On the Information Bottleneck Theory of Deep Learning

Authors: A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B.D. Tracey, and D.D. Cox

Presenter: Zhongyuan Zhao

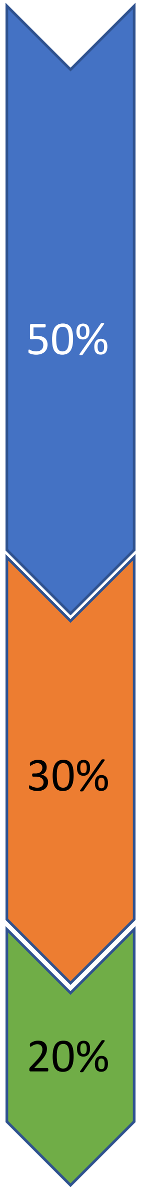
zhzhao@cse.unl.edu

Oct. 5th, 2018

A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B.D. Tracey, and D.D. Cox, "On the Information Bottleneck Theory of Deep Learning", ICLR 2018, [Online] https://openreview.net/forum?id=ry_WPG-A-

Outline

- Part 1: Information Bottleneck Theory of Deep Learning
  - Information Theory Recap
  - Information Plane
  - IB Theory
- Part 2: Attacks on the IB theory
  - Counter Example
  - Summary
- Discussions



Part 1: Information Bottleneck Theory of Deep Neuron Networks

The Hebrew University of Jerusalem

Information Bottleneck Theory of DNN

- Who
  - Ravid Schwartz-Ziv, Naftali Tishby, etc. The Hebrew University of Jerusalem
- What – an attempt to explain DNN
  - Training Dynamics
  - Learning Processes
  - Internal Representation
- How
  - Information Theory

[1] R. Schwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. arXiv preprint arXiv:1703.00810, 2017
[2] Michal Moshkovich and Naftali Tishby. Mixing complexity and its applications to neural networks. 2017. URL https://arxiv.org/abs/1703.00729
[3] Naftali Tishby and Noga Zaslavsky. Deep Learning and the information Bottleneck Principle. In Information Theory Workshop (ITW), 2015 IEEE, Pages 1-5. IEEE, 2015
[4] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck Method. In Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing, 1999.

Information Theory: Entropy

A Discrete Random Variable X with possible values $\{x_1, ..., x_n\}$

Information $$I(x_i) = \log_2 \frac{1}{P(x_i)} = -\log_2 P(x_i) \geq 0$$

Entropy $$H(X) = E[I(X)] = \sum_{i=1}^{n} P(x_i) I(x_i) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

Joint Entropy $$H(X, Y) = -\sum_{x \in \mathscr{X}} \sum_{y \in \mathscr{Y}} p(x, y) \log p(x, y).$$

Conditional Entropy $$H(X|Y) = -\sum_{i,j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(y_j)}$$

X : {0, 1}

P(0)=0.5, P(1)=0.5

$$I(0) = \log_2 \frac{1}{P(0)} = 1$$

$$H(X) = 0.5\times 1 + 0.5\times 1 = 1$$

P(0)=0.9, P(1)=0.1

I(0)=0.152, I(1) = 3.3219

H(X) = 0.469

[5] Carl Bergstrom, Lecture 3: Joint entropy, conditional entropy, relative entropy, and mutual information
http://octavia.zoology.washington.edu/teaching/429/lecturenotes/lecture3.pdf

Information Theory: Mutual Information

Mutual Information $I(X;Y) = \sum_{x \in X, y \in Y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) = H(X) - H(X|Y)$

Definition 5 *The mutual information $I(X;Y)$ measures how much (on average) the realization of random variable Y tells us about the realization of X, i.e., how by how much the entropy of X is reduced if we know the realization of Y.*

$$I(X;Y) = H(X) - H(X|Y)$$

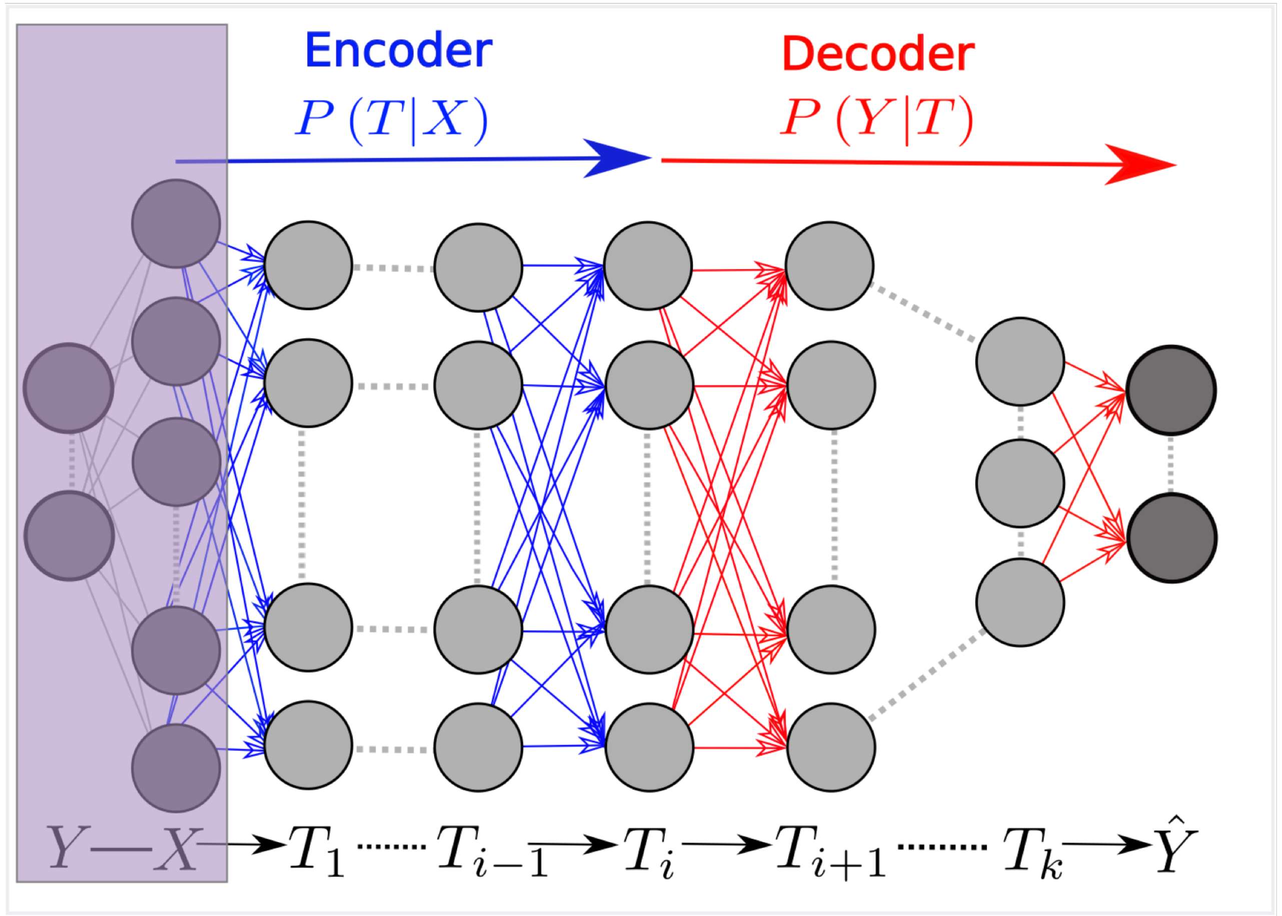
Surprisingly, mutual information is symmetric; X tells us exactly as much about Y as Y tells us about X.

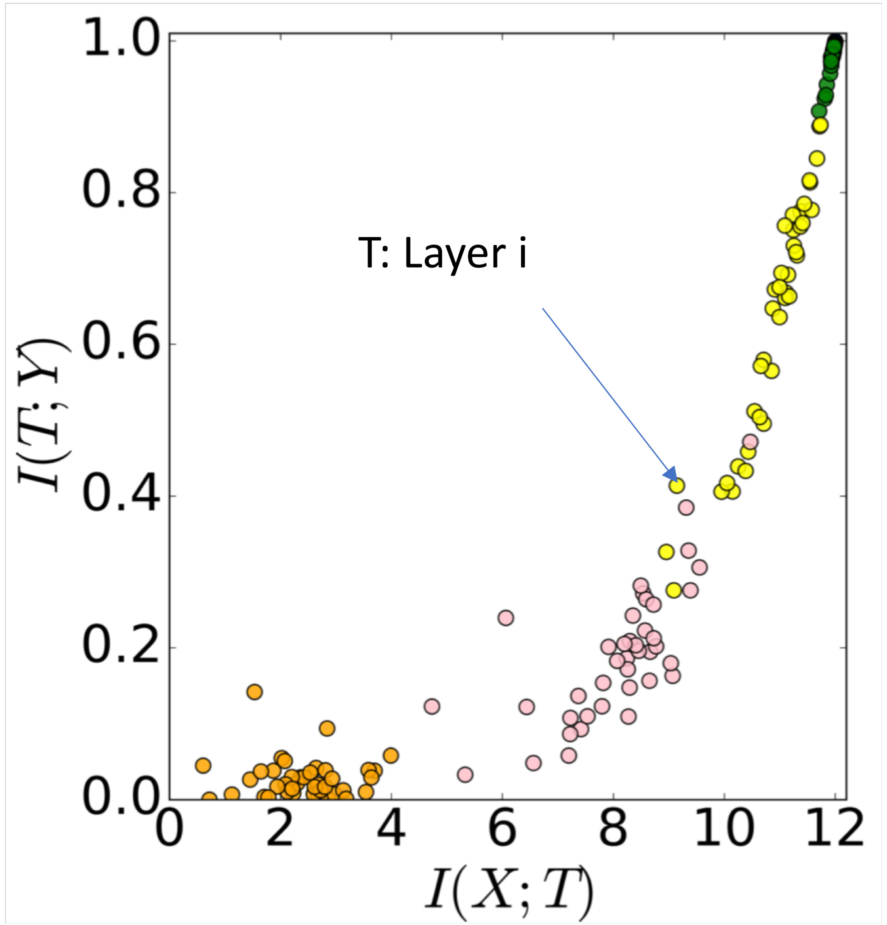
Theorem 2 *Symmetry of mutual information*

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = I(Y,X).$$

[5] Carl Bergstrom, Lecture 3: Joint entropy, conditional entropy, relative entropy, and mutual information
http://octavia.zoology.washington.edu/teaching/429/lecturenotes/lecture3.pdf

Information Plane of DNN





Information Plane (Cont.)

- Invariance to Invertible Transformation

$$I\left(X;Y\right)=I\left(\psi(X);\phi(Y)\right)$$

- Markov Chain

$$X\rightarrow Y\rightarrow Z$$

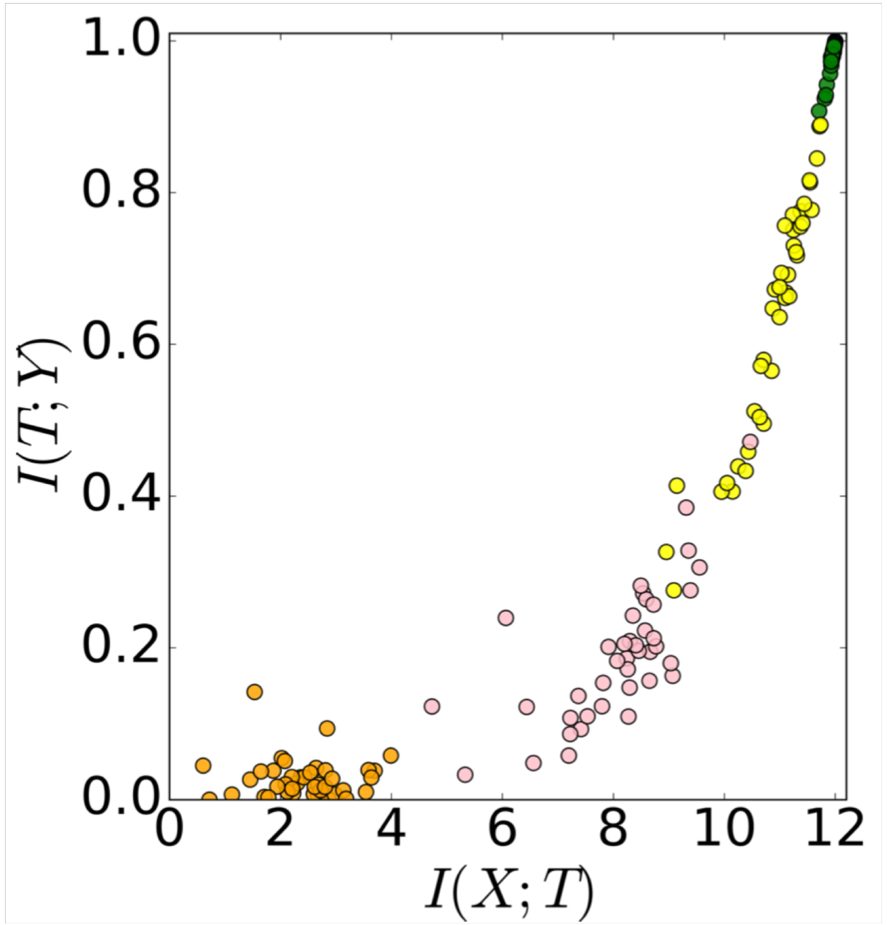
$$I(X;Y)\geq I(X;Z)$$

- Mutual Information Over Layers of DNN

$$I(X;Y)\geq I(T_1;Y)\geq I(T_2;Y)\geq ...\geq I(T_k;Y)\geq I(\hat{Y};Y)$$
$$H(X)\geq I(X;T_1)\geq I(X;T_2)\geq ...\geq I(X;T_k)\geq I(X;\hat{Y}).$$

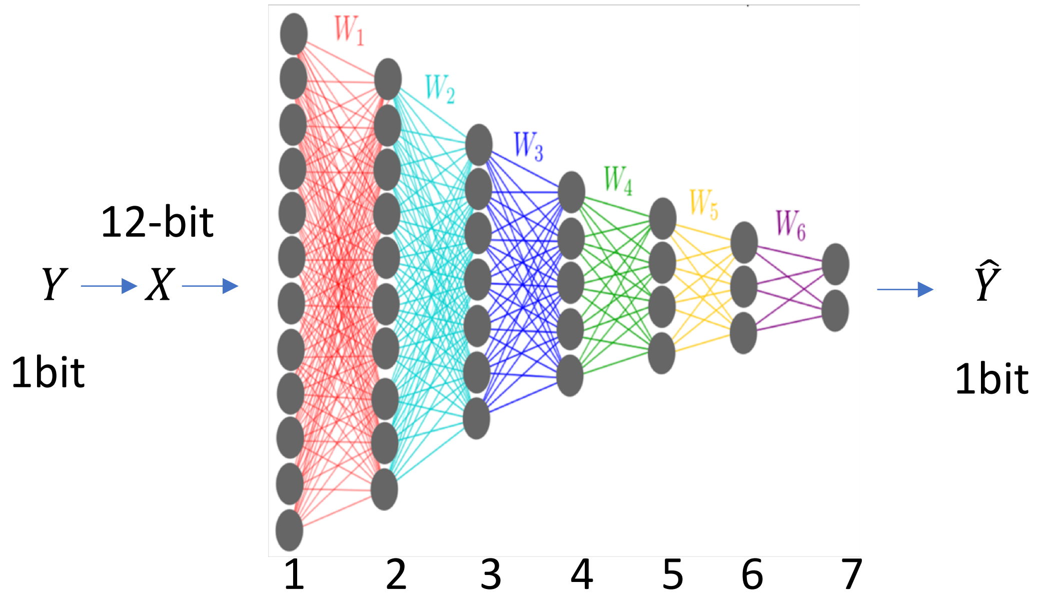
A **Markov chain** is "a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event".

--Wikipedia



Experimental Setup

- 7 Fully Connected Hidden Layers
  - 12-10-7-5-4-3-2
- Label Y: {0,1}
- Input X: 12-binary vector, total 4096, equally distributed



Mutual Information of Hidden Layers

- 30-bin for each neuron's output

$$t \in T_i$$

- Joint Distribution

$$P(T_i, X)$$

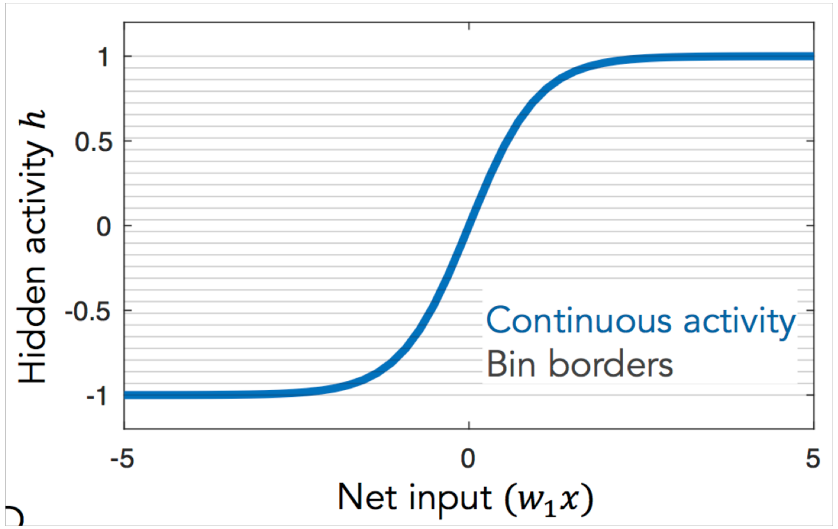
$$P(T_i, Y) = \sum_x P(x, Y) P(T_i|x)$$

- Mutual Information

$$I(X; T_i)$$

$$I(T_i; Y)$$

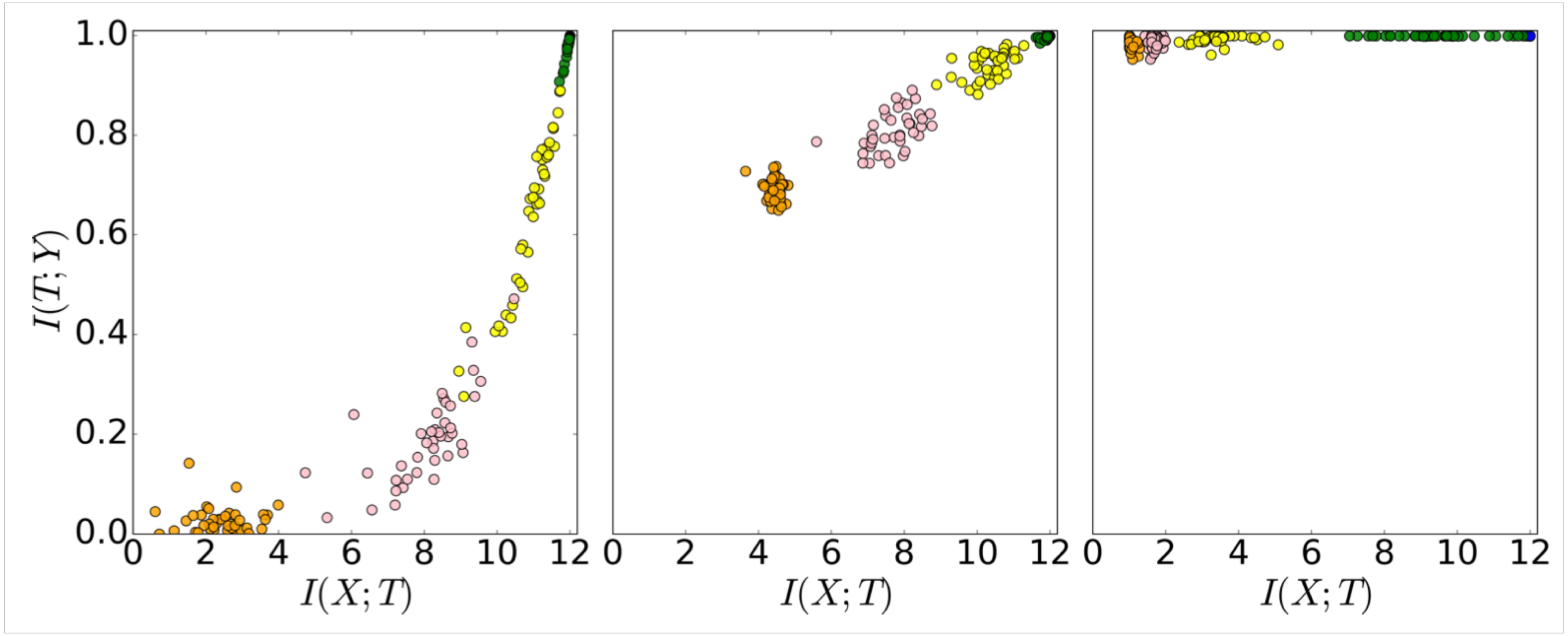
30-bin for Each Neuron's output



Training Process

1. Color-coded hidden Layer
2. 50 randomized networks

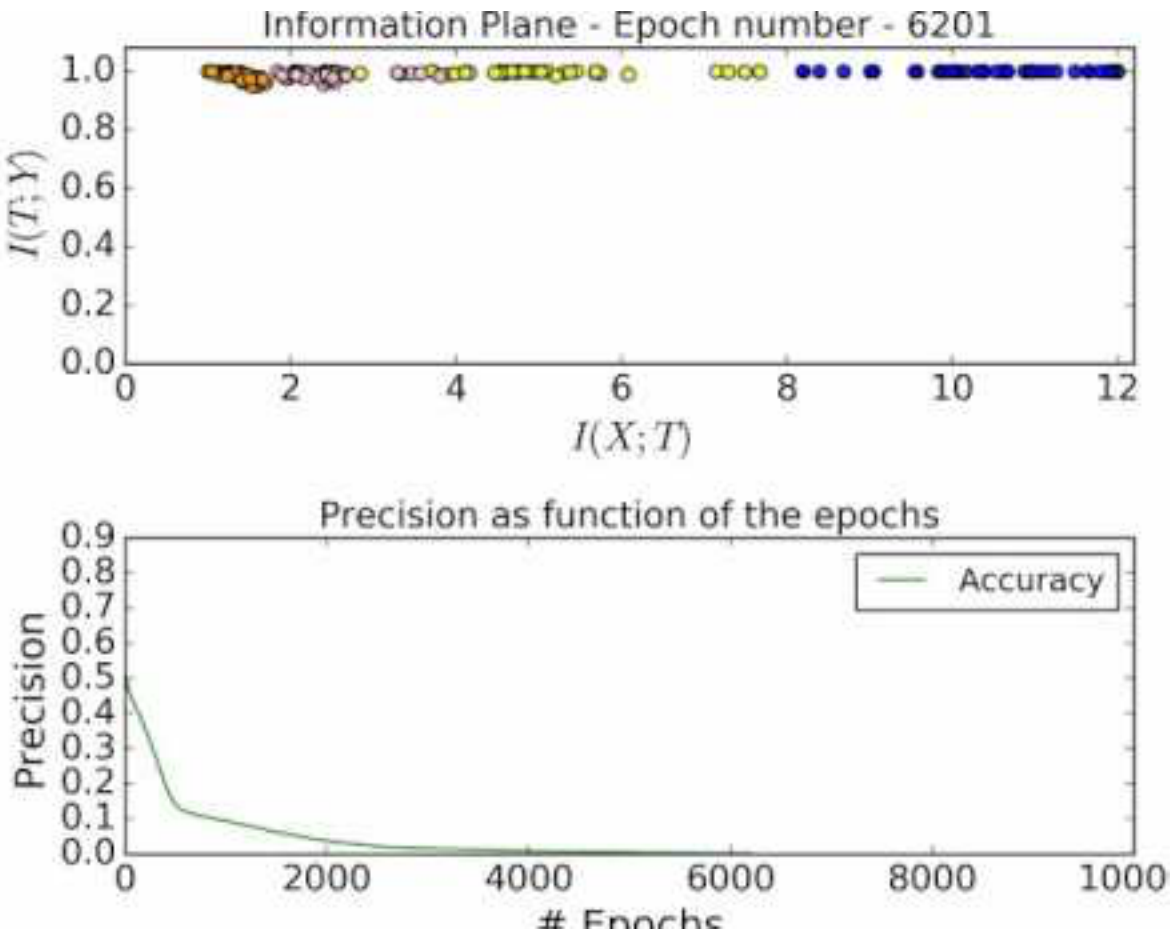
https://goo.gl/rygyIT
https://goo.gl/DQWuDD



Initial Weights 400 Epoch 9000 Epoch

Training Process

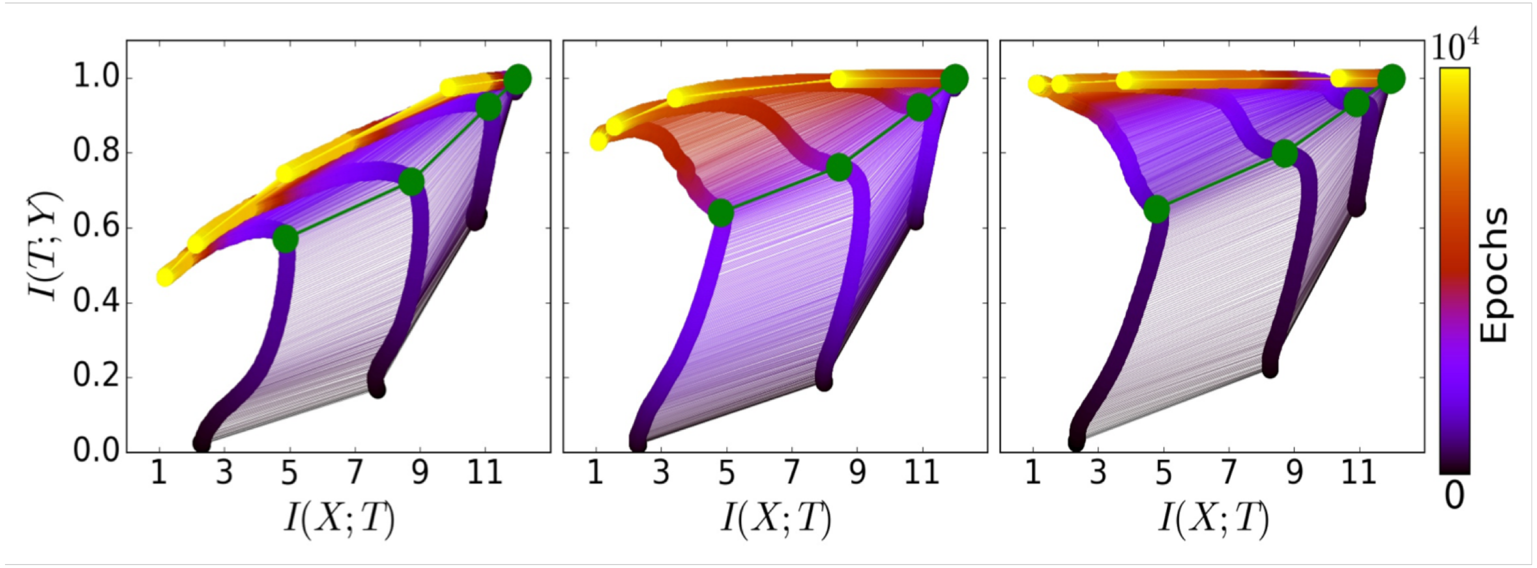
- Fitting
- Compression



Two Phase Training Process

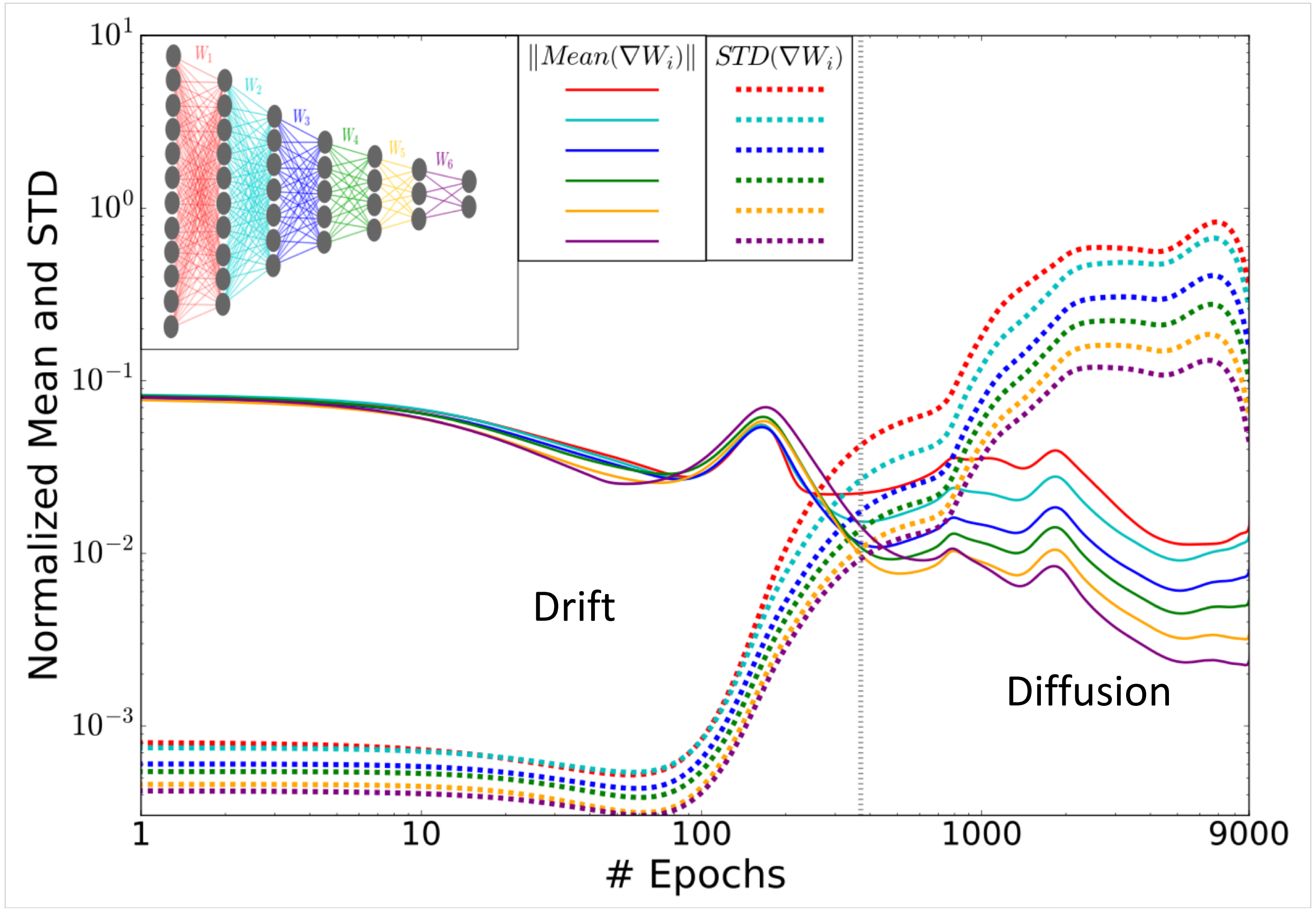
- Fitting (ERM)
  - Increase I(X;T) — Get relevant information from X
- Representation-Compression
  - Decrease I(X;T) — Drop Irrelevant Information from X

5% Data 45% Data 85% Data



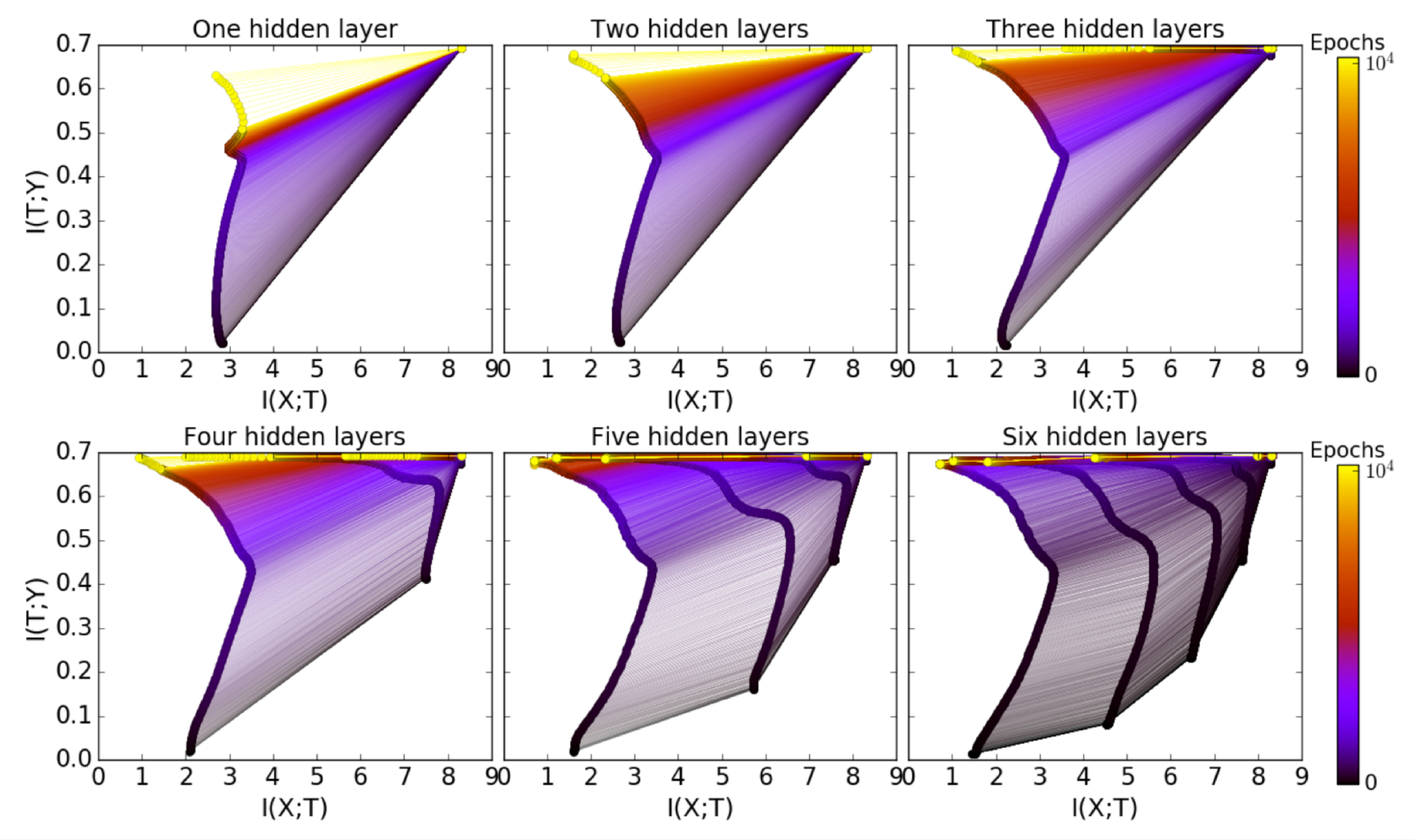
Drift and Diffusion Phases of SGD Optimization

- Weights' Stochastic Gradients

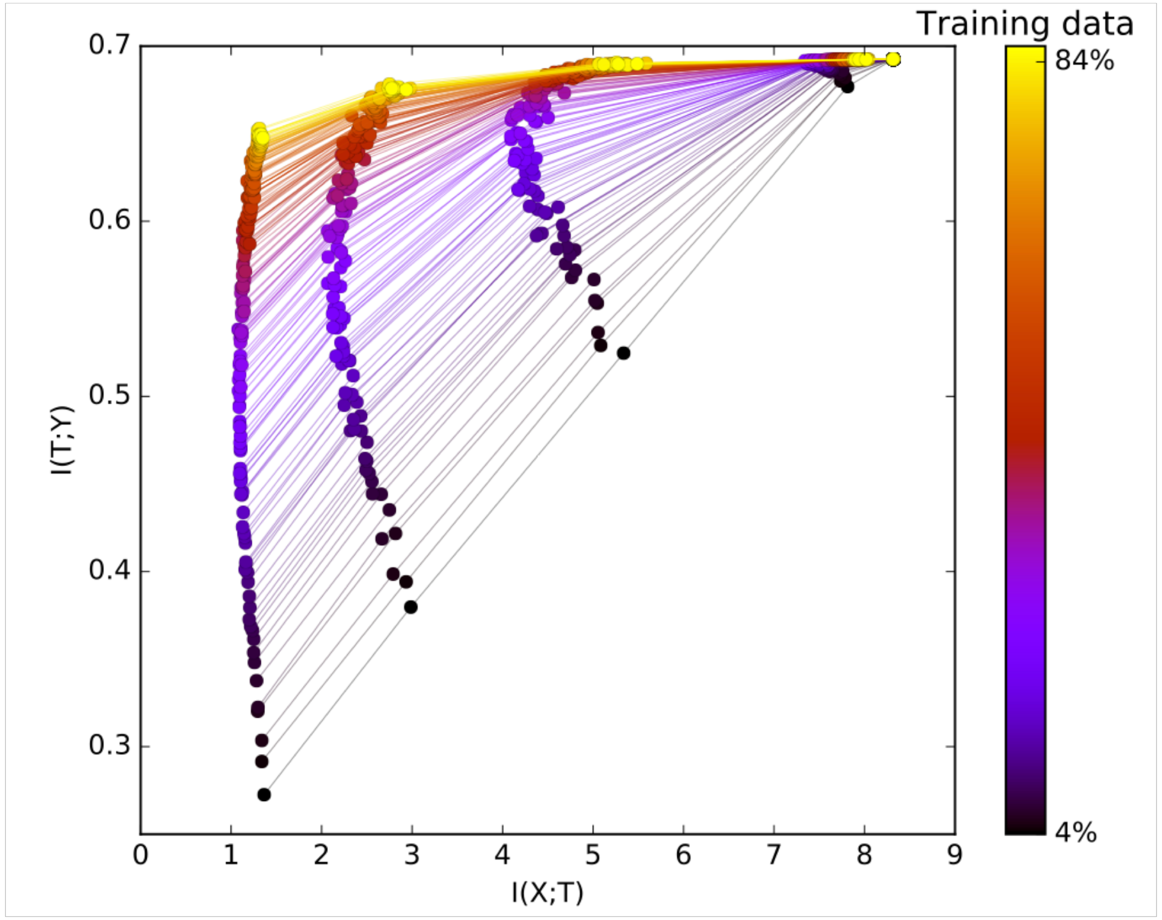


Computational Benefit of Hidden Layers

1. More Layers, less epoch
2. Compression phase is shorter if a layer start from a compressed layer
3. Compression is faster for deeper layers



Evolution of layers with training sample sizes



Converged Network with different training sample sizes

Summary of IB Theory on DNN

- Model
  - See DNN as a Markov Chain
  - Mutual Information of hidden layers with X, Y
- Learning Processes
  - Learning, Compression
- Training Dynamics
  - Drift, Diffusion
- Internal Representation
  - Information bottleneck set by training sample size
  - Hidden Layers brings computational benefit

Part 2: The Attack on IB Theory

Harvard University, Santa Fe Institute, MIT-IBM Watson AI Lab

Attacks on IB Theory on Deep Learning

- Who
  - A.M. Saxe, Y. Bansal, J. Dapello, M. Advani, (Harvard)
  - A. Kolchinsky, B.D. Tracey (Santa Fe Institute)
  - D.D. Cox (Harvard, MIT-IBM Watson AI Lab)
- What
  - Two Training Phases: Fitting, Compression
  - Compression phase related to generalization
  - Compression occurs due to diffusion-like behavior of SGD
- How
  - Counter Examples

[6] Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, and David Daniel Cox. "On the Information Bottleneck Theory of Deep Learning", ICLR 2018, https://openreview.net/forum?id=ry_WPG-A-

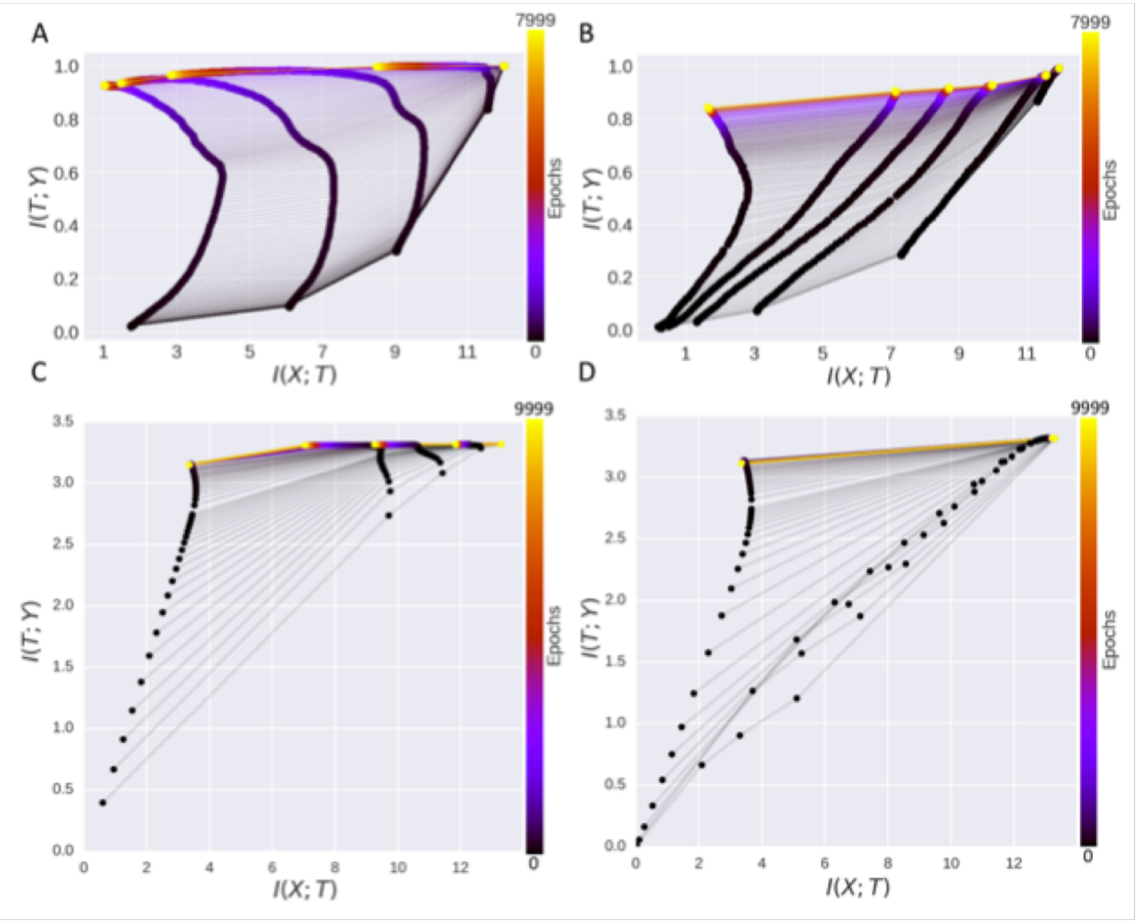
Example 1: Relu v.s. tanh

A: Replication of [1]
B: Replace tanh with Relu

X: 12-binary Vector
(12-10-7-5-4-3-2)

C: tanh network
D: Relu network

X: MNIST
(784-1024-20-20-10)



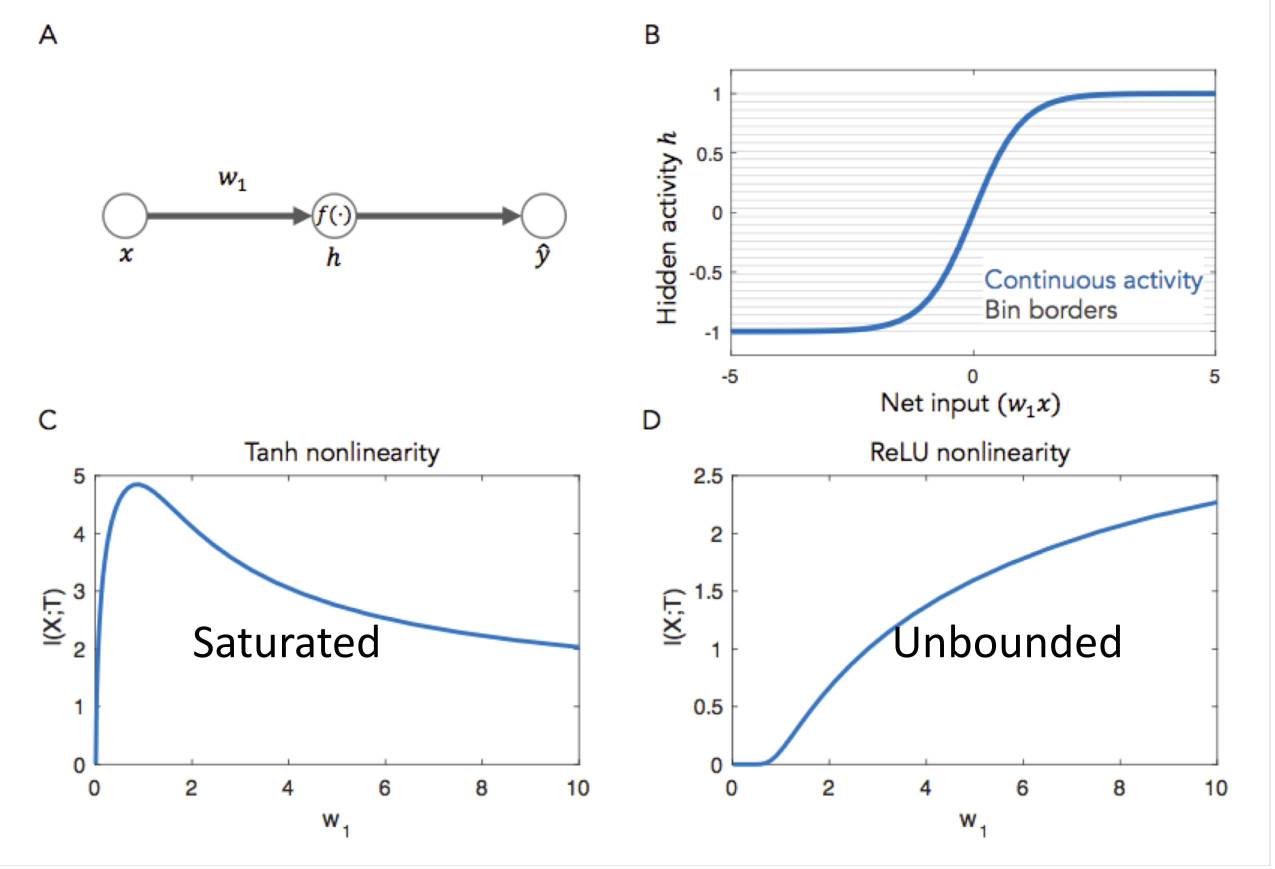
[1] R. Schwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. arXiv preprint arXiv:1703.00810, 2017

Minimal Model Illustration

1 hidden layer
Scalar Gaussian Input X ~N(0,1)

$$\begin{aligned} I(T;X) &= H(T) - H(T|X) \\ &= H(T) \\ &= -\sum_{i=1}^{N} p_i \log p_i \end{aligned}$$

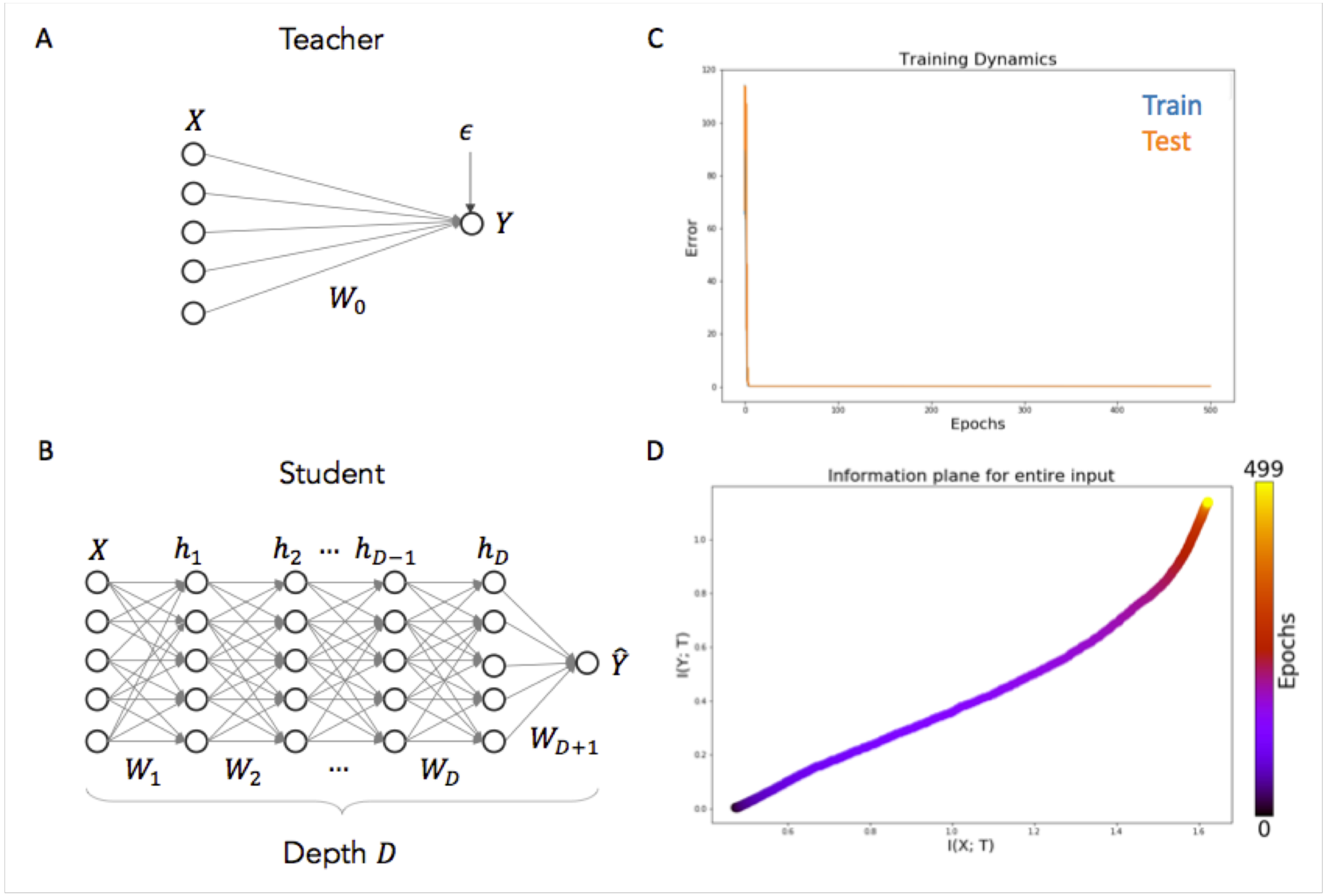
$$p_i = P(X \geq f^{-1}(b_i)/w_1 \text{ and } X < f^{-1}(b_{i+1})/w_1),$$



tanh Relu

Example 2: Deep Linear Network

A: Linear Teacher Network
  Gaussian input X
  Add noise
B: Deep Linear Student Network
  1 hidden layer



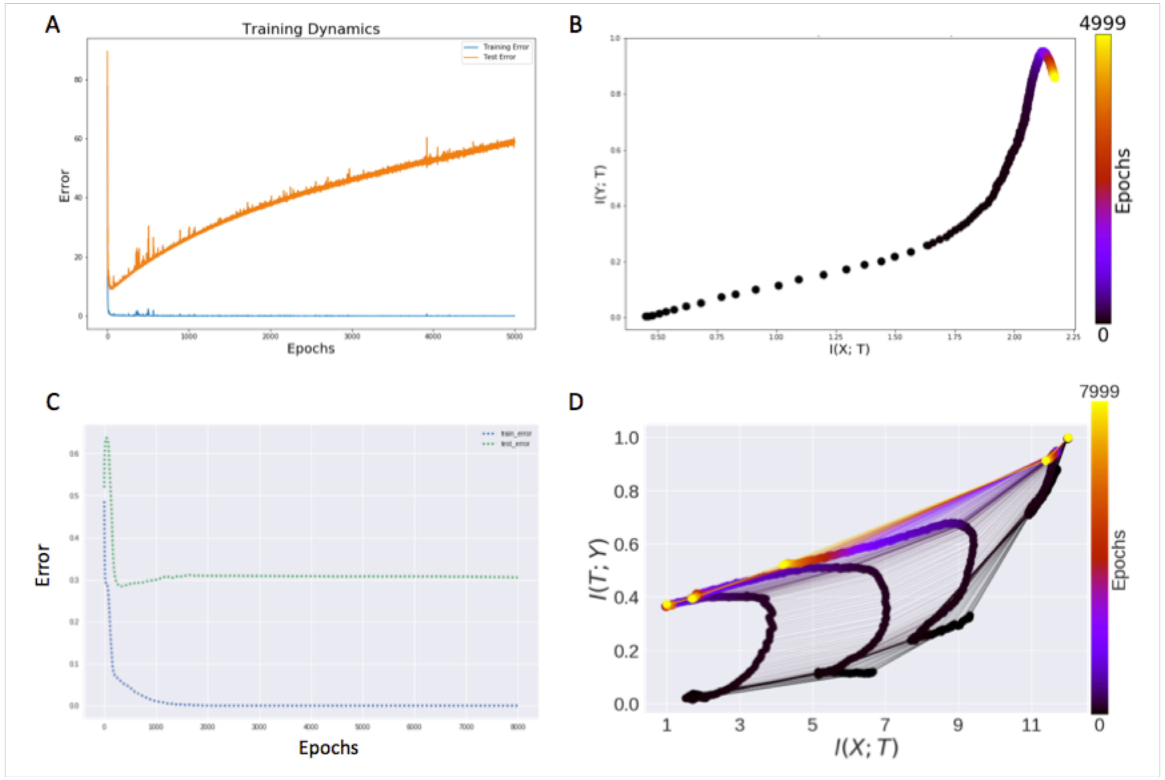
100 units   100 units   1 units

Example 3: Over-Training (over-fitting)

Deep Linear network

30% data

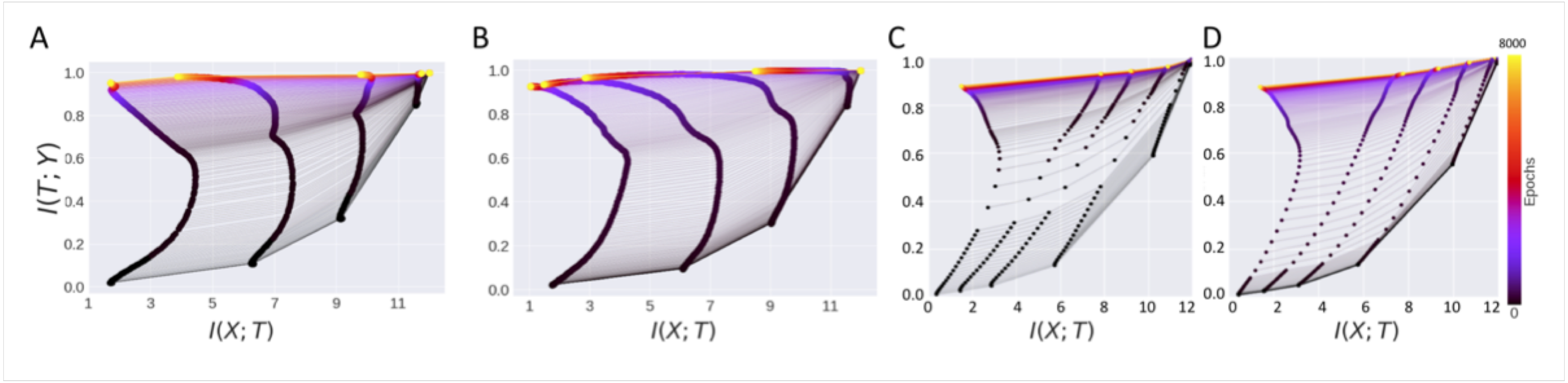
Tanh network



Cause of Compression: SGD v.s. BGD

Theoretical Claim:
Diffusion phase→ small training error→

- Stochastic Gradient Descent
  - random input samples, the weights evolve in a stochastic way during training
- Batch Gradient Descent
  - Full training dataset, no randomness or diffusion-like behavior in its updates

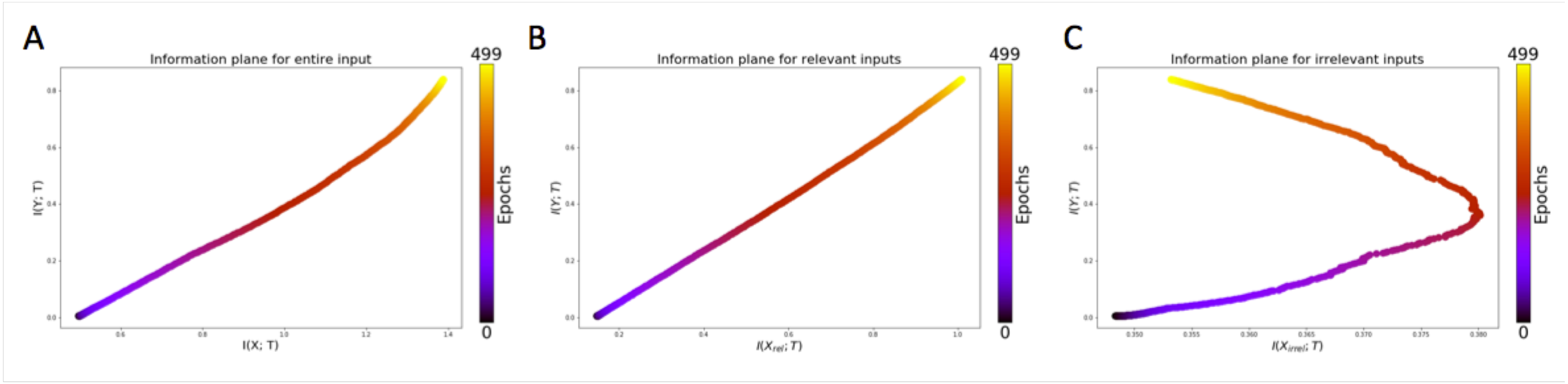


Tanh, SGD | Tanh, BGD | Relu, SGD | Relu, BGD

Simultaneous Fitting and Compression

- Deep Linear Network
  - Task Relevant Inputs: X_{rel} — Signal + Noise — 30 inputs
  - Taks Irrelevant Inputs: X_{irrel} — Only Noise — 70 inputs

SGD
(5 samples/Batch)



Entire Inputs | Task-Relevant Subspace | Task-Irrelevant Subspace

Summary

- Compression dynamics
  - not general feature of deep networks
  - Critically influenced by the nonlinearities
    - Double Saturating nonlinearities lead to compression
- Compression
  - not caused by Stochasticity in training process
  - May not casually linked to generalization
- Compression
  - may still occurs in a subset of inputs
  - May not be visible from trace of information metrics
- Information Bottleneck Principle
  - May still offer important insight into deep networks
  - Could yield new training algo for inherently stochastic network, and regularization

Response from Naftali Tishby

https://www.reddit.com/r/MachineLearning/comments/79efus/r_on_the_information_bottleneck_theory_of_deep/