Lottery Ticket Hypothesis through the lens of RMT

Aman Mehra (2017017)

Aim

- The goal of this work is to study the Lottery Ticket Hypothesis [1] to identify properties that characterize sparse subnetworks which train to accuracies at par with the unpruned network.
- Doing so, the hope is to identify ways to find classes of subnetworks which -
  - Can be found in a **zero-shot** manner. That is do **NOT** require iterative pruning.
  - Are **NOT** restricted to the weight initialization used by the dense unpruned network.

Setup

- **Architecture**: Two classes of networks are studied using Tanh activations:
  - **FCNet**: A **8** layered fully connected network with a uniform hidden dimension of **256**.
  - **ConvNet**: A **8** layered CNN with **128 3x3** convolutions per layer.

- **Dataset:**
  - All models are trained for **50** epochs on the **CIFAR-10** dataset.

Setup

- **Initialization**: Two initializations are studied:
  - **Normal Initialization:** This is the random normal initialization used on both the **FCNet** and **ConvNet**. The weights and biases are sampled iid from a gaussian distribution of **mean 0** and standard deviation $\boldsymbol{\sigma_w}$**(=1.4 /** *sqrt(***layer width***))* and $\boldsymbol{\sigma_b}$**(=0.3 /** *sqrt(***layer width***))* respectively.
    - Moreover, the standard deviation is chosen such that the network lies on the ***edge of chaos*** [2]. A point where the correlation of any two random inputs passes unchanged in expectation throughout the network and avoiding exploding and vanishing gradients [3]
  - **Orthogonal Initialization**: The orthogonal initialization performed for **FCNet** converts the random matrix **W** with weights drawn iid from the standard normal distribution, to an orthogonal matrix ($\mathbf{W^TW=I}$) created using the QR decomposition of W.

Experiments

Overview

- To find characteristic properties the following are studied
  - Layerwise Weight statistics at different levels of sparsity
    - Weight mean
    - Weight standard deviation
  - Normalized squared length of activations (QMap) [2] - Explained on next slide
  - Layerwise Weight distribution
  - Chi value (χ) which characterizes transient dynamics of order and chaos in networks
  - QMap by incorporating true empirical weight distribution into the QMap recurrence relation

Mean Field Theory for Neural Networks

- Given pre-activations **h** and post-activations **x** for an activation function **ϕ**. The following equations describe a **D** layered network and its **Q-Map** (**q**)

$$\mathbf{x}^l = \phi(\mathbf{h}^l) \qquad \mathbf{h}^l = \mathbf{W}^l \mathbf{x}^{l-1} + \mathbf{b}^l \quad \text{for } l = 1, \ldots, D,$$

$$q^l = \frac{1}{N_l} \sum_{i=1}^{N_l} (\mathbf{h}_i^l)^2 \quad = \quad <(\mathbf{h}_i^l)^2>$$

$$q^l = \mathcal{V}(q^{l-1} \,|\, \sigma_w, \sigma_b) \equiv \sigma_w^2 \int \mathcal{D}z \, \phi\left(\sqrt{q^{l-1}} z\right)^2 + \sigma_b^2, \quad \text{for} \quad l = 2, \ldots, D,$$

For an overview on mean field theory and its connection to expressivity and trainability of deep networks checkout my slides

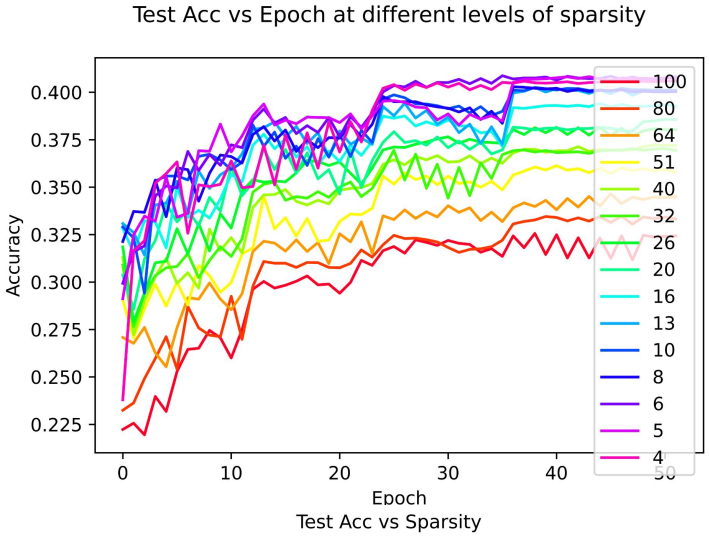
Baselines:
Iterative Magnitude Pruning

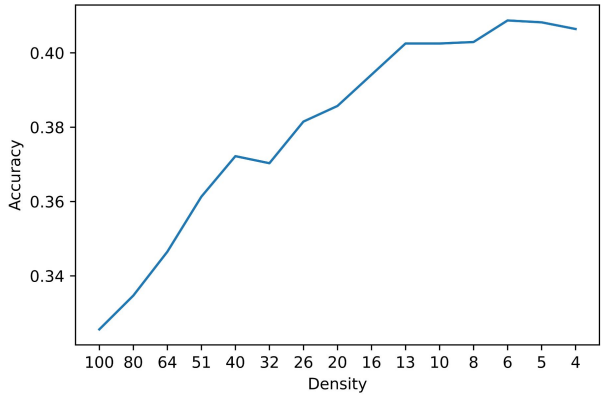
Setup

- Here we establish baselines of accuracy for each architecture by performing iterative magnitude pruning for each initialization and architecture.
- The algorithm is as follows -
  a. Randomly initialize the given network with some parameters θ_0
  b. Train for some number of epochs **j** (50 in our experiments)
  c. Prune $p^{1/n}\%$ of the unpruned weights with the lowest **l1** norm per layer by setting them to **0**
  d. Reset unpruned weights to their value in parameterization θ_0
  e. Repeat steps b-d **n** times.
- For our experiments we set the algorithm hyperparameters to:
  a. **j** = 50
  b. **n** =15
  c. $p^{1/n}$ **=** 10% for last layer and 20% for the rest

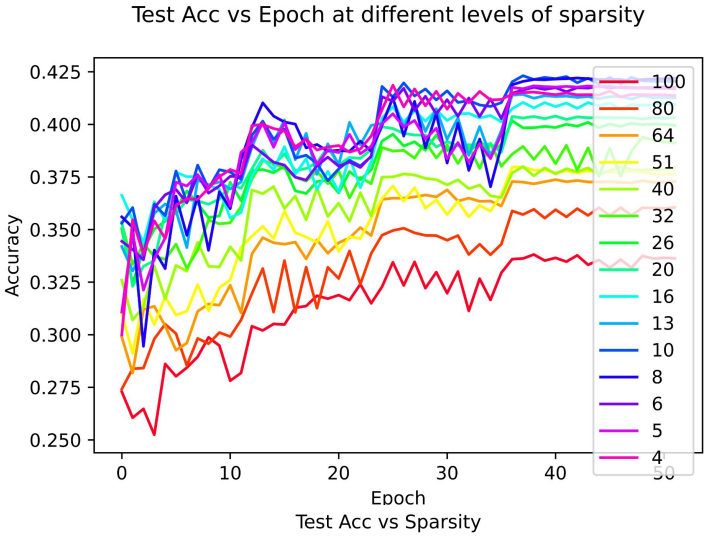
Test Accuracies vs Density (1-sparsity) - FCNet

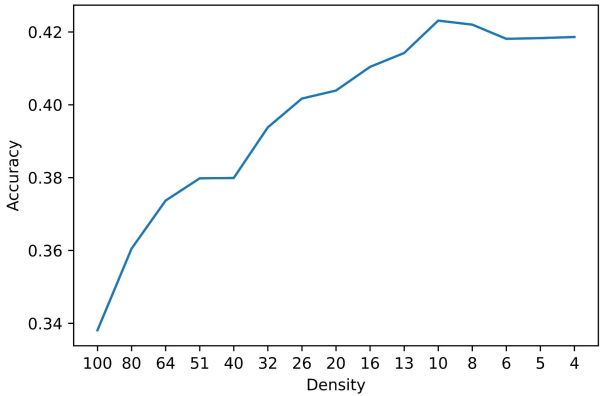
Normal (0, 1.4)



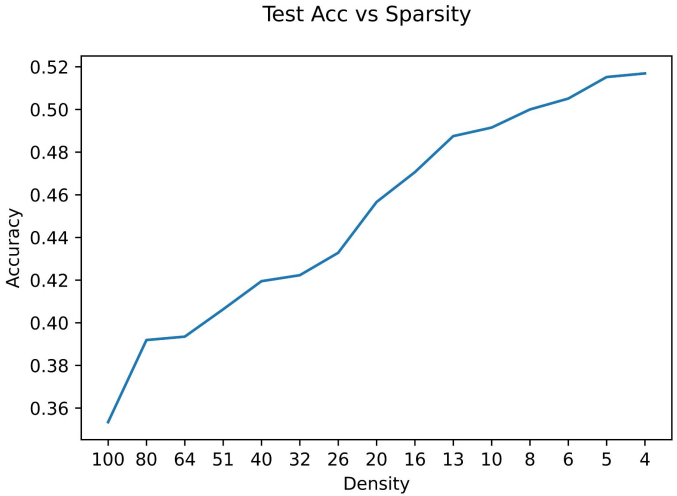


Orthogonal

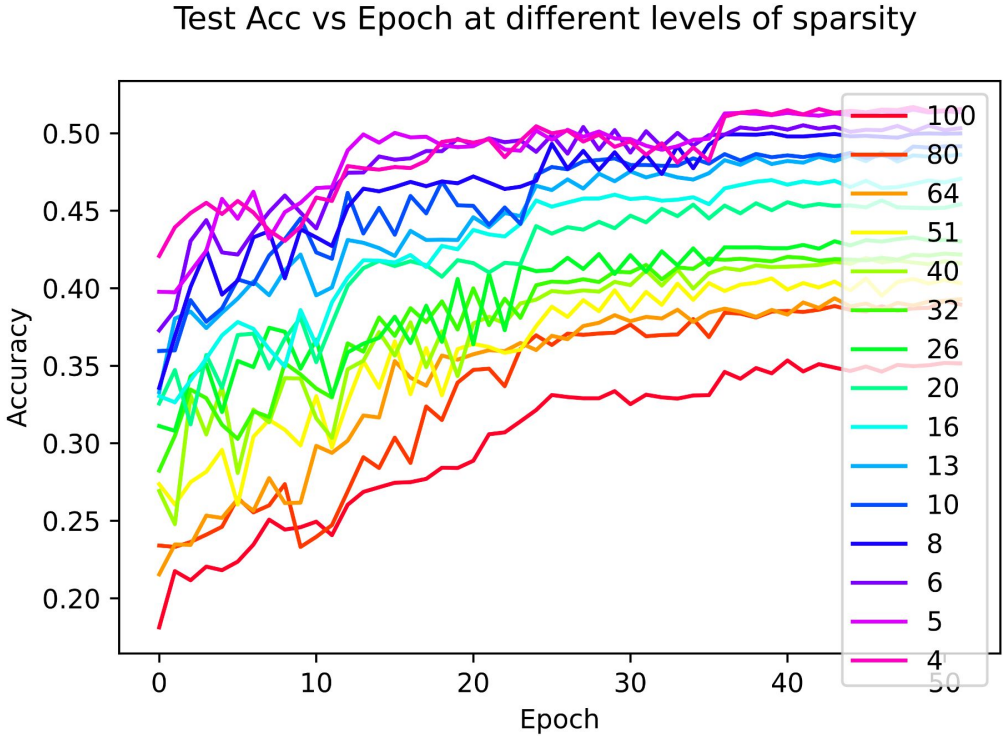




Test Accuracies vs Density - ConvNet



Normal (0, 1.4)



*Note: The plots incorrectly state the labels to represent sparsity. They refer to density (= 1- sparsity)

Implications

- We validate the Lottery Ticket hypothesis for our test setup across architectures and initialization schemes. The hypothesis states -

  The Lottery Ticket Hypothesis. *A randomly-initialized, dense neural network contains a subnetwork that is initialized such that—when trained in isolation—it can match the test accuracy of the original network after training for at most the same number of iterations.*

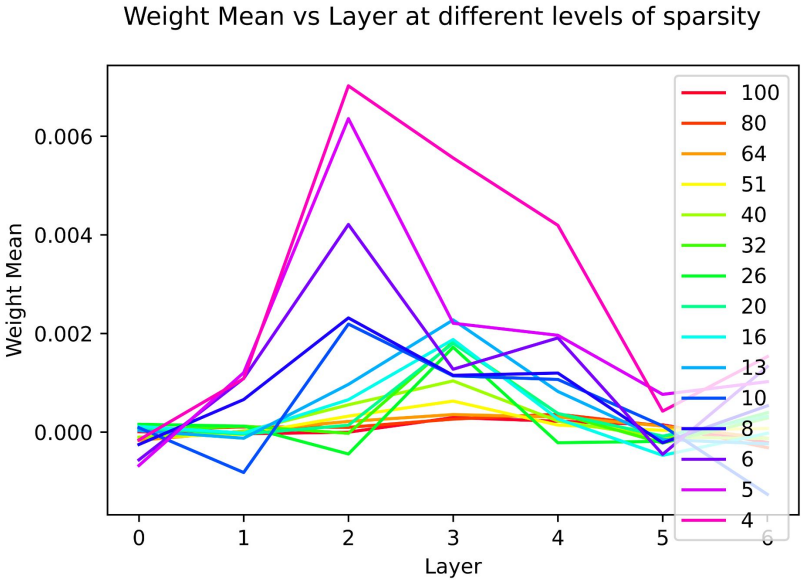
- We in-fact see a rise in accuracy upon pruning.
- With this experiment we set baselines that we would like to match

Layerwise Weight Statistics

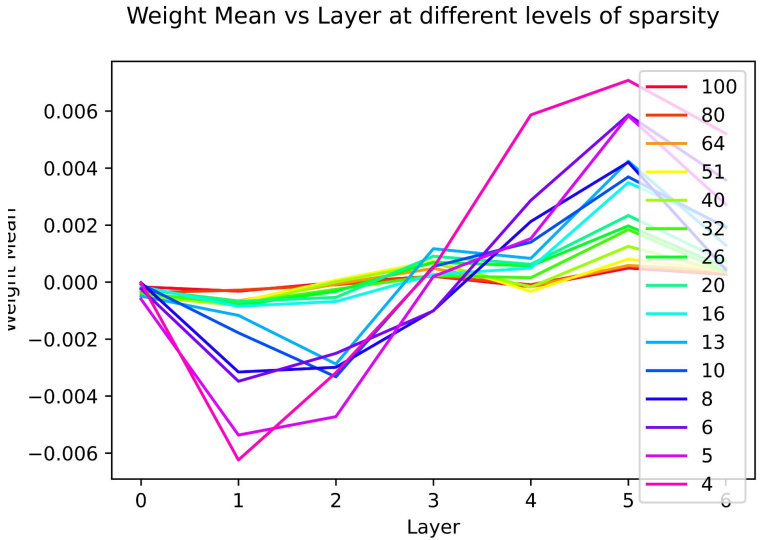
Motivation

- Here we study the weight mean and standard deviation at different sparsity (density) levels to see if any systematic trends emerge when the network is subjected to iterative magnitude pruning.

Weight Mean (FCNet)

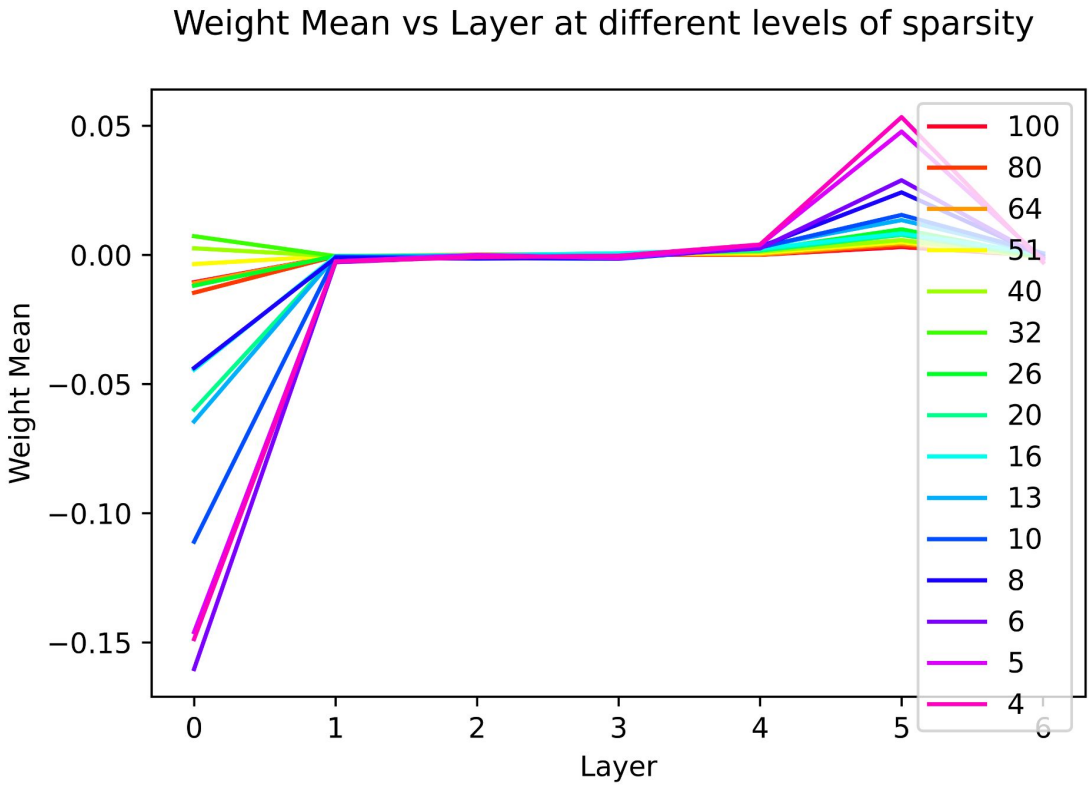


Normal

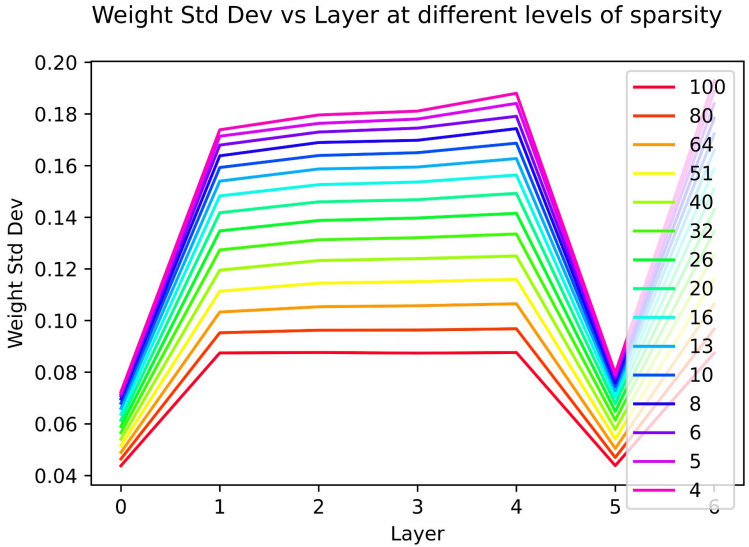


Orthogonal

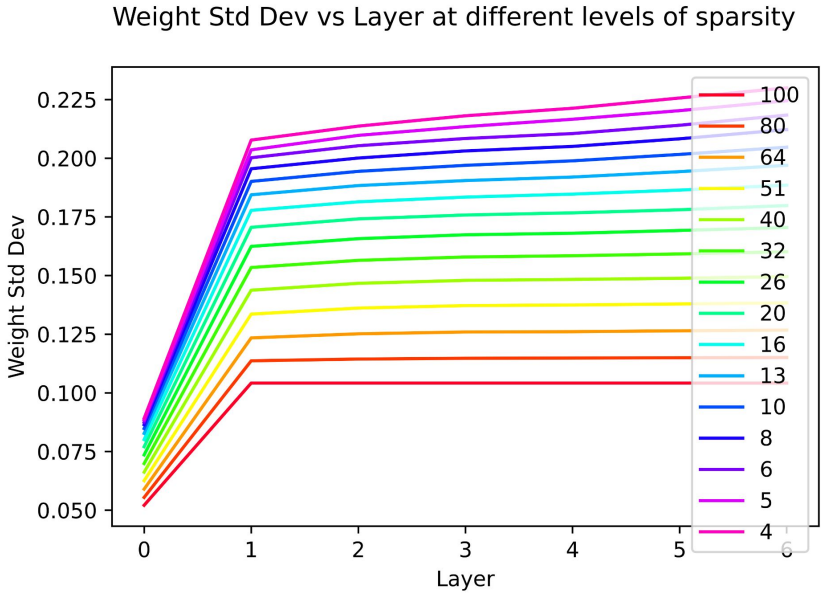
Weight Mean (ConvNet Normal Init)



Weight Std Dev(FCNet)

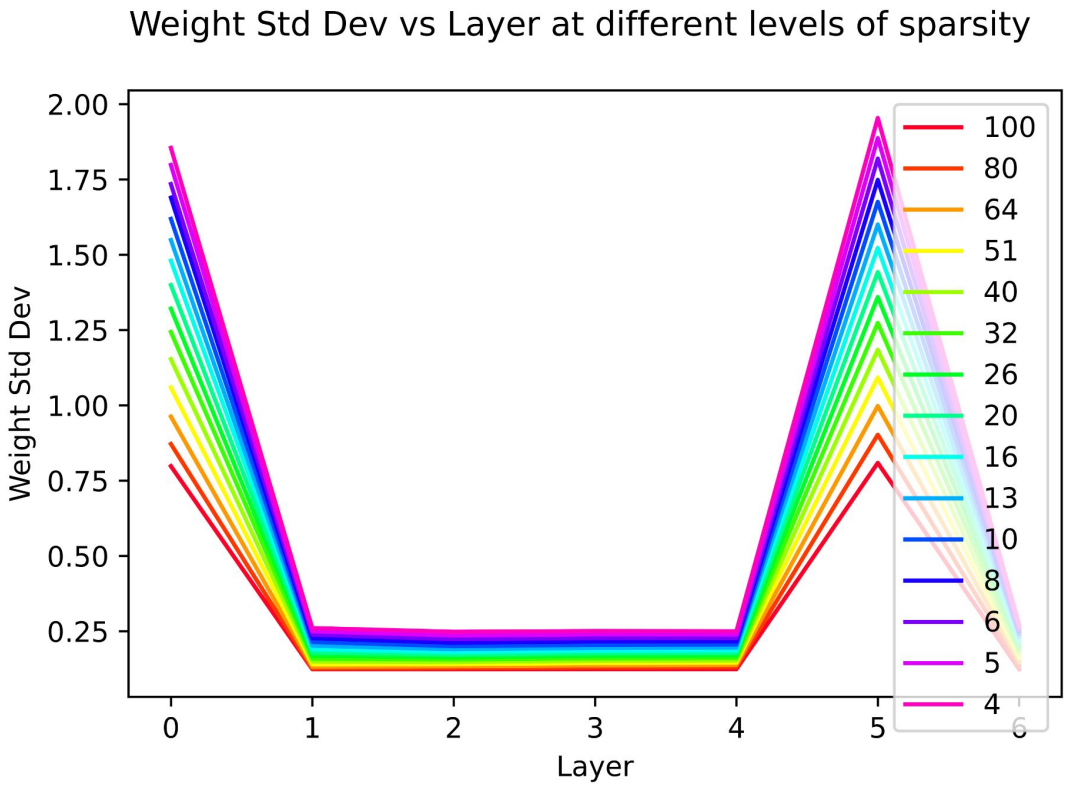


Normal



Orthogonal

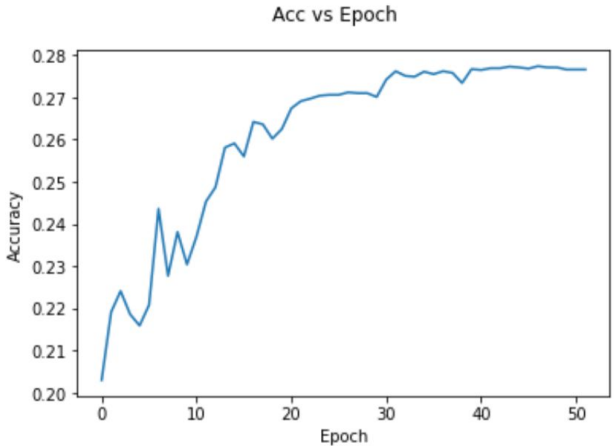
Weight Std Dev (ConvNet Normal Init)



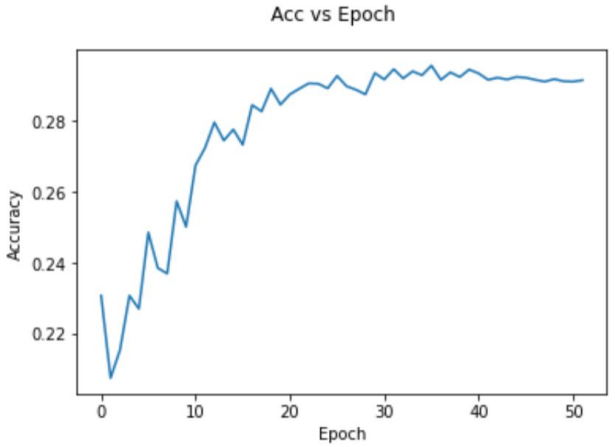
Implications

- Some insights drawn from the above plots:
  - **Weight MEAN:** The mean barely deviates from its original values. The slight variation seen at higher sparsity levels can be attributed to the inaccuracy of the sample mean in estimating the true mean of the distribution due to a lack of samples.
  - **Weight STD DEV:**
    - These plots present an interesting trend -
      - ***Increasing (decreasing) sparsity (density) increases weight std deviation***
    - Moreover, this variations in the std dev. are in the same order of magnitude as its original value. This indicates that this isn't a statistical inaccuracy due to lack of samples.

A Small Experiment (FCNet - Normal - 20% Density)



Std = 1.4



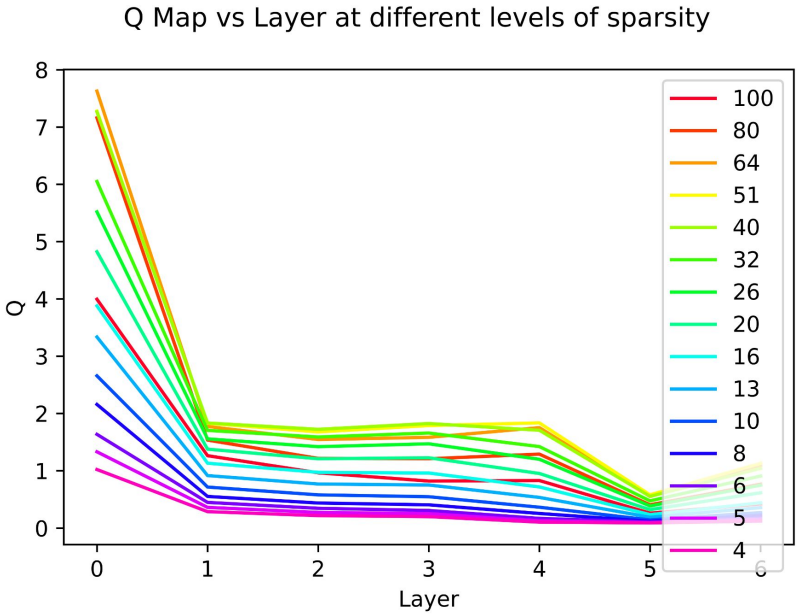
Std = 2.24

Layerwise Q Maps

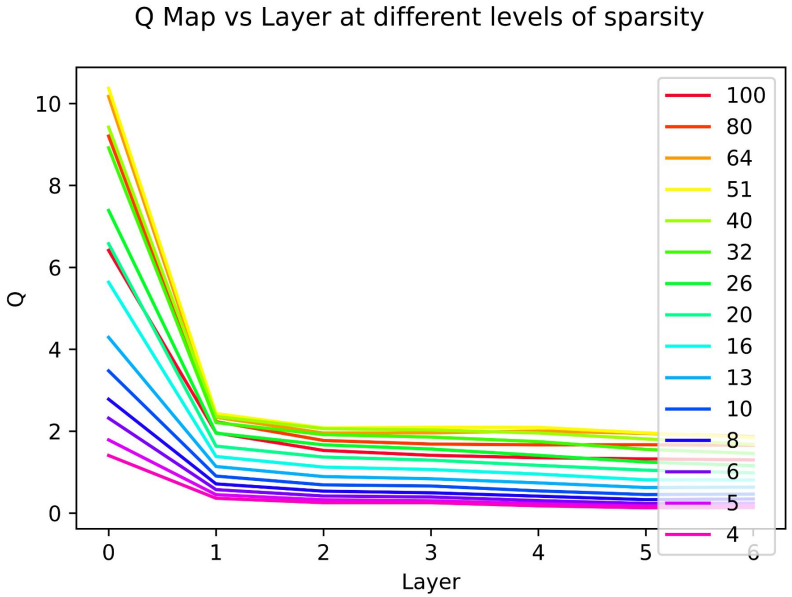
Motivation

- In this section we take use tools from mean field theory to understand signal propagation in random networks as their sparsity changes.
- We study the iterative Q Map and observe how its fixed point value varies across sparsity levels and how this compares with our theoretical observations.

Layerwise Q Maps (FCNet)

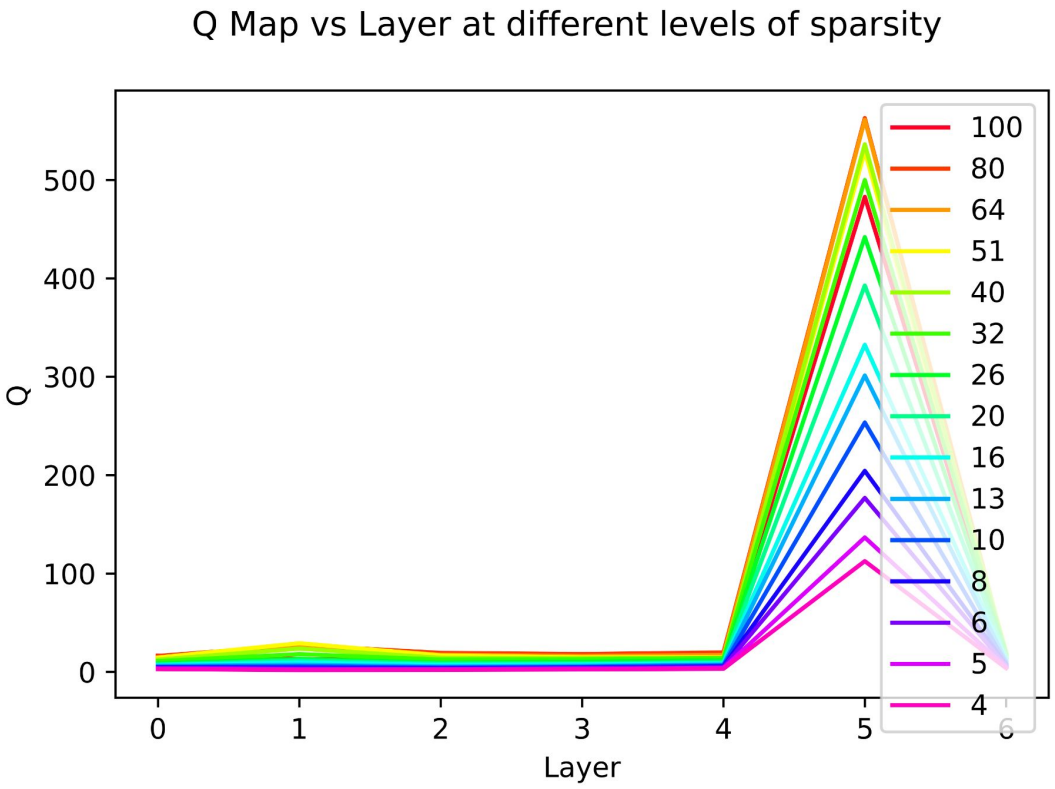


Normal



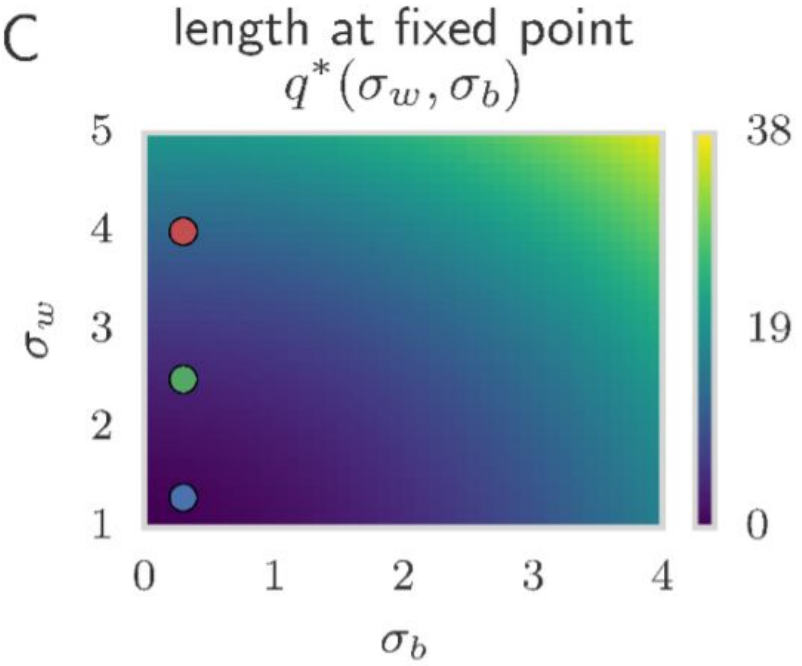
Orthogonal

Layerwise Q Maps (ConvNet Normal Init)



Observation

- It can be observed that there are two trends that emerge from the empirically observed Q-Maps
  a. From density 100% to roughly 51%, the fixed point of the expected normalized square length ($\mathbf{q}^*$) is increasing
  b. Subsequently from 51% to 4% there is a downward trend as ($\mathbf{q}^*$) decreases with increased sparsity
- This is curious as
  a. It had been observed [2] (See fig on next slide) that $\mathbf{q}^*$ increase as the weight and bias std dev increases.
  b. We know that these statistics increase with sparsity from our previous experiment, hence the downward trend below 51% density indicates a discrepancy between theory and practice.



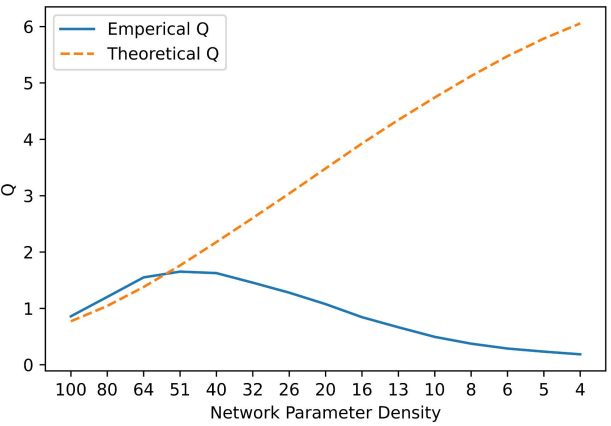
Iterative Q Map Fixed point as a function of weight and Bias Std Dev. (Taken from Poole et al. [2])

Theory vs Empirical Q Maps

FCNet

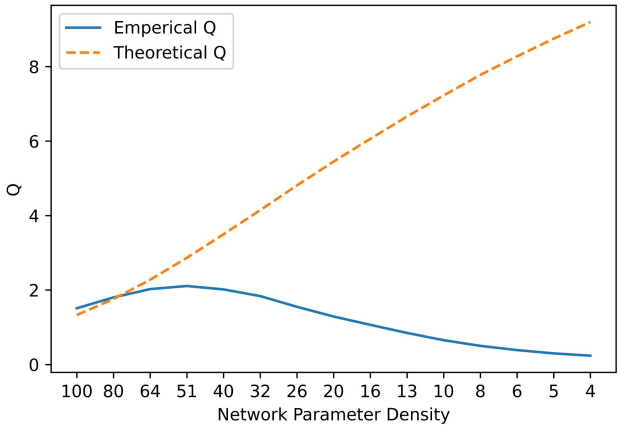
Normal Init

Q Map at different levels of sparsity



Orthogonal Init

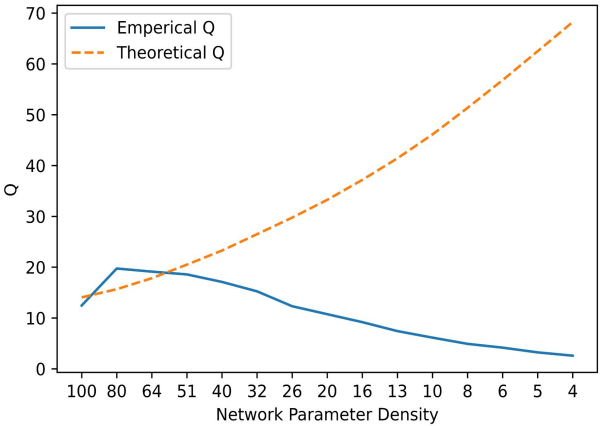
Q Map at different levels of sparsity



ConvNet

Normal Init

Q Map at different levels of sparsity



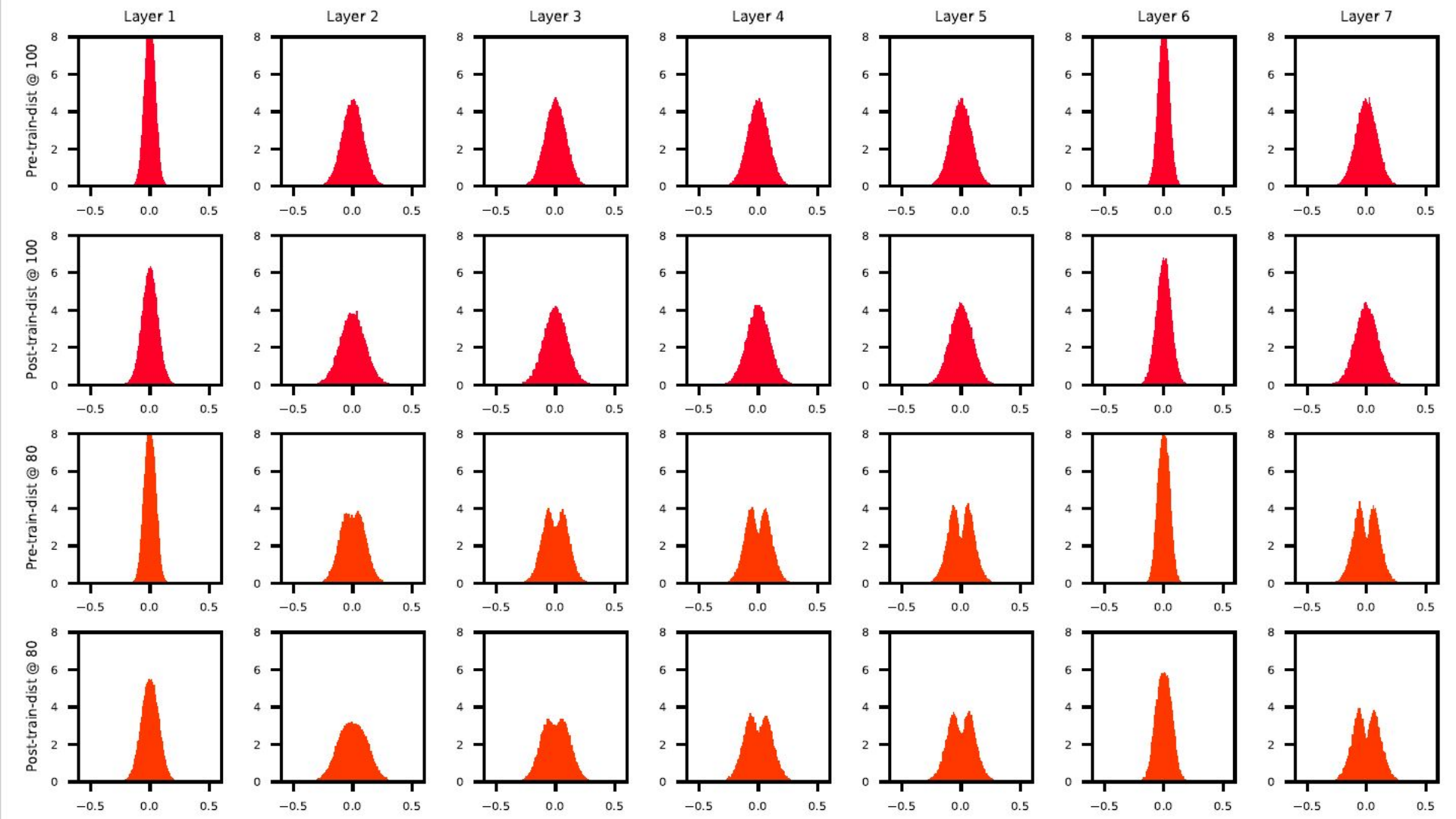
Observation

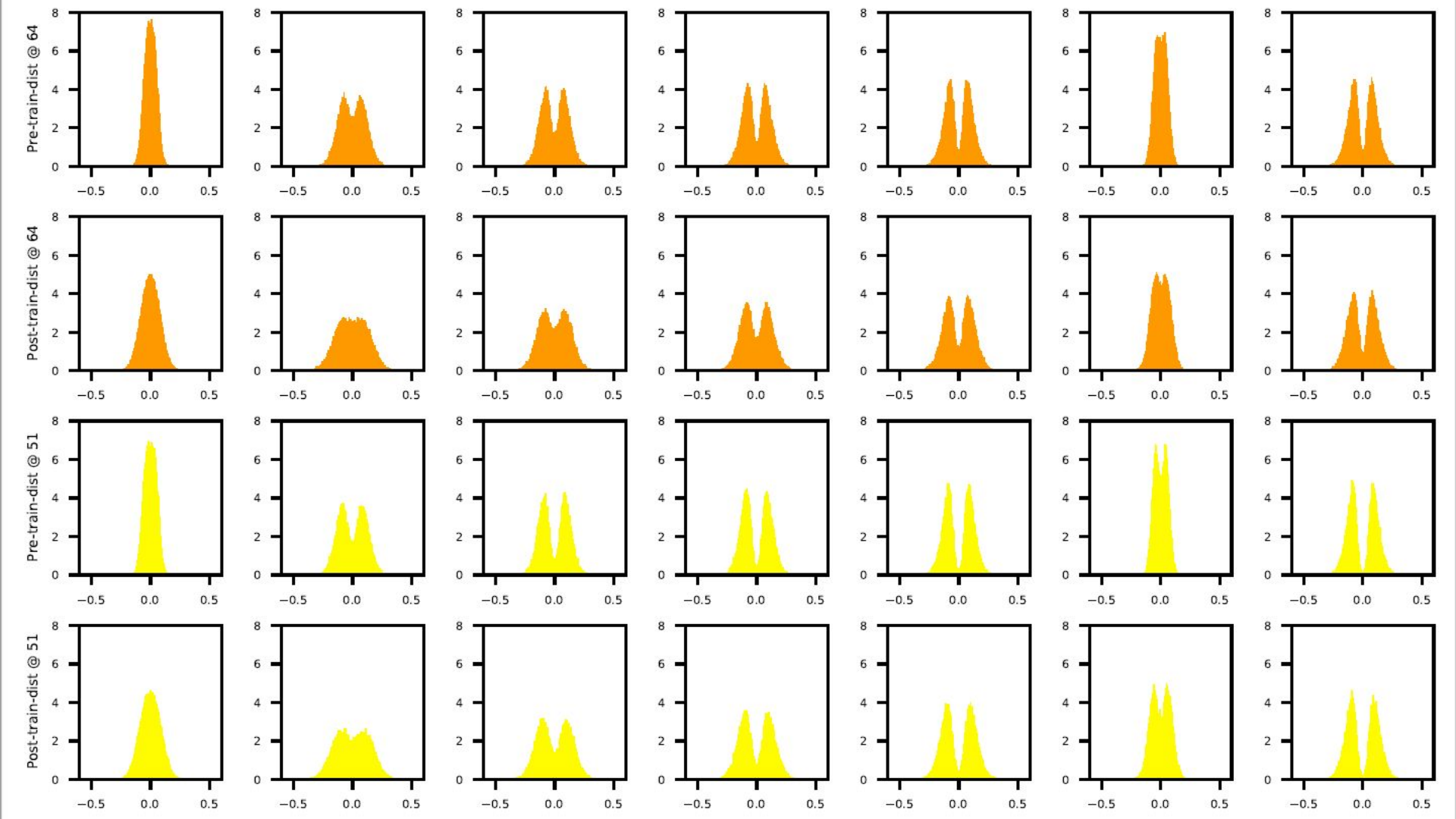
- Now on comparing the theoretically computed iterative Q-Map with the practically observed value we see clear discrepancy beginning at around the **51% density** mark.
- This indicates that one or more assumptions employed while modeling the signal have broken due to the sparsity introduced.
- The first culprit is the gaussian assumption over the distribution of weights which we investigate in the next section

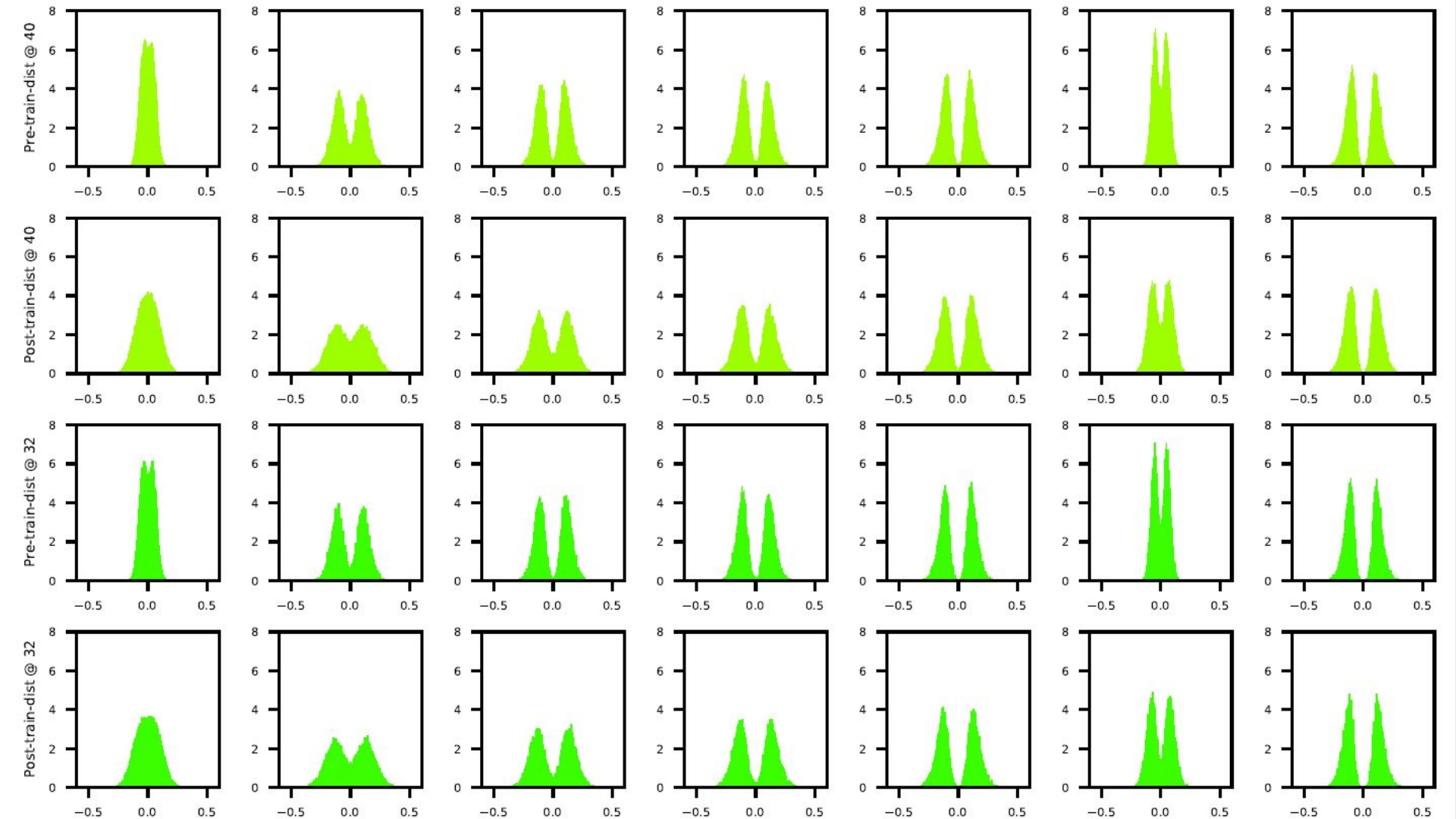
Layerwise Weight Distribution

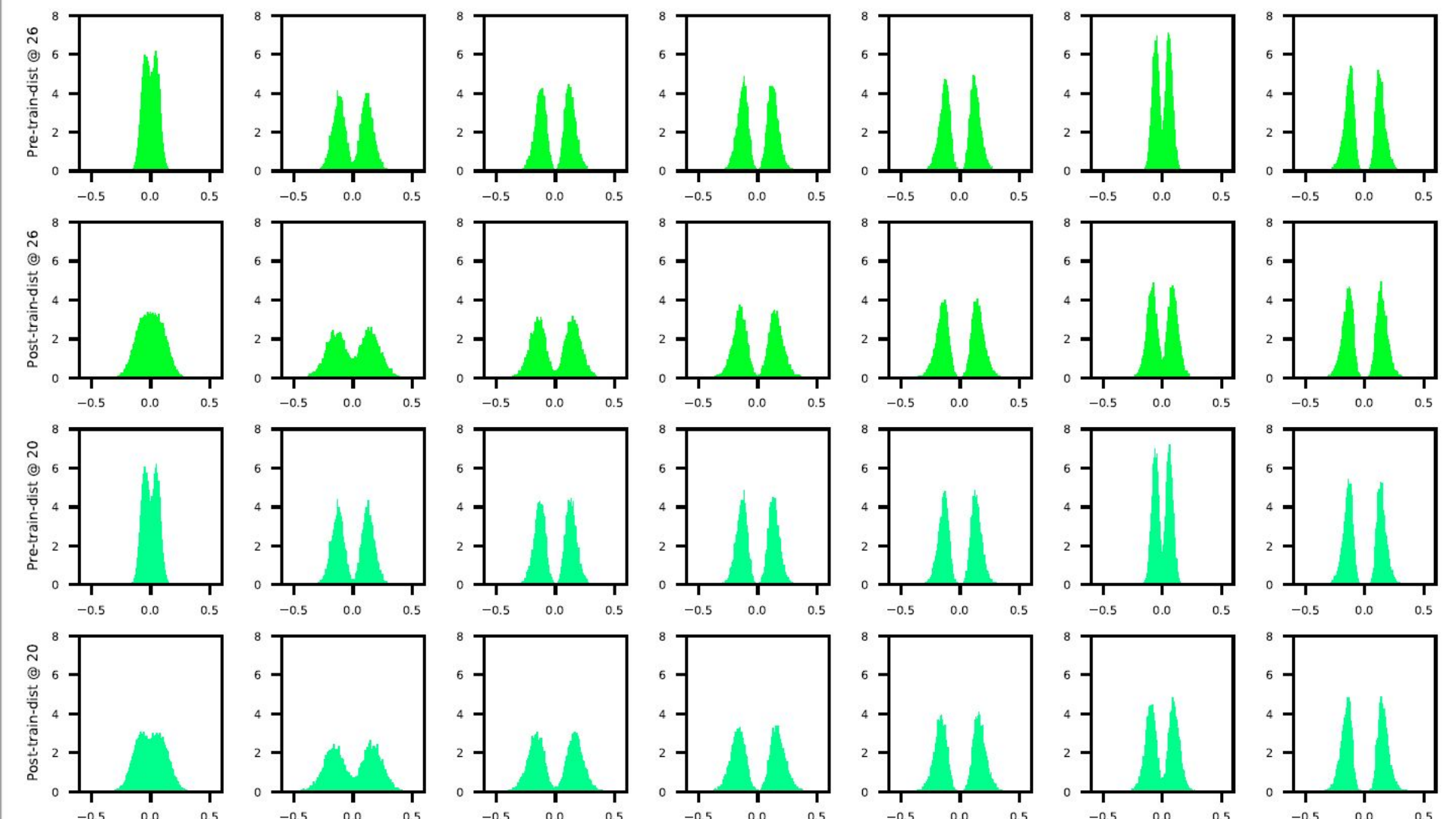
Motivation

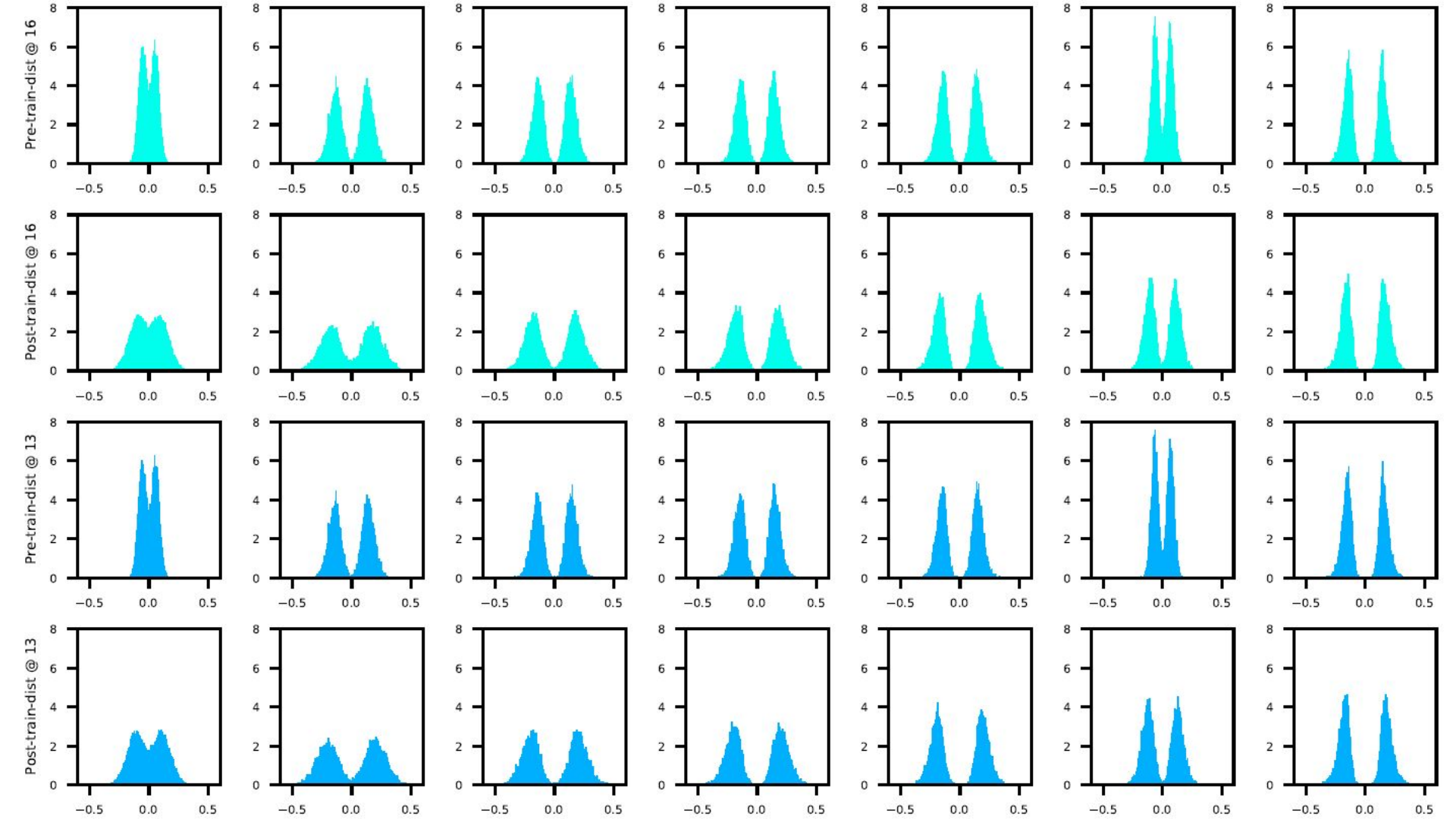
- In the previous section we saw that the theoretically computed Q-Maps did not match the observed theoretical results.
- One of the factors could be the gaussian assumption over the distribution of random weights.
- We investigate this be observing the evolution of the weight distribution over layers as sparsity is increased.
- Two instances are observed -
  a. The distribution **pre-training**
  b. The distribution **post-training**
- The next 5 slides show the distributions of the FCNet with Normal init.


Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6
Layer 7
Pre-train-dist @ 100
Post-train-dist @ 100
Pre-train-dist @ 80
Post-train-dist @ 80
0
2
4
6
8
−0.5
0.0
0.5


Pre-train-dist @ 64
Post-train-dist @ 64
Pre-train-dist @ 51
Post-train-dist @ 51
0
2
4
6
8
−0.5
0.0
0.5


Pre-train-dist @ 40
Post-train-dist @ 40
Pre-train-dist @ 32
Post-train-dist @ 32
0
2
4
6
8
−0.5
0.0
0.5


Pre-train-dist @ 26
Post-train-dist @ 26
Pre-train-dist @ 20
Post-train-dist @ 20
0
2
4
6
8
−0.5
0.0
0.5

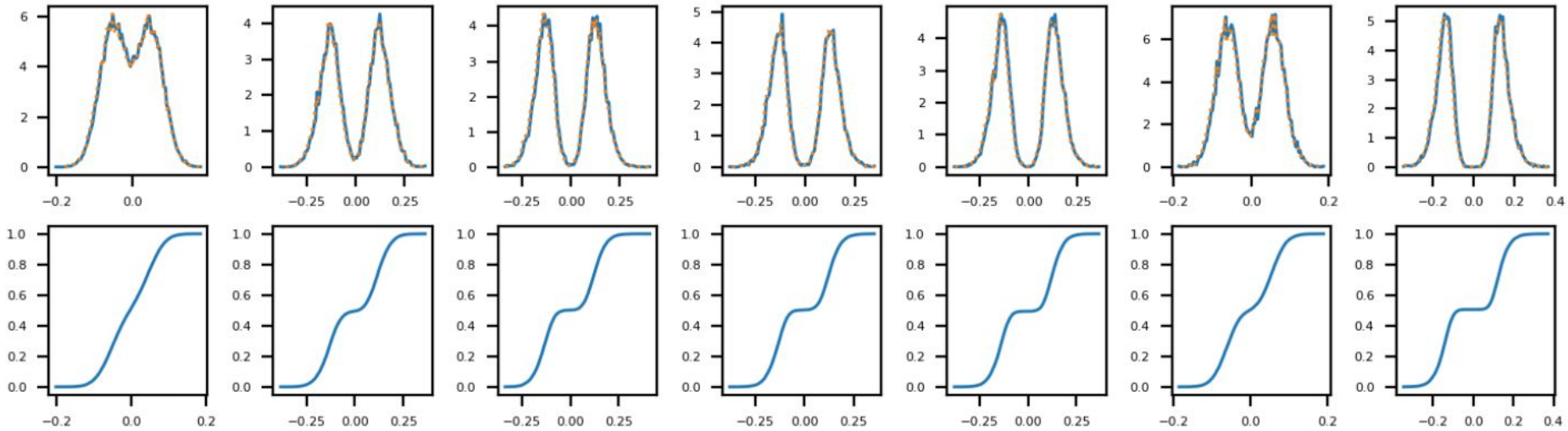

Pre-train-dist @ 16
Post-train-dist @ 16
Pre-train-dist @ 13
Post-train-dist @ 13
0
2
4
6
8
−0.5
0.0
0.5

Observations

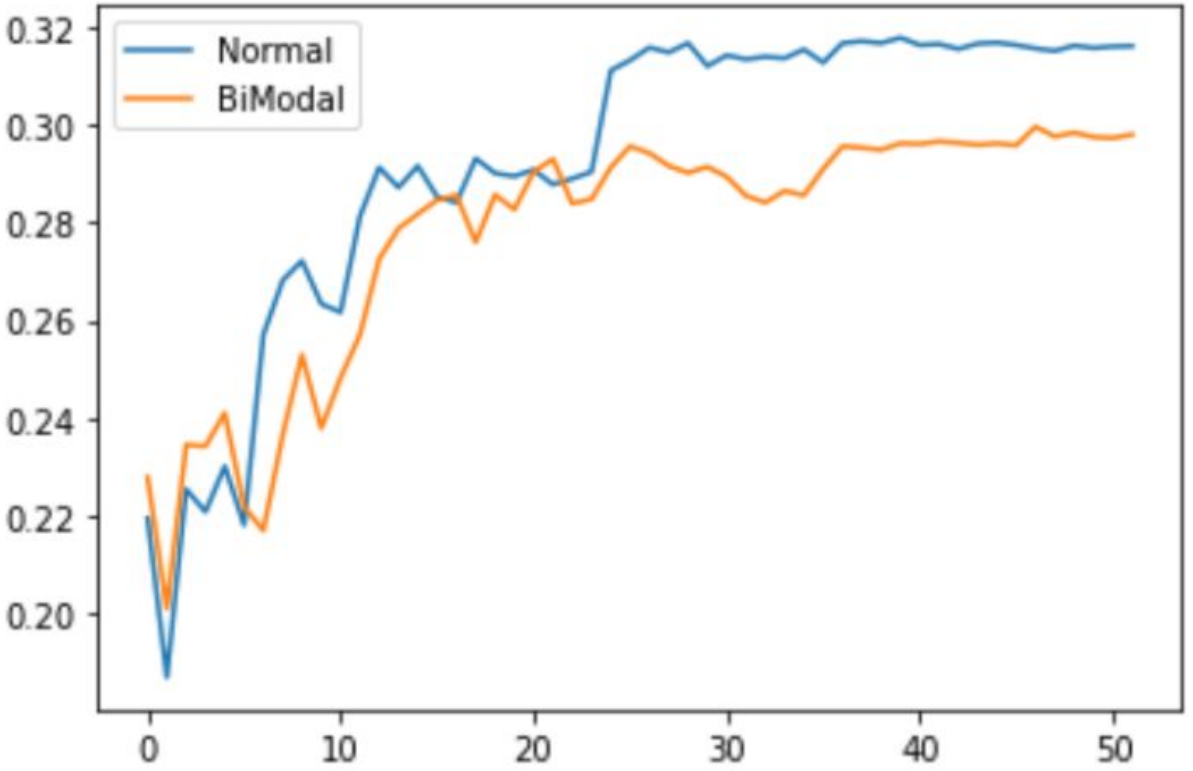
- As we can see from the above plots, as sparsity increases, the weight distribution increasingly shifts from a gaussian to a bi-modal distribution.
- One could then ask the natural question, would initializing a smaller network (#params = # pruned params) with such a bimodal distribution yield in test performance akin to an iteratively prune network?

Bi-Modal Initialization

- To model such an initialization, we capture the weight distribution through histograms at a certain level of density (say 40%) and compute the CDF as seen below. This is then used to compute an inverse CDF and sample weights s



Does accuracy improve on random bi-modal subnets?



Observations

- The answer unfortunately is **NO**

  Accuracy doesn’t increase with random bi-modal initialization.

Chi at Varying Sparsity

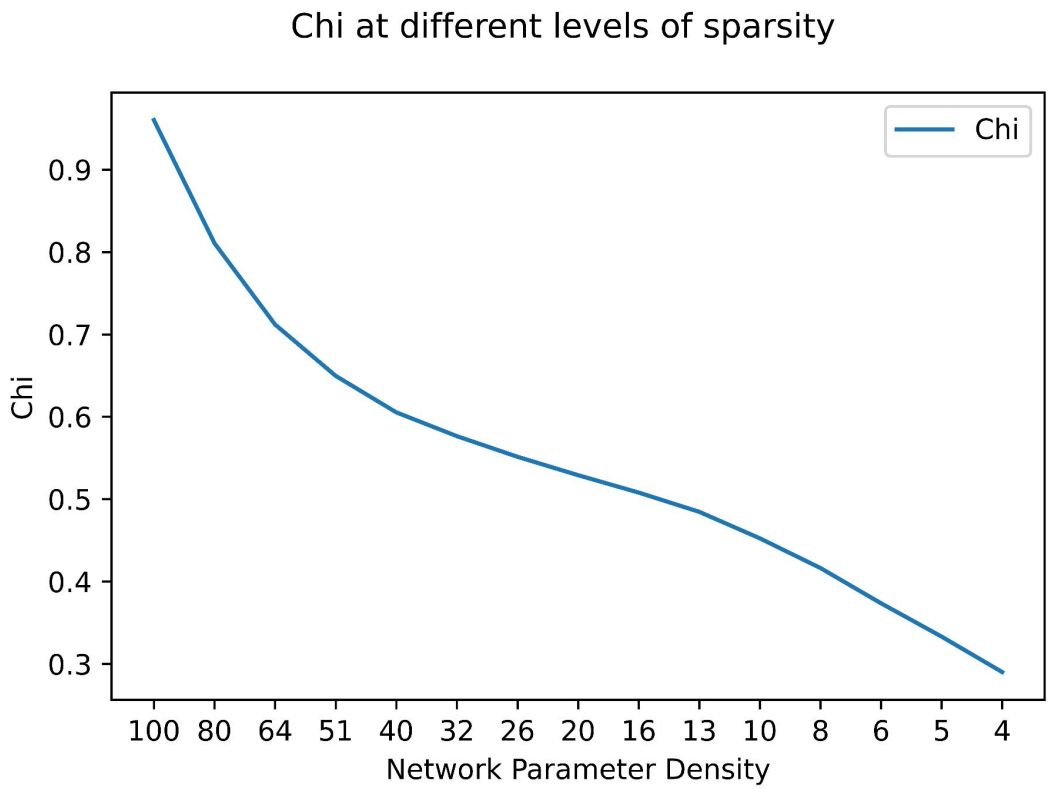
Overview

- The Chi value for a DNN is the derivative of the correlation map (C-Map) [2] at the universal fixed point of the C-Map, **c** = 1.
- The C-Map captures the correlation between two random signal as they propagate through the network.
- **Chi** captures the stability of the fixed point at **c = 1.**
  - If **Chi > 1** , then **c<1** is unstable. Points decorrelate in this regime.
  - If **Chi < 1** , then **c=1** is unstable. Points converge to the same region
  - If **Chi = 1** , the **c** is stable. Correlation between points remains fixed and propagates to exponential depths
- It is ideal to have a weight initialization with **Chi=1** (called edge of chaos)

Motivation

- We study the Chi value for the FCNet under Normal initialization.
- At full density the network is initialized at the edge of chaos (**Chi=1**)
- We then observe how the Chi value changes with sparsity.

Chi Value at Varying Density (1- Sparsity)



Implications

- The results are extremely curious. As sparsity increases Chi decreases, as a result points should decorrelate and converge as they propagate through the network.
- This reduced ability to propagate signal however, doesn't seem to affect accuracy contrary to theory.
- Thus a hypothesis is -
    - *The characterizing feature of a winning subnetwork may lie in its ability to back-propagate gradients rather than forward propagate signals.*
- To verify the above the first line of study will be investigate the spectrum of the input-output Jacobian to study **dynamic isometry** in sparse subnets

Subsequent Lines of Inquiry

- Mean and Std Dev. of in degree of activations at each layer (are they uniformly distributed?) (This could explain discrepancy in Q-Maps)
- Subsequently adjust variance according to the graphs to see if smaller networks without iterative pruning can achieve similar performance.
- Validate that the improved accuracy observed in LTH after pruning is not an attribute of over fitting. To do so test using ResNet.
    - I am working on the recursive Q Map formulation for a ResNet with skip connections across 2 layers.
- Study the spectrum of the Jacobian at different levels of pruning. This will shed light into the backprop dynamics of the signal under sparsity.

References

[1] J. Frankle and M. Carbin. The Lottery Ticket Hypothesis: Finding Sparse Trainable Neural Networks. ICLR 2019

[2] B. Poole and S. Lahiri and M. Raghu and J. Sohl-Dickstein and S. Ganguli. Exponential Expressivity in Deep Neural Networks through Transient Chaos. NIPS 2016.

[3] S. S Schoenholz and J. Gilmer and S. Ganguli and J. Sohl-Dickstein. Deep Information Propagation. ICLR 2017.