NEURAL COLLAPSE IN DEEP HOMOGENEOUS CLASSIFIERS AND THE ROLE OF WEIGHT DECAY

Akshay Rangamani, Andrzej Banburski-Fahey

Center for Brains, Minds, and Machines, MIT

ABSTRACT

Neural Collapse is a phenomenon recently discovered in deep classifiers where the last layer activations collapse onto their class means, while the means and last layer weights take on the structure of dual equiangular tight frames. In this paper we present results showing the role of weight decay in the emergence of Neural Collapse in deep homogeneous networks. We show that certain near-interpolating minima of deep networks satisfy the Neural Collapse condition, and this can be derived from the gradient flow on the regularized square loss. We also show that weight decay is necessary for neural collapse to occur. We support our theoretical analysis with experiments that confirm our results.

Index Terms— Deep Learning, Inductive Bias, Neural Collapse, Weight Decay, Classification

1. INTRODUCTION

In a recent paper Papyan, Han and Donoho [1] described four empirical properties of the terminal phase of training (TPT) in deep networks, using the cross-entropy loss function. TPT begins at the epoch where training error first vanishes. During TPT, the training error stays effectively zero, while training loss is pushed toward zero. Direct empirical measurements expose an inductive bias they call Neural Collapse (NC), involving four interconnected phenomena. (NC1) Cross-example within-class variability of last-layer training activations collapses to zero, as the individual activations themselves collapse to their class means. This means that the classification margins of all data in the training set converge to the same value. (NC2) The class means collapse to the vertices of a simplex equiangular tight frame (ETF). (NC3) Up to rescaling, the last-layer classifiers collapse to the class means or in other words, to the simplex ETF (i.e., to a self-dual configuration). (NC4) For a given activation, the classifier's decision collapses to simply choosing whichever class has the closest train class mean (i.e., the nearest class center [NCC] decision rule). Together, these empirical properties suggest a dramatic simplification of the complex training dynamics during this terminal phase.

The natural question is whether it is possible to derive Neural Collapse and characterize under which conditions we should expect it to appear. In this article we study the dynamics of learning for homogeneous deep networks (ones without the bias term) trained with square loss. We show that, in this setting, weight decay is necessary for Neural Collapse to occur and that solutions with weight decay do not interpolate the data exactly. Our main result is that for global minima of the loss that are close to interpolation (in a specific sense described below), Neural Collapse emerges. We then go on to numerically test our predictions and we show that while it is possible to achieve same training loss with and without weight decay while training a deep network with square loss, NC properties only emerge when weight decay is present.

1.1. Related Work

There has been much recent work on the analysis of deep networks and linear models trained using exponential-type losses for classification [2, 3, 4, 5, 6, 7, 8, 9, 10]. Recent interest in using the square loss for classification has been spurred by the experiments in [11], though the practice of using the square loss is much older [12]. Since the empirical discovery of Neural Collapse [1], there have been a few papers attempting to derive its emergence. In [13] it was shown that in the case of unbalanced datasets, the inductive bias of NC transforms to that of Minority Collapse, in which minority classes collapse to the same predictor and become indistinguishable. [14, 15, 16] show NC in the regime of "unconstrained features". Other papers have shown the emergence of NC when using the cross entropy loss [17, 13, 18]. While preparing this paper, we became aware of recent results by Ergen and Pilanci [19] (see also [20, 21]) who derived neural collapse for the square loss, through a convex dual formulation of deep networks. Our independent derivation is different and uses simple properties of the dynamics of learning.

2. THEORETICAL ANALYSIS

2.1. Deep Homogeneous Networks

In this paper we consider the problem of supervised deep learning for multiclass classification. We have a training dataset $\mathcal{S} = \{(\boldsymbol{x}_{n(c)}, y_{n(c)}\}_{n=1,c=1}^{N,C}$. Our dataset comes from C classes and is balanced, meaning it contains N training examples per class. We use $\boldsymbol{x}_{n(c)} \in \mathbb{R}^d$ to denote the features

of the n^{th} training example from the c^{th} class, and $y_{n(c)} \in \{1, \dots, C\}$ to denote its label. We will also use $\boldsymbol{y}_{n(c)} \in \mathbb{R}^C$ to denote the one-hot vector corresponding to the class label $y_{n(c)}$. We use deep homogeneous neural networks to perform our classification task. Deep homogeneous networks are denoted as $f_{\boldsymbol{W}} : \mathbb{R}^d \to \mathbb{R}^C$

$$f_{\boldsymbol{W}}(\boldsymbol{x}) = \boldsymbol{W}_L \sigma(\boldsymbol{W}_{L-1} \dots \boldsymbol{W}_2 \sigma(\boldsymbol{W}_1 \boldsymbol{x}) \dots)$$

where $\sigma(z) = \max(z, 0)$ is the ReLU nonlinearity, and $\boldsymbol{W}_l$ is the weight matrix in layer l of the deep network. We write the C-dimensional output of our network as $f_{\boldsymbol{W}}(\boldsymbol{x}) = [f^{(1)}_{\boldsymbol{W}}(\boldsymbol{x}) \dots f^{(C)}_{\boldsymbol{W}}(\boldsymbol{x})]^\top$. We will also use $\boldsymbol{h}(\boldsymbol{x}) \in \mathbb{R}^p$ to denote the last layer features of the classifier, ie, $f_{\boldsymbol{W}}(\boldsymbol{x}) = \boldsymbol{W}_L \boldsymbol{h}(\boldsymbol{x})$. The networks are trained by minimizing the regularized square loss:

$$\begin{aligned} \mathcal{L}(\boldsymbol{W}) = \frac{1}{2NC} \sum_{n=1,c=1}^{N,C} ||f_{\boldsymbol{W}}(\boldsymbol{x}_{n(c)}) - \boldsymbol{y}_{n(c)}||_2^2 \\ + \frac{\lambda}{2} \sum_{k=1}^{L} ||\boldsymbol{W}_k||_F^2 \end{aligned} \tag{1}$$

We analyze the training dynamics of the deep homogeneous networks under gradient flow (a limit of gradient descent when the learning rate is infinitesimally small): $\frac{\partial \boldsymbol{W}}{\partial t} = -\nabla_{\boldsymbol{W}} \mathcal{L}$

2.2. Neural Collapse

Neural Collapse is a phenomenon that emerges at the terminal phase of training deep networks that is characterized by 4 conditions. The precise formulation for the conditions are given below:

(NC1) Variability collapse: $\boldsymbol{h}(\boldsymbol{x}_{n(c)}) = \boldsymbol{\mu}_c$

(NC2) Equinorm, Equiangularity of the class means:

$$||\boldsymbol{\mu}_c - \boldsymbol{\mu}_G||_2 = ||\boldsymbol{\mu}_{c'} - \boldsymbol{\mu}_G||_2, \forall c, c'$$
$$\langle \boldsymbol{\mu}_c, \boldsymbol{\mu}_{c'} \rangle \propto \frac{C}{C-1} \delta_{c,c'} - \frac{1}{C}$$

(NC3) Self Duality: $\frac{\boldsymbol{W}_L^\top}{||\boldsymbol{W}_L||_F} = \frac{\dot{\boldsymbol{M}}}{||\dot{\boldsymbol{M}}||_F}$, where $\dot{\boldsymbol{M}} \in \mathbb{R}^{p \times C}$ is a matrix whose columns are $\boldsymbol{\mu}_c - \boldsymbol{\mu}_G$

(NC4) Nearest Class Center Classification:

$$\text{argmax}_c \langle \boldsymbol{W}_L^c, \boldsymbol{h}(\boldsymbol{x}) \rangle = \text{argmin}_c ||\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{\mu}_c||_2$$

Under these conditions the class means and the classifier are said to exhibit a Simplex Equiangular Tight Frame (ETF) structure, which means $\boldsymbol{W}_L \propto \boldsymbol{K}^*$ (up to a rotation), where $\boldsymbol{K}^* = \sqrt{\frac{C}{C-1}} \left(\boldsymbol{I}_C - \frac{1}{C} \boldsymbol{1}\boldsymbol{1}^\top \right)$ is the canonical simplex ETF.

2.3. Interpolation and Weight Decay

First, we will present a simple result that shows that for any homogeneous model trained using the unregularized square loss, solutions that exhibit Neural Collapse do not interpolate.

Lemma 2.1. *Consider a homogeneous model $f(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{h}(\boldsymbol{x})$ on a multiclass classification problem trained using the unregularized square loss. The solutions that exhibit Neural Collapse are not interpolating solutions, and hence are not global minima of the unregularized square loss.*

Proof. Let $\boldsymbol{H} \in \mathbb{R}^{p \times NC}$ be a matrix of features of a homogeneous model evaluated on a training dataset $\mathcal{S}$. Under condition **(NC1)**, we know that $\boldsymbol{H}$ can be factorized as $\boldsymbol{H} = \boldsymbol{M}\boldsymbol{Y}$ where $\boldsymbol{M} \in \mathbb{R}^{p \times C}$ is a matrix whose columns are the class means (let us assume the global mean $\boldsymbol{\mu}_G = 0$), and the matrix $\boldsymbol{Y} \in \mathbb{R}^{C \times NC}$ collects the one-hot vectors corresponding to the training labels. Under conditions **(NC2),(NC3)** we have that $\boldsymbol{W}\boldsymbol{M} = \frac{\alpha C}{C-1} \left(\boldsymbol{I} - \frac{1}{C} \boldsymbol{1}\boldsymbol{1}^\top \right)$, where $\alpha > 0$ is a proportionality constant. If we consider the MSE loss function, any solution that exhibits Neural Collapse will have:

$$\begin{aligned} \mathcal{L}(\boldsymbol{W}) &= \frac{1}{2NC} ||\boldsymbol{W}\boldsymbol{H} - \boldsymbol{Y}||_F^2 \\ &= \frac{1}{2NC} ||\boldsymbol{W}\boldsymbol{M}\boldsymbol{Y} - \boldsymbol{Y}||_F^2 \\ &= \frac{1}{2NC} \sum_{i=1}^{NC} || \frac{\alpha C}{C-1} (\boldsymbol{e}_{y_i} - \frac{1}{C} \boldsymbol{1}) - \boldsymbol{e}_{y_i} ||_2^2 \\ &= \frac{1}{2} \left((1-\alpha)^2 + \frac{\alpha^2}{C-1} \right) \end{aligned}$$

This shows that the Neural Collapse solution cannot interpolate the training dataset. □

The previous result suggests that regularization such as weight decay may be necessary to find solutions that exhibit Neural Collapse. Adding weight decay to our square loss function results in global minima that do not interpolate. We show this in the following result.

Lemma 2.2. *The global minima of the regularized squared loss of a deep homogeneous ReLU network do not interpolate the data.*

Proof. We consider the continuous dynamics of training a deep homogeneous ReLU network using gradient flow (for more details about gradient flow see [21]). Using the loss in equation (1), we have $\frac{\partial \boldsymbol{W}_l}{\partial t} = -\nabla_{\boldsymbol{W}_l} \mathcal{L}$, as well as $\frac{\partial ||\boldsymbol{W}_l||_F^2}{\partial t} = -2 \langle \boldsymbol{W}_l, \nabla_{\boldsymbol{W}_l} \mathcal{L} \rangle$. Let us define $\eta_{n(c)} = \langle \boldsymbol{y}_{n(c)}, f_{\boldsymbol{W}}(\boldsymbol{x}_{n(c)}) \rangle$. This means:

$$\begin{aligned} -\frac{\partial ||\boldsymbol{W}_l||^2}{\partial t} &= \frac{2}{NC} \sum_{n=1,c=1}^{N,C} ||f_{\boldsymbol{W}}(\boldsymbol{x}_{n(c)})||_2^2 - \eta_{n(c)} \\ &\quad + 2\lambda ||\boldsymbol{W}_l||_F^2 \\ \implies \mathcal{L}(\boldsymbol{W}) &= -\frac{1}{4} \frac{\partial ||\boldsymbol{W}_l||^2}{\partial t} + \frac{\lambda}{2} \sum_{k=1,k \neq l}^{L} ||\boldsymbol{W}_k||_F^2 \\ &\quad + \frac{1}{2NC} \sum_{n=1,c=1}^{N,C} ||\boldsymbol{y}_{n(c)}||_2^2 - \eta_{n(c)} \end{aligned}$$

Now at a critical point of the regularized loss that is interpolating we simultaneously have $\eta_{n(c)} = ||\boldsymbol{y}_{n(c)}||_2^2 = 1$ and $\frac{\partial ||\boldsymbol{W}_l||_F^2}{\partial t} = 0$, which means $\mathcal{L} = \frac{\lambda}{2}\sum_{k=1,k\neq l}^{L} ||\boldsymbol{W}_k||_F^2$. This is a contradiction since it is smaller than $\frac{\lambda}{2}\sum_{k=1}^{L} ||\boldsymbol{W}_k||_F^2$, which is the value of the loss at an interpolating solution if evaluated directly. Hence we cannot have a global minimum of a deep homogeneous ReLU network that also interpolates. □

2.4. Neural Collapse under Weight Decay and Symmetric Quasi-Interpolation

The results in the previous section tell us that for the unregularized problem, solutions that exhibit Neural Collapse do not interpolate the training dataset. Moreover, the minima for the regularized square loss minimization problem also do not interpolate the training dataset. Having established that weight decay seems to be necessary to finding solutions with Neural Collapse, we now show that it is sufficient, under one key assumption. We assume that the error of the solutions found during training is evenly distributed among all the components of the classifier.

Assumption 1 (Symmetric Quasi-interpolation)**.** *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 $S = \{(x_n, y_n)\}$ if for all training examples $x_{n(c)}$ in class c, $f^{(c)}(x_{n(c)}) = 1 - \epsilon$, and $f^{(c')}(x_{n(c)}) = \frac{\epsilon}{C-1}$.*

Under this assumption of symmetric quasi-interpolation, we can show that deep homogeneous ReLU networks exhibit all four conditions of Neural Collapse.

Theorem 2.3. *For a ReLU deep network trained on a balanced dataset using Gradient Flow on the square loss with Weight Decay λ, critical points of Gradient Flow that satisfy Assumption 1 also satisfy the* **(NC1-4)** *conditions for Neural Collapse.*

Proof. Our training objective is

$$\mathcal{L} = \frac{1}{2NC}\sum_{n,c,i=1}^{N,C,C}\left(\boldsymbol{y}_{n(c)}^{(i)} - f_{\boldsymbol{W}}^{(i)}(\boldsymbol{x}_{n(c)})\right)^2 + \frac{\lambda}{2}\sum_k ||W_k||_F^2$$

We use gradient flow to train the network: $\frac{\partial \boldsymbol{W}}{\partial t} = -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}}$. Let us analyze the dynamics of the last layer, considering each classifier vector $\boldsymbol{W}_L^c$ of $\boldsymbol{W}_L$ separately:

$$\begin{aligned}\frac{\partial \boldsymbol{W}_L^c}{\partial t} &= \frac{\sum_{n,c'}(\boldsymbol{y}_{n(c')}^c - \langle \boldsymbol{W}_L^c, \boldsymbol{h}(\boldsymbol{x}_{n(c')})\rangle)\boldsymbol{h}(\boldsymbol{x}_{n(c')})}{NC} - \lambda \boldsymbol{W}_L^c \\ &= \frac{1}{NC}\sum_n (1 - \langle \boldsymbol{W}_L^c, \boldsymbol{h}(\boldsymbol{x}_{n(c)})\rangle)\boldsymbol{h}(\boldsymbol{x}_{n(c)}) \\ &\quad + \frac{1}{NC}\sum_{n,c'\neq c}(-\langle \boldsymbol{W}_L^c, \boldsymbol{h}(\boldsymbol{x}_{n(c')})\rangle)\boldsymbol{h}(\boldsymbol{x}_{n(c')}) - \lambda \boldsymbol{W}_L^c\end{aligned}$$

Let us consider solutions that achieve *symmetric quasi-interpolation*, with $f_{\boldsymbol{W}}^{(c)}(\boldsymbol{x}_{n(c)}) = 1 - \epsilon$, and $f_{\boldsymbol{W}}^{(c')}(\boldsymbol{x}_{n(c)}) = \frac{\epsilon}{C-1}$. It is fairly straightforward to see that since $f_{\boldsymbol{W}}^{(c)}$ and $\boldsymbol{W}_L^c$ do not depend on n, neither does $\boldsymbol{h}(\boldsymbol{x}_{n(c)})$, which shows NC1. Under the conditions of NC1 we know that all feature vectors in a class collapse to the class mean, i.e., $\boldsymbol{h}(\boldsymbol{x}_{n(c)}) = \boldsymbol{\mu}_c$. Let us denote the global feature mean by $\boldsymbol{\mu}_G = \frac{1}{C}\sum_c \boldsymbol{\mu}_c$. This means we have:

$$\frac{\partial \boldsymbol{W}_L^c}{\partial t} = 0 \quad \implies \boldsymbol{W}_L^c = \frac{\epsilon}{\lambda(C-1)} \times (\boldsymbol{\mu}_c - \boldsymbol{\mu}_G) \quad (2)$$

This implies that the last layer parameters $\boldsymbol{W}_L$ are a scaled version of the centered class-wise feature matrix $\boldsymbol{M} = [\ldots \boldsymbol{\mu}_c - \boldsymbol{\mu}_G \ldots]$. Thus at equilibrium, with quasi interpolation of the training labels, we obtain $\frac{\boldsymbol{W}_L}{||\boldsymbol{W}_L||_F} = \frac{\boldsymbol{M}^\top}{||\boldsymbol{M}||_F}$. This is the condition for NC3.

From the gradient flow equations, we can also see that at equilibrium, with quasi interpolation, all classifier vectors in the last layer ($\boldsymbol{W}_L^c$, and hence $\boldsymbol{\mu}_c - \boldsymbol{\mu}_G$) have the same norm:

$$\begin{aligned}&\left\langle \boldsymbol{W}_L^c, \frac{\partial \boldsymbol{W}_L^c}{\partial t}\right\rangle = -\lambda ||\boldsymbol{W}_L^c||_2^2 \\ &+ \frac{1}{NC}\sum_{n,c'}(\boldsymbol{y}_{n(c')}^c - f_{\boldsymbol{W}}^{(c)}(\boldsymbol{x}_{n(c')}))f_{\boldsymbol{W}}^{(c)}(\boldsymbol{x}_{n(c')}) \quad (3) \\ &\frac{\partial \boldsymbol{W}_L^c}{\partial t} = 0 \implies ||\boldsymbol{W}_L^c||_2^2 = \frac{1}{C\lambda}\left(\epsilon - \frac{C}{C-1}\epsilon^2\right)\end{aligned}$$

From the quasi-interpolation of the correct class label we have that $\langle \boldsymbol{W}_L^c, \boldsymbol{\mu}_c\rangle = 1-\epsilon$ which means $\langle \boldsymbol{W}_L^c, \boldsymbol{\mu}_G\rangle + \langle \boldsymbol{W}_L^c, \boldsymbol{\mu}_c - \boldsymbol{\mu}_G\rangle = 1 - \epsilon$. Now using (2)

$$\begin{aligned}\langle \boldsymbol{W}_L^c, \boldsymbol{\mu}_G\rangle &= 1 - \epsilon - \frac{\lambda(C-1)}{\epsilon}||\boldsymbol{W}_L^c||^2 \\ &= 1 - \epsilon - \frac{\lambda(C-1)}{\epsilon}\frac{1}{C\lambda}\left(\epsilon - \frac{C}{C-1}\epsilon^2\right) = \frac{1}{C}.\end{aligned}$$

From the quasi-interpolation of the incorrect class labels, we have that $\langle \boldsymbol{W}_L^c, \boldsymbol{\mu}_{c'}\rangle = \frac{\epsilon}{C-1}$, which means $\langle \boldsymbol{W}_L^c, \boldsymbol{\mu}_{c'} - \boldsymbol{\mu}_G\rangle + \langle \boldsymbol{W}_L^c, \boldsymbol{\mu}_G\rangle = \frac{\epsilon}{C-1}$. Plugging in the previous result and using (3) yields

$$\begin{aligned}\frac{\lambda(C-1)}{\epsilon} \times \langle \boldsymbol{W}_L^c, \boldsymbol{W}_L^{c'}\rangle &= \frac{\epsilon}{C-1} - \frac{1}{C} \\ \langle \boldsymbol{V}_L^c, \boldsymbol{V}_L^{c'}\rangle &= \frac{1}{||\boldsymbol{W}_L^c||_2^2}\frac{\epsilon}{\lambda(C-1)}\left(\frac{\epsilon}{C-1} - \frac{1}{C}\right) \quad (4) \\ &= -\frac{1}{C-1}\end{aligned}$$

Here $\boldsymbol{V}_L^c = \frac{\boldsymbol{W}_L^c}{||\boldsymbol{W}_L^c||_2}$, and we use the fact that all the norms $||\boldsymbol{W}_L^c||_2$ are equal. This completes the proof that the normalized classifier parameters form an ETF. Moreover since $\boldsymbol{W}_L^c \propto \boldsymbol{\mu}_c - \boldsymbol{\mu}_G$ and all the proportionality constants are independent of c, we obtain $\sum_c \boldsymbol{W}_L^c = 0$. This completes the proof of the NC2 condition. NC4 follows then from NC1-NC2, as shown by theorems in [1]. □

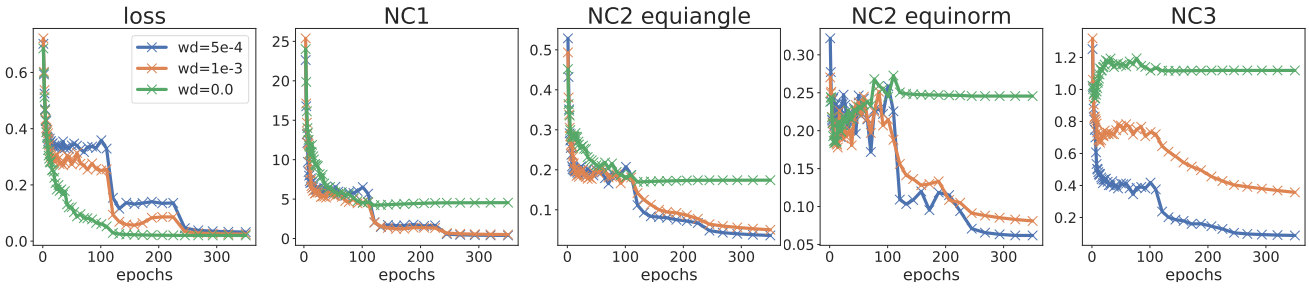

Fig. 1: Measurement of Neural Collapse (NC) on CIFAR10 indicates that it emerges in the presence of weight decay. Exact description of quantities measured can be found in Section 3. Green lines (no weight decay) do not exhibit NC while orange and blue (weight decay) do.

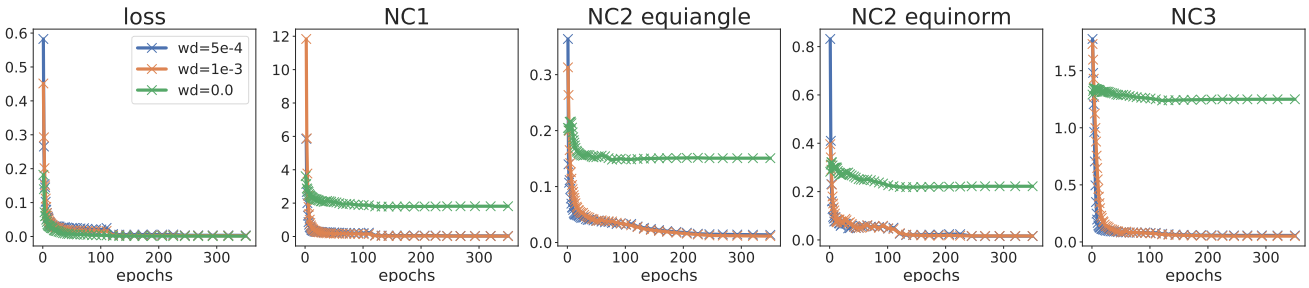

Fig. 2: Measurement of Neural Collapse on MNIST. Neural Collapse is even more clear in the presence of weight decay.

3. EXPERIMENTS

In this section we present the results of experiments using deep homogeneous networks to empirically verify the impact of weight decay on Neural Collapse. We trained a 5 layer deep convolutional network (without biases) with and without weight decay on two datasets - CIFAR10 and MNIST, until they achieved 100% accuracy and near zero square loss on the training dataset. We measured the Neural Collapse conditions in all cases and plot them in figures 1, 2. **(NC1)** was measured by computing $\text{Tr}(\boldsymbol{\Sigma}_W \boldsymbol{\Sigma}_B^{-1})$ where $\boldsymbol{\Sigma}_W, \boldsymbol{\Sigma}_B$ are the within and between class covariance matrices. **(NC2)** was measured by computing the variance of the norms of the class means $(\text{Var}(||\boldsymbol{\mu}_c - \boldsymbol{\mu}_G||_2))$ and the shifted cosine distances $(\cos(\boldsymbol{\mu}_c, \boldsymbol{\mu}_G) + 1/(C-1))$. **(NC3)** was measured as $\left|\left|\frac{\boldsymbol{W}}{||\boldsymbol{W}||_F} - \frac{\dot{\boldsymbol{M}}}{||\dot{\boldsymbol{M}}||_F}\right|\right|_F$. In both figures, the orange and blue lines indicate the trajectories with different values of weight decay, while the green line indicates the trajectory without weight decay.

In all cases, we observe that with and without weight decay we are able to achieve similar squared loss levels, so the solutions are not distinguished by the level of their interpolation. However in both datasets, we see that the NC conditions are achieved in the presence of weight decay. In all plots the green line (without weight decay) is well above the orange and blue lines (different weight decay values), indicating that neural collapse has not been reached in the absence of weight decay.

In the previous section, we were able to establish that under Assumption 1, weight decay is sufficient for Neural Collapse. We also saw that without weight decay, Neural Collapse solutions are not optimal. These experimental results further support the case that weight decay is necessary for Neural Collapse in the case of deep homogeneous networks.

4. CONCLUSIONS AND FUTURE WORK

We summarize here the main predictions of our analysis of Neural Collapse in deep homogeneous ReLU networks. Our analysis predicts that Neural Collapse emerges not only for the case of cross-entropy, for which it was empirically found in [1], but also for the square loss for deep homogeneous networks. We also show that solutions that exhibit Neural Collapse are not interpolating solutions for the unregularized square loss, indicating that weight decay may be necessary for deep homogeneous networks. However open questions still remain - including testing whether our symmetric quasi-interpolation assumption is satisfied. We would also like to be able to characterize NC in deep networks with biases.

While Neural Collapse leads to greatly simplified dynamics for deep classifiers, its connection to the question of generalization remains open, i.e., : is Neural Collapse desirable for generalization? Our analysis suggests that NC1 to NC4 should take place for any quasi-interpolating solutions, including solutions that do not have large margin, as well as for randomly labeled datasets. However our proof also shows that the length of the Simplex ETF classifier vectors is proportional to the training margin. This motivates further experiments testing the connection between Neural Collapse and generalization.

5. REFERENCES

[1] Vardan Papyan, XY Han, and David L Donoho, "Prevalence of neural collapse during the terminal phase of deep learning training," *Proceedings of the National Academy of Sciences*, vol. 117, no. 40, pp. 24652–24663, 2020.

[2] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro, "The implicit bias of gradient descent on separable data," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 2822–2878, 2018.

[3] Kaifeng Lyu and Jian Li, "Gradient descent maximizes the margin of homogeneous neural networks," *CoRR*, vol. abs/1906.05890, 2019.

[4] Tomaso Poggio, Andrzej Banburski, and Qianli Liao, "Theoretical issues in deep networks," *PNAS*, 2020.

[5] Tomaso Poggio, Qianli Liao, and Andrzej Banburski, "Complexity control by gradient descent in deep networks," *Nature Communications*, 2020.

[6] Lenaic Chizat and Francis Bach, "Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss," in *Conference on Learning Theory*. PMLR, 2020, pp. 1305–1338.

[7] Tengyu Xu, Yi Zhou, Kaiyi Ji, and Yingbin Liang, "When will gradient methods converge to max-margin classifier under relu models?," *Stat*, vol. 10, no. 1, pp. e354, 2021.

[8] Song Mei, Andrea Montanari, and Phan-Minh Nguyen, "A mean field view of the landscape of two-layer neural networks," *Proceedings of the National Academy of Sciences*, vol. 115, no. 33, pp. E7665–E7671, 2018.

[9] Zhengdao Chen, Grant M Rotskoff, Joan Bruna, and Eric Vanden-Eijnden, "A dynamical central limit theorem for shallow neural networks," *arXiv preprint arXiv:2008.09623*, 2020.

[10] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang, "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 322–332.

[11] Like Hui and Mikhail Belkin, "Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks," *arXiv preprint arXiv:2006.07322*, 2020.

[12] Ryan M. Rifkin, *Everything Old Is New Again: A Fresh Look at Historical Approaches to Machine Learning*, Ph.D. thesis, Massachusetts Institute of Technology, 2002.

[13] Cong Fang, Hangfeng He, Qi Long, and Weijie J. Su, "Layer-peeled model: Toward understanding well-trained deep neural networks," *CoRR*, vol. abs/2101.12699, 2021.

[14] Dustin G. Mixon, Hans Parshall, and Jianzong Pi, "Neural collapse with unconstrained features," *CoRR*, vol. abs/2011.11619, 2020.

[15] X. Y. Han, Vardan Papyan, and David L. Donoho, "Neural collapse under MSE loss: Proximity to and dynamics on the central path," *CoRR*, vol. abs/2106.02073, 2021.

[16] Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu, "A geometric analysis of neural collapse with unconstrained features," 2021.

[17] Jianfeng Lu and Stefan Steinerberger, "Neural collapse with cross-entropy loss," *CoRR*, vol. abs/2012.08465, 2020.

[18] Stephan Wojtowytsch et al., "On the emergence of tetrahedral symmetry in the final and penultimate layers of neural network classifiers," *arXiv preprint arXiv:2012.05420*, 2020.

[19] Tolga Ergen and Mert Pilanci, "Revealing the structure of deep neural networks via convex duality," *arXiv preprint arXiv:2002.09773*, 2020.

[20] T. Poggio and Q. Liao, "Generalization in deep network classifiers trained with the square loss," *Center for Brains, Minds and Machines (CBMM) Memo No. 112*, 2021.

[21] A. Banburski Q. Liao A.Rangamani, M. Xu and T. Poggio, "Dynamics and neural collapse in deep classifiers trained with the square loss," *Center for Brains, Minds and Machines (CBMM) Memo No. 117*, 2021.