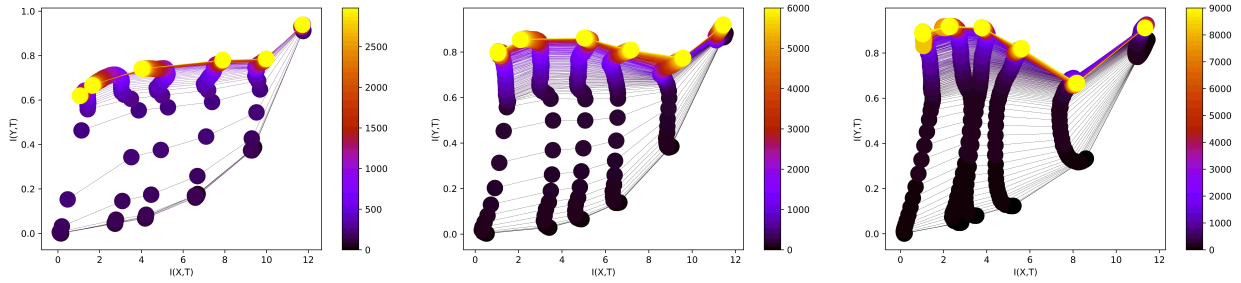
This appendix contains figures that visualize the Kernel Density Estimation (KDE) discrepancies when using activation function tanh and ReLu.

- tanh - see Figure 1, Figure 2: Almost every figure that visualizes a NN with the tanh activation function has a pronounced dip in label information, this observation seems to exists regardless of network shape or training size
- ReLu - see Figure 3, Figure 4: Only some of the ReLu NNs experience the same dip in information in the second layer.

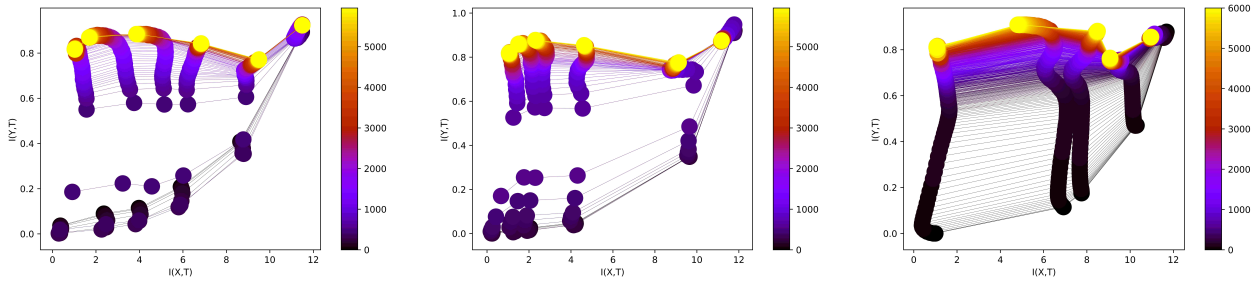
I was not able to find the source of this error. In the paper by Saxe their method does not seem to suffer from the same issue. In order to not introduce potentially wrong data into the project I decided to leave out information planes that were measured with the KDE MIE.

In order to implement the KDE MIE I used the same code that was used by Saxe in his paper.



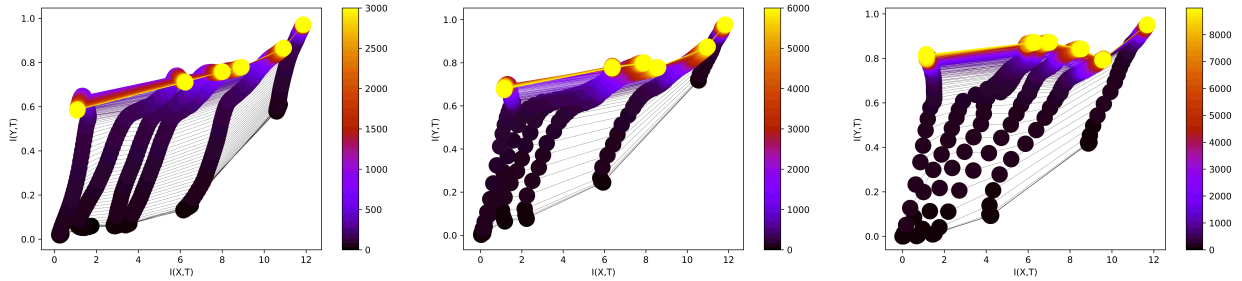
(a) Training size - 20% (b) Training size - 40% (c) Training size - 70%

Figure 1: tanh: Demonstrating KDE for different training sizes. Tweaking training size for Tishby's KDE MIE. Hyperparameters: Dataset - Tishby's, activation function - tanh, batch size - 512, network shape 12,10,8,6,4,2.



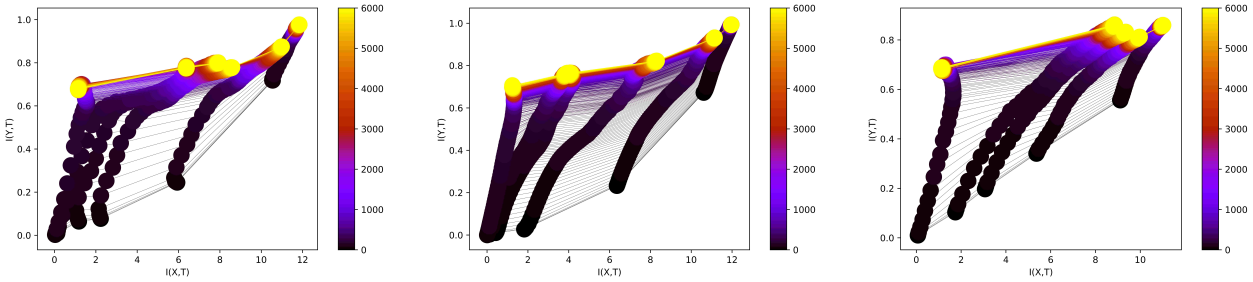
(a) Network Shape - 12,10,8,6,4,2, Default (b) Network Shape - 12,10,4,2,2,2 (c) Network Shape - 12,12,12,12,2

Figure 2: tanh: Demonstrating KDE for different network shapes. Tweaking training size for Tishby's KDE MIE. Hyperparameters: Dataset - Tishby's, activation function - tanh, batch size - 512, training size - 40%.



(a) Training size - 20% (b) Training size - 40% (c) Training size - 70%

Figure 3: ReLu: Demonstrating KDE for different training sizes. Tweaking training size for Tishby's KDE MIE. Hyperparameters: Dataset - Tishby's, activation function - tanh, batch size - 512, network shape 12,10,8,6,4,2.



(a) Network Shape - 12,10,8,6,4,2, Default (b) Network Shape - 12,10,4,2,2,2 (c) Network Shape - 12,12,12,12,2

Figure 4: ReLu: Demonstrating KDE for different network shapes. Tweaking training size for Tishby's KDE MIE. Hyperparameters: Dataset - Tishby's, activation function - ReLu, batch size - 512, training size - 40%.