On the Information Bottleneck Theory of Deep Learning

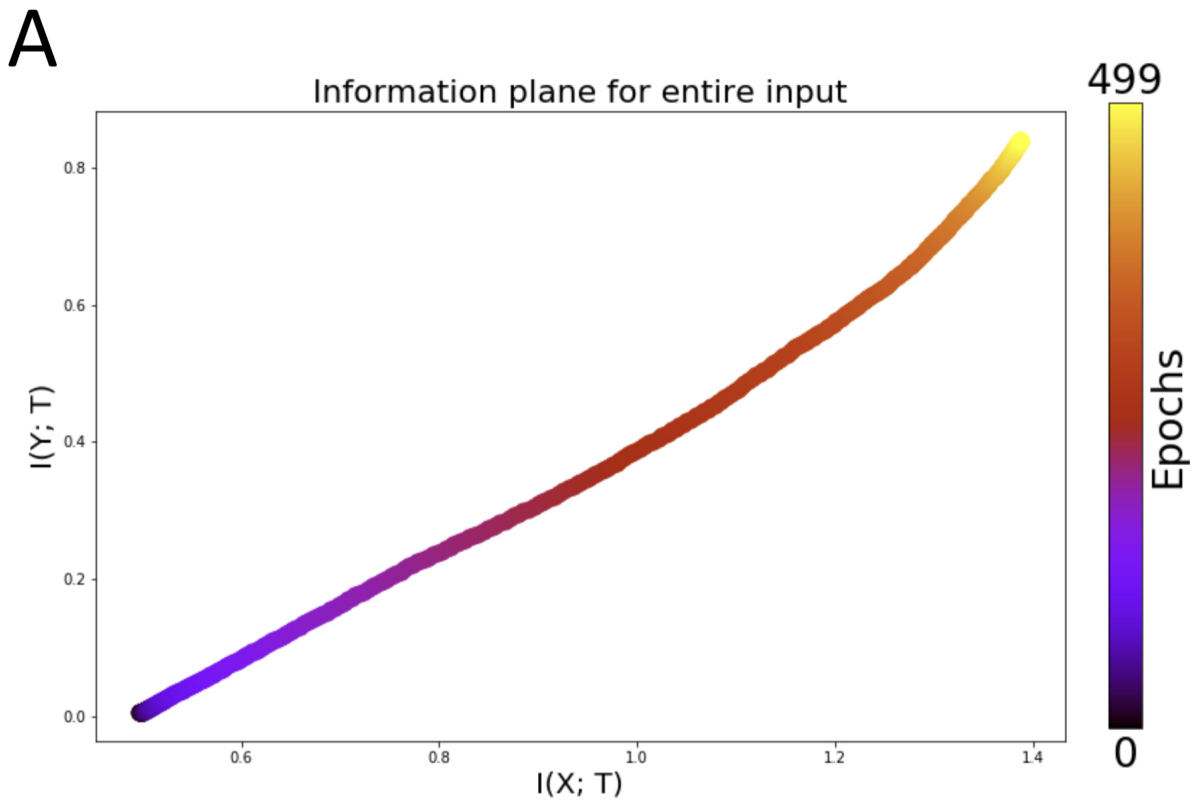
Polina Barabanshchikova

MIPT

November 27, 2022

Motivation

Trajectory in the information plane is the amount of information in a hidden layer regarding the input and output measured over the course of learning



Motivation

Observation(Shwartz-Ziv, Tishby)
Trajectories in the information plane consist of two distinct phases:

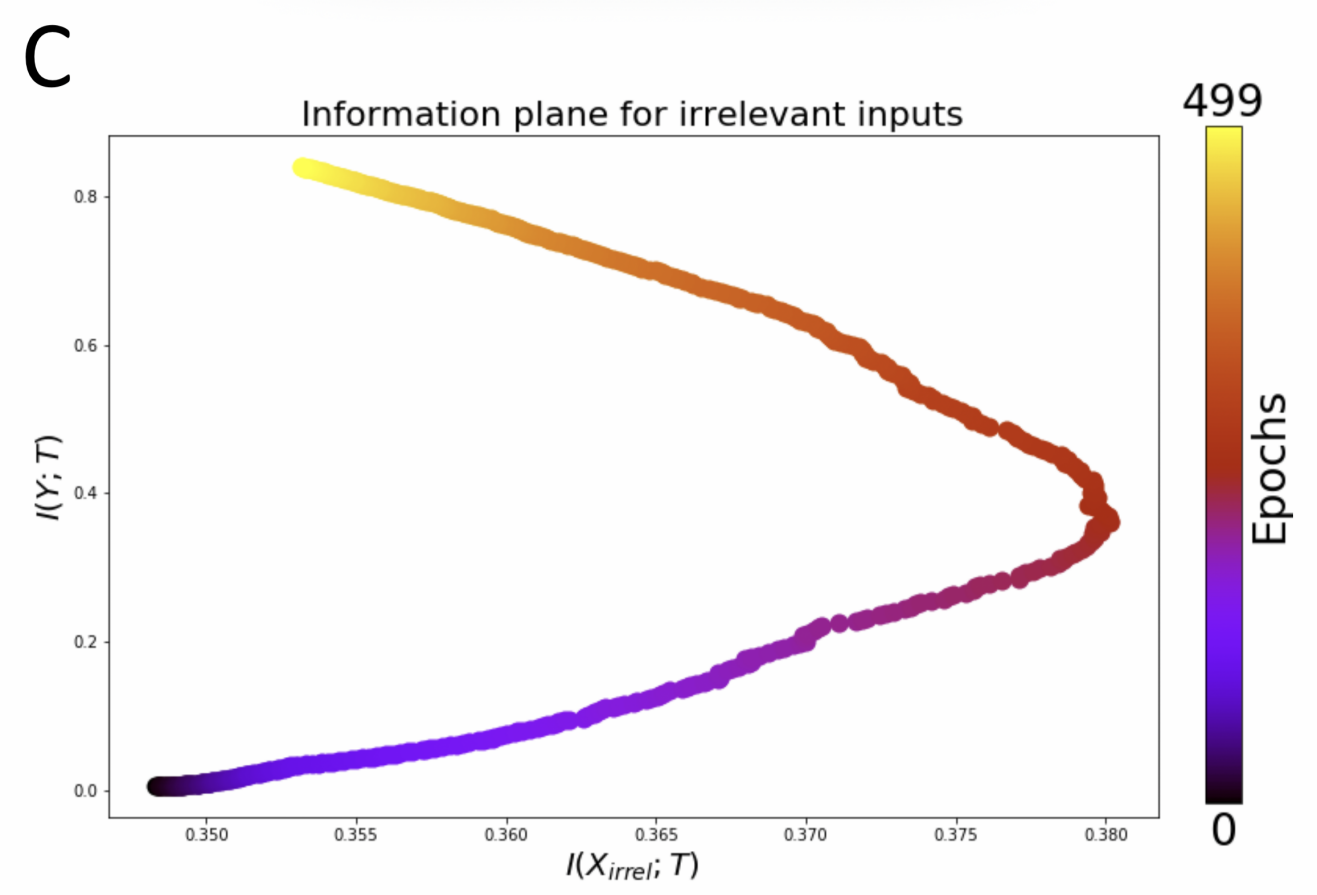
- a "fitting" phase where mutual information between the hidden layers and both the input and output increases
- a subsequent "compression" phase where mutual information between the hidden layers and the input decreases

Hypothesis(Shwartz-Ziv, Tishby)

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

Motivation

Fitting and compression phases



Results

- The information plane trajectory predominantly depends on the neural nonlinearity employed: double-sided saturating nonlinearities yield a compression phase, but others do not
- There is no evident causal connection between compression and generalization
- The compression phase, when it exists, does not arise from stochasticity in training

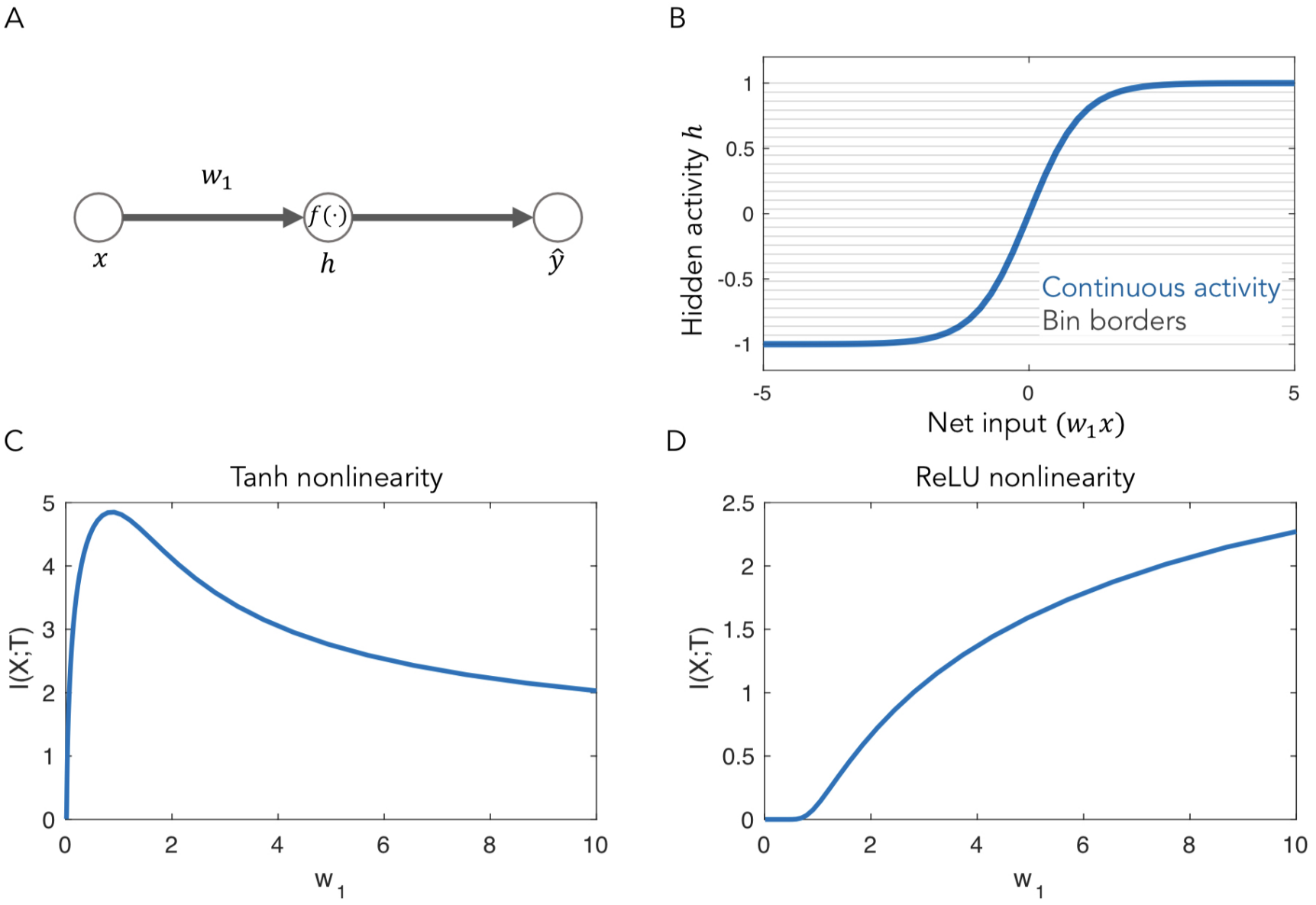
Method description. Minimal model

Consider the simple three neuron network with first layer weight w_1 and a nonlinearity f. Input $X \sim \mathcal{N}(0, 1)$ is fed through the net, yielding the hidden unit activity $h = f(w_1 X)$.
To calculate the mutual information with the input, this hidden unit activity is binned $T = \text{bin}(h)$ (into evenly spaced bins: from -1 to 1 for tanh and between the minimum and maximum activity values for relu).

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

where $p_i = P(h > b_i \text{ and } h < b_{i+1})$.
The actual $I(h; X)$ is infinite.

Experiments. Minimal model



Experiments. MLP

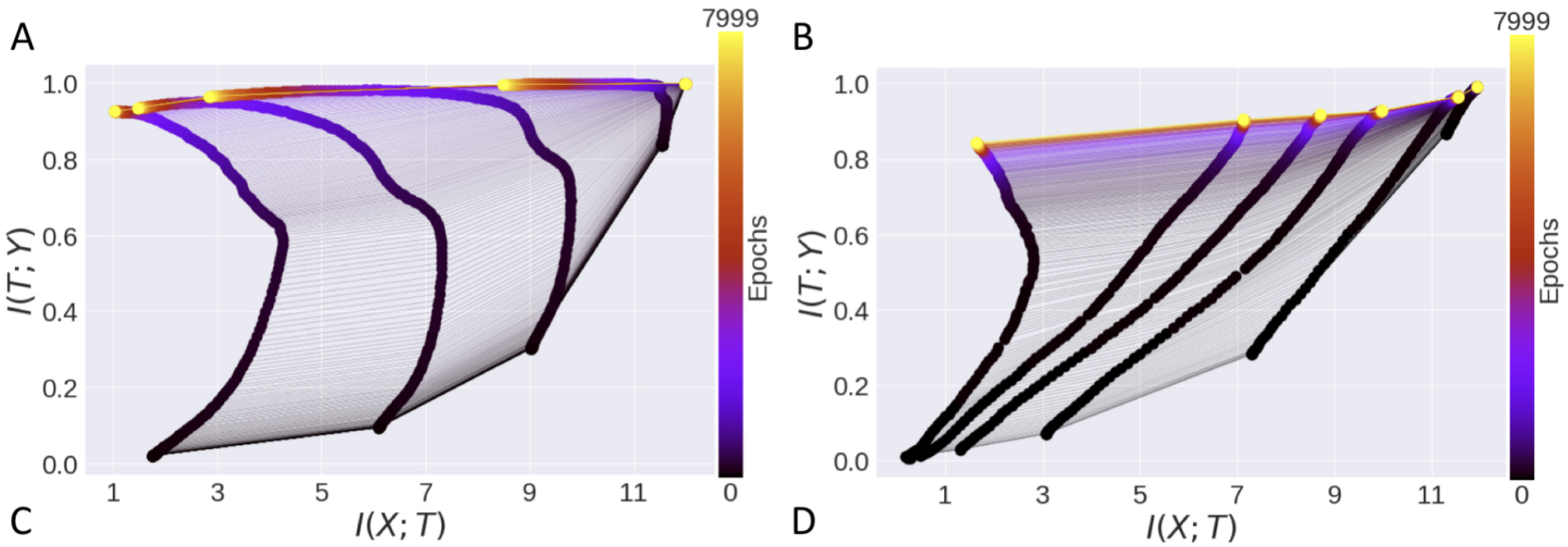
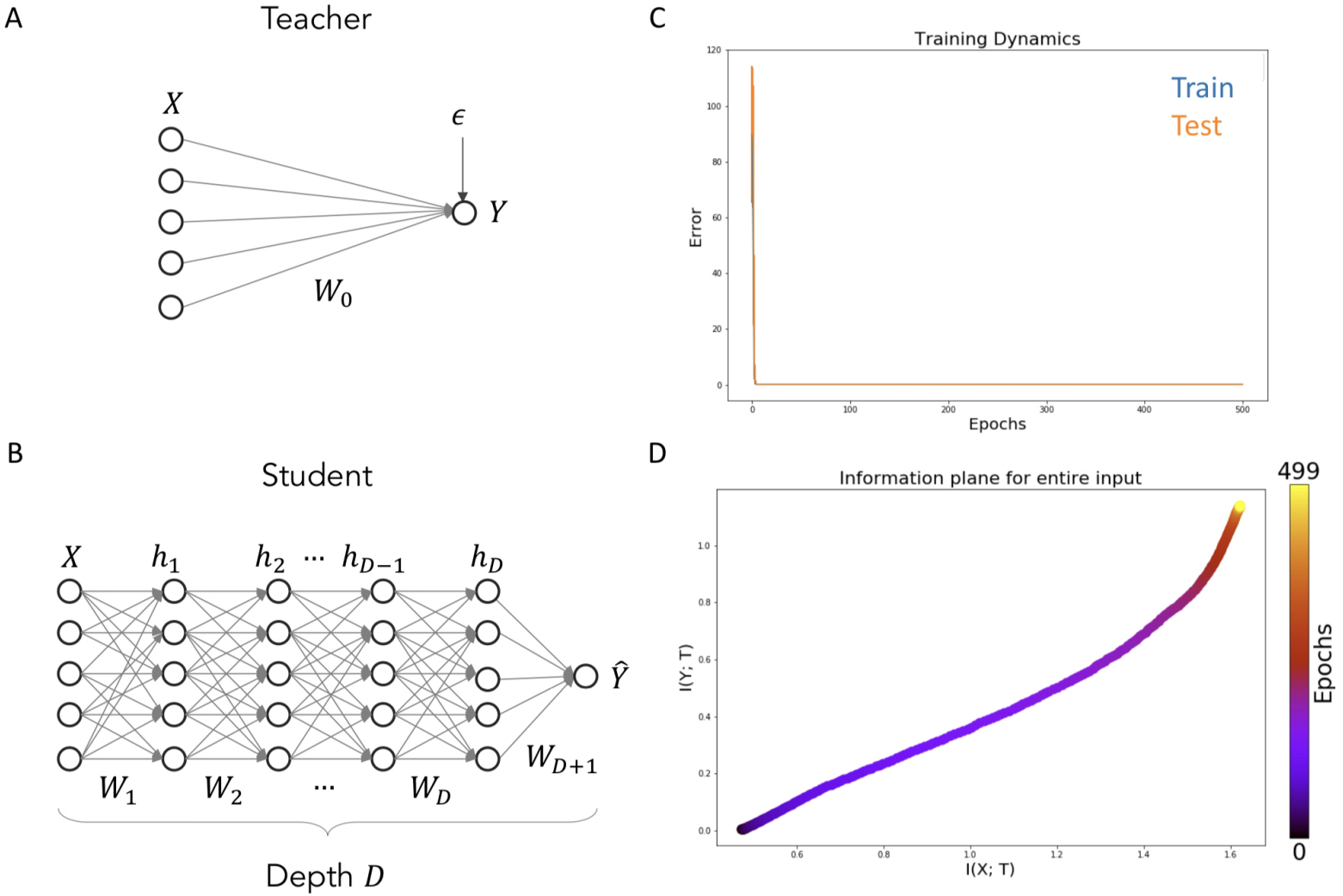


Figure 1: Information plane dynamics and neural nonlinearities. (A) Replication of Shwartz-Ziv and Tishby for a network with tanh nonlinearities. The x-axis plots information between each layer and the input, while the y-axis plots information between each layer and the output. The color scale indicates training time in epochs. Each of the six layers produces a curve in the information plane with the input layer at far right, output layer at the far left. Different layers at the same epoch are connected by fine lines. (B) Information plane dynamics with ReLU nonlinearities (except for the final layer of 2 sigmoidal neurons).

Method description. Generalization

Consider a scenario where a linear teacher network generates input and output examples which are then fed to a deep linear student network to learn.
Assume $X \sim \mathcal{N}(0,1)$ and $Y = W_0 X + \varepsilon_0$, where $\varepsilon_0 \sim \mathcal{N}(0, \sigma_0^2)$ and the weights W_0 are drawn independently from $\mathcal{N}(0, \sigma_w^2)$.
The student network is trained to minimize the mean squared error between its output and the target. Here the student is a deep linear neural network consisting of potentially many layers, but without nonlinearities, that is $\hat{Y} = W_{D+1} \dots W_1 X = W_{tot} X$.
For the purpose of calculating the mutual information, assume that Gaussian noise is added to the hidden layer activity, $T = W_{tot} X + \varepsilon$.

Experiments. Good generalization without compression



Experiments. Overtraining

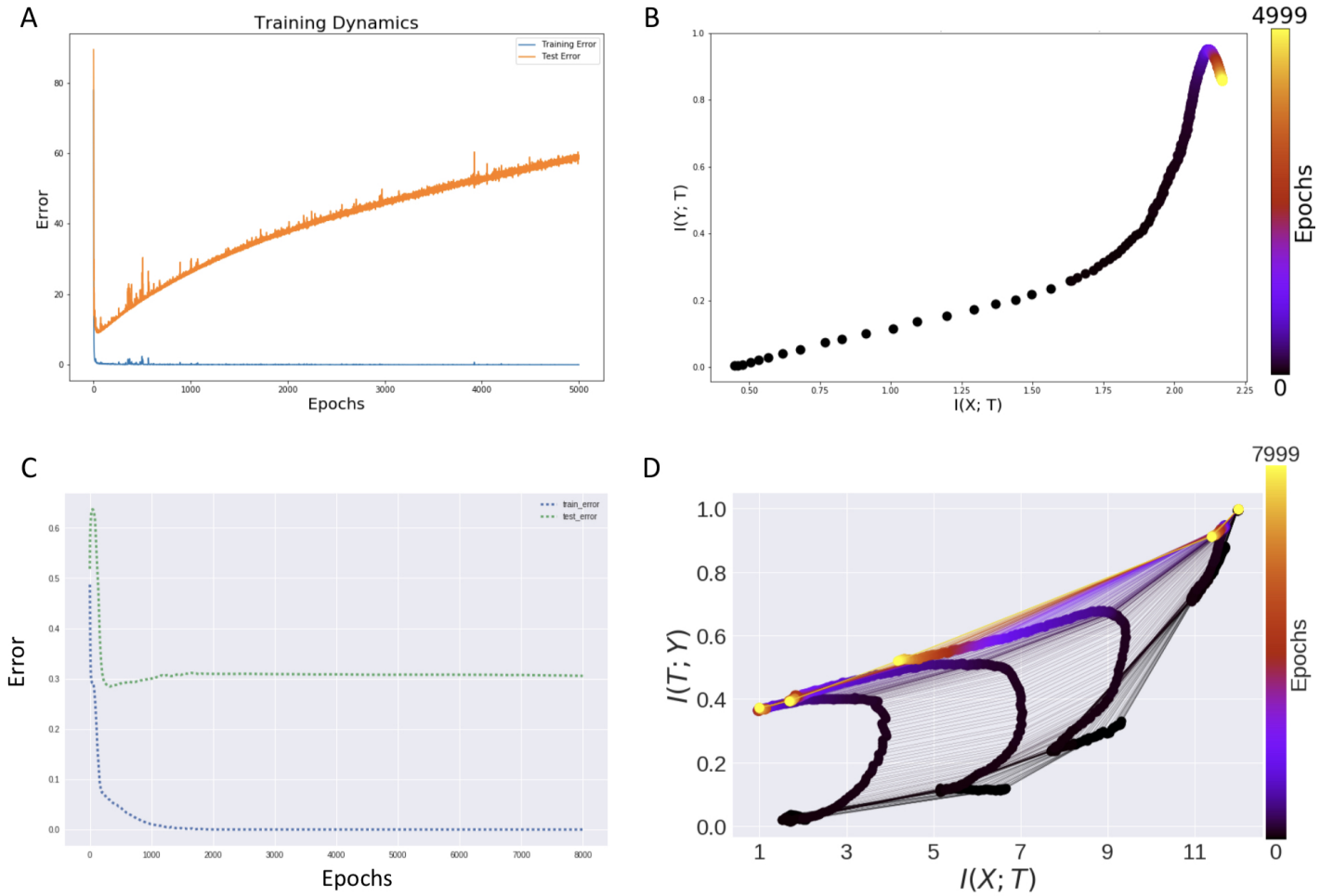


Figure 2: (A), (B) Overtitting with ReLU. (C), (D) Overfitting with Tanh.

Experiments. Stochastic training

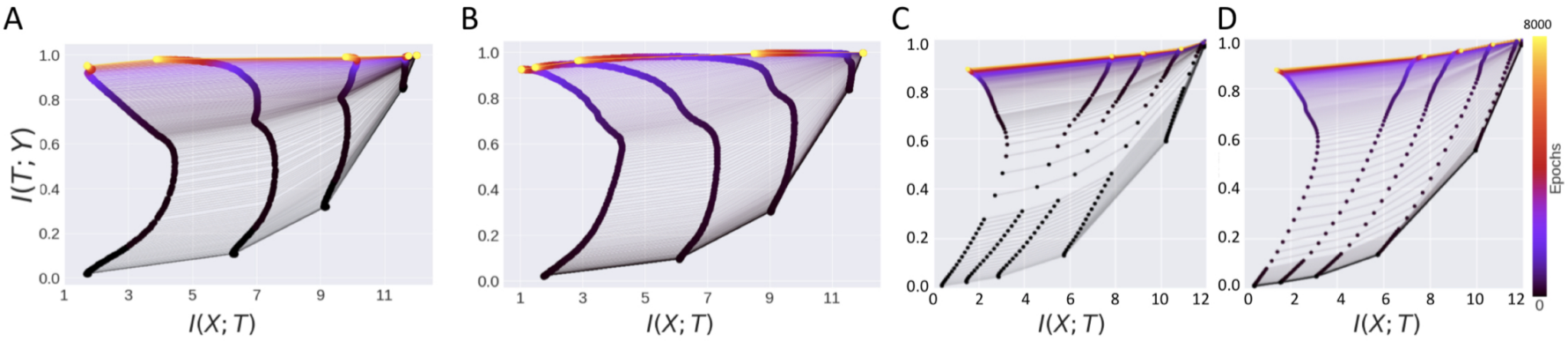


Figure 3: (A) tanh network trained with SGD. (B) tanh network trained with BGD. (C) ReLU network trained with SGD. (D) ReLU network trained with BGD. Both random and non-random training procedures show similar information plane dynamics.

Experiments. Simultaneous fitting and compression

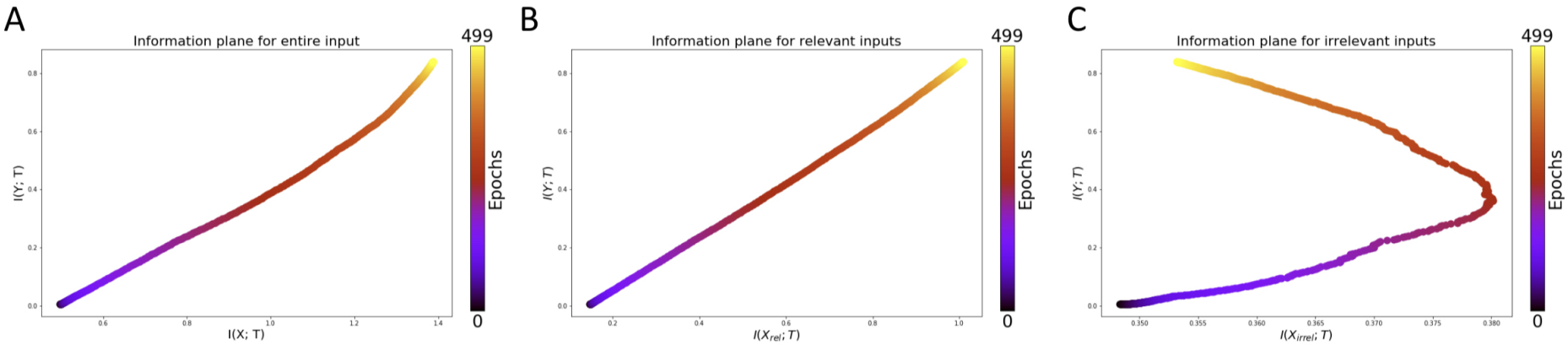


Figure 4: (A) For a task with a large task-irrelevant subspace in the input, a linear network shows no overall compression of information about the input. (B) The information with the task-relevant subspace increases robustly over training. (C) However, the information specifically about the task-irrelevant subspace does compress after initially growing as the network is trained.

Literature

[1] Andrew Michael Saxe et al. “On the Information Bottleneck Theory of Deep Learning”. In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=ry_WPG-A-.

Questions

1. What is the main factor that influence the information plane trajectory of a model?

2. The softsign activation function is defined by $f(x) = \frac{x}{1+|x|}$. Does it yield a compression phase or not? Does it cause more compression than tanh / relu? Explain why