Linear Convergence Analysis of Neural Collapse with Unconstrained Features

Peng Wang* PENGWA@UMICH.EDU
University of Michigan, Ann Arbor

Huikang Liu* LIUHUIKANG@SHUFE.EDU.CN
Shanghai University of Finance and Economics, Shanghai

Can Yaras* CJYARAS@UMICH.EDU
Laura Balzano GIRASOLE@UMICH.EDU
Qing Qu QINGQU@UMICH.EDU
University of Michigan, Ann Arbor

Abstract

In this work, we study the recently discovered neural collapse (NC) phenomenon, which is prevalent in training over-parameterized deep neural networks for classification tasks. Existing work has shown that any optimal solution of the trained problem for classification tasks is an NC solution and has a benign landscape under the unconstrained feature model. However, these results do not provide an answer to the question of how quickly gradient descent can find an NC solution. To answer this question, we prove an error bound property of the trained problem, which refers to the inequality that bounds the distance of a point to the optimal solution set by the norm of its gradient, under the unconstrained feature model. Using this error bound, we show linear convergence of gradient descent for finding an NC solution.

1. Introduction

Over the past years, a large amount of work is devoted to attempting to understand the underlying mechanism of deep neural networks (DNNs) from a theoretical point of view; see, e.g., [1, 6, 12]. Towards the goal of understanding the representations learned by deep neural networks, a recent line of seminal works [3, 14] presents an intriguing phenomenon named *neural collapse* (NC) that occurs pervasively across a range of canonical classification problems during the terminal phase of training. Specifically, the authors observed that the last-layer features and the last-layer linear classifiers of a trained DNN exhibit the following simple but elegant structure: (i) *Variability Collapse:* the individual features of each class concentrate to their class-means; (ii) *Convergence to Simplex ETF:* the class-means have the same length and are maximally distant. In other words, they form a Simplex Equiangular Tight Frame (ETF); (iii) *Convergence to Self-Duality:* the last-layer linear classifiers perfectly match their class-means up to rescaling.

Study of NC under unconstrained feature model. Recently, a line of works demystifies the NC phenomenon in theory based on the so-called *unconstrained feature model* [13, 23], which is also called *layer-peeled model* [3]; see, e.g., [3, 4, 14, 15, 18, 20, 21, 23] and the references therein.

*. These authors contributed equally to this work.

© P. Wang*, H. Liu*, C. Yaras*, L. Balzano & Q. Qu.

Specifically, consider an L-layer fully connected neural network of the form

$$\psi_{\boldsymbol{\Theta}}(\boldsymbol{x}) = \boldsymbol{W}_L \underbrace{\sigma\left(\boldsymbol{W}_{L-1} \ldots \sigma\left(\boldsymbol{W}_1 \boldsymbol{x} + \boldsymbol{b}_1\right) + \boldsymbol{b}_{L-1}\right)}_{\phi_{\boldsymbol{\theta}}(\boldsymbol{x})} + \boldsymbol{b}_L,$$

where $\sigma(\cdot)$ is an activation function, $\boldsymbol{\Theta} = \{\boldsymbol{W}_k, \boldsymbol{b}_k\}_{k=1}^{L}$ denotes all the network parameters, and $\boldsymbol{\theta} = \{\boldsymbol{W}_k, \boldsymbol{b}_k\}_{k=1}^{L-1}$ denotes the network parameters up to the last layer. In particular, the output of the penultimate layer, denoted by $\phi_{\boldsymbol{\theta}}(\boldsymbol{x})$, is referred to as the *feature* of the sample $\boldsymbol{x}$ learned by the neural network. Given training samples $\{(\boldsymbol{x}_{k,i}, \boldsymbol{y}_k)\} \subseteq \mathbb{R}^d \times \mathbb{R}^K$ drawn from the same data distribution, they studied the multi-class (e.g., K classes) classification problem by minimizing the empirical risk over these samples,

$$\min_{\boldsymbol{\Theta}} \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{n} \mathcal{L}\left(\psi_{\boldsymbol{\Theta}}(\boldsymbol{x}_{k,i}), \boldsymbol{y}_k\right) + \frac{\lambda}{2} \|\boldsymbol{\Theta}\|_F^2, \tag{1}$$

where $N = nK$ is the total number of samples, $\boldsymbol{y}_k \in \mathbb{R}^K$ is an one-hot vector with only the k-th entry being 1 and the remaining ones being 0 for all $k \in [K]$, $\boldsymbol{x}_{k,i} \in \mathbb{R}^d$ is the i-th sample in the k-th class, n denotes the number of training samples in each class, $\lambda > 0$ is the regularization parameter, and $\mathcal{L} : \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}$ is a loss function. Note that the unconstrained feature model assumes that the last-layer features of the neural network are free optimization variables, which simplifies the non-linear interaction across layers. Intuitively, this simplification is reasonable because modern deep neural networks are often highly over-parameterized such that last-layer features can approximate or interpolate any point in the feature space [5, 10]. We introduce an auxiliary variable $\boldsymbol{h}_{k,i} = \phi_{\boldsymbol{\theta}}(\boldsymbol{x}_{k,i})$, which denotes the last-layer feature corresponding to the sample $\boldsymbol{x}_{k,i}$. Using the unconstrained feature model and letting $\boldsymbol{W} = \boldsymbol{W}_L^T$ and $\boldsymbol{b} = \boldsymbol{b}_L$, we consider a variant of Problem (1) by treating $\boldsymbol{h}_{k,i}$ for all k, i as free optimization variables, i.e.,

$$\min_{\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}} \frac{1}{nK} \sum_{k=1}^{K} \sum_{i=1}^{n} \mathcal{L}\left(\boldsymbol{W}^T \boldsymbol{h}_{k,i} + \boldsymbol{b}, \boldsymbol{y}_k\right) + \frac{\lambda_W}{2} \|\boldsymbol{W}\|_F^2 + \frac{\lambda_H}{2} \|\boldsymbol{H}\|_F^2. \tag{2}$$

Recently, it has been shown that any global optimal solution of Problem (2) and its constrained counterparts satisfies the NC properties when the function $\mathcal{L}$ is the CE loss, MSE loss, or supervised contrastive loss function; see, e.g., [3, 4, 14, 15, 18, 20, 23]. In particular, Yaras et al. [18], Zhou et al. [20], Zhu et al. [23] showed that their studied training function has a benign landscape with no spurious local minimum. This implies that the NC solutions can be reached by gradient descent methods with random initialization [2, 7]. However, these results still cannot answer the question that how fast gradient descent can reach an NC solution.

Linear convergence analysis under the error bound. A powerful approach to analyzing the convergence behavior of gradient descent type methods for optimizing convex optimization problems is to use an error bound property; see, e.g., [11, 16, 22] and the references therein. Recently, Liu et al. [8, 9], Wang et al. [17], Zheng et al. [19] extended this approach to study non-convex optimization problems. Consider an optimization problem

$$v^* = \min_{\boldsymbol{x} \in \mathcal{E}} F(\boldsymbol{x}), \tag{3}$$

where $\mathcal{E}$ is a finite-dimensional Euclidean space and $F : \mathcal{E} \to (-\infty, +\infty)$ is a continuously differentiable function. Let $\mathcal{X} \subseteq \mathcal{E}$ denote the set of optimal solutions of Problem (3). Then, we say that an error bound holds for Problem (3) if there exist constants $\delta, \kappa > 0$ such that for all $\boldsymbol{x} \in \mathcal{E}$ with $\operatorname{dist}(\boldsymbol{x}, \mathcal{X}) \leq \delta$,

$$\operatorname{dist}(\boldsymbol{x}, \mathcal{X}) \leq \kappa \|\nabla F(\boldsymbol{x})\|_2. \tag{4}$$

Suppose that the error bound holds for Problem (3). A unified approach to analyzing convergence rate of first-order iterative methods have been established based on the error bound; see, e.g., [11, 16, 22].

Fact 1 *Suppose that the optimal solution of Problem (3) is non-empty, i.e., $\mathcal{X} \neq \emptyset$, and the error bound (4) holds for Problem (3). Suppose in addition that the sequence $\{\boldsymbol{x}^k\}_{k \geq k_1}$ for an index $k_1 \geq 0$ satisfies the following conditions:*
(i). (Sufficient Decrease) *There exists a constant $\kappa_1 > 0$ such that*

$$F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) \leq -\kappa_1 \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2.$$

(ii). (Cost-to-Go Estimate) *There exists a constant $\kappa_2 > 0$ such that*

$$F(\boldsymbol{x}^{k+1}) - v^* \leq \kappa_2 \left(\operatorname{dist}^2(\boldsymbol{x}^k, \mathcal{X}) + \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2 \right).$$

(iii). (Safeguard) *There exists a constant $\kappa_3 > 0$ such that*

$$\|\nabla F(\boldsymbol{x}^k)\| \leq \kappa_3 \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|.$$

Then, the sequence $\{F(\boldsymbol{x}^k)\}_{k \geq 0}$ converges Q-linearly to v^* and $\{\boldsymbol{x}^k\}_{k \geq 0}$ converges R-linearly to some $\boldsymbol{x}^* \in \mathcal{X}$.

Contributions of this work. Motivated by above discussions, we are devoted to establishing an error bound for Problem (2) and proving linear convergence of gradient descent for solving Problem (2) in this work. Specifically, we study Problem (2) with the MSE loss and CE loss, respectively. We first characterize the optimal solution set of Problem (2) and then prove the error bound using the structure of the optimal solution set. Our experimental results on synthetic datasets complement and support our theoretical developments.

2. Main Results

Before we proceed, we introduce some essential notation. Given a point $\boldsymbol{y} \in \mathbb{R}^n$ and a non-empty and closed set $\mathcal{X} \subseteq \mathbb{R}^n$, we denote the Euclidean distance of $\boldsymbol{y}$ to $\mathcal{X}$ by $\operatorname{dist}(\boldsymbol{y}, \mathcal{X}) = \min\{\|\boldsymbol{x} - \boldsymbol{y}\|_2 : \boldsymbol{x} \in \mathcal{X}\}$. We use $\mathbf{1}_n$ to denote the n-dimensional all-one vector. Let $\boldsymbol{P} := \boldsymbol{I} - \mathbf{1}_K \mathbf{1}_K^T / K$ and $\boldsymbol{P}^\perp := \mathbf{1}_K \mathbf{1}_K^T / K$. Given a positive integer n, we denote by $[n]$ the set $\{1, \ldots, n\}$. We write all the features and the classifiers in the matrix form

$$\boldsymbol{H} := \begin{bmatrix} \boldsymbol{H}_1 & \ldots & \boldsymbol{H}_n \end{bmatrix} \in \mathbb{R}^{d \times N},\ \boldsymbol{H}_i = \begin{bmatrix} \boldsymbol{h}_{1,i}, \ldots, \boldsymbol{h}_{K,i} \end{bmatrix} \in \mathbb{R}^{d \times K},\ \boldsymbol{W} := \begin{bmatrix} \boldsymbol{w}_1 & \ldots & \boldsymbol{w}_K \end{bmatrix} \in \mathbb{R}^{d \times K}.$$

For ease of exposition, let $\mathcal{X}$ denote the optimal solution set of Problem (2). Without loss of generality, we assume the label matrix to be $\boldsymbol{Y} = \mathbf{1}_n^T \otimes \boldsymbol{I}_K$, where $\otimes$ denotes the Kronecker product.

2.1. Mean Squared Error Loss

Suppose that the loss function $\mathcal{L}$ is the MSE loss of the form $\mathcal{L}(\boldsymbol{z}, \boldsymbol{y}_k) = \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{y}_k\|_2^2$. Substituting this into Problem (2) yields

$$\min_{\boldsymbol{W},\boldsymbol{H},\boldsymbol{b}} F(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}) := \frac{1}{2N}\left\|\boldsymbol{W}^T\boldsymbol{H} + \boldsymbol{b}\mathbf{1}_N^T - \boldsymbol{Y}\right\|_F^2 + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{\lambda_H}{2}\|\boldsymbol{H}\|_F^2. \tag{5}$$

Then, we characterize the optimal solution set of Problem (5) as follows.

Proposition 1 *For the matrix $\boldsymbol{P} = \boldsymbol{I}_K - \mathbf{1}_K\mathbf{1}_K^T/K$, let $\boldsymbol{P} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be its eigenvalue decomposition such that $\boldsymbol{V} = [\bar{\boldsymbol{V}}\ \boldsymbol{v}] \in \mathcal{O}^K$ with $\bar{\boldsymbol{V}}\bar{\boldsymbol{V}}^T = \boldsymbol{P}$, $\boldsymbol{v}\boldsymbol{v}^T = \boldsymbol{P}^\perp$, and $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{I}_{K-1} & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix}$. The optimal solution set $\mathcal{X}$ of Problem (5) can be characterized as follows:*
(i) If $\lambda_W\lambda_H \geq 1/(nK^2)$, we have $\mathcal{X} = \left\{\left(\mathbf{0}, \mathbf{0}, \frac{1}{K}\mathbf{1}_K\right)\right\}$.
(ii) If $\lambda_W\lambda_H < 1/(nK^2)$, we have

$$\mathcal{X} = \left\{\left(\boldsymbol{W}, \boldsymbol{H}, \frac{1}{K}\mathbf{1}_K\right) : \boldsymbol{W} = \frac{\sqrt[4]{n\lambda_H}}{\sqrt[4]{\lambda_W}}\begin{bmatrix}\sqrt{\gamma}\boldsymbol{U}\bar{\boldsymbol{V}}^T \\ \mathbf{0}\end{bmatrix}, \boldsymbol{H}_i = \frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}}\boldsymbol{W}, \forall i \in [n], \boldsymbol{U} \in \mathcal{O}^{d\times(K-1)}\right\},$$

where $\gamma := 1 - K\sqrt{n\lambda_W\lambda_H}$.

This result indicates that when $\lambda_W\lambda_H < 1/K^2$, any global optimal solution of Problem (5) satisfies the NC properties. Compared to [20, Theorem 3.1], we provide all the parameters that depend on λ_W and λ_H explicitly. Based on the above result, we prove the error bound of Problem (5).

Theorem 2 *For Problem (5), if $\lambda_W\lambda_H \neq 1/K^2$, there exist constants $\delta_1, \kappa_1 > 0$ that depend on λ_W and λ_H such that for any $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ satisfying $\operatorname{dist}((\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}), \mathcal{X}) \leq \delta_1$, it holds that*

$$\operatorname{dist}((\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}), \mathcal{X}) \leq \kappa_1\|\nabla F(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})\|_F. \tag{6}$$

Before we proceed, some remarks are in order. First, when $\lambda_W\lambda_H \neq 1/K^2$, we show that the error bound holds for Problem (5) for any point in the neighborhood of the optimal solution set. Second, using this error bound and Fact 1, we can prove the linear convergence of gradient descent for solving Problem (5).

2.2. Cross-Entropy Loss

Suppose that the loss function $\mathcal{L}$ is the CE loss of the form

$$\mathcal{L}_{\mathrm{CE}}(\boldsymbol{z}, \boldsymbol{y}_k) = -\log\left(\frac{\exp(z_k)}{\sum_{\ell=1}^K \exp(z_\ell)}\right). \tag{7}$$

As done in [3, 4], we consider Problem (2) without the bias term when we use the CE loss. Substituting (7) into Problem (2) yields

$$\min_{\boldsymbol{W},\boldsymbol{H}} F(\boldsymbol{W}, \boldsymbol{H}) := \frac{1}{N}\sum_{k=1}^K\sum_{i=1}^n \mathcal{L}_{\mathrm{CE}}(\boldsymbol{W}^T\boldsymbol{h}_{k,i}, \boldsymbol{y}_k) + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{\lambda_H}{2}\|\boldsymbol{H}\|_F^2. \tag{8}$$

Then, we characterize the optimal solution set of Problem (8) as follows.

Proposition 3 *Suppose that the dimension of features is no smaller than the number of classes, i.e., $d \geq K$. Then, the global optimal solution set $\mathcal{X}$ of Problem (8) can be characterized as follows:*
(i) If $\lambda_W \lambda_H \geq 1/(nK^2)$, it holds that $\mathcal{X} = \{(\mathbf{0}, \mathbf{0})\}$.
(ii) If $\lambda_W \lambda_H < 1/(nK^2)$, it holds that

$$\mathcal{X} = \left\{ (\boldsymbol{W}, \boldsymbol{H}) : \boldsymbol{W} = \frac{\sqrt[4]{n\lambda_H}}{\sqrt[4]{\lambda_W}} \sqrt{\log \gamma} \boldsymbol{U}\boldsymbol{P},\ \boldsymbol{H}_i = \frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}} \boldsymbol{W}, \forall i \in [n],\ \boldsymbol{U} \in \mathcal{O}^{d \times K} \right\},$$

where $\gamma := 1/\sqrt{n\lambda_W \lambda_H} - K + 1$.

Based on this result, we prove the error bound for Problem (8).

Theorem 4 *Suppose that the number of classes $K = 2$ and the dimension of features is no smaller than the number of classes, i.e., $d \geq K$. For Problem (8), if $\lambda_W \lambda_H \neq 1/K^2$, there exist constants $\delta_2, \kappa_2 > 0$ that depend on λ_W and λ_H such that for any $(\boldsymbol{W}, \boldsymbol{H})$ satisfying $\operatorname{dist}((\boldsymbol{W}, \boldsymbol{H}), \mathcal{X}) \leq \delta_2$, it holds that*

$$\operatorname{dist}((\boldsymbol{W}, \boldsymbol{H}), \mathcal{X}) \leq \kappa_2 \|\nabla F(\boldsymbol{W}, \boldsymbol{H})\|_F.$$

Remark that since the structure of the CE loss is more complicated than that of the MSE loss, we can only establish the error bound for Problem (8) when $K = 2$. We leave the error bound when $K \geq 2$ for future work.

3. Experimental Results

In this section, we corroborate our theory with experimental results by solving Problem (2) and by solving Problem (1) with $\psi_{\boldsymbol{\Theta}}$ being a two-layer neural network on synthetic data. First, we employ batch gradient descent using backtracking line search for optimizing Problem (2) with both MSE and CE losses and different weight and feature regularization parameters λ_W, λ_H, respectively. In the tests, we set the number of classes as $K = 10$, the number of samples in each class as $n = 10$, and the dimension of features as $d = 20$. Then, we report the convergence performance of gradient descent in Figure 1 with different λ_W and λ_H. One can observe that the optimality gap measured by $F^k - F^*$, where F^k is the function value at the k-th iteration and F^* is the optimal function value computed by the NC solutions, converges linearly *independent* of the regularization parameters λ_W, λ_H. We refer the readers to Appendix for the convergence rate of the NC metrics.

Next, we consider training a two-layer neural network on synthetic data by solving Problem (1) using gradient descent with constant step size. We generate the training samples as follows. The samples $\boldsymbol{x}_{k,i}$ are drawn *i.i.d.* from a zero-mean D-dimensional Gaussian distribution with $D = 1000$ and covariance $\boldsymbol{I}/4$, and the labels are drawn uniformly from $\{1, \ldots, K\}$ (so that we have balanced classes). In the tests, we set the number of classes as $K = 10$, the number of samples in each class as $n = 100$, width of the first layer as $m = 256$, and dimension of features as $d = 20$. Then, we report the results in Figure 2. Unlike the UFM, we found that the regularization parameter is important in the convergence rate and requires careful tuning - for particular settings of λ_W, λ_H, we can achieve linear convergence when training a shallow network with explicit regularization. We refer the readers to Appendix for the convergence rate of the NC metrics.

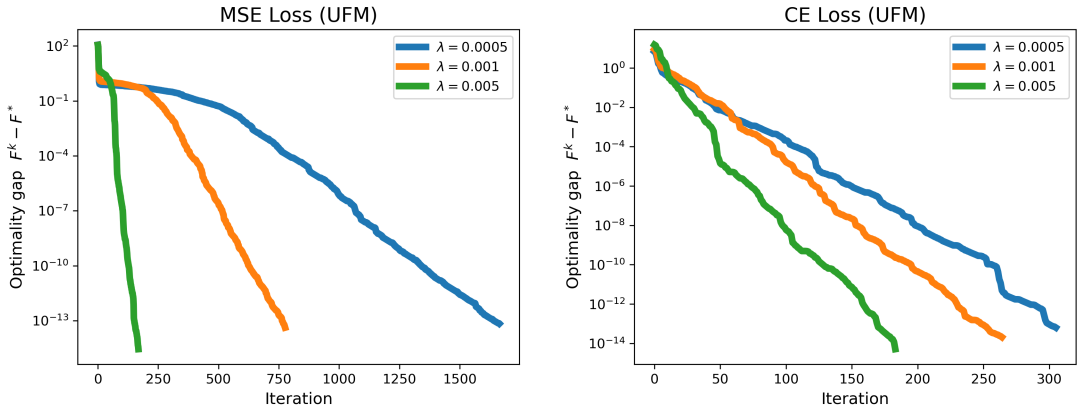

Figure 1: $K = 10$ classes, $n = 10$ samples per class, $d = 20$, $\lambda_W = n\lambda$, $\lambda_H = \lambda$.

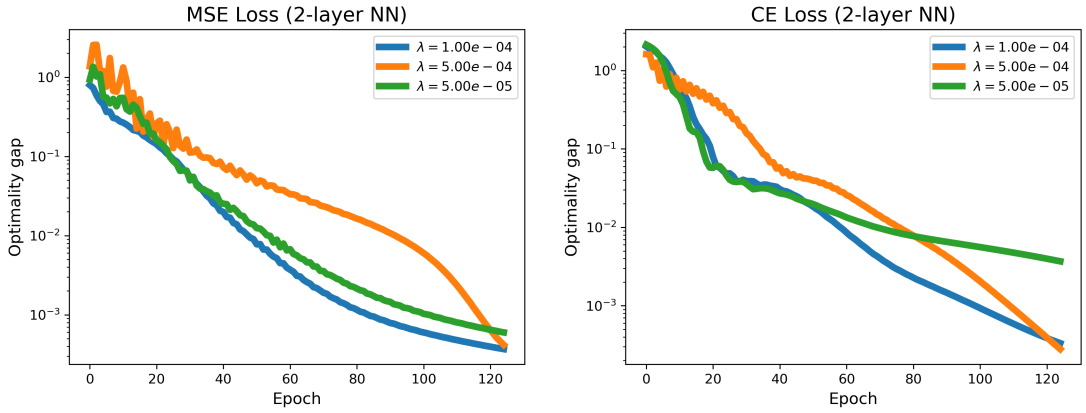

Figure 2: We use $K = 10$ classes, $n = 100$ samples per class, first-layer width $m = 256$, $d = 20$ dimensional feature space, with regularization $\lambda_W = n\lambda$, $\lambda_H = \lambda$.

4. Conclusions

In this work, we studied the NC phenomenon under the unconstrained feature model. We first characterized the set of optimal solutions of the trained problem. Based on this characterization, we showed that error bound holds for the trained problem. Using this, we further established the linear convergence of gradient descent for optimizing the trained problem. Finally, we supported our theoretical results by experimental results. As a future work, we would like to extend our analysis to the case of $K \geq 2$ for the CE loss. Another interesting direction is to analyze the error bound of the constrained counterparts of Problem (2); see, e.g., [3, 18].

References

[1] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

[2] Christopher Criscitiello and Nicolas Boumal. Efficiently escaping saddle points on manifolds. *Advances in Neural Information Processing Systems*, 32, 2019.

[3] Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43), 2021.

[4] Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised constrastive learning. In *International Conference on Machine Learning*, pages 3821–3830. PMLR, 2021.

[5] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

[6] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[7] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732. PMLR, 2017.

[8] Huikang Liu, Anthony Man-Cho So, and Weijie Wu. Quadratic optimization with orthogonality constraint: explicit łojasiewicz exponent and linear convergence of retraction-based line-search and stochastic variance-reduced gradient methods. *Mathematical Programming*, 178(1):215–262, 2019.

[9] Huikang Liu, Man-Chung Yue, and Anthony Man-Cho So. A unified approach to synchronization problems over subgroups of the orthogonal group. *arXiv preprint arXiv:2009.07514*, 2020.

[10] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. *Advances in neural information processing systems*, 30, 2017.

[11] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

[12] 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*, 115(33): E7665–E7671, 2018.

[13] Dustin G Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *arXiv preprint arXiv:2011.11619*, 2020.

[14] 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*, 117(40):24652–24663, 2020.

[15] Tom Tirer and Joan Bruna. Extended unconstrained features model for exploring deep neural collapse. *arXiv preprint arXiv:2202.08087*, 2022.

[16] Paul Tseng and Sangwoon Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1):387–423, 2009.

[17] Peng Wang, Huikang Liu, and Anthony Man-Cho So. Linear convergence of a proximal alternating minimization method with extrapolation for ℓ_1-norm principal component analysis. *arXiv preprint arXiv:2107.07107*, 2021.

[18] Can Yaras, Peng Wang, Zhihui Zhu, Laura Balzano, and Qing Qu. Neural collapse with normalized features: A geometric analysis over the riemannian manifold. *arXiv preprint arXiv:2209.09211*, 2022.

[19] Taoli Zheng, Peng Wang, and Anthony Man-Cho So. A linearly convergent algorithm for rotationally invariant ℓ_1 norm principal component analysis. *arXiv preprint arXiv:2210.05066*, 2022.

[20] Jinxin Zhou, Xiao Li, Tianyu Ding, Chong You, Qing Qu, and Zhihui Zhu. On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features. *arXiv preprint arXiv:2203.01238*, 2022.

[21] Jinxin Zhou, Chong You, Xiao Li, Kangning Liu, Sheng Liu, Qing Qu, and Zhihui Zhu. Are all losses created equal: A neural collapse perspective. *arXiv preprint arXiv:2210.02192*, 2022.

[22] Zirui Zhou and Anthony Man-Cho So. A unified approach to error bounds for structured convex optimization problems. *Mathematical Programming*, 165(2):689–728, 2017.

[23] Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34, 2021.

Appendix

Linear Convergence of the NC Metrics

In Section 3, we demonstrated that gradient descent achieves linear convergence in terms of *function value* for training Problem (1) under the UFM and Problem (2) with shallow networks, respectively. A natural question is that whether we can empirically observe linear convergence of metrics that measure the various aspects of neural collapse. For measuring different aspects of neural collapse as introduced in Section 1, we adopt similar NC metrics as those in [14, 20, 23],

$$\mathcal{NC}_1 := \frac{1}{K}\text{trace}(\boldsymbol{\Sigma}_W \boldsymbol{\Sigma}_B^{\dagger})$$
$$\mathcal{NC}_2 := \left\| \frac{\boldsymbol{W}^\top \boldsymbol{W}}{\|\boldsymbol{W}^\top \boldsymbol{W}\|_F} - \frac{1}{\sqrt{K-1}}\left(\boldsymbol{I}_K - \boldsymbol{1}_K \boldsymbol{1}_K^\top\right)\right\|_F$$
$$\mathcal{NC}_3 := \left\| \frac{\boldsymbol{W}^\top \overline{\boldsymbol{H}}}{\|\boldsymbol{W}^\top \overline{\boldsymbol{H}}\|_F} - \frac{1}{\sqrt{K-1}}\left(\boldsymbol{I}_K - \boldsymbol{1}_K \boldsymbol{1}_K^\top\right)\right\|_F,$$

where $\boldsymbol{\Sigma}_W$ and $\boldsymbol{\Sigma}_B$ are the within-class and between-class covariance matrices (see [14, 23] for more details), $\boldsymbol{\Sigma}_B^{\dagger}$ denotes pseudo inverse of $\boldsymbol{\Sigma}_B$, and $\overline{\boldsymbol{H}}$ is the centered class mean matrix. More specifically, $\mathcal{NC}_1$ measures within class variability collapse, $\mathcal{NC}_2$ measures convergence to the simplex ETF, and $\mathcal{NC}_3$ measures duality collapse.

We first investigate the convergence rate of the NC metrics under the UFM, using the same setup as that in Section 3. We report the result in Figure 3. One can observe that all three metrics converge linearly to 0 for both MSE and CE losses. In particular, $\mathcal{NC}_1$ converges substantially faster than the other two metrics.

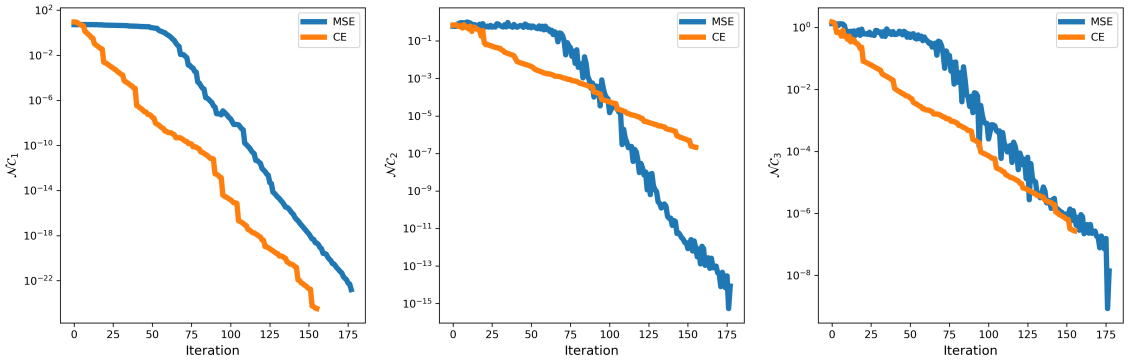

Figure 3: $K = 10$ classes, $n = 10$ samples per class, $d = 20$, $\lambda_W = 5 \times 10^{-4}$, $\lambda_H = 5 \times 10^{-5}$.

Analogously, we investigate the convergence rate of the NC metrics for a two-layer neural network using the same setup and data as those in Section 3. We report the result in Figure 4. We again observe that all three metrics converge linearly to 0 for both MSE and CE losses using appropriate regularization.

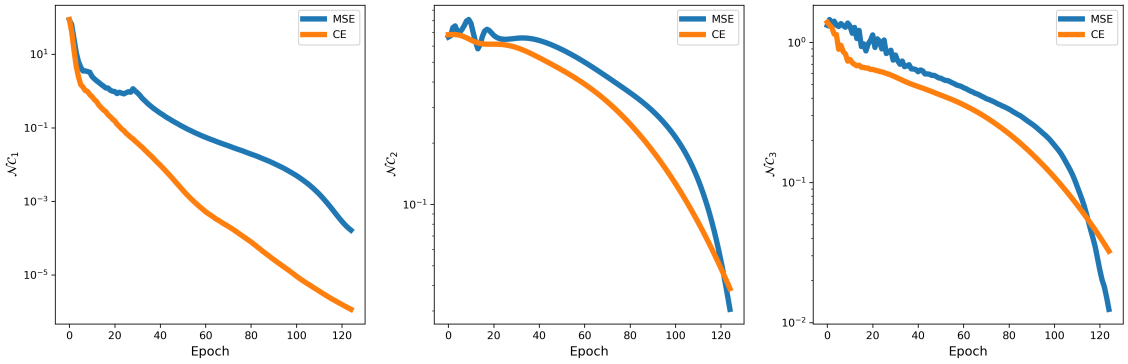

Figure 4: We use $K = 10$ classes, $n = 100$ samples per class, first-layer width $m = 256$, $d = 20$ dimensional feature space, with regularization $\lambda_W = n\lambda$, $\lambda_H = \lambda$.

Motivated by the above observations, an interesting direction is to prove linear convergence of these NC metrics using the error bound condition. This is left as a future work.