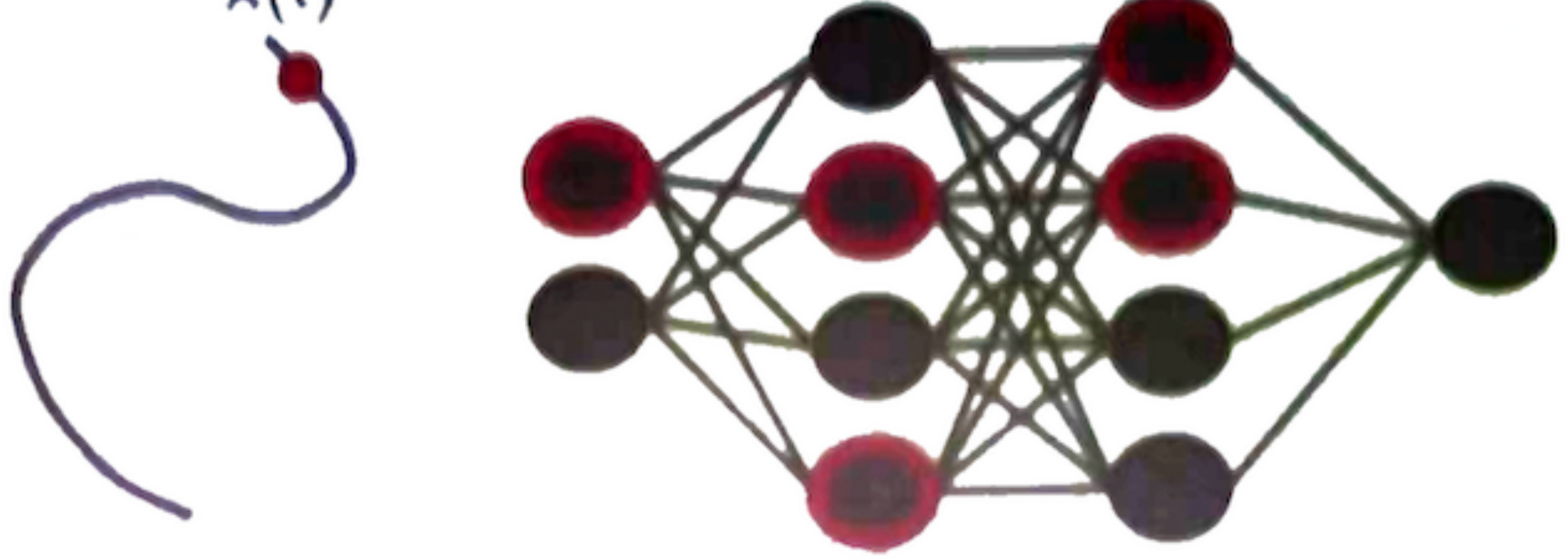
The number of linear regions

Zaslavsky's theorem: a shallow network (i.e. one hidden layer), with the same number of parameters as a deep network, has a much smaller number of linear regions than the number achieved by their choice of weights W_0 *for the deep network.*

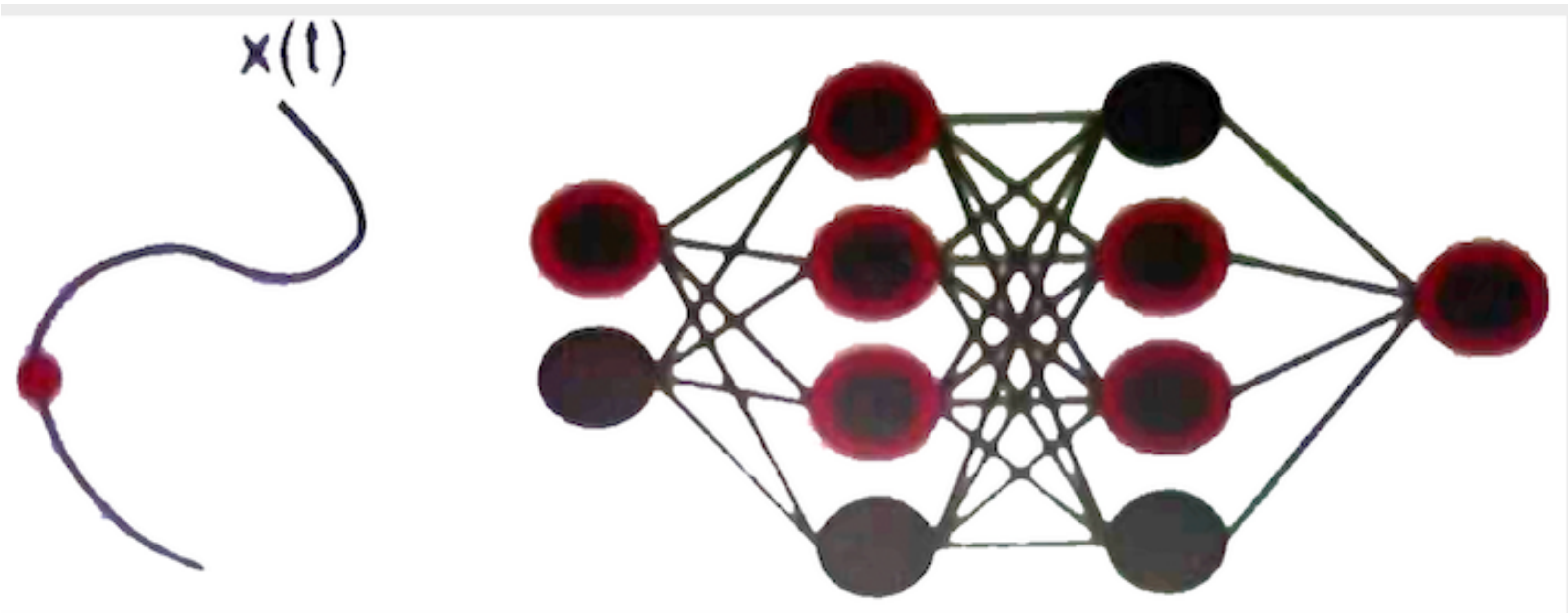
$$\forall W \quad \mathcal{T}\left(F_{A_1}([0,1];W)\right) < \mathcal{T}\left(F_{A_l}([0,1];W_0)\right)$$

Activation Pattern

How to measure?



10010111000



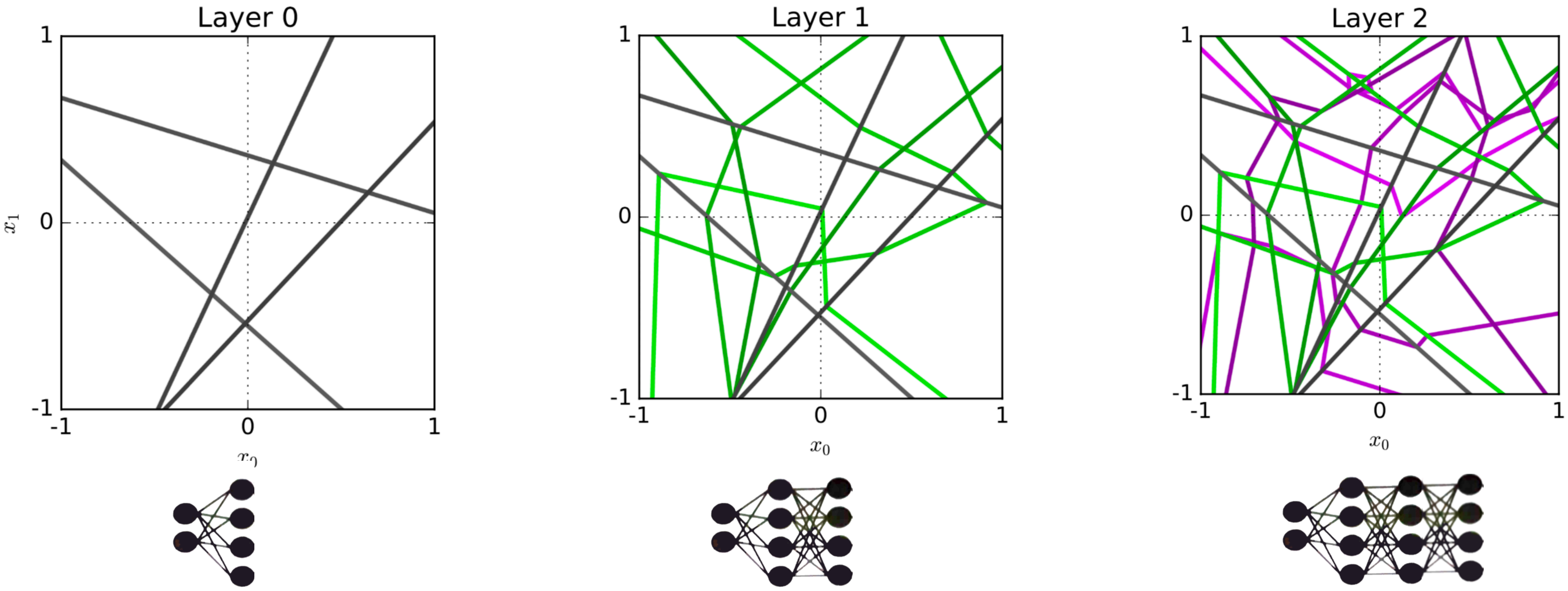
10111001101

number of Activation Patterns

Activation Pattern

Metric: number of Activation Patterns

Activation patterns are in one-one correspondence with linear regions in input space.



Activation Pattern

What determines number of activation patterns?

Upper bound grows linearly with depth and input dimension:

Given a network with:

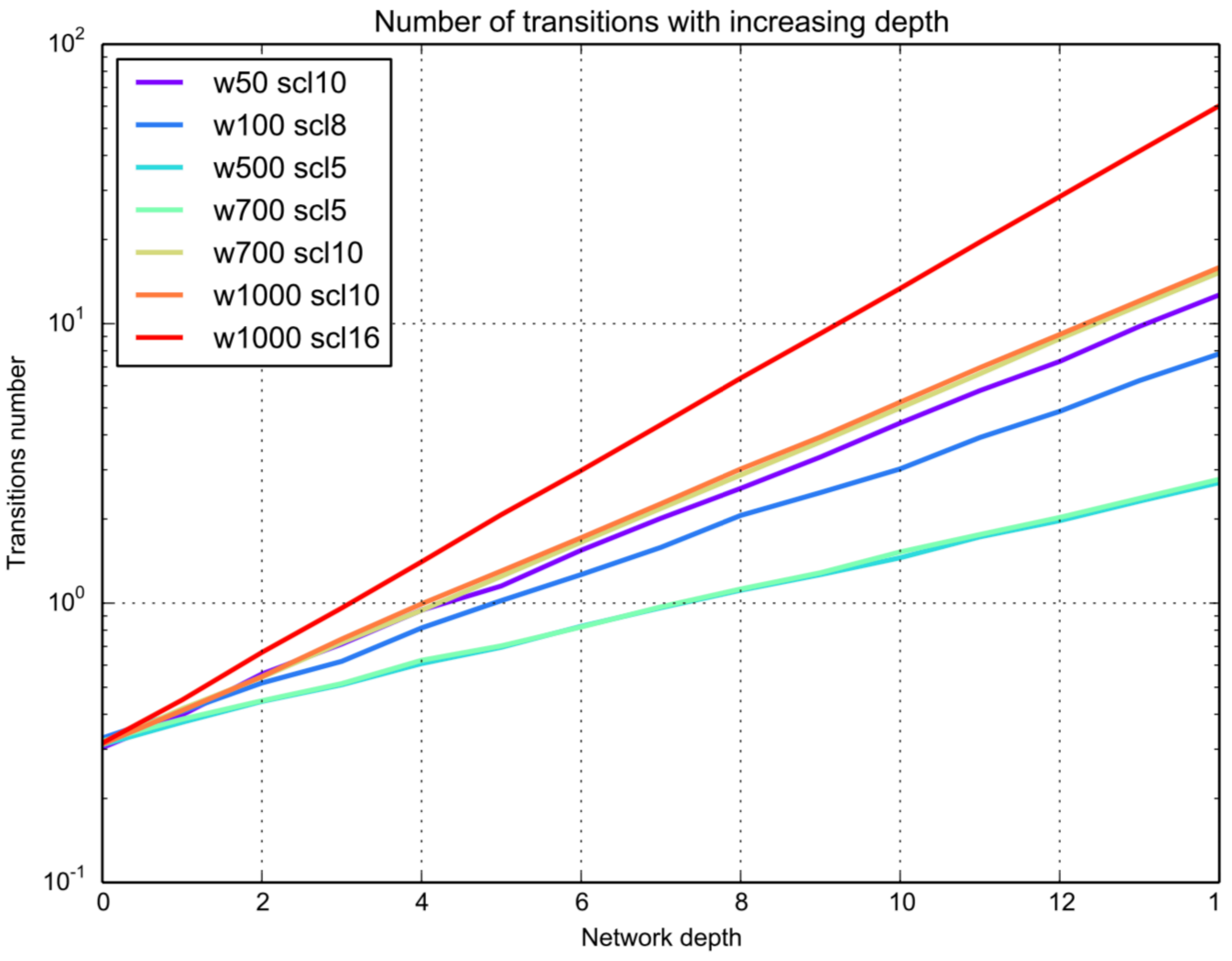
- Depth n
- Width k
- Input dimension m

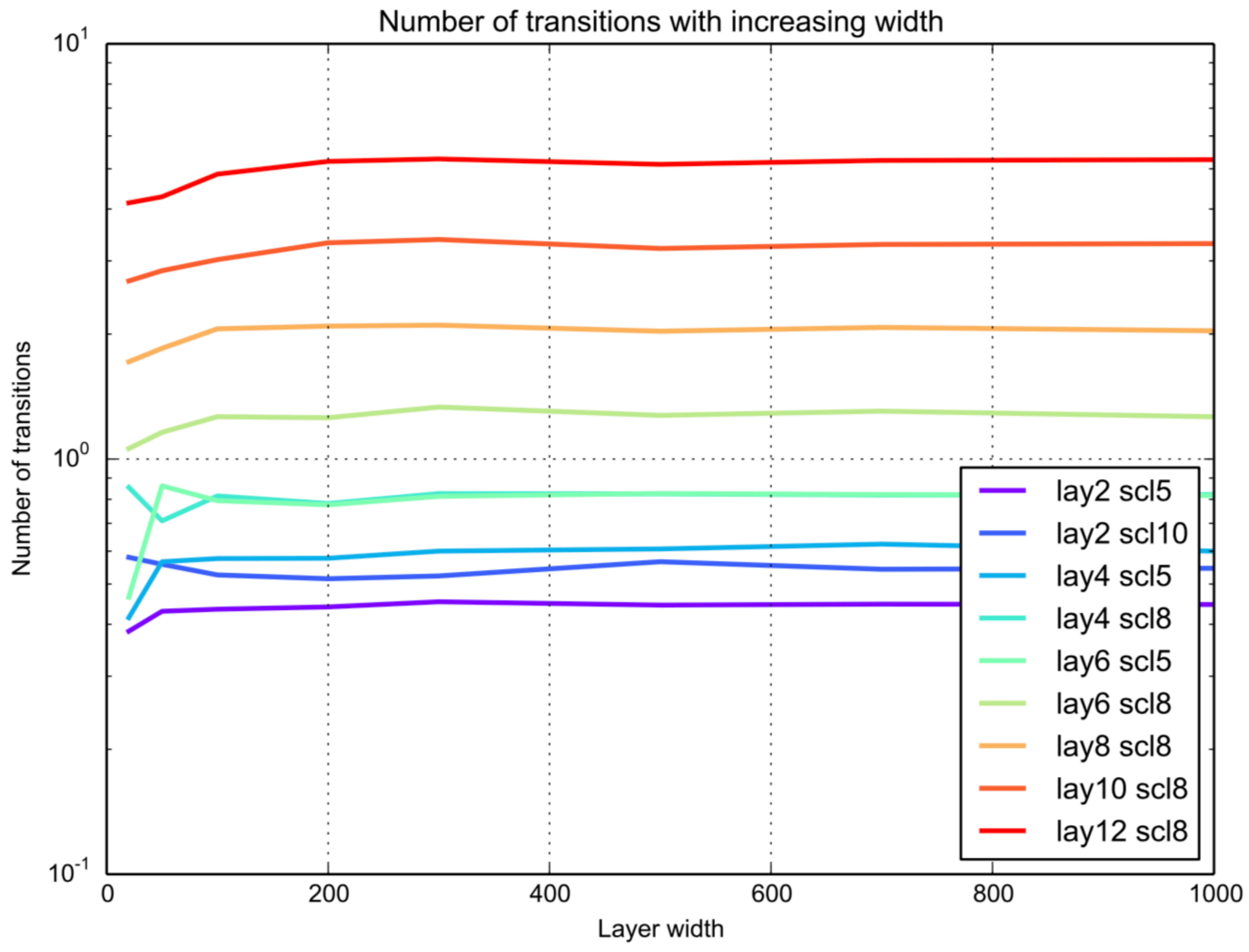
$$O(k^{mn})$$ **(for ReLU)**

$$O((2k)^{mn})$$ **(for hard tanh)**

Activation Pattern

Experiment Verification





Activation Pattern

Insight: A Derivation

Question: When number of total neurons are fixed, how to arrange them to get best expressive power?

Answer: Total N neurons, N/k depth, k width. Input dimension m.

Maximize $O(k^{m\frac{N}{k}})$ ==> k = e

Conclusion: When k >= 3, APs decrease when k increases.

Activation Pattern

Review

1. What is it?
2. How to measure?
3. What determines it?
4. Usage?

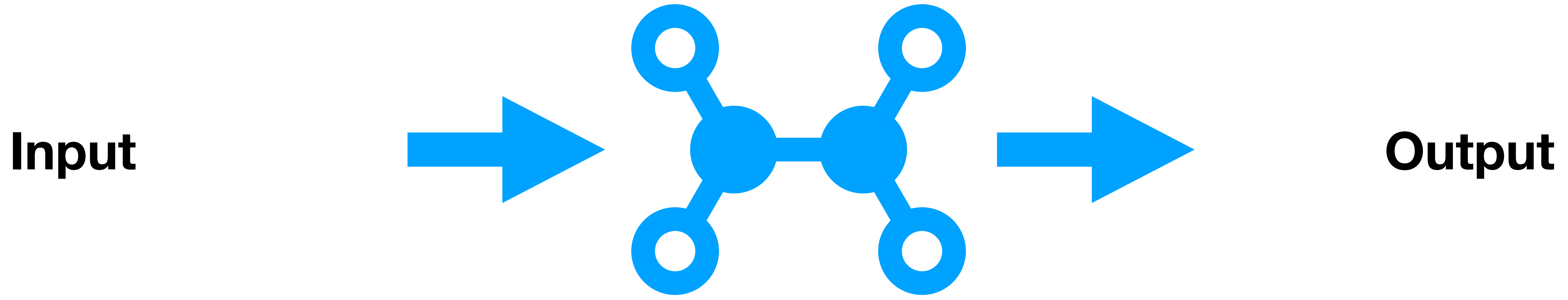
1. Number of states
2. Activation Patterns, 10111001101
3. $O(k^{mn})$ (ReLU), $O((2k)^{mn})$ (hard tanh)
4. k = 3

Question?

Contents

Trajectory Length

Rethink: What is Expressive Power?



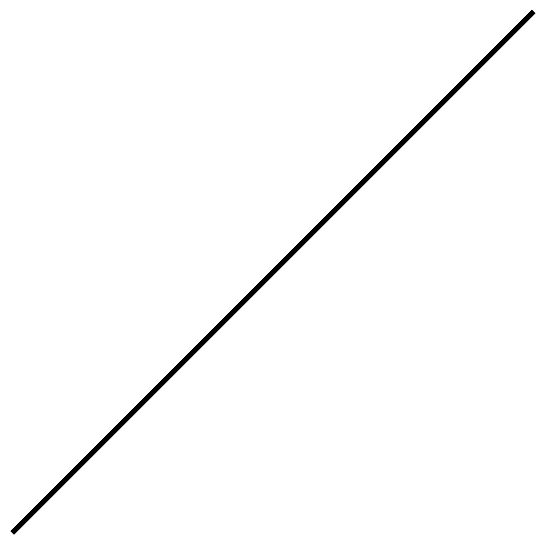
Neural network: A mapping from input to output.

Expressive Power: How complex the mapping is.

Trajectory Length

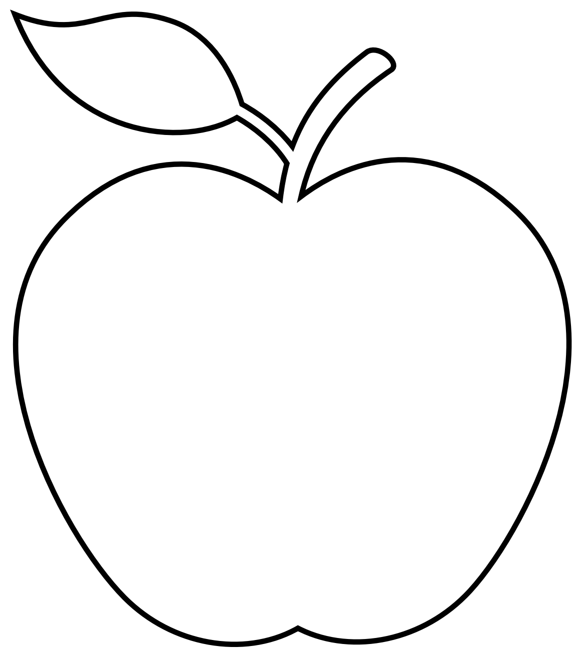
Rethink: What is Expressive Power?

- Consider an (one-dimensional) input trajectory



$x(t) = tx_1 + (1-t)x_0$

$x(t) = \cos(\pi t/2)x_0 + \sin(\pi t/2)x_1$

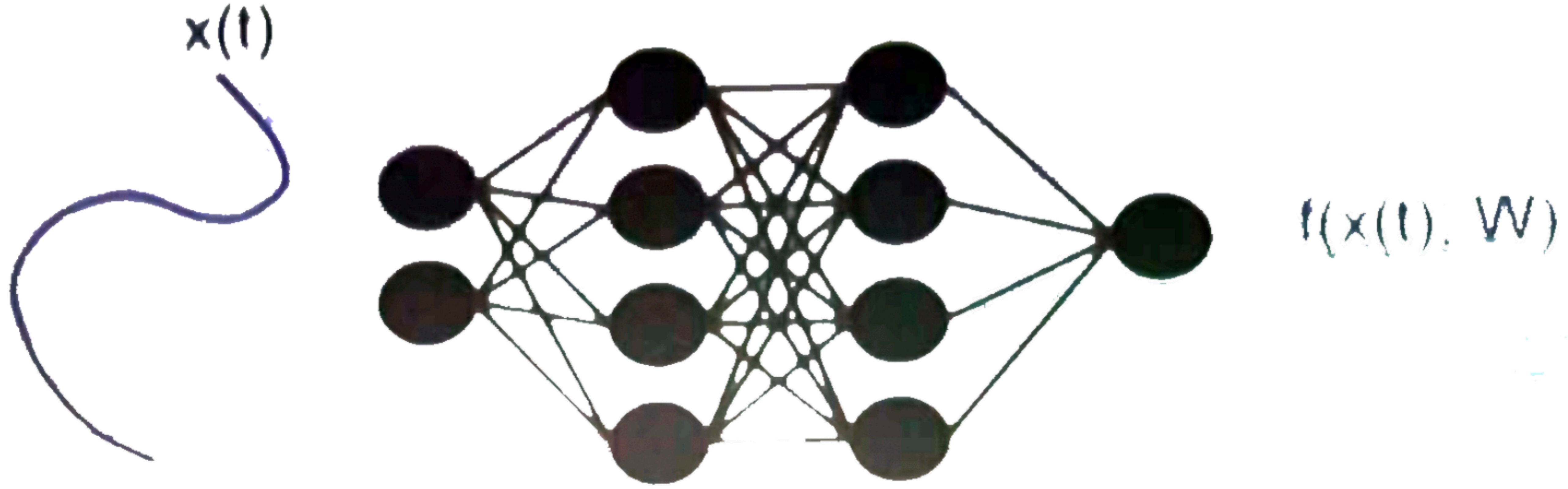


???

Trajectory Length

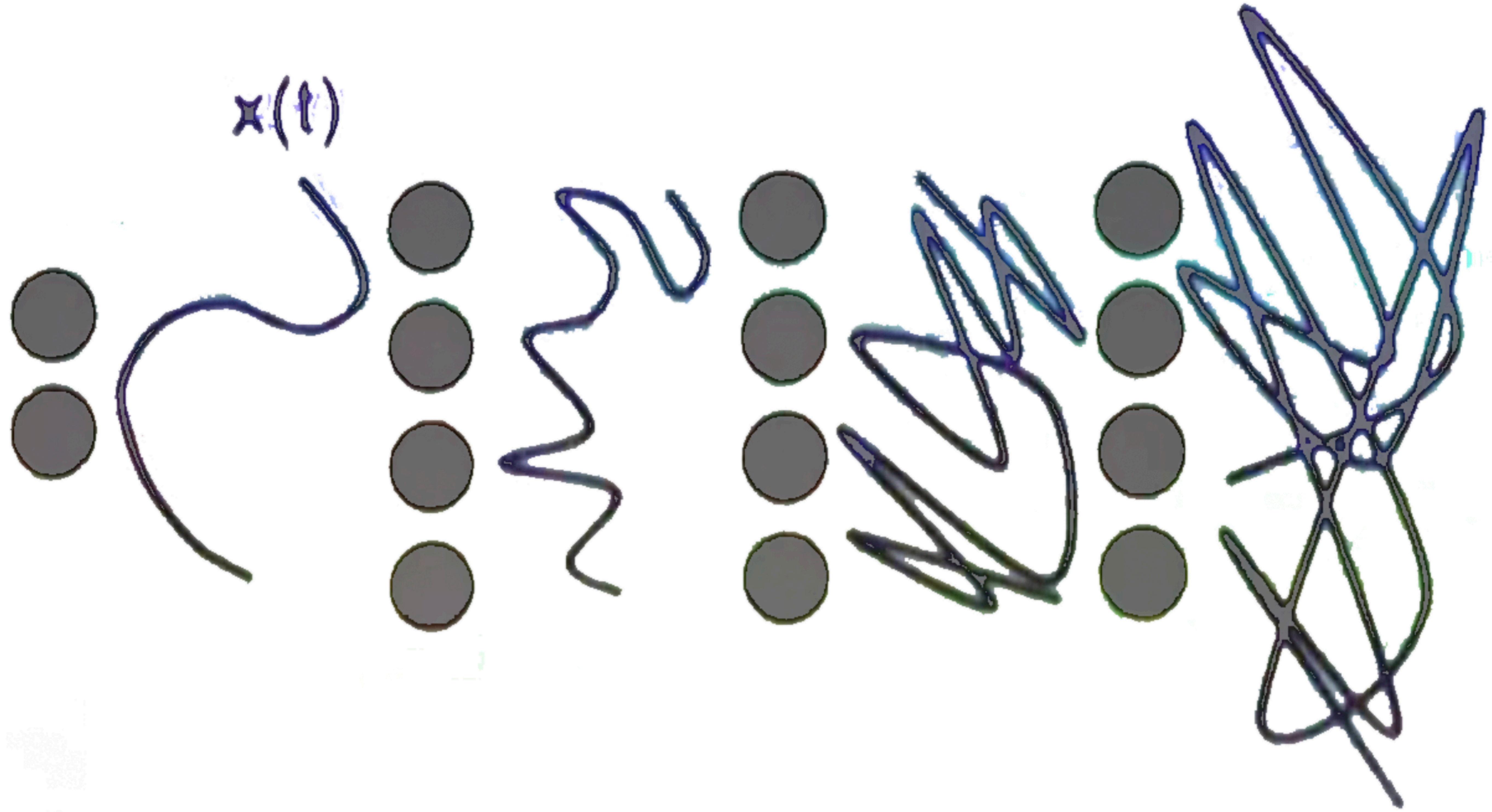
Rethink: What is Expressive Power?

- Consider an (one-dimensional) input trajectory



Trajectory Length

Metric: Trajectory Length



How does the trajectory length increase?

Trajectory Length

What determines Trajectory Length?

For a network with depth d, width k, weights ~ $\mathcal{N}\left(0, \sigma_w^2/k\right)$, *bias* ~ $\mathcal{N}\left(0, \sigma_b^2\right)$, *we have*

(a)

$$\mathbb{E}\left[l(z^{(d)}(t))\right] \geq O\left(\frac{\sigma_w\sqrt{k}}{\sqrt{k+1}}\right)^d l(x(t))$$

for ReLUs

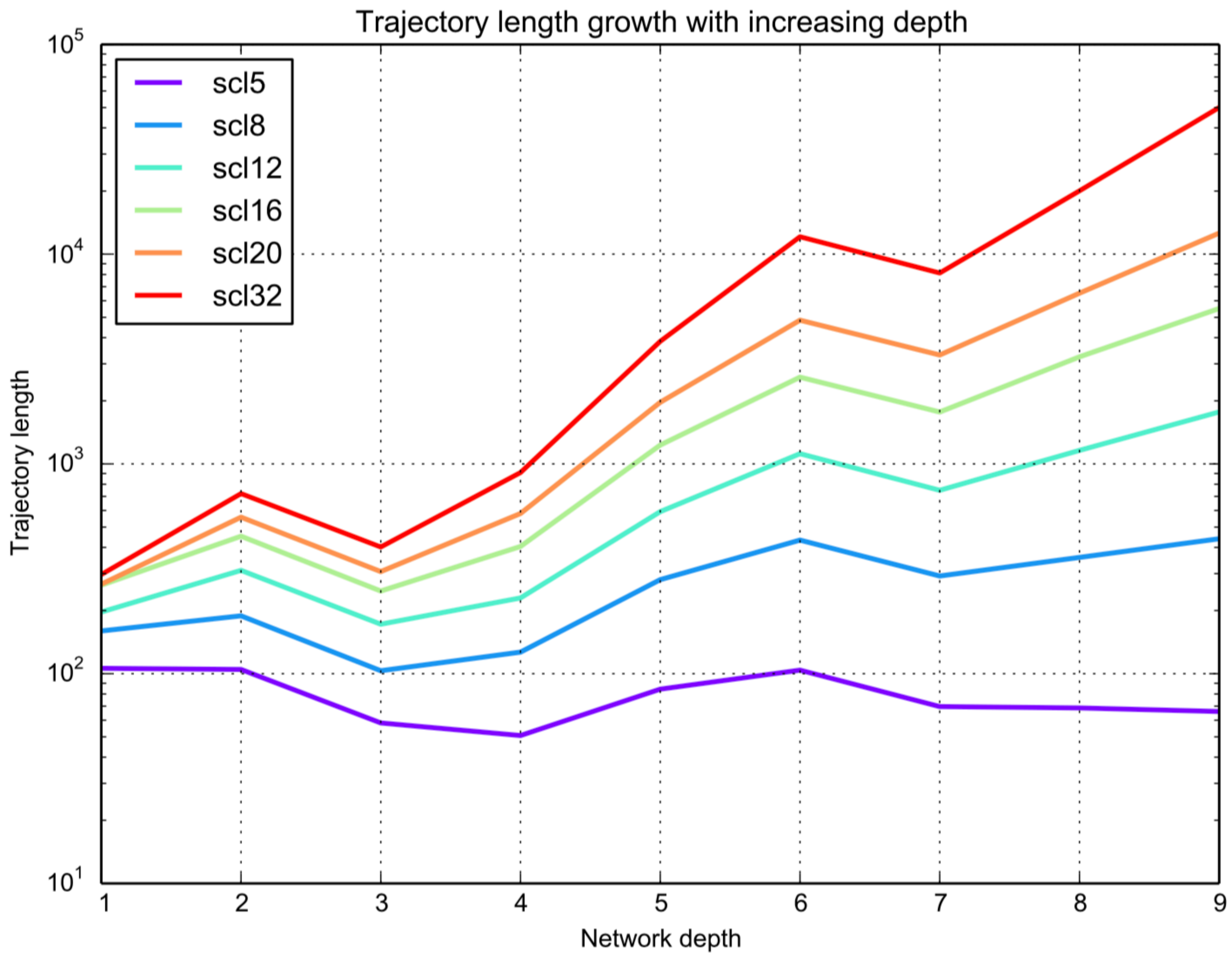
(b)

$$\mathbb{E}\left[l(z^{(d)}(t))\right] \geq O\left(\frac{\sigma_w\sqrt{k}}{\sqrt{\sigma_w^2+\sigma_b^2+k\sqrt{\sigma_w^2+\sigma_b^2}}}\right)^d l(x(t))$$

for hard tanh

Trajectory Length

Experiment Verification



Conv net on CIFAR-10, ReLU:

- Growth with depth exponentially
- Growth with σ_w

Trajectory Length

Relationship with number of Linear Regions

For a hard tanh network with depth d, n hidden layers, width k, weights ~ $\mathcal{N}\left(0,\sigma_w^2/k\right)$, bias ~ $\mathcal{N}\left(0,\sigma_b^2\right)$, we have

Review: Trajectory Length

(b)

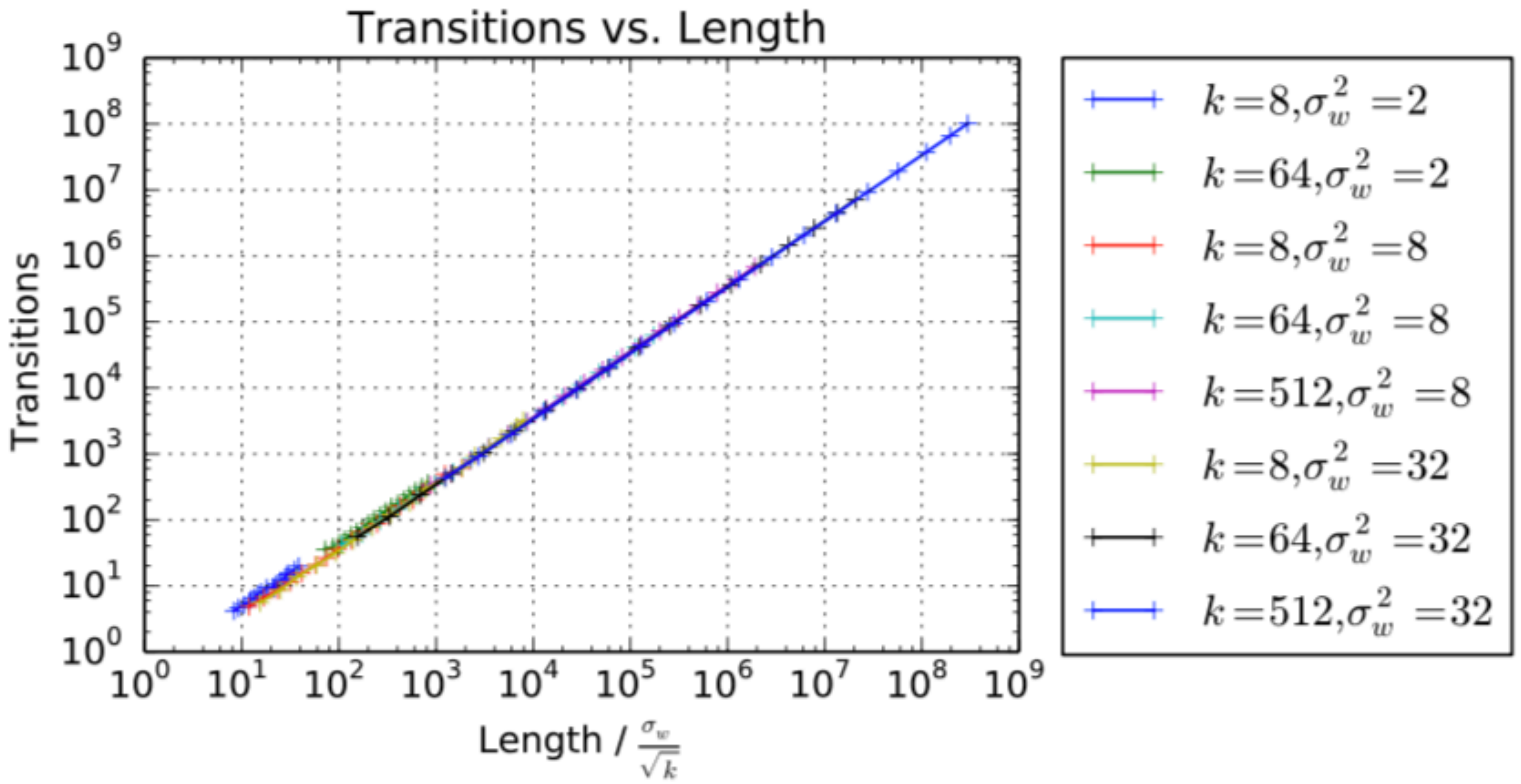
$$\mathbb{E}\left[l(z^{(d)}(t))\right] \geq O\left(\frac{\sigma_w\sqrt{k}}{\sqrt{\sigma_w^2+\sigma_b^2+k\sqrt{\sigma_w^2+\sigma_b^2}}}\right)^d l(x(t))$$

for hard tanh

Transitions:

$$g(k,\sigma_w,\sigma_b,n) = O\left(\frac{\sqrt{k}}{\sqrt{1+\frac{\sigma_b^2}{\sigma_w^2}}}\right)^n$$

Then $\mathcal{T}(F_{A_{n,k}}(x(t);W)) = O(g(k,\sigma_w,\sigma_b,n))$.

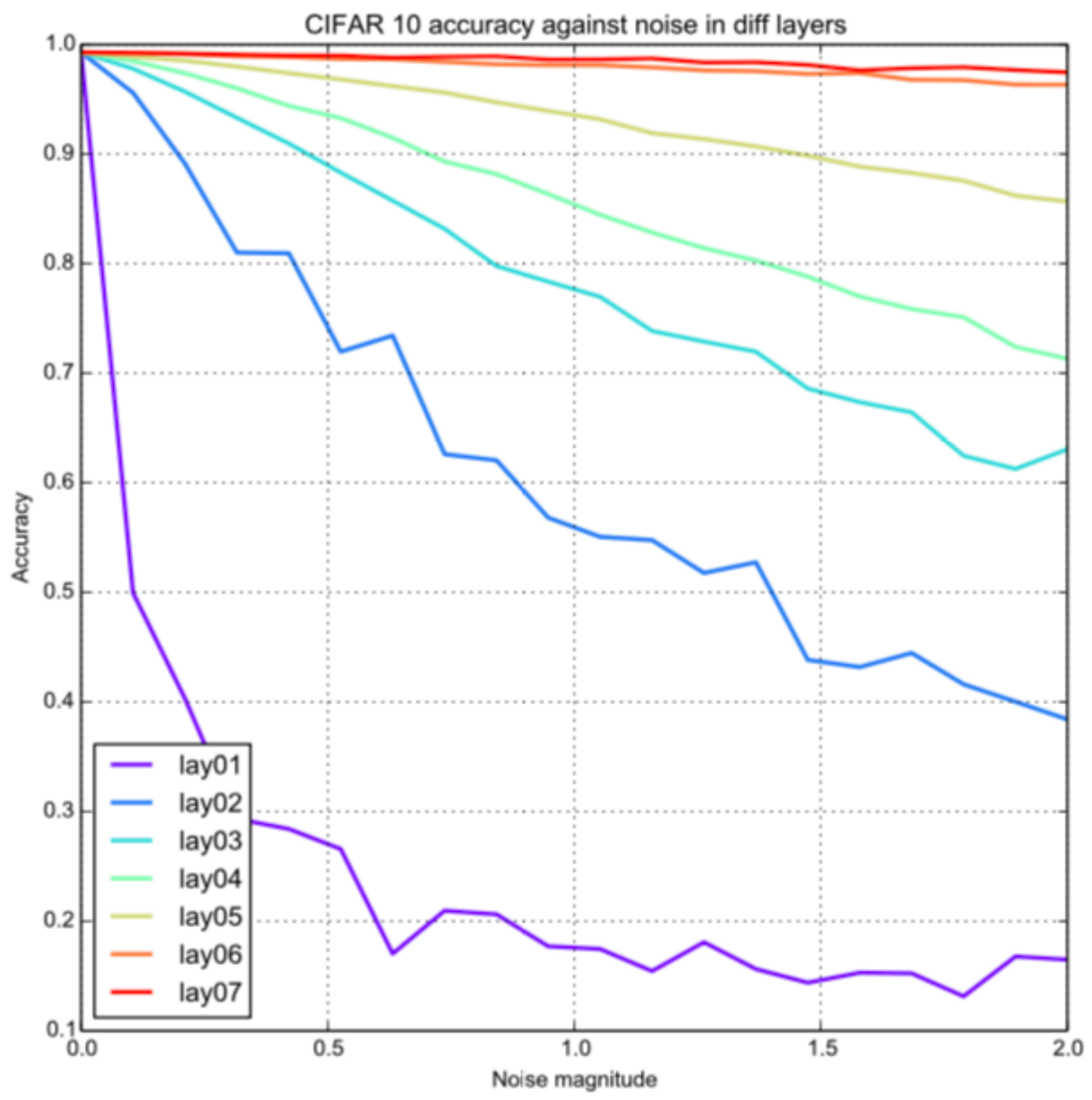


Trajectory Length

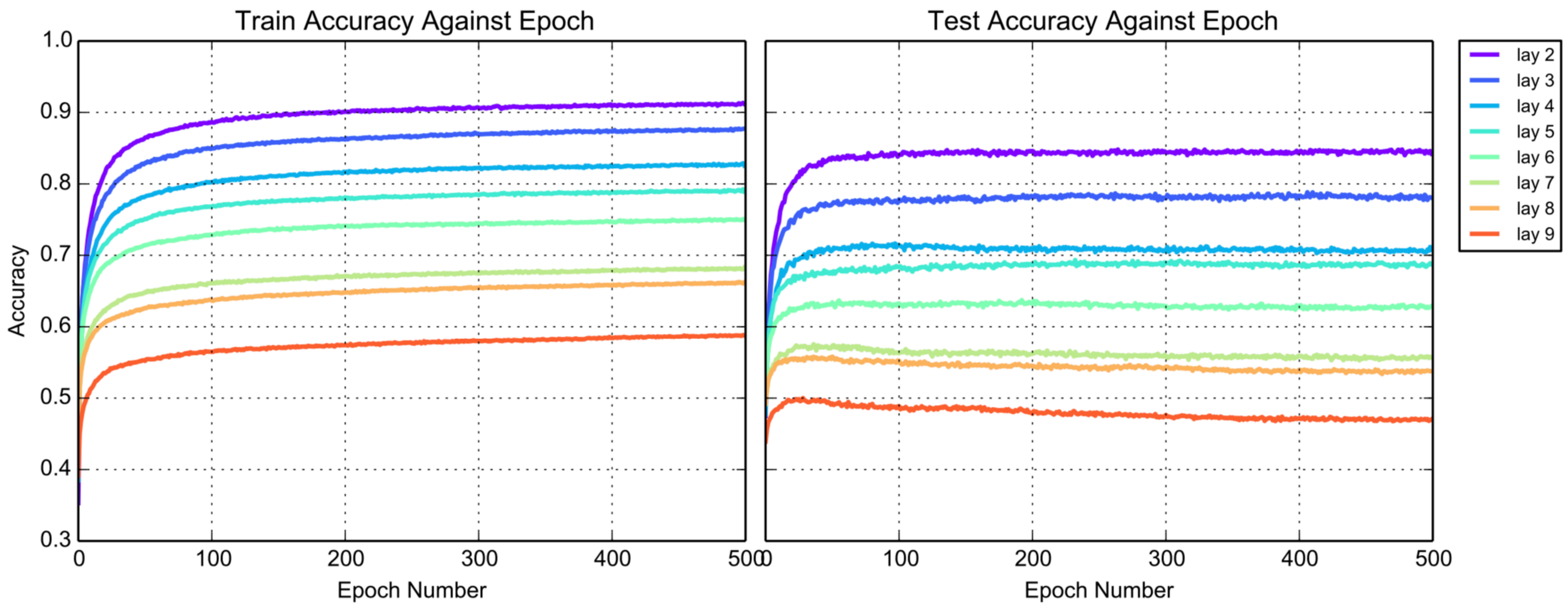
Insights: Trajectory and Stability

A perturbation at a layer grows exponentially in the *remaining depth* after that layer.

Add noise on different layers:

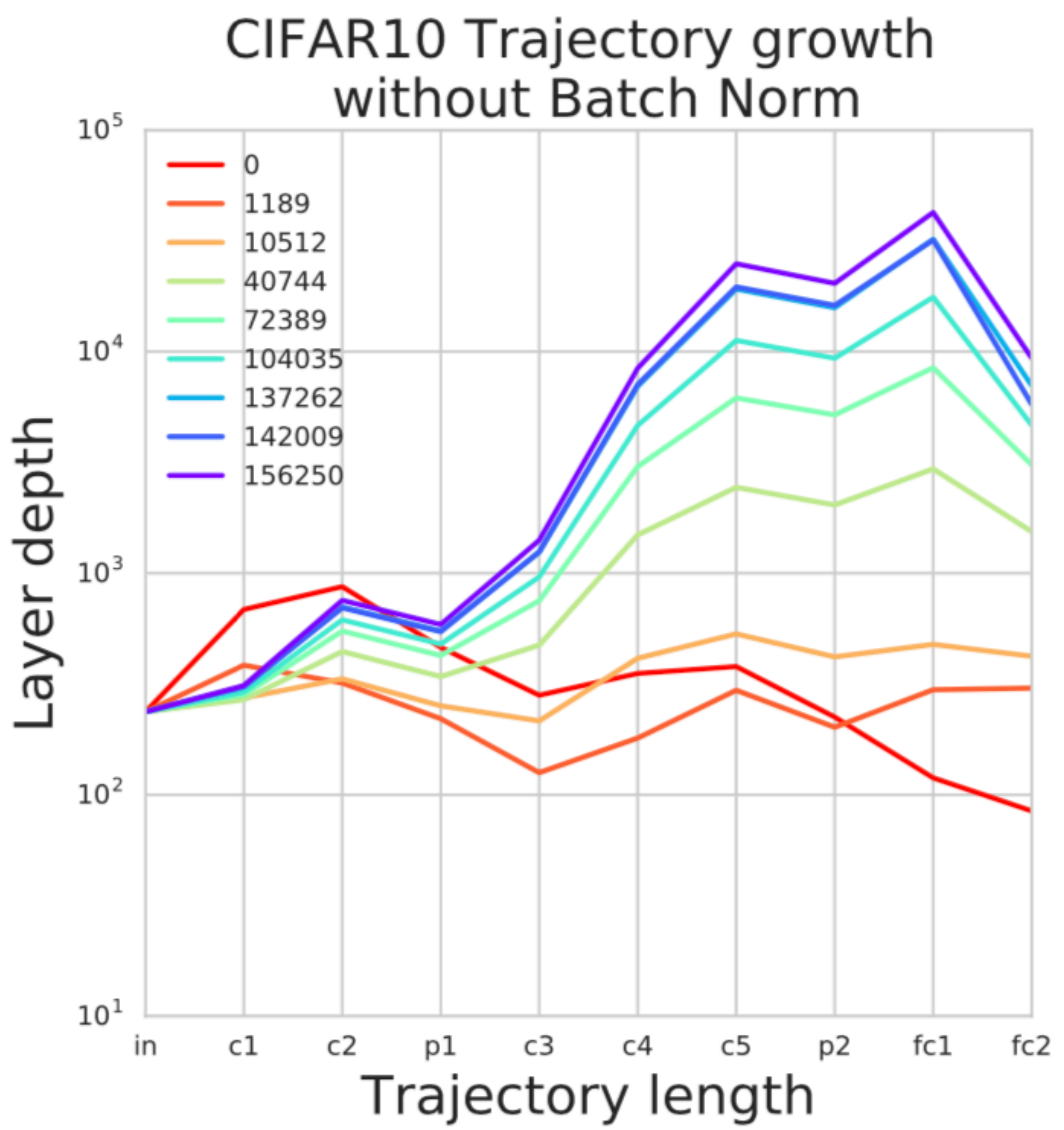


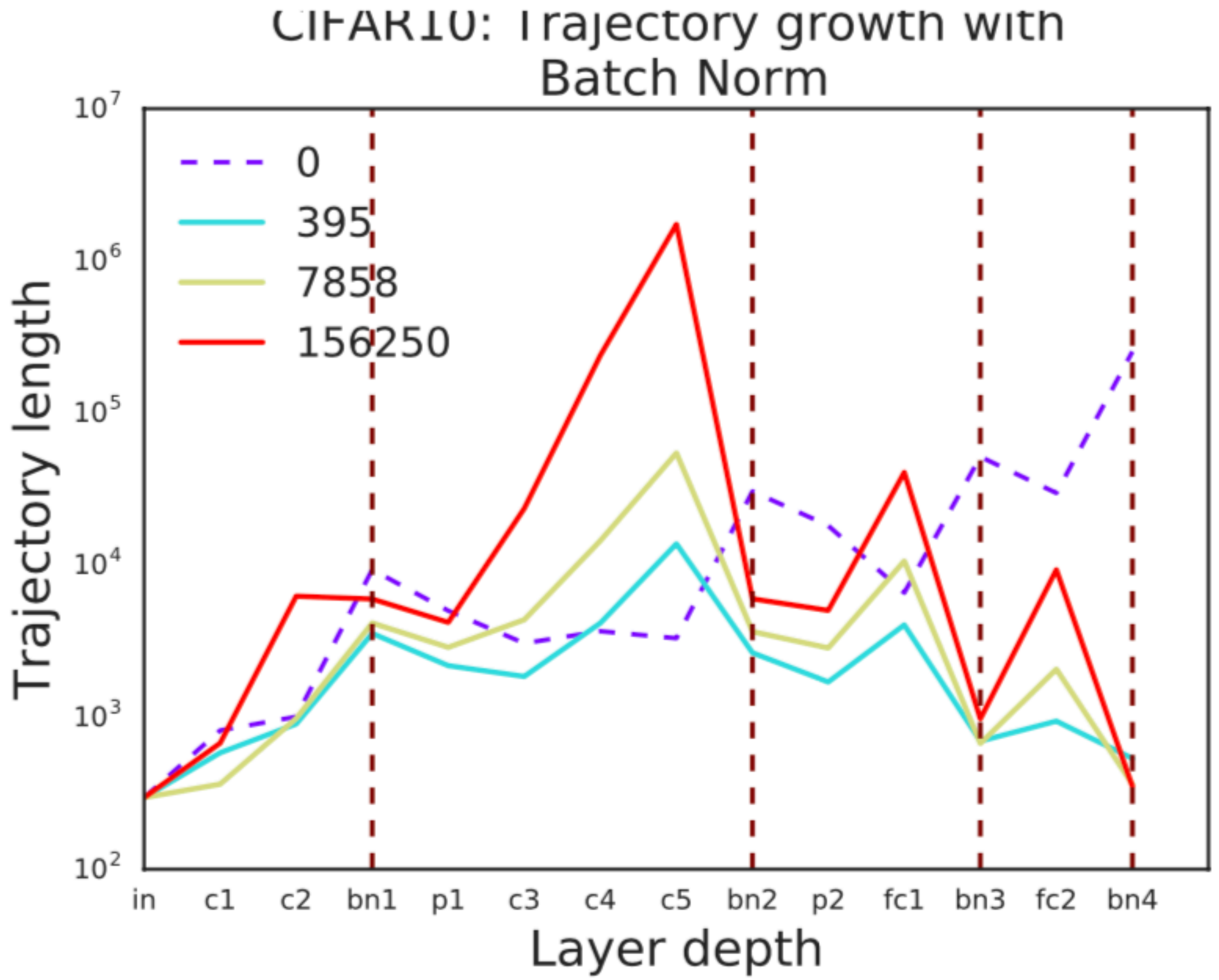
Only one layer trainable:



Trajectory Length

Insights: Trajectory and Batch Normalization





Trajectory Length

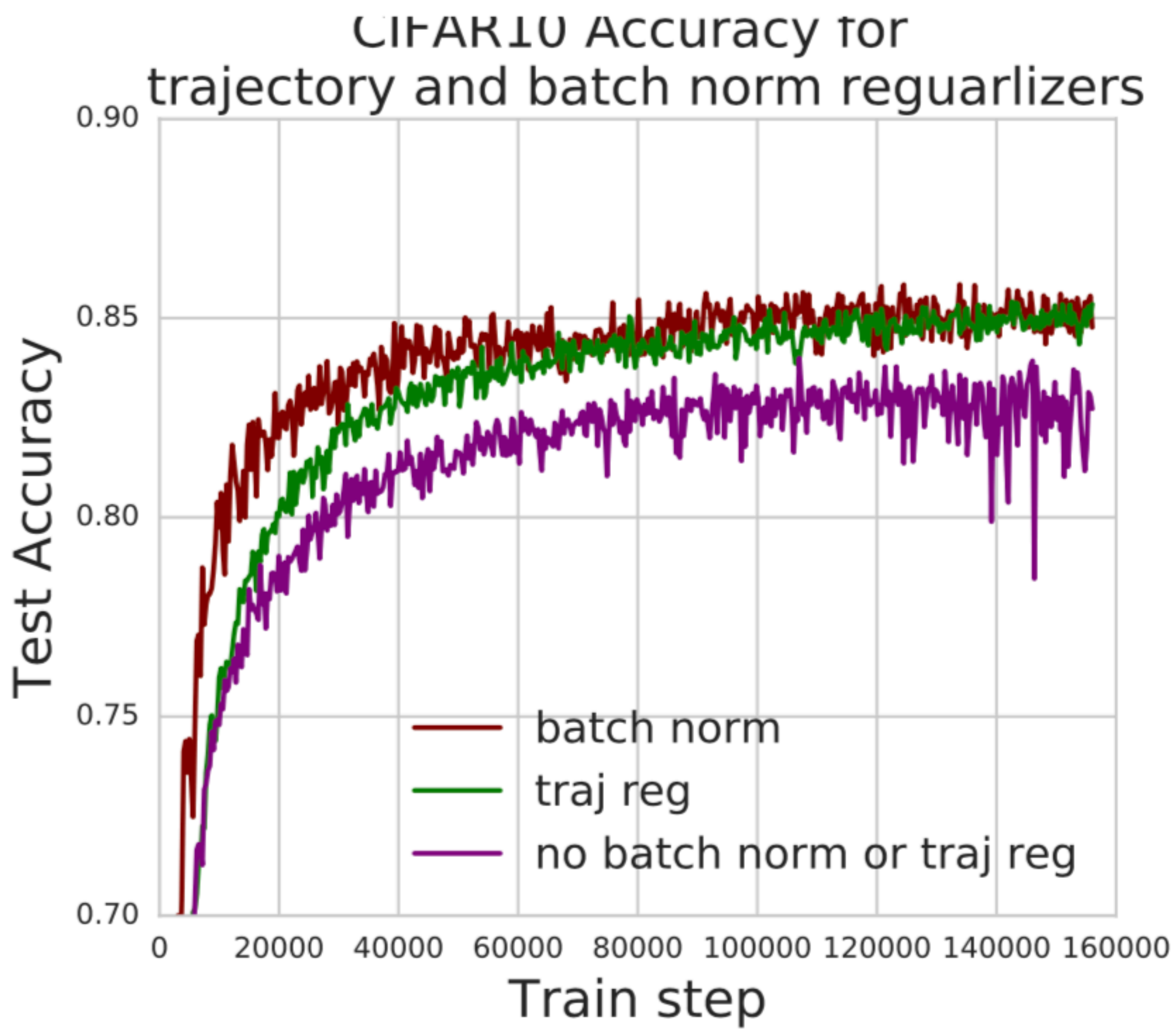
Insights: Trajectory Normalization

Trajectory Normalization: Scale trajectory length directly.

Trajectory regularization layers: add to the loss

$$\lambda \frac{currentlength}{originallength}$$

In practice, compute the sum of distances between adjacent points in the mini-batch.



Trajectory Length

Review

1. What is it?
2. How to measure?
3. What determines it?
4. Usage?

1. Mapping complexity
2. Trajectory length
3. $\mathbb{E}\left[l(z^{(d)}(t))\right] \geq O\left(\frac{\sigma_w\sqrt{k}}{\sqrt{k+1}}\right)^d l(x(t))$, $\mathbb{E}\left[l(z^{(d)}(t))\right] \geq O\left(\frac{\sigma_w\sqrt{k}}{\sqrt{\sigma_w^2+\sigma_b^2+k\sqrt{\sigma_w^2+\sigma_b^2}}}\right)^d l(x(t))$
4. Stability, Batch-norm, Traj-reg

Question?

Contents

Conclusion

Related Materials

- Paper and Supplementary Link: http://proceedings.mlr.press/v70/raghu17a.html
- Presentation Video: https://vimeo.com/237276052
- ICLR 2017 discussion: https://openreview.net/forum?id=B1TTpYKgx

Conclusion

Further Work

- What about other activation functions?
  - Their previous paper [1] talked about it. It combines Riemannian geometry with the mean field theory of high dimensional chaos to study it. (What are they? :-( )
- What about setting the input as a plane or other hyper-space, instead of only trajectory?
- What if the network is not regular (i.e. the width is not the same?)
- Is there any other proper metric for expressive power?

[1] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein and S. Ganguli. Exponential expressivity in deep neural networks through transient chaos. In NeurIPS, 2016

Conclusion

Thank you!

Any questions or evaluations?