ON THE INFORMATION BOTTLENECK THEORY OF DEEP LEARNING

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

15 Feb. 2018

Agenda

1. Introduction
2. Compression and Neural Nonlinearities
3. Information Plane Dynamics in Deep Linear Networks
4. Compression in Batch Gradient Descent and SGD
5. Simultaneous Fitting and Compression
6. Discussion

Introduction

- analyzes and responds to (Tishby & Zaslavsky, 2015; Shwartz-Ziv & Tishby, 2017)
- Tishbyproposed his theory can be used to compare different architectures
- information bottleneck (IB) theory provides a fundamental bound on the amount of input compression and target output information that any representation can achieve (Tishby et al., 1999)

Introduction

- "fitting" phase:
  mutual information between the hidden layers and both the input and output increases

- "compression" phase :
  mutual information between the hidden layers and the input decreases

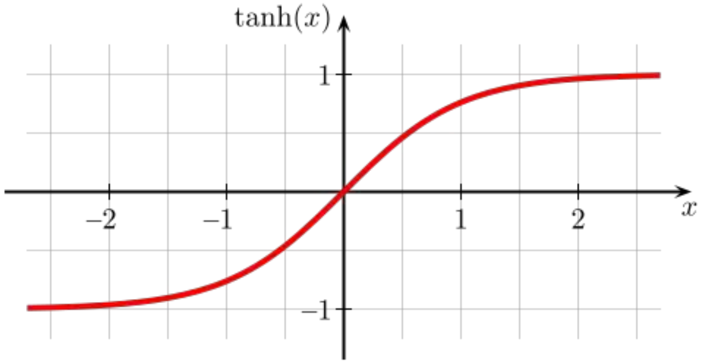
- Hypothesize:
  - compression phase is responsible for the excellent generalization performance of deep networks
  - occurs due to the random diffusion-like behavior of stochastic gradient descent.

Introduction

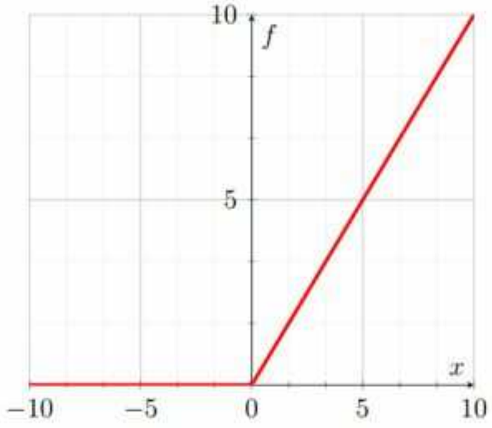
Aim is to study these phenomena using a combination of analytical methods and simulation

Compression and Neural Nonlinearities

tanh nonlinearity activation function:

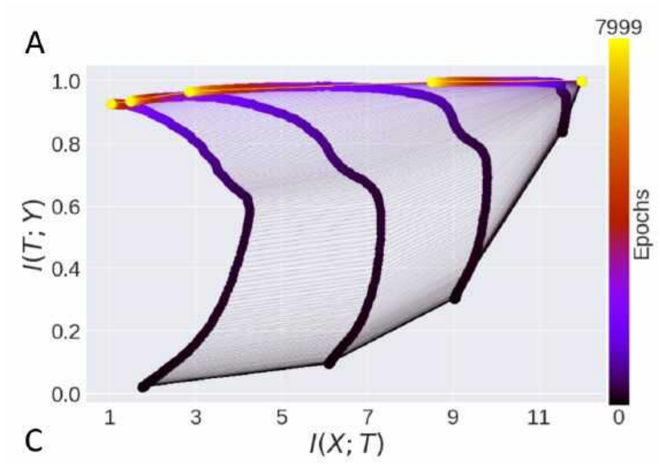


rectified linear activation function:



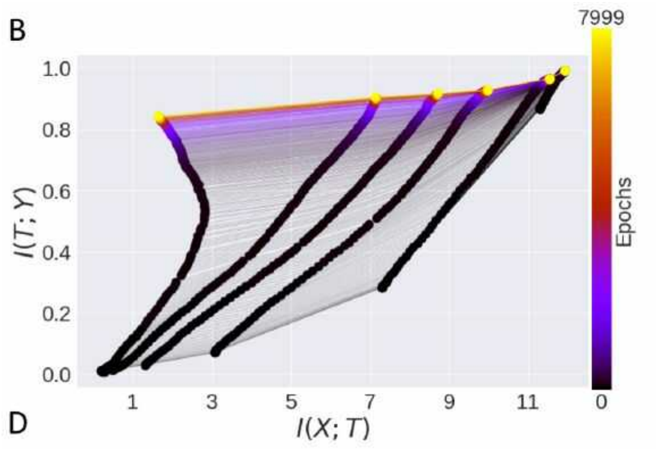
Compression and Neural Nonlinearities

Tanh nonlinearity activation function:



compression

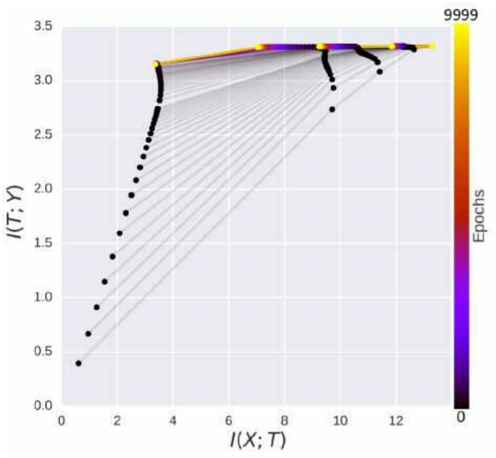
Rectified Linear activation function:



no compression

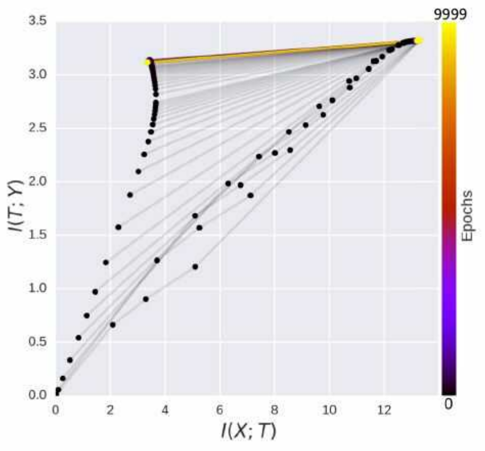
Compression and Neural Nonlinearities

Tanh trained on MNIST:



compression

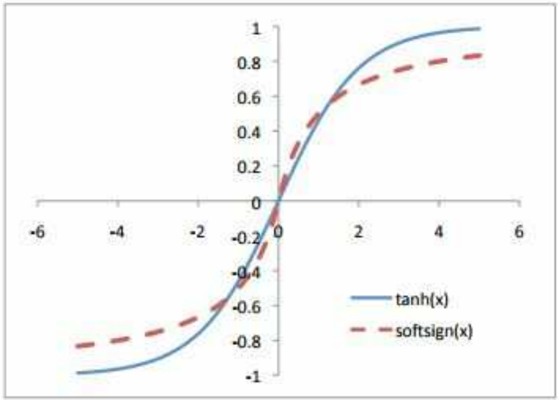
ReLU trained on MNIST:



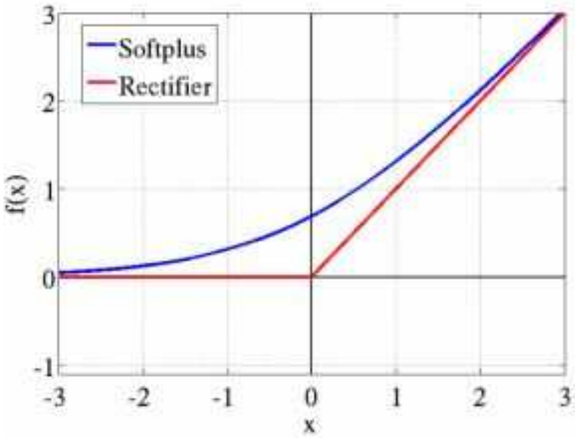
no compression

Compression and Neural Nonlinearities

Soft-sign activation function:

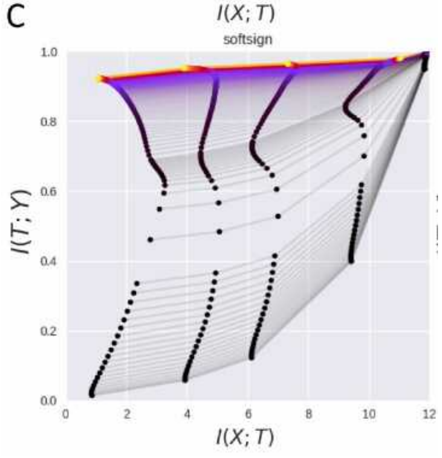


Soft-plus activation function:



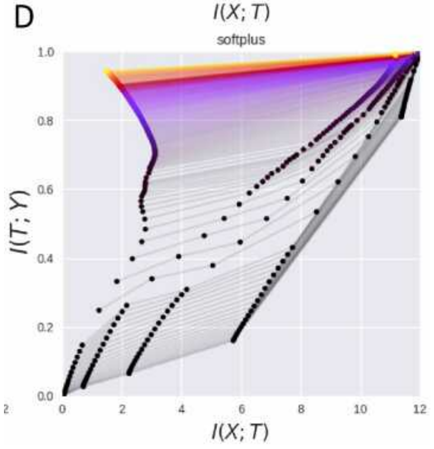
Compression and Neural Nonlinearities

Soft-sign activation function:



modest compression

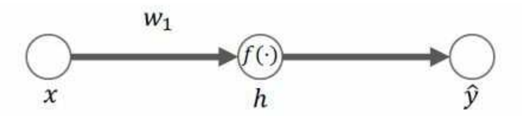
Soft-plus activation function:

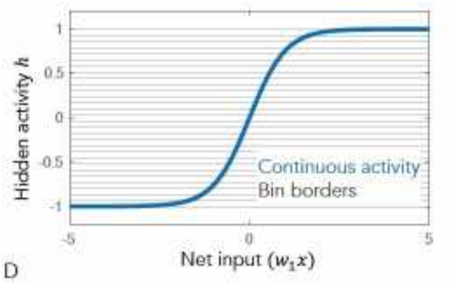


no compression

Compression and Neural Nonlinearities

Three Neuron Model:





Mutual Information:

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

$$H(T|X) = 0$$

since T is a deterministic function of X

Compression and Neural Nonlinearities

input X produces a hidden unit activity that lands in bin i, defined by lower and upper bin limits bi and bi+1:

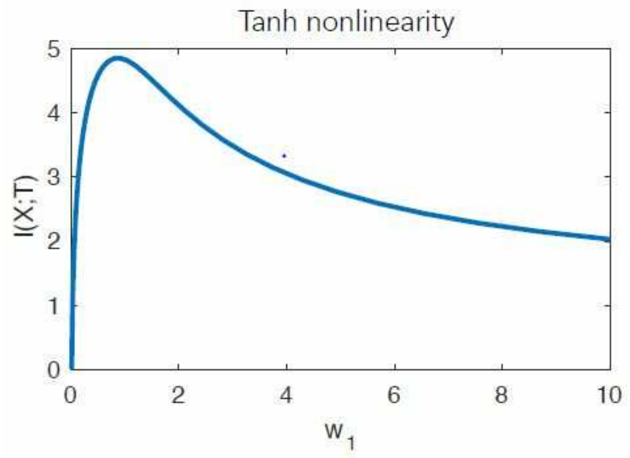
$$p_i = P(h \geq b_i \text{ and } h < b_{i+1})$$

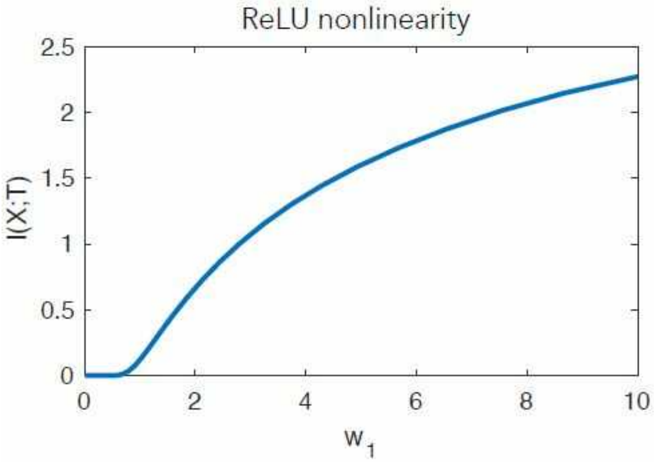
for monotonic nonlinearities f() using the cumulative density of X:

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

Compression and Neural Nonlinearities

mutual information as a function of weight



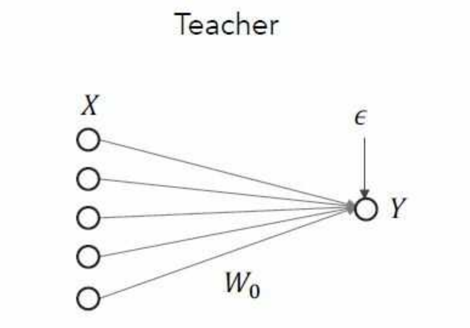


Compression and Neural Nonlinearities

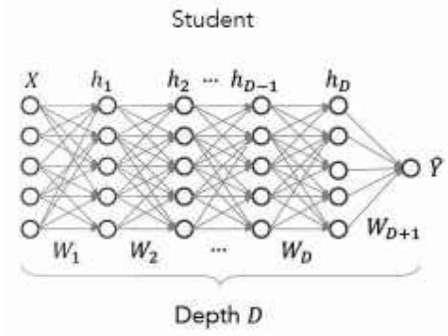
binning procedure can be viewed as implicitly adding noise to the hidden layer activity:

$$T = h + \epsilon$$

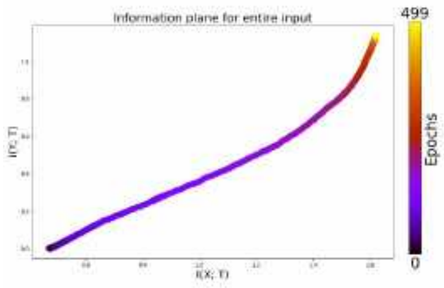
Information Plane Dynamics in Deep Linear Networks



generates a dataset by passing Gaussian inputs X through its weights and adding noise



A deep linear student network is trained on the dataset



no compression

Compression and Neural Nonlinearities

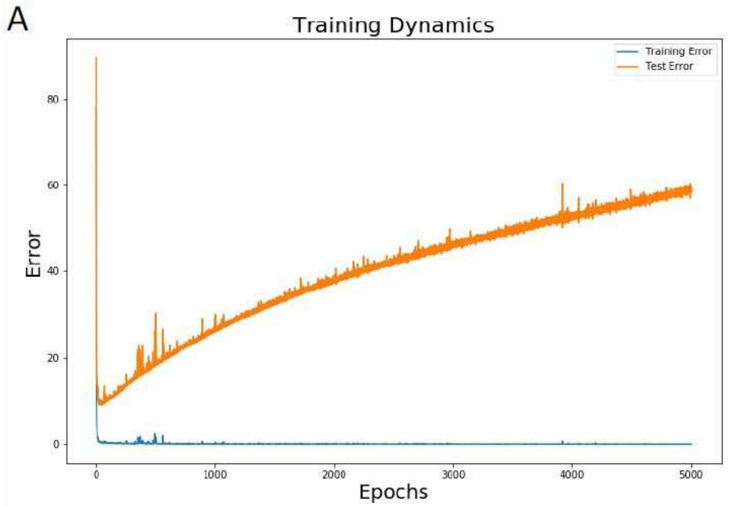
student network is then trained to minimize the mean squared error:

$$E_g(t) = ||W_o - W_{tot}(t)||_F^2 + \sigma_o^2$$

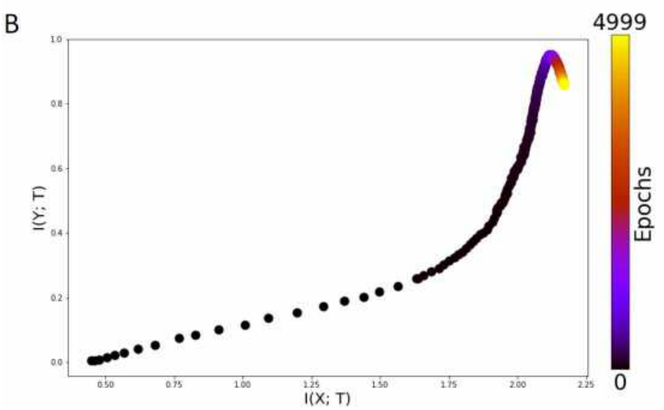
calculating the mutual information:

$$I(T;X) = \log|\bar{W}\bar{W}^T + \sigma_{MI}^2 I_{N_h}| - \log|\sigma_{MI}^2 I_{N_h}|$$

Information Plane Dynamics in Deep Linear Networks

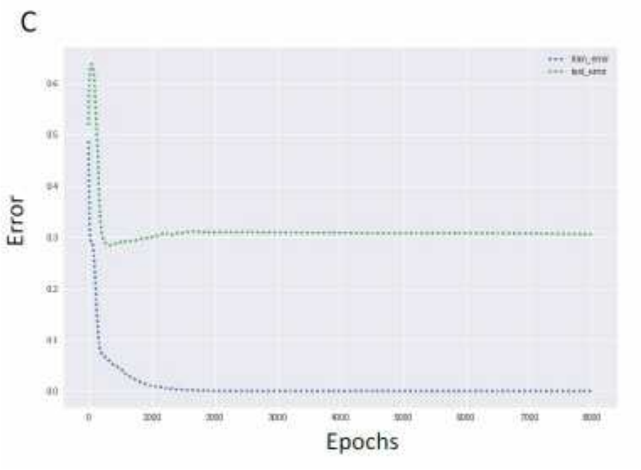


Average training and test mean square error for a deep linear network trained with SGD

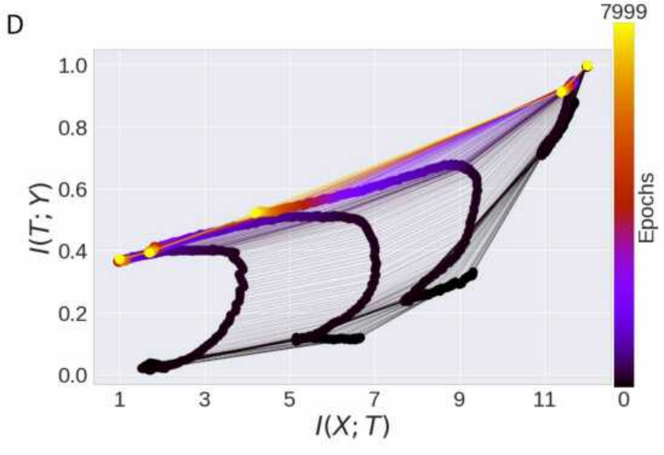


No compression is observed

Information Plane Dynamics in Deep Linear Networks

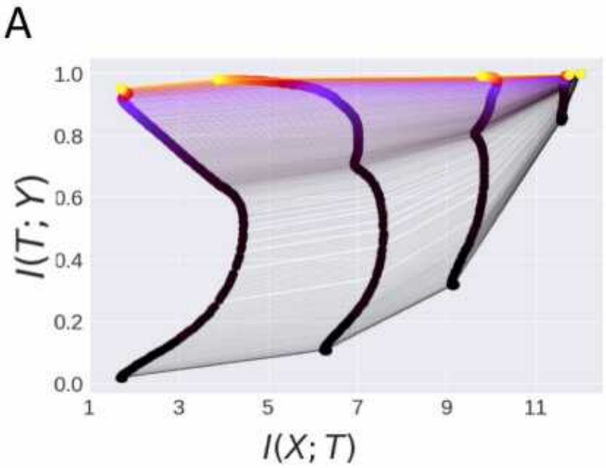


Average train and test accuracy (% correct) for nonlinear tanh networks

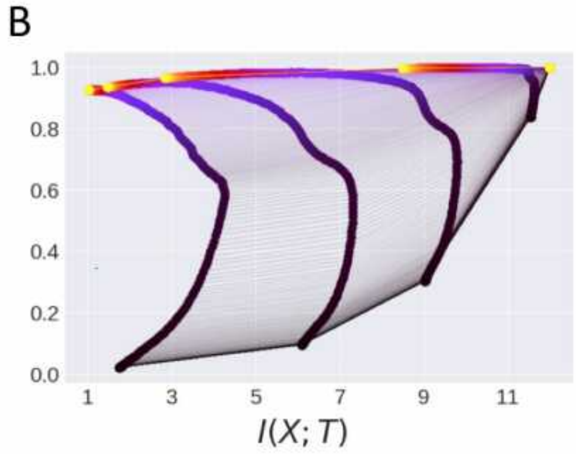


Overfitting occurs despite continued compression

Information Plane Dynamics in Deep Linear Networks

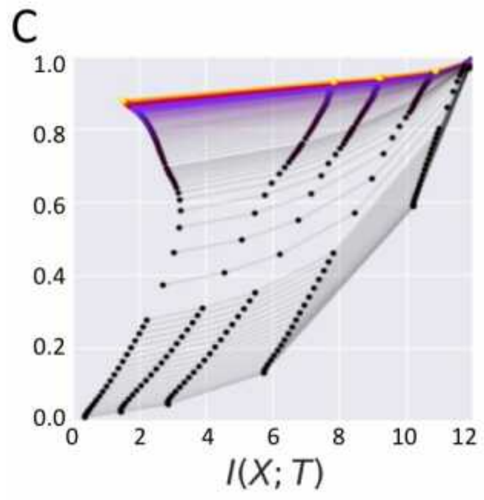


tanh network trained with SGD

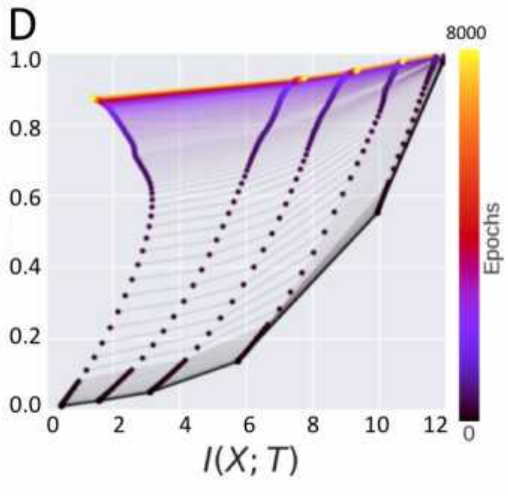


tanh network trained with BGD

Information Plane Dynamics in Deep Linear Networks



ReLU network trained with SGD



ReLU network trained with BGD

Compression in Batch Gradient Descent and SGD

Stochastic gradient descent is responsible for the compression phase?

“drift” phase:

mean of the gradients over training samples is large relative to the standard deviation of the gradients

“diffusion” phase:

the mean becomes smaller than the standard deviation of the gradients

Compression in Batch Gradient Descent and SGD

Stochastic gradient descent is responsible for the compression phase?

-> Explanation does not hold up to either theoretical or empirical

theoretical:

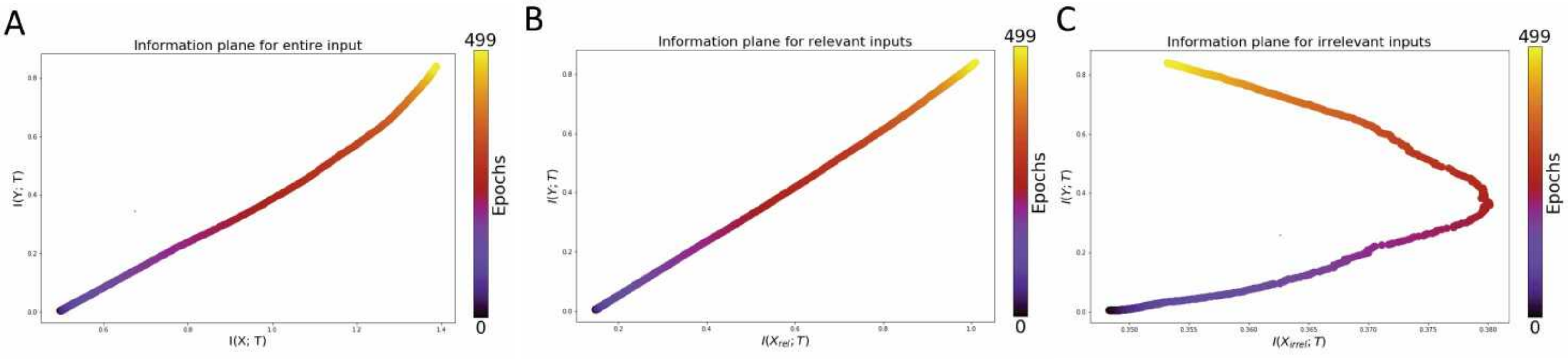
There is no general reason that a given set of weights sampled from this distribution (i.e., the weight parameters found in one particular training run) will maximize H(XjT)

empirical:

- stochasticity of the SGD is not necessary for compression

- showed by training tanh and ReLU networks with SGD and BGD

Simultaneous Fitting and Compression



For a large task-irrelevant subspace in the input, a linear network shows no overall compression

Information with the task-relevant subspace increases robustly over training

Information about the task-irrelevant subspace does compress over training

Discussion

- compression dynamics in the information plane are not a general feature of deep networks
- stochasticity in the training process does not contribute to compression
- generalization performance may not clearly track information plane behavior (link between compression and generalization?)
- link the information bottleneck principle with current practice in deep networks

Thank you for your attention!

Any questions?