An investigation into Symmetric Quasi-Interpolation in deep homogenous networks with weight-decay

Marius Lindegaard
MIT
lindegrd@mit.edu

Abstract

In a recent paper [Papyan et al., 2020] the discovery of *neural collapse* (NC) in deep classifiers has provided insight into an emergent structure of the activations in the last layer of neural networks. Using mse-loss in deep homogenous networks, Rangamani and Banburski-Fahey [2021] gives some theoretical ground for the NC phenomenon, resting on a central assumption of *Symmetric Quasi-Interpolation* (SQI). This work empirically shows that the condition is met in average over the datapoints, but not in the most general form stated in Rangamani and Banburski-Fahey [2021]. A short preliminary analysis of the activations in the second to last hidden layer is also presented, not finding any structure resembling NC nor low dimensionality of activations as observed in NC.

1 Introduction

The use of deep neural networks in machine learning has over the past decade been instrumental in building the state-of-the-art methods for challenges such as image classification, object-detection and many other image processing tasks. In many cases, the models used are often highly over-parametrized, having orders of magnitude higher number of parameters than datapoints used to train them. According to the classical theories of machine learning, this could lead to substantial generalization error as the models are trained to interpolate the training data. Despite the classical theory, overparametrized models trained to interpolation are often able to generalize well [Belkin et al., 2019] when training past the point of interpolating the data achieving no error on predicition.

In Papyan et al. [2020] this terminal phase of training (TPT) was shown to induce a particular structure in the final hidden layer of the neural network coined neural collapse (NC): The in-class variability of activations in the last hidden layer converges to 0 as TPT progresses (NC1), and the class means converge to a simplex equiangluar tight frame forming equiangular and equinorm class-mean-vectors in relation to the global mean (NC2). Furthermore, the weights of the last linear classifier converge to the class means up to rescaling (NC3) and the final classification decision simplifies to a nearest-class-center classification in the final hidden layer activations (NC4). A more thorough mathematical definition is given in sec. 2.1.

This discovery of NC has prompted theoretical investigations into the phenomenon with Rangamani and Banburski-Fahey [2021] providing a theoretical justification for the NC phenomenon given "Symmetric Quasi-interpolation" (SQI) (see sec. 2.2) of the dataset in deep homogenous networks with ReLU activations and weight-decay. The main contribution presented in this submission is to show that the SQI-condition is not met as stated in Rangamani and Banburski-Fahey [2021] in empirical experiments, but is true when averaging over all samples.

Another contribution is a very brief look at the properties of the second-to-last hidden layer observing whether the same NC phenomenon appears in this layer. In short, while there might be some structure to the activations and corresponding weights, the NC phenomenon does not appear in the form described.

Preprint. Under review.

2 Theoretical background

In the experiments and following theory a homogenous neural network, one without bias-terms, for classificaion is used with ReLU activation at each layer except the final output layer. The network has P in the last hidden layer and an output layer of width C, the number of classes. The dataset is assumed to be balanced, having an equal number of examples N of each class c.

2.1 Neural collapse and NC metrics

Closely following the definitions in Han et al. [2021], for sample i of class c we define the activations in the last hidden layer as $h_{i,c} \in \mathbb{R}^P$ after the ReLU activation. The linear classifier from the last hidden layer to the output layer consists of vectors $w_{c'} \in \mathbb{R}^P$ for each class $c' \in 1, ..., C$ so the output activation for class c' for sample i of class c is $\langle h_{i,c}, w_{c'} \rangle$. Then, using the global mean $\mu_G = \frac{1}{NC} \sum_{i,c} h_{i,c}$, we have relative feature class-means $\mu_c = \mu'_c - \mu_G = \frac{1}{N} \sum_i h_{i,c} - \mu_G$.

Neural collapse is then characterized by the following properties of convergence over increasing number of training epochs. We have
(NC1) Within-class veriabillity collapse:

$$\Sigma_W = \frac{1}{NC} \sum_{i,c} (h_{i,c} - \mu_c)(h_{i,c} - \mu_c)^T \to \mathbf{0} \tag{1}$$

(NC2) Convergence to simplex ETF:

$$\frac{\langle \mu_c, \mu_{c'} \rangle}{\|\mu_c\|_2 \|\mu_{c'}\|_2} \to \delta_{c,c'} \frac{C}{C-1} - \frac{1}{C-1} \quad \text{where } \delta_{j,k} \text{ is the Kronecker delta} \tag{2}$$

$$\|\mu_c\|_2 - \|\mu_{c'}\|_2 \to 0 \tag{3}$$

(NC3) Convergence to self-duality:

$$\frac{w_c}{\|w_c\|_2} - \frac{\mu_c}{\|\mu_c\|_2} \to 0 \tag{4}$$

(NC4) Simplification to nearest class center classification:

$$\arg\max_c \langle w_c, h \rangle \to \arg\min_c \|h - \mu_c\|_2 \tag{5}$$

Further intuition for the phenomenon is given in the original paper on NC.[Papyan et al., 2020]

2.2 Symmetric Quasi-interpolation

The assumption given in Rangamani and Banburski-Fahey [2021] from which (**NC1-4**) is derived is

Assumption 1: *Consider a C-class classification problem with inputs in $\mathbb{R}^d$. A classifier $f : \mathbb{R}^d \to \mathbb{R}^C$ symmetrically quasi-interpolates a training dataset $S = \{(x_{i,c}, y_{i,c})\}$ if there exists an ϵ such that for all examples we have*

$$f^{(c')}(x_{i,c}) = \begin{cases} 1 - \epsilon & c = c' \\ \frac{\epsilon}{C-1} & c \neq c' \end{cases} \tag{6}$$

where $f^{(c')}(x_{i,c}) \in \mathbb{R}$ is the output of the network in component c' for sample input $x_{i,c}$.

For analysis the values

$$\begin{aligned} \epsilon_{i,c} &= 1 - f^{(c)}(x_{i,c}) \\ \tilde{\epsilon}_{i,c,c'} &= f^{(c')}(x_{i,c}) \quad \text{where } c \neq c' \end{aligned} \tag{7}$$

are used, where if **Assumption 1** is true we expect there to be some ϵ such that $\epsilon_{i,c} = \epsilon \ \forall i, c$ and $(C-1)\tilde{\epsilon}_{i,c,c'} = \epsilon \ \forall i, c, c'$.

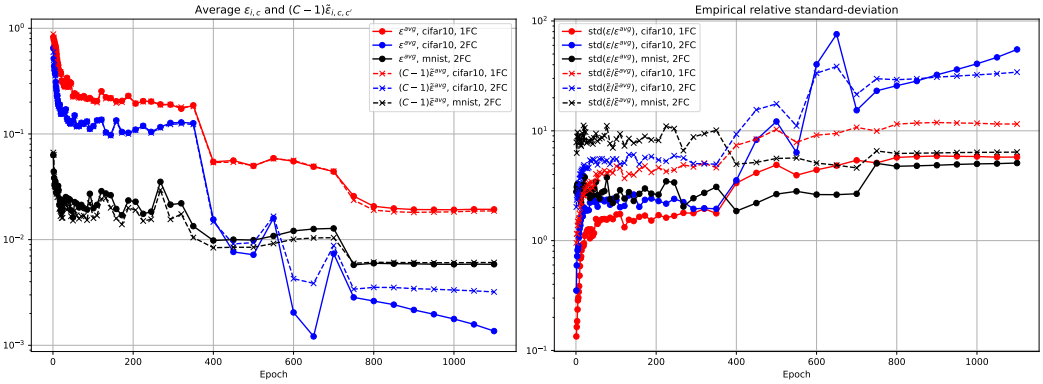


Figure 1: Error metrics for the SQI conditions. The terminal phase of training, when the model has effectively 0 prediction error, starts at around 400 epochs for *cifar10* and very early for *mnist*.

3 Experiments and results

3.1 Methodology

The models used in the experiment are all homogenous convolutional neural networks with ReLU activation functions on all hidden layers. The final layers referred to as fully connected (FC) layers are single-pixed convolutional layers of a high number of channels, functionally equivalent to FC layers. The model used referred to as `1FC` consists of 5 convolutional layers, the last one being a FC equivalent. The model used referred to as `2FC` consists of 6 convolutional layers, the last 2 being a FC equivalent. The datasets used are the `cifar10` and `mnist` 10-class classification tasks. Training is done, importantly, with weight-decay. An implementation may be found on Github [1], with further configurations and training details.

3.2 Evaluation of Symmetric Quasi-interpolation

In order to evaluate the adherence to the SQI condition the mean and variance of $\epsilon_{i,c}$ and $\bar{\epsilon}_{i,c,c'}$ as given in eq. 7. Specifically, define

$$\epsilon^{avg} = \frac{1}{NC}\sum_{i,c}\epsilon_{i,c} \quad \text{and} \quad \tilde{\epsilon}^{avg} = \frac{1}{NC(C-1)}\sum_{\substack{i,c,c'\\ c\neq c'}}\tilde{\epsilon}_{i,c,c'}. \tag{8}$$

From figure 1 it can be observed that the condition of **Assumption 1** is satisfied *on average*. For each of the different training sessions $\epsilon^{avg} \simeq (C-1)\tilde{\epsilon}^{avg}$ for most of the run in the terminal phase of training, and most seem to converge to a constant value. An exception is the final epochs of training on `cifar10` with the `2FC` network. Interestingly, **Assumption 1** seems to hold true on average through all stages of training, even before TPT.

On the other hand, the relative standard deviation of the $\epsilon_{i,c}$ and $\tilde{\epsilon}_{i,c,c'}$,

$$\text{std}\left(\frac{\epsilon}{\epsilon^{avg}}\right) = \frac{1}{\epsilon^{avg}}\sqrt{\frac{1}{NC}\sum_{i,c}\left(\epsilon_{i,c}-\epsilon^{avg}\right)^2} \tag{9}$$

$$\text{std}\left(\frac{\tilde{\epsilon}}{\tilde{\epsilon}^{avg}}\right) = \frac{1}{\tilde{\epsilon}^{avg}}\sqrt{\frac{1}{NC(C-1)}\sum_{\substack{i,c,c'\\ c\neq c'}}\left(\tilde{\epsilon}_{i,c,c'}-\tilde{\epsilon}^{avg}\right)^2} \tag{10}$$

is significantly higher than 1. Figure 1 indicates that after convergence when ϵ^{avg} and $\tilde{\epsilon}^{avg}$ is constant, the cross-sample variance does not converge to 0. This breaks with **Assumption 1**.

An inspection of a few of the samples from `cifar10, 2FC` in table 1 confirms the high variance and interestingly even has datapoints with $\epsilon_{i,c}$ and $\tilde{\epsilon}_{i,c,c'}$ smaller than 0.

[1] If allowed by the authors of the codebase on which this project is built.

$\epsilon_{i,c} \cdot 10^3$	$\tilde{\epsilon}_{i,c,c'} \cdot 10^3, \quad c \neq c'$								
80	2.1	7.2	6.6	-2.7	18	4.2	31	-1.1	-1.6
53	-2.7	15	5.8	11	6.9	-0.89	-5.0	12	-6.9
-55	-5.3	-7.8	-5.5	-4.8	-5.6	4.7	-8.6	-9.2	-3.1

Table 1: $\epsilon_{i,c}$ and $\tilde{\epsilon}_{i,c,c'}$ of a few datapoints in the *CIFAR10* dataset evaluated on the fully trained model *2FC*.

This leads to a restatement of **Assumption 1** in terms of the average:

Assumption 2: *Consider a C-class classification problem with inputs in $\mathbb{R}^d$. A classifier $f : \mathbb{R}^d \to \mathbb{R}^C$ is said to on average quasi-interpolate a training dataset $S = \{(x_{i,c}, y_{i,c})\}$ if there exists an ϵ such that*

$$\epsilon = \frac{1}{NC} \sum_{i,c} 1 - f^{(c)}(x_{i,c}) \quad \text{and} \quad \epsilon = \frac{1}{NC} \sum_{\substack{i,c,c' \\ c \neq c'}} f^{(c')}(x_{i,c}) \tag{11}$$

where $f^{(c')}(x_{i,c}) \in \mathbb{R}$ is the output of the network in component c' for sample input $x_{i,c}$.

Notably, the variance also indicates that **NC1-3** does not occur exactly as the training converges to a solution. If they did, we would expect the convergence to a class mean in the last hidden layer and the class-means obeying eq. 2 and 3 as well as eq. 4 to impose a symmetry on the last layer activations in terms of equal within-class activations and equal out-of-class activations, essentially no variance in $\epsilon_{i,c}$ and in $\tilde{\epsilon}_{i,c,c'}$. This symmetry is broken, as shown by the non-zero and, towards the end of TPT, non-decreasing variance.

3.3 Experimental observations in second-to-last layer

3.3.1 Evaluating neural collapse

In the `2FC` models there are two hidden layers fully-connected to the next one before the output layer, allowing for an analysis of whether NC happens in the second-to-last hidden layer.

The experiments conclusively show that NC does not occur in the second to last hidden layer, even as NC is progresses in the last hidden layer. The results are presented in figure 2.

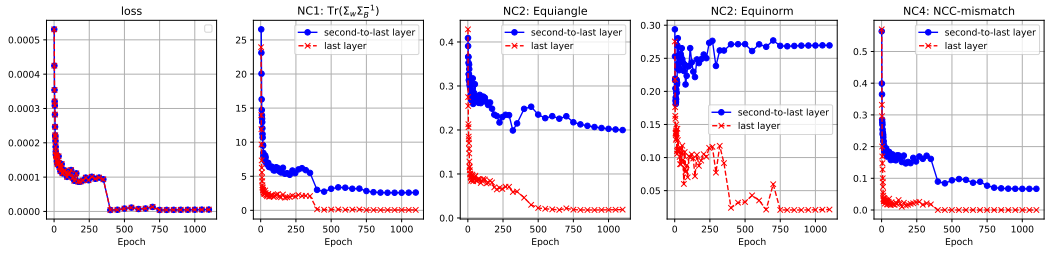
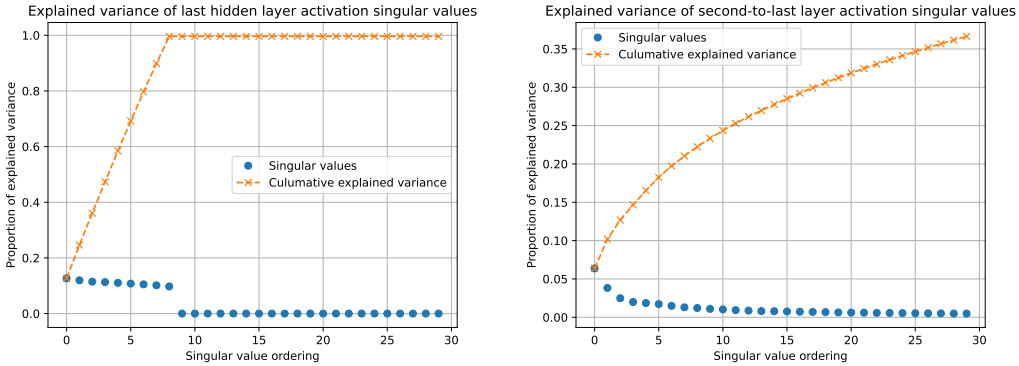


Figure 2: Lack of neural collapse in the second-to-last layer on the *CIFAR10* dataset with the *2FC* model. After near-interpolation around epoch 400 the loss goes to 0 and we continue TPT. For NC1, Σ_w is as defined in 1 and $\Sigma_B = \frac{1}{C}\sum_c (\mu_c - \mu_G)(\mu_c - \mu_G)^T$ the between-class covariance. For NC2 Equiangle, the average over $c \neq c'$ of (eq. 2)+$\frac{1}{C-1}$ is used. For NC2 Equinorm the relative standard deviation of $\|\mu_c\|_2$ corresponding to eq. 3 is used. Finally, the NC4-condition the misclassification error when using nearest-class-center classification in the respective layer is calculated.

Notably, there is some structure to the activations, especially the NCC-mismatch. This structure of classification being approximated (with some error) by nearest-class-center classification is not too unexpected even disregarding neural collapse. The feature engineering done in the previous parts of the neural network should amount to some problem simplification, and while the nature of the problem simplification is not obvious, observing this particular structure does not necessarily imply any NC structure.



(a) Last hidden layer

(b) Second-to-last hidden layer

Figure 3: Singular values of the activations of the 50000 datapoints in the *cifar10* dataset for the respective FC layers in the same *2FC* network at the end of training. The singular values are normalized in order for the total explained variance by all 256 or 1024 (respectively) singular values to total 1.

3.3.2 Singular values and explained variance

As a key component of neural collapse is the convergence of hidden layer activations to a $C-1$ dimensional equiangular tight frame (by eq. 2 and 3), an investigation into the dimensionality of the data in the second-to-last layer was performed. The results of investigating the singular values of the training data activation matrix is shown in figure 3.

The singular values of the last layer, exhibiting neural collapse, are extremely regular and with a sharp cutoff after $C-1=9$ singular values of approximately equal size explaining practically all the data variance. The second-to-last hidden layer activations, however, do not have anything resembling the same regularity. While the first singular value is of course the largest, there are no observable irregularities in the proportion of total variance explained. A key point is that the cumulative explained variance of fig. 3b converges to 1 relatively slowly, indicating little to no bias towards low dimensionality of activations in this layer.

4 Conclusions

Summarizing the results there is not observed a convergence to **Assumption 1** of Rangamani and Banburski-Fahey [2021] as stated in the original paper. A restatement of their assumption is presented, **Assumption 2**, which empirically does hold true. This second assumption, together with a bound on the variance of output activation errors, could lead to a bound on deviation from the NC conditions. This extended theoretical analysis is left as future work.

Interestingly, the analysis also indicates that **NC1-3** are not all fullfilled excactly as training progresses. This might hint at a problem with the optimization in the final stages of TPT. This result could possibly be an artifact of optimization parameters, and remedies such as increasing weight-decay, changing step-size or introducing more or less stochasticity might lead to a different result. This is also left as future work.

Lastly, the empirical investigations into the properties of activations in the second to last hidden layer shows no indication of neural collapse and no indication of low data-dimensionality. While the activations in this layer may be characterized by some regularity, more work is required in order to determine what those regularities are.

Acknowledgments

Thank you to the staff and lecturers of MIT subject 9.520 Fall 2021 for both the course inspiring this project and their patience. An additional thanks to Akshay Rangamani and Andrzej Benburski-Fahey for providing a base from which to further develop both theory and empirical experiments.

References

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *CoRR*, abs/2008.08186, 2020. URL `https://arxiv.org/abs/2008.08186`.

Akshay Rangamani and Andrzej Banburski-Fahey. Neural collapse in deep homogenous classifiers and the role of weight decay. *preprint*, 2021.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning practice and the bias-variance trade-off, 2019.

X. Y. Han, Vardan Papyan, and David L. Donoho. Neural collapse under MSE loss: Proximity to and dynamics on the central path. *CoRR*, abs/2106.02073, 2021. URL `https://arxiv.org/abs/2106.02073`.