Linear Convergence Analysis of Neural Collapse with Unconstrained Features

author names withheld

Under Review for OPT 2022

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, under the unconstrained feature model we prove an error bound property of the trained loss, which refers to the inequality that bounds the distance of a point in the optimal solution set by the norm of its gradient. Using this error bound, we can show linear convergence of gradient descent for finding an NC solution.

1. Introduction

Towards the goal of understanding the representations learned by deep neural networks (DNNs), a recent line of seminal works [2, 10] 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.

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* [9, 17], which is also called *layer-peeled model* [2]; see, e.g., [2, 3, 10, 11, 14, 15, 17] and the references therein. Specifically, consider an L-layer fully connected neural network of the form

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

where $\sigma(\cdot)$ is an activation function and $\boldsymbol{\Theta} = \{\boldsymbol{W}_k, \boldsymbol{b}_k\}_{k=1}^L$ denotes the network parameters. 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,

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

where $\boldsymbol{y}_k \in \mathbb{R}^K$ 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 [4, 7]. We introduce an auxiliary variables $\boldsymbol{h}_{k,i} = \psi_{\boldsymbol{\Theta}}(\boldsymbol{x}_{k,i})$, which denotes the last-layer features 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 counterpart satisfies the NC properties when the function $\mathcal{L}$ is the CE loss, MSE loss, or supervised contrastive loss function; see, e.g., [2, 3, 10, 11, 14, 15, 17]. In particular, Yaras et al. [14], Zhou et al. [15], Zhu et al. [17] showed that their studied training loss 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 [1, 5]. 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 is to use an error bound property; see, e.g., [8, 12, 16] and the references therein. 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 F is a closed proper convex function and the error bound holds for Problem (3). It follows from [16, Fact 1] that if the sequence $\{\boldsymbol{x}^k\}$ generated by the studied method for solving Problem (3) satisfies sufficient descent, cost-to-go estimate, and safegurad properties, the sequence $\{F(\boldsymbol{x}^k)\}$ converges Q-linearly to v^* and $\{\boldsymbol{x}^k\}$ converges R-linearly to some $\boldsymbol{x}^* \in \mathcal{X}$. Recently, Liu et al. [6], Wang et al. [13] proved the error bound for non-convex problems with orthogonality constraint and further established linear convergence of their studied methods.

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 optimal solution set. Our experimental results on synthetic datasets complement and support our theoretical developments.

2. Main Results

To simplify our analysis, we assume that the number of training samples in each class is 1, i.e., $n = 1$. Before we proceed, we introduce some essential notation. Given a point $\boldsymbol{y} \in \mathbb{R}^n$ and a non-empty closed set $\mathcal{Y} \subseteq \mathbb{R}^n$, we denote the Euclidean distance of $\boldsymbol{y}$ to $\mathcal{Y}$ by $\operatorname{dist}(\boldsymbol{y}, \mathcal{Y}) = \min\{\|\boldsymbol{x} - \boldsymbol{y}\|_2 : \boldsymbol{x} \in \mathcal{Y}\}$. 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

$$\boldsymbol{H} := \begin{bmatrix} \boldsymbol{h}_1 & \boldsymbol{h}_2 & \ldots & \boldsymbol{h}_K \end{bmatrix} \in \mathbb{R}^{d\times K}, \quad \boldsymbol{W} := \begin{bmatrix} \boldsymbol{w}_1 & \boldsymbol{w}_2 & \ldots & \boldsymbol{w}_K \end{bmatrix} \in \mathbb{R}^{d\times K}.$$

For ease of exposition, let $\mathcal{X}$ denote the optimal solution set of Problem (2).

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 and $n = 1$ into Problem (2) yields

$$\min_{\boldsymbol{W},\boldsymbol{H}\in\mathbb{R}^{d\times K},\boldsymbol{b}\in\mathbb{R}^K} F(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}) := \frac{1}{2K}\left\|\boldsymbol{W}^T\boldsymbol{H} + \boldsymbol{b}\mathbf{1}_K^T - \boldsymbol{Y}\right\|_F^2 + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{\lambda_H}{2}\|\boldsymbol{H}\|_F^2. \quad (5)$$

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

Proposition 1 *For the matrix $\boldsymbol{P}$, 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 $\lambda_W\lambda_H \geq 1/K^2$, we have $\mathcal{X} = \left\{\left(\mathbf{0}, \mathbf{0}, \frac{1}{K}\mathbf{1}_K\right)\right\}$.
(ii) If $\lambda_W\lambda_H < 1/K^2$, we have

$$\mathcal{X} = \left\{\left(\boldsymbol{W}, \boldsymbol{H}, \frac{1}{K}\mathbf{1}_K\right) : \boldsymbol{W} = \frac{\sqrt[4]{\lambda_H}}{\sqrt[4]{\lambda_W}}\begin{bmatrix} \sqrt{\gamma}\boldsymbol{U}\bar{\boldsymbol{V}}^T \\ \mathbf{0} \end{bmatrix}, \boldsymbol{H} = \frac{\sqrt{\lambda_W}}{\sqrt{\lambda_H}}\boldsymbol{W}, \boldsymbol{U} \in \mathcal{O}^{d\times(K-1)}\right\}, \quad (6)$$

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

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 [15, Theorem 3.1], we provide all the parameters that depend on λ_W and λ_H in (6) 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 $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}) \in \mathbb{R}^{d\times K} \times \mathbb{R}^{d\times K} \times \mathbb{R}^K$ satisfying $\operatorname{dist}((\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}), \mathcal{X}) \leq \delta_1$, it holds that*

$$\operatorname{dist}((\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}), \mathcal{X}) \leq \kappa_1\|\nabla F(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})\|_F. \quad (7)$$

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 [16, 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

$$\mathcal{L}_{\mathrm{CE}}(\boldsymbol{z}, \boldsymbol{y}_k) = -\log\left(\frac{\exp(z_k)}{\sum_{\ell=1}^{K}\exp(z_\ell)}\right). \tag{8}$$

As done in [2, 3], we consider Problem (2) without the bias term when we use the CE loss. Substituting (8) and $n = 1$ into Problem (2) yields

$$\min_{\boldsymbol{W},\boldsymbol{H}\in\mathbb{R}^{d\times K}} F(\boldsymbol{W},\boldsymbol{H}) := \frac{1}{K}\mathcal{L}_{\mathrm{CE}}(\boldsymbol{W}^T\boldsymbol{h}_k, \boldsymbol{y}_k) + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{\lambda_H}{2}\|\boldsymbol{H}\|_F^2. \tag{9}$$

Then, we characterize the optimal solution set of Problem (9) 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 (9) can be characterized as follows:*
(i) If $\lambda_W\lambda_H \geq 1/K^2$, it holds that $\mathcal{X} = \{(\mathbf{0}, \mathbf{0})\}$.
(ii) If $\lambda_W\lambda_H < 1/K^2$, it holds that

$$\mathcal{X} = \left\{(\boldsymbol{W},\boldsymbol{H}) : \boldsymbol{W} = \frac{\sqrt[4]{\lambda_H}}{\sqrt[4]{\lambda_W}}\sqrt{\log\left(\frac{1}{\sqrt{\lambda_W\lambda_H}} - K + 1\right)}\boldsymbol{U}\boldsymbol{P},\ \boldsymbol{H} = \frac{\sqrt{\lambda_W}}{\sqrt{\lambda_H}}\boldsymbol{W}, \boldsymbol{U} \in \mathcal{O}^{d\times K}\right\}.$$

Based on this result, we prove the error bound of Problem (9).

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 (9), 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 $(\boldsymbol{W},\boldsymbol{H}) \in \mathbb{R}^{d\times K} \times \mathbb{R}^{d\times K}$ 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 (9) 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 gradient descent with 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. Then, we report the results in Figure 1. One can observe that the optimality gap measured by $F^k - F^*$,

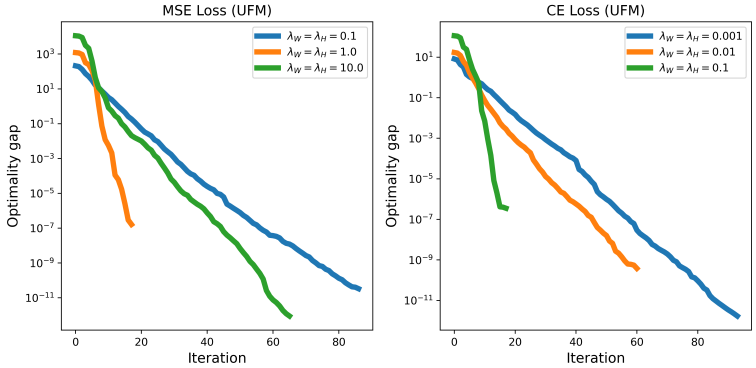

Figure 1: We use $K = 10$ classes, $n = 10$ samples per class, and $d = 20$ feature dimensions.

where F^k is the function value at the k-th iteration and F^* is the optimal value, converges linearly *independent* of the regularization parameters λ_W, λ_H.

Next, we consider training a two-layer neural network on synthetic data with batch gradient descent using 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). 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.

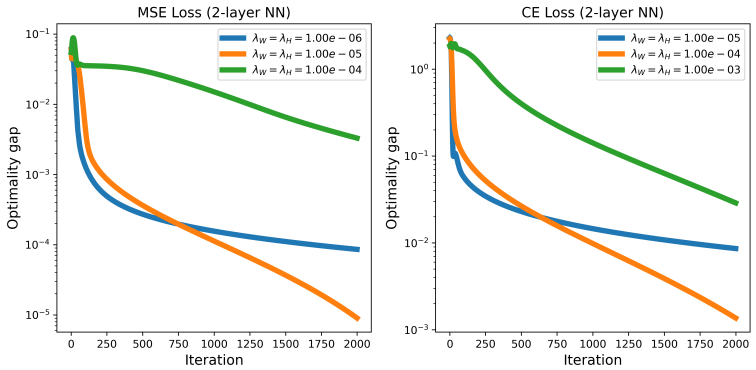

Figure 2: We use $K = 10$ classes, $n = 100$ samples per class, first-layer width $m = 256$, and a $d = 20$ dimensional feature space.

4. Conclusions

In this work, we studied the NC phenomenon under the unconstrained feature model. We showed that error bound holds for the trained problem by characterizing its optimal solution set. Using this, we can further establish the linear convergence of gradient descent for optimizing the trained problem. We support our theorems on practical deep neural network training. As a future work, we would like to extend our analysis to the case of $n \geq 2$. Another interesting direction is to analyze the error bound of the constrained counterpart of Problem (2); see, e.g., [2, 14].

References

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

[2] 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.

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

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

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.