Convergence Analysis for Learning Orthonormal Deep Linear Neural Networks

Zhen Qin, Xuwei Tan, and Zhihui Zhu, *Member, IEEE*

***Abstract*—Enforcing orthonormal or isometric property for the weight matrices has been shown to enhance the training of deep neural networks by mitigating gradient exploding/vanishing and increasing the robustness of the learned networks. However, despite its practical performance, the theoretical analysis of orthonormality in neural networks is still lacking; for example, how orthonormality affects the convergence of the training process. In this letter, we aim to bridge this gap by providing convergence analysis for training orthonormal deep linear neural networks. Specifically, we show that Riemannian gradient descent with an appropriate initialization converges at a linear rate for training orthonormal deep linear neural networks with a class of loss functions. Unlike existing works that enforce orthonormal weight matrices for all the layers, our approach excludes this requirement for one layer, which is crucial to establish the convergence guarantee. Our results shed light on how increasing the number of hidden layers can impact the convergence speed. Experimental results validate our theoretical analysis.**

***Index Terms*—Convergence analysis, deep neural networks, orthonormal structure, Riemannian optimization.**

I. Introduction

ENFORCING orthonormal or isometric properties of the weight matrices has numerous advantages for the practice of deep learning: (i) it provides a better initialization [1], [2], (ii) it mitigates the problem of exploding/vanishing gradients during training [3], [4], [5], [6], [7], (iii) the resulting *orthonormal neural networks* [2], [8], [9], [10], [11], [12], [13], [14], [15], [16] exhibit improved robustness [17] and reduced overfitting issues [18].

Various approaches have been proposed for training neural networks, mainly falling into two categories: soft orthonormality and hard orthonormality. The first category of methods, such as those in [10], [12], [17], [19], adds an additional orthonormality regularization term to the training loss, resulting in weight matrices that are approximately orthonormal. In contrast, the other methods, as found in [9], [11], [20], learn weight matrices that are exactly orthonormal through the use of Riemannian optimization algorithms on the Stiefel manifold.

While orthonormal neural networks demonstrate strong practical performance, there remains a gap in the theoretical analysis of orthonormality in neural networks. For example, convergence analysis for training neural networks has been extensively studied [21], [22], [23], [24], [25], [26], [27], [28]. However, all these results focus on standard training without orthonormal constraints, making them inapplicable to the training of orthonormal neural networks. To the best of our knowledge, there is a lack of rigorous convergence analysis even for orthonormal deep linear neural networks (ODLNNs). Despite its linear structure, a deep linear neural network still presents a non-convex training problem and has served as a testbed for understanding deep neural networks [21], [22], [23], [24]. In this letter, we aim to understand the effect of the orthonormal structure on the training process by studying ODLNN.

Our contribution: Specifically, we provide a local convergence rate of Riemannian gradient descent (RGD) for training the ODLNN. To achieve this, unlike existing works [9], [11], [17] that impose orthonormal constraints on all the weight matrices, we exclude such a constraint for one layer (say the weight matrix in the first hidden layer). The exclusion of a specific layer plays a crucial role in analyzing the convergence rate. Our findings demonstrate that within a specific class of loss functions, adhering to the restricted correlated gradient condition [29], the RGD algorithm exhibits linear convergence speed when appropriately initialized. Notably, our results also indicate that as the number of layers in the network increases, the rate of convergence only experiences a polynomial decrease. The validity of our theoretical analysis has been confirmed by experiments.

Notation: We use bold capital letters (e.g., $\boldsymbol{A}$) to denote matrices, bold lowercase letters (e.g., $\boldsymbol{a}$) to denote vectors, and italic letters (e.g., a) to denote scalar quantities. The superscript $(\cdot)^\top$ denotes the transpose. $\|\boldsymbol{A}\|$ and $\|\boldsymbol{A}\|_F$ respectively represent the spectral norm and Frobenius norm of $\boldsymbol{A}$. $\sigma_{\min}(\boldsymbol{A})$ is the smallest singular value of $\boldsymbol{A}$. The condition number of $\boldsymbol{A}$ is defined as $\kappa(\boldsymbol{A}) = \frac{\|\boldsymbol{A}\|}{\sigma_{\min}(\boldsymbol{A})}$. $\|\boldsymbol{a}\|_2$ is the l_2 norm of $\boldsymbol{a}$. For a positive integer K, $[K]$ denotes the set $\{1, \ldots, K\}$. $b = \Omega(a)$ represents $b \geq ca$ for some universal constant c.

II. Riemannian Gradient Descent for Orthonormal Deep Linear Neural Networks

Problem statement: Given a training set $\{(\boldsymbol{x}_i, \boldsymbol{y}_i^\star)\}_{i=1}^n \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, our goal is to estimate a hypothesis (predictor) from a parametric family $\mathcal{H} := \{h_\theta : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y} | \theta \in \boldsymbol{\Theta}\}$ by minimizing the following empirical risk:

$$\min_{\theta \in \boldsymbol{\Theta}} g(\theta) = \frac{1}{n} \sum_{i=1}^{n} l(h_\theta(\boldsymbol{x}_i); \boldsymbol{y}_i^\star), \tag{1}$$

where $l(h_\theta(\boldsymbol{x}_i); \boldsymbol{y}_i^\star)$ is a suitable loss that captures the difference between the network prediction $h_\theta(\boldsymbol{x}_i)$ and the label $\boldsymbol{y}_i^\star$. For

Manuscript received 24 November 2023; revised 23 February 2024; accepted 28 February 2024. Date of publication 6 March 2024; date of current version 21 March 2024. This work was supported by NSF under Grant CCF-2240708 and Grant CCF-2241298. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jianxin Li. *(Corresponding author: Zhihui Zhu.)*

The authors are with the Department of Computer Science and Engineering, Ohio State University, Columbus, OH 43210 USA (e-mail: qin.660@osu.edu; tan.1206@osu.edu; zhu.3440@osu.edu).

Digital Object Identifier 10.1109/LSP.2024.3374085

convenience, we stack all the training samples together as $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{x}_1 & \cdots & \boldsymbol{x}_n\end{bmatrix}$ and $\boldsymbol{Y}^\star = \begin{bmatrix}\boldsymbol{y}_1^\star & \cdots & \boldsymbol{y}_n^\star\end{bmatrix}$.

Our main focus is on orthonormal deep linear neural networks (ODLNNs), which are fully-connected neural networks of form $h_\theta(\boldsymbol{x}_i) = \boldsymbol{W}_N \cdots \boldsymbol{W}_1 \boldsymbol{x}_i$ with $\boldsymbol{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}$ for $i \in [N]$, where $d_0 = d_x$ and $d_N = d_y$. In ODLNNs, we further assume that the weight matrices to be row orthogonal or column orthogonal depending on the dimension. Without loss of generality, we assume that all the matrices $\{\boldsymbol{W}_i\}_{i\geq 2}$ are column orthonormal, except for $\boldsymbol{W}_1$. This is different to the previous works [9], [11], [17] which impose orthonormal constraints on all the weight matrices. Allowing $\mathbf{W}_1$ to be unstructured offers more flexibility, as otherwise $\mathbf{W}_N \cdots \mathbf{W}_1$ can only represent an orthonormal matrix, which restricts the output the same energy as input (i.e., $\|\boldsymbol{y}_i^\star\|_2 = \|\boldsymbol{x}_i\|_2$). The choice of free weight matrix can vary, and the following analysis would still hold. Now the training loss can be written as

$$\min_{\substack{\boldsymbol{W}_i \in \mathbb{R}^{d_i \times d_{i-1}} \\ i\in[N]}} g(\boldsymbol{W}_N, \ldots, \boldsymbol{W}_1) = L\left(\boldsymbol{W}_N \cdots \boldsymbol{W}_1 \boldsymbol{X}; \boldsymbol{Y}^\star\right),$$
$$\text{s. t.} \qquad \boldsymbol{W}_i^\top \boldsymbol{W}_i = \mathbf{I}_{d_{i-1}}, \ \ i = 2, \ldots N, \tag{2}$$

where L denotes a loss function encompassing all samples.

Definition 1 (Data model): Following the previous work on deep linear neural networks [21], [22], we assume that the dataset $\boldsymbol{X}$ is whitened, i.e., its empirical covariance matrix is an identity matrix as $\boldsymbol{X}\boldsymbol{X}^\top = \mathbf{I}_{d_x}$. Also assume that the output is generated by a teacher ODLNN model,[1] i.e., $\boldsymbol{Y}^\star = \boldsymbol{W}_N^\star \cdots \boldsymbol{W}_1^\star \boldsymbol{X}$, where $\boldsymbol{W}_i^\star \in \mathbb{R}^{d_i \times d_{i-1}}$ and $\{\boldsymbol{W}_i^\star\}_{i\geq 2}$ are column orthonormal matrices.

Stiefel manifold: The Stiefel manifold $\mathrm{St}(m,n) = \{\boldsymbol{C} \in \mathbb{R}^{m\times n} : \boldsymbol{C}^\top \boldsymbol{C} = \mathbf{I}_n\}$ is a Riemannian manifold that is composed of all $m \times n$ orthonormal matrices. We can regard $\mathrm{St}(m,n)$ as an embedded submanifold of a Euclidean space and further define $\mathrm{T}_{\boldsymbol{C}}\mathrm{St} := \{\boldsymbol{A} \in \mathbb{R}^{m\times n} : \boldsymbol{A}^\top \boldsymbol{C} + \boldsymbol{C}^\top \boldsymbol{A} = 0\}$ as its tangent space at the point $\boldsymbol{C} \in \mathrm{St}(m,n)$. For any $\boldsymbol{B} \in \mathbb{R}^{m\times n}$, the projection of $\boldsymbol{B}$ onto $\mathrm{T}_{\boldsymbol{C}}\mathrm{St}$ is given by [30]

$$\mathcal{P}_{\mathrm{T}_{\boldsymbol{C}}\mathrm{St}}(\boldsymbol{B}) = \boldsymbol{B} - \frac{1}{2}\boldsymbol{C}(\boldsymbol{B}^\top \boldsymbol{C} + \boldsymbol{C}^\top \boldsymbol{B}), \tag{3}$$

and its orthogonal complement is $\mathcal{P}^\perp_{\mathrm{T}_{\boldsymbol{C}}\mathrm{St}}(\boldsymbol{B}) = \boldsymbol{B} - \mathcal{P}_{\mathrm{T}_{\boldsymbol{C}}\mathrm{St}}(\boldsymbol{B}) = \frac{1}{2}\boldsymbol{C}(\boldsymbol{B}^\top \boldsymbol{C} + \boldsymbol{C}^\top \boldsymbol{B})$. When we have a gradient $\boldsymbol{B}$ defined in the Hilbert space, we can use the projection operator (3) to compute the Riemannian gradient $\mathcal{P}_{\mathrm{T}_{\boldsymbol{C}}\mathrm{St}}(\boldsymbol{B})$ on the tangent space of the Stiefel manifold. To project $\widehat{\boldsymbol{C}} = \boldsymbol{C} - c\mathcal{P}_{\mathrm{T}_{\boldsymbol{C}}\mathrm{St}}(\boldsymbol{B})$ with any positive constant c back onto the Stiefel manifold, we can utilize the polar decomposition-based retraction, i.e.,

$$\mathrm{Retr}_{\boldsymbol{C}}(\widehat{\boldsymbol{C}}) = \widehat{\boldsymbol{C}}(\widehat{\boldsymbol{C}}^\top \widehat{\boldsymbol{C}})^{-\frac{1}{2}}. \tag{4}$$

Riemannian gradient descent (RGD): Given the gradient $\nabla_{\boldsymbol{W}_i} g(\boldsymbol{W}_N^{(t)}, \ldots, \boldsymbol{W}_1^{(t)})$, we can compute the Riemannian gradient $\mathcal{P}_{\mathrm{T}_{\boldsymbol{W}_i}\mathrm{St}}(\nabla_{\boldsymbol{W}_i} g(\boldsymbol{W}_N^{(t)}, \ldots, \boldsymbol{W}_1^{(t)}))$ on the Stiefel manifold via (3). To streamline the notation, let us represent $\nabla_{\boldsymbol{W}_i} g(\boldsymbol{W}_N^{(t)}, \ldots, \boldsymbol{W}_1^{(t)})$ as $\nabla_{\boldsymbol{W}_i^{(t)}} g$. Now the weight matrices can be updated via the following RGD:

$$\boldsymbol{W}_1^{(t+1)} = \boldsymbol{W}_1{}^{(t)} - \mu\gamma \nabla_{\boldsymbol{W}_1^{(t)}} g,$$
$$\boldsymbol{W}_i^{(t+1)} = \mathrm{Retr}_{\boldsymbol{W}_i}(\boldsymbol{W}_i^{(t)} - \mu \mathcal{P}_{\mathrm{T}_{\boldsymbol{W}_i}\mathrm{St}}(\nabla_{\boldsymbol{W}_i^{(t)}} g)), i \geq 2, \tag{5}$$

where $\mu > 0$ is the learning rate for $\{\boldsymbol{W}_i\}$ and $\gamma > 0$ controls the ratio between the learning rates for $\boldsymbol{W}_1$ and $\{\boldsymbol{W}_i\}_{i\geq 2}$. The discrepant learning rates in (5) are used to accelerate the convergence rate of $\boldsymbol{W}_1$ since the energy of $\boldsymbol{W}_1^\star$ and $\{\boldsymbol{W}_i^\star\}_{i\geq 2}$ are unbalanced, i.e., $\|\boldsymbol{W}_1^\star\|^2 = \|\boldsymbol{Y}^\star\|^2$ and $\|\boldsymbol{W}_i^\star\|^2 = 1$ when the dataset $\boldsymbol{X}$ is whitened, i.e. $\boldsymbol{X}\boldsymbol{X}^\top = \mathbf{I}_{d_x}$.

III. Convergence Analysis

In this section, we will delve into the convergence rate analysis of RGD for training ODLNNs. Towards that goal, we will use the teacher model introduced in Definition 1 that the training samples $(\boldsymbol{X}, \boldsymbol{Y}^\star)$ are generated according to $\boldsymbol{Y}^\star = \boldsymbol{W}_N^\star \cdots \boldsymbol{W}_1^\star \boldsymbol{X}$. Given the nonlinear nature of the retraction operation in the RGD, we will study the convergence in terms of the weight matrices $\mathrm{W} = \{\boldsymbol{W}_i\}$ and $\mathrm{W}^\star = \{\boldsymbol{W}_i^\star\}$. But we will show that the convergence can be equivalently established in terms of the outputs.

Distance measure: Consider that the factors in $\mathrm{W}^\star$ are identifiable up to orthonormal transforms since $\boldsymbol{W}_N^\star \cdots \boldsymbol{W}_1^\star \boldsymbol{X} = \boldsymbol{W}_N^\star \boldsymbol{R}_{N-1} \boldsymbol{R}_{N-1}^\top \boldsymbol{W}_{N-1}^\star \boldsymbol{R}_{N-2} \cdots \boldsymbol{W}_1^\star \boldsymbol{X}$ for any orthonormal matrices $\boldsymbol{R}_i \in \mathbb{O}^{d_i \times d_i}, i \in [N-1]$. Also, $\boldsymbol{W}_1^\star$ and $\boldsymbol{W}_i^\star$ could be imbalanced as $\boldsymbol{W}_i^\star$ are orthonormal for $i \geq 2$. This discrepancy can be quantified by observing that $\|\boldsymbol{W}_1^\star\| = \|\boldsymbol{Y}^\star\|$ (since the input matrix $\boldsymbol{X}$ is whitened) and $\|\boldsymbol{W}_i^\star\| = 1$ for all $i \geq 2$. Thus, we propose the following measure to capture the distance between two sets of factors:

$$\mathrm{dist}^2(\mathrm{W}, \mathrm{W}^\star) = \min_{\substack{\boldsymbol{R}_i \in \mathbb{O}^{d_i \times d_i}, \\ i\in[N-1]}} \sum_{i=2}^{N} \|\boldsymbol{Y}^\star\|^2 \|\boldsymbol{W}_i - \boldsymbol{R}_i^\top \boldsymbol{W}_i^\star \boldsymbol{R}_{i-1}\|_F^2 + \|\boldsymbol{W}_1 - \boldsymbol{R}_1^\top \boldsymbol{W}_1^\star\|_F^2. \tag{6}$$

Here the coefficient $\|\boldsymbol{Y}^\star\|^2$ is to harmonize the energy levels between $\mathrm{W} = \{\boldsymbol{W}_i\}_{i\geq 2}$ and $\boldsymbol{W}_1$. The following result elucidates the connection between $\mathrm{dist}^2(\mathrm{W}, \mathrm{W}^\star)$ and $\|\boldsymbol{Y} - \boldsymbol{Y}^\star\|_F^2$, guaranteeing the convergence of $\boldsymbol{Y}$ as W approaches the global minima.

Lemma 1: Assume a whitened input $\boldsymbol{X} \in \mathbb{R}^{d_x \times n}$, i.e. $\boldsymbol{X}\boldsymbol{X}^\top = \mathbf{I}_{d_x}$. Let $\boldsymbol{Y} = \boldsymbol{W}_N \cdots \boldsymbol{W}_1 \boldsymbol{X}$ and $\boldsymbol{Y}^\star = \boldsymbol{W}_N^\star \cdots \boldsymbol{W}_1^\star \boldsymbol{X}$, where $\boldsymbol{W}_i, \boldsymbol{W}_i^\star \in \mathbb{R}^{d_i \times d_{i-1}}$ are orthonormal for $i = 2, \ldots, N$. Given that $\|\boldsymbol{W}_1\|^2 \leq \frac{9\|\boldsymbol{Y}^\star\|^2}{4}$ and $\|\boldsymbol{W}_1^\star\|^2 = \|\boldsymbol{Y}^\star\|^2$, we can get

$$\|\boldsymbol{Y} - \boldsymbol{Y}^\star\|_F^2 \geq \frac{1}{(16N-8)\kappa^2(\boldsymbol{Y}^\star)} \mathrm{dist}^2(\mathrm{W}, \mathrm{W}^\star), \tag{7}$$

$$\|\boldsymbol{Y} - \boldsymbol{Y}^\star\|_F^2 \leq \frac{9N}{4} \mathrm{dist}^2(\mathrm{W}, \mathrm{W}^\star). \tag{8}$$

Proof: Using the result [29, eq. (E.4)], for any $j \geq 2$, we have that $\|\boldsymbol{W}_N \cdots \boldsymbol{W}_j - \boldsymbol{W}_N^\star \cdots \boldsymbol{W}_j^\star \boldsymbol{R}_{j-1}\|_F^2 \leq \frac{4\|\boldsymbol{Y} - \boldsymbol{Y}^\star\|_F^2}{\sigma_{\min}^2(\boldsymbol{Y}^\star)}$ for any $\boldsymbol{R}_i \in \mathbb{O}^{d_i \times d_i}$. It follows that

$$\begin{aligned}
&\|\boldsymbol{W}_{N-1} - \boldsymbol{R}_{N-1}^\top \boldsymbol{W}_{N-1}^\star \boldsymbol{R}_{N-2}\|_F^2 \\
&= \|\boldsymbol{W}_N^\star \boldsymbol{R}_{N-1} \boldsymbol{W}_{N-1} - \boldsymbol{W}_N^\star \boldsymbol{W}_{N-1}^\star \boldsymbol{R}_{N-2}\|_F^2 \\
&\leq 2\|\boldsymbol{W}_N^\star \boldsymbol{R}_{N-1} - \boldsymbol{W}_N\|_F^2 \|\boldsymbol{W}_{N-1}\|^2 \\
&\quad + 2\|\boldsymbol{W}_N \boldsymbol{W}_{N-1} - \boldsymbol{W}_N^\star \boldsymbol{W}_{N-1}^\star \boldsymbol{R}_{N-2}\|_F^2 \\
&\leq \frac{16\|\boldsymbol{Y} - \boldsymbol{Y}^\star\|_F^2}{\sigma_{\min}^2(\boldsymbol{Y}^\star)}.
\end{aligned} \tag{9}$$

[1] Here, for the sake of simplifying the subsequent analysis, we designate the teacher model as the ODLNN, which can encompass any linear model.

Similarly, we get $\|\boldsymbol{W}_i - \boldsymbol{R}_i^\top \boldsymbol{W}_i^\star \boldsymbol{R}_{i-1}\|_F^2 \le \frac{16\|\boldsymbol{Y}-\boldsymbol{Y}^\star\|_F^2}{\sigma_{\min}^2(\boldsymbol{Y}^\star)}$ for $i = 2, \ldots, N-2$. We now bound $\|\boldsymbol{W}_1 - \boldsymbol{R}_1^\top \boldsymbol{W}_1^\star\|_F^2$ by

$$\begin{aligned}
&\|\boldsymbol{W}_1 - \boldsymbol{R}_1^\top \boldsymbol{W}_1^\star\|_F^2 \\
&= \|\boldsymbol{W}_N^\star \cdots \boldsymbol{W}_2^\star \boldsymbol{R}_1 \boldsymbol{W}_1 - \boldsymbol{W}_N^\star \cdots \boldsymbol{W}_1^\star\|_F^2 \\
&\le 2\|\boldsymbol{W}_1\|^2 \|\boldsymbol{W}_N \cdots \boldsymbol{W}_2 - \boldsymbol{W}_N^\star \cdots \boldsymbol{W}_2^\star \boldsymbol{R}_1\|_F^2 \\
&\quad + 2\|\boldsymbol{W}_N \cdots \boldsymbol{W}_1 - \boldsymbol{W}_N^\star \cdots \boldsymbol{W}_1^\star\|_F^2 \\
&\le \frac{20\|\boldsymbol{Y}^\star\|^2 \|\boldsymbol{Y} - \boldsymbol{Y}^\star\|_F^2}{\sigma_{\min}^2(\boldsymbol{Y}^\star)}, \qquad (10)
\end{aligned}$$

where the second inequality uses the fact that $\boldsymbol{Y} = \boldsymbol{W}_N \cdots \boldsymbol{W}_1 \boldsymbol{X}$ and $\|\boldsymbol{A}\boldsymbol{X}\|_F = \|\boldsymbol{A}\|_F$ for any $\boldsymbol{A}$ since $\boldsymbol{X}$ is whitened. Based on the preceding discussion and the definition of $\mathrm{dist}^2(\mathrm{W}, \mathrm{W}^\star)$, we can conclude (7).

Finally, we can prove the other direction by

$$\begin{aligned}
&\|\boldsymbol{Y} - \boldsymbol{Y}^\star\|_F^2 \\
&= \left\| \sum_{i=1}^N \boldsymbol{W}_N^\star \cdots \boldsymbol{W}_{i+1}^\star \boldsymbol{R}_i (\boldsymbol{W}_i - \boldsymbol{R}_i^\top \boldsymbol{W}_i^\star \boldsymbol{R}_{i-1}) \boldsymbol{W}_{i-1} \cdots \boldsymbol{X} \right\|_F^2 \\
&\le N \Bigg(\sum_{i=2}^N \frac{9\|\boldsymbol{Y}^\star\|^2}{4} \|\boldsymbol{W}_i - \boldsymbol{R}_i^\top \boldsymbol{W}_i^\star \boldsymbol{R}_{i-1}\|_F^2 \\
&\quad + \|\boldsymbol{W}_1 - \boldsymbol{R}_1^\top \boldsymbol{W}_1^\star\|_F^2 \Bigg) \le \frac{9N}{4} \mathrm{dist}^2(\mathrm{W}, \mathrm{W}^\star). \qquad (11)
\end{aligned}$$

■

Main results: To establish the convergence rate of RGD, we require the loss function L to satisfy a certain property. Given that our primary focus is the analysis of the local convergence, we will assume that the loss function behaves well only in a local region. Specifically, we will consider a category of loss functions that satisfies the so-called restricted correlated gradient (RCG) condition [29]:

Definition 2: We say the loss function $L(\cdot; \boldsymbol{Y}^\star)$ satisfies RCG$(\alpha, \beta, \mathcal{C})$ condition for $\alpha, \beta > 0$ and the set $\mathcal{C}$ if

$$\begin{aligned}
&\langle \nabla L(\boldsymbol{Y}_1; \boldsymbol{Y}^\star) - \nabla L(\boldsymbol{Y}_2; \boldsymbol{Y}^\star), \boldsymbol{Y}_1 - \boldsymbol{Y}_2 \rangle \\
&\ge \alpha \|\boldsymbol{Y}_1 - \boldsymbol{Y}_2\|_F^2 + \beta \|\nabla L(\boldsymbol{Y}_1; \boldsymbol{Y}^\star) - \nabla L(\boldsymbol{Y}_2; \boldsymbol{Y}^\star)\|_F^2 \qquad (12)
\end{aligned}$$

for any $\boldsymbol{Y}_1, \boldsymbol{Y}_2 \in \mathcal{C}$.

The RCG condition is a generalization of the strong convexity. When L represents the MSE loss, i.e., $L(\boldsymbol{Y}, \boldsymbol{Y}^\star) = \|\boldsymbol{Y} - \boldsymbol{Y}^\star\|_F^2$, which is commonly used in the convergence analysis of training deep linear networks [22], [23], [31], it satisfies the RCG condition with $\alpha = \beta = 1$ and $\mathcal{C} = \mathbb{R}^{d_y \times n}$. The RCG condition may also accommodate other loss functions such as the cross entropy (CE) loss.

Based on Definition 2, we can initially deduce the Riemannian regularity condition as an extension of the regularity condition found in matrix factorization [32], [33], ensuring that gradients remain well-behaved within a defined region. Specifically, we have

Lemma 2: (Riemannian regularity condition) Suppose the training data $(\boldsymbol{X}, \boldsymbol{Y}^\star)$ is generated according to the data model in Definition 1. Also assume that the loss function L in (2) adheres to the RCG$(\alpha, \beta, \mathcal{C})$ condition where $\mathcal{C} \triangleq \{\boldsymbol{Y} : \|\boldsymbol{Y} - \boldsymbol{Y}^\star\|_F^2 \le \frac{\alpha\beta\sigma_{\min}^2(\boldsymbol{Y}^\star)}{72(2N-1)^2(N^2-1)\kappa^2(\boldsymbol{Y}^\star)}\}$. Under this assumption, for any $\mathrm{W} \in \{\mathrm{W} : \mathrm{dist}^2(\mathrm{W}, \mathrm{W}^\star) \le \frac{\alpha\beta\sigma_{\min}^2(\boldsymbol{Y}^\star)}{9(2N-1)(N^2-1)}\}$, the function g in (2) satisfies the Riemannian regularity condition as following:

$$\begin{aligned}
&\sum_{i=2}^N \langle \boldsymbol{W}_i - \boldsymbol{R}_i^\top \boldsymbol{W}_i^\star \boldsymbol{R}_{i-1}, \mathcal{P}_{\mathrm{T}_{\boldsymbol{W}_i}\mathrm{St}}(\nabla_{\boldsymbol{W}_i} g) \rangle \\
&\quad + \langle \boldsymbol{W}_1 - \boldsymbol{R}_1 \boldsymbol{W}_1^\star, \nabla_{\boldsymbol{W}_1} g \rangle \\
&\ge \frac{\alpha}{16(2N-1)\kappa^2(\boldsymbol{Y}^\star)} \mathrm{dist}^2(\mathrm{W}, \mathrm{W}^\star) + \frac{\beta}{(9N-5)\|\boldsymbol{Y}^\star\|^2} \\
&\quad \times \left(\sum_{i=2}^N \|\mathcal{P}_{\mathrm{T}_{\boldsymbol{W}_i}\mathrm{St}}(\nabla_{\boldsymbol{W}_i} g)\|_F^2 + \|\boldsymbol{Y}^\star\|^2 \|\nabla_{\boldsymbol{W}_1} g\|_F^2 \right). \qquad (13)
\end{aligned}$$

Proof: To begin with, we can derive $\|\boldsymbol{W}_1\|^2 \le 2\|\boldsymbol{W}_1^\star\|^2 + 2\|\boldsymbol{W}_1 - \boldsymbol{R}_1^\top \boldsymbol{W}_1^\star\|^2 \le 2\|\boldsymbol{Y}^\star\|^2 + 2\mathrm{dist}^2(\mathrm{W}, \mathrm{W}^\star) \le 2\|\boldsymbol{Y}^\star\|^2 + \frac{2\alpha\beta\sigma_{\min}^2(\boldsymbol{Y}^\star)}{9(2N-1)(N^2-1)} \le \frac{9\|\boldsymbol{Y}^\star\|^2}{4}$ where $\alpha\beta \le \frac{1}{4}$ [29] is used.

Next, through the gradients $\nabla_{\boldsymbol{W}_i} g = \boldsymbol{W}_{i+1}^\top \cdots \boldsymbol{W}_N^\top \nabla L(\boldsymbol{Y}; \boldsymbol{Y}^\star) \boldsymbol{X}^\top \cdots \boldsymbol{W}_{i-1}^\top, i \in [N]$, we need to derive

$$\begin{aligned}
\|\nabla_{\boldsymbol{W}_1} g\|_F^2 &\le \|\nabla L(\boldsymbol{Y}; \boldsymbol{Y}^\star) - \nabla L(\boldsymbol{Y}^\star; \boldsymbol{Y}^\star)\|_F^2, \qquad (14) \\
\|\nabla_{\boldsymbol{W}_i} g\|_F^2 &\le \|\boldsymbol{W}_1\|^2 \|\nabla L(\boldsymbol{Y}; \boldsymbol{Y}^\star)\|_F^2 \\
&\le \frac{9\|\boldsymbol{Y}^\star\|^2}{4} \|\nabla L(\boldsymbol{Y}; \boldsymbol{Y}^\star) - \nabla L(\boldsymbol{Y}^\star; \boldsymbol{Y}^\star)\|_F^2, \qquad (15)
\end{aligned}$$

where we employ $\nabla L(\boldsymbol{Y}^\star; \boldsymbol{Y}^\star) = 0$. Combing (14) and (15), we can obtain

$$\begin{aligned}
&\sum_{i=2}^N \|\mathcal{P}_{\mathrm{T}_{\boldsymbol{W}_i}\mathrm{St}}(\nabla_{\boldsymbol{W}_i} g)\|_F^2 + \|\boldsymbol{Y}^\star\|^2 \|\nabla_{\boldsymbol{W}_1} g\|_F^2 \\
&\le \sum_{i=2}^N \|\nabla_{\boldsymbol{W}_i} g\|_F^2 + \|\boldsymbol{Y}^\star\|^2 \|\nabla_{\boldsymbol{W}_1} g\|_F^2 \\
&\le \frac{9N-5}{4} \|\boldsymbol{Y}^\star\|^2 \|\nabla L(\boldsymbol{Y}; \boldsymbol{Y}^\star) - \nabla L(\boldsymbol{Y}^\star; \boldsymbol{Y}^\star)\|_F^2, \qquad (16)
\end{aligned}$$

in which the first inequality follows from the fact that for any matrix $\boldsymbol{B} = \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}(\boldsymbol{B}) + \mathcal{P}^\perp_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}(\boldsymbol{B})$ where $\mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}(\boldsymbol{B})$ and $\mathcal{P}^\perp_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}(\boldsymbol{B})$ are orthogonal, we have $\|\mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}(\boldsymbol{B})\|_F^2 \le \|\boldsymbol{B}\|_F^2$.

Before analyzing the lower bound of cross term in (12), we need to establish the upper bound for the inner product between the orthogonal complement and the gradient as following:

$$\begin{aligned}
T &= \sum_{i=2}^N \langle \mathcal{P}^\perp_{\mathrm{T}_{\boldsymbol{W}_i}\mathrm{St}}(\boldsymbol{W}_i - \boldsymbol{R}_i^\top \boldsymbol{W}_i^\star \boldsymbol{R}_{i-1}), \nabla_{\boldsymbol{W}_i} g \rangle \\
&\le \sum_{i=2}^N \frac{1}{2} \|\boldsymbol{W}_i\| \|\boldsymbol{W}_i - \boldsymbol{R}_i^\top \boldsymbol{W}_i^\star \boldsymbol{R}_{i-1}\|_F^2 \\
&\quad \times \|\nabla L(\boldsymbol{Y}; \boldsymbol{Y}^\star) - \nabla L(\boldsymbol{Y}^\star; \boldsymbol{Y}^\star)\|_F \|\boldsymbol{W}_1\| \\
&\le \frac{\beta}{4} \|\nabla L(\boldsymbol{Y}; \boldsymbol{Y}^\star) - \nabla L(\boldsymbol{Y}^\star; \boldsymbol{Y}^\star)\|_F^2 \\
&\quad + \frac{9(N-1)\|\boldsymbol{Y}^\star\|^2}{16\beta} \sum_{i=2}^N \|\boldsymbol{W}_i - \boldsymbol{R}_i^\top \boldsymbol{W}_i^\star \boldsymbol{R}_{i-1}\|_F^4 \\
&\le \frac{\beta}{4} \|\nabla L(\boldsymbol{Y}; \boldsymbol{Y}^\star) - \nabla L(\boldsymbol{Y}^\star; \boldsymbol{Y}^\star)\|_F^2 \\
&\quad + \frac{9(N-1)}{16\beta\|\boldsymbol{Y}^\star\|^2} \mathrm{dist}^4(\mathrm{W}, \mathrm{W}^\star), \qquad (17)
\end{aligned}$$

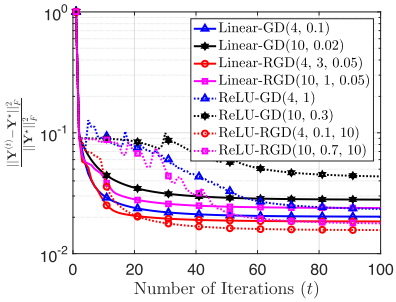

Fig. 1. Convergence analysis for GD(N, μ) and RGD(N, μ, γ) with different activation functions and N.

where $\mathcal{P}^{\perp}_{\mathrm{T}_{\boldsymbol{W}_i}\mathrm{St}}(\boldsymbol{W}_i - \boldsymbol{R}_i^\top \boldsymbol{W}_i^\star \boldsymbol{R}_{i-1}) = \frac{1}{2}\boldsymbol{W}_i(\boldsymbol{W}_i - \boldsymbol{R}_i^\top \boldsymbol{W}_i^\star \boldsymbol{R}_{i-1})^\top(\boldsymbol{W}_i - \boldsymbol{R}_i^\top \boldsymbol{W}_i^\star \boldsymbol{R}_{i-1})$ [30] and $\nabla L(\boldsymbol{Y}^\star;\boldsymbol{Y}^\star) = 0$.

Let us now introduce the notation $\boldsymbol{H} = \boldsymbol{Y}^\star - \boldsymbol{W}_N \cdots \boldsymbol{W}_2 \boldsymbol{R}_1^\top \boldsymbol{W}_1^\star \boldsymbol{X} + \sum_{i=1}^{N} \boldsymbol{W}_N \cdots \boldsymbol{W}_{i+1}(\boldsymbol{W}_i - \boldsymbol{R}_i^\top \boldsymbol{W}_i^\star \boldsymbol{R}_{i-1})\boldsymbol{W}_{i-1} \cdots \boldsymbol{W}_1 \boldsymbol{X}$, enabling us to simplify the expression of the cross term within (13). Then (13) can be rewritten as

$$
\begin{aligned}
&\sum_{i=2}^{N} \left\langle \boldsymbol{W}_i - \boldsymbol{R}_i^\top \boldsymbol{W}_i^\star \boldsymbol{R}_{i-1}, \mathcal{P}_{\mathrm{T}_{\boldsymbol{W}_i}\mathrm{St}}(\nabla_{\boldsymbol{W}_i} g) \right\rangle \\
&\quad + \langle \boldsymbol{W}_1 - \boldsymbol{R}_1^\top \boldsymbol{W}_1^\star, \nabla_{\boldsymbol{W}_1} g \rangle \\
&= \langle \nabla L(\boldsymbol{Y};\boldsymbol{Y}^\star) - \nabla L(\boldsymbol{Y}^\star;\boldsymbol{Y}^\star), \boldsymbol{Y} - \boldsymbol{Y}^\star + \boldsymbol{H} \rangle - T \\
&\geq \alpha \|\boldsymbol{Y} - \boldsymbol{Y}^\star\|_F^2 + \frac{\beta}{4} \|\nabla L(\boldsymbol{Y};\boldsymbol{Y}^\star) - \nabla L(\boldsymbol{Y}^\star;\boldsymbol{Y}^\star)\|_F^2 \\
&\quad - \frac{9(N^2-1)}{16\|\boldsymbol{Y}^\star\|^2} \mathrm{dist}^4(\mathrm{W}, \mathrm{W}^\star) \\
&\geq \frac{\alpha}{16(2N-1)\kappa^2(\boldsymbol{Y}^\star)} \mathrm{dist}^2(\mathrm{W}, \mathrm{W}^\star) + \frac{\beta}{(9N-5)\|\boldsymbol{Y}^\star\|^2} \\
&\quad \times \left(\sum_{i=2}^{N} \|\mathcal{P}_{\mathrm{T}_{\boldsymbol{W}_i}\mathrm{St}}(\nabla_{\boldsymbol{W}_i} g)\|_F^2 + \|\boldsymbol{Y}^\star\|^2 \|\nabla_{\boldsymbol{W}_1} g\|_F^2 \right), \quad (18)
\end{aligned}
$$

where the first inequality follows the RCG condition, (17) and $\|\boldsymbol{H}\|_F^2 \leq \frac{9N(N-1)}{8\|\boldsymbol{Y}^\star\|^2} \mathrm{dist}^4(\mathrm{W}, \mathrm{W}^\star)$ which is established through a mathematical transformation and the use of norm inequalities. The detailed proof for the upper bound of $\|\boldsymbol{H}\|_F^2$ has been omitted here due to space limitations. In the last line, we leverage (16), Lemma 1 and $\mathrm{dist}^2(\mathrm{W}, \mathrm{W}^\star) \leq \frac{\alpha\beta\sigma^2_{\min}(\boldsymbol{Y}^\star)}{9(2N-1)(N^2-1)}$. ■

We note that according to (7) in Lemma 1, the set $\{\mathrm{W} : \mathrm{dist}^2(\mathrm{W}, \mathrm{W}^\star) \leq \frac{\alpha\beta\sigma^2_{\min}(\boldsymbol{Y}^\star)}{9(2N-1)(N^2-1)}\}$ implies the region $\mathcal{C}$. By leveraging Riemannian regularity condition in Lemma 2 and the nonexpansiveness property of the polar decomposition-based retraction in [30, Lemma 1], we ultimately reach the following conclusion:

Theorem 1: In accordance with the identical conditions outlined in Lemma 2, we assume the initialization satisfies $\mathrm{dist}^2(\mathrm{W}^{(0)}, \mathrm{W}^\star) \leq \frac{\alpha\beta\sigma^2_{\min}(\boldsymbol{Y}^\star)}{9(2N-1)(N^2-1)}$. When employing the learning rate $\mu \leq \frac{2\beta}{(9N-5)\|\boldsymbol{Y}^\star\|^2}$ and $\gamma = \|\boldsymbol{Y}^\star\|^2$ in RGD (5), we have

$$
\mathrm{dist}^2(\mathrm{W}^{(t+1)}, \mathrm{W}^\star) \leq \left(1 - \frac{\alpha\sigma^2_{\min}(\boldsymbol{Y}^\star)\mu}{8(2N-1)}\right) \mathrm{dist}^2(\mathrm{W}^{(t)}, \mathrm{W}^\star).
$$

Our results reveal that the RGD demonstrates a linear convergence rate with polynomial decay concerning N. In addition, by Lemma 1, we can easily obtain $\|\boldsymbol{Y}^{(t)} - \boldsymbol{Y}^\star\|_F^2 \leq (1 - \frac{\alpha\sigma^2_{\min}(\boldsymbol{Y}^\star)\mu}{8(2N-1)})^t \frac{\alpha\beta\sigma^2_{\min}(\boldsymbol{Y}^\star)}{8N^2-12N+4}$. It is worth noting that, through analogous analysis, Theorem 1 can be extended to a broader scenario wherein $\boldsymbol{W}_j^\star$ takes on an arbitrary matrix, and matrices $\boldsymbol{W}_i^\star$ exhibit row orthogonality for $i < j$ and column orthogonality for $i > j$. Moreover, we emphasize that our focus is primarily on the local convergence property of the RGD, and does not cover initialization methods extensively investigated in prior research, such as those discussed in [2], [21], [34], [35].

The research most closely related to our work is the convergence analysis of gradient descent in deep linear neural models across multiple layers, as demonstrated in [22], wherein the MSE loss function is taken into account. It is established that $\|\boldsymbol{Y}^{(t)} - \boldsymbol{Y}^\star\|_F^2 \leq \epsilon$ can be deduced when $t \geq \Omega(\frac{N^3\|\boldsymbol{Y}^\star\|_F^6}{c^2} \log(\frac{1}{\epsilon}))$, provided that $c \leq \sigma_{\min}(\boldsymbol{Y}^\star)$. Applying a similar derivation as presented in [22, Theorem 1], utilizing both Theorem 1 and Lemma 1, we can infer that $t \geq \Omega(N^2\kappa^2(\boldsymbol{Y}^\star)\log(\frac{1}{\epsilon}))$ in the RGD is sufficient to meet the same requirement. This further highlights the convergence advantage of the RGD.

IV. Experiments

In this section, we conduct experiments to compare the performance of the RGD with gradient descent (GD). Specifically, we concentrate on the multi-class classification task using the MNIST dataset, where the input feature dimension is 784, while the output feature is 10, represented as $\boldsymbol{y}_i \in \mathbb{R}^{10}$. In this representation, each $\boldsymbol{y}_i$ is designed such that it holds a value of 1 solely at the position that aligns with its categorical label, leaving the other positions assigned to 0. For our model architecture, we can deploy an multi-layer perceptron (MLP) with N layers where $\boldsymbol{W}_1 \in \mathbb{R}^{100\times784}$, $\{\boldsymbol{W}_i\}_{i=2}^{N-2} \in \mathbb{R}^{100\times100}$, $\boldsymbol{W}_{N-1} \in \mathbb{R}^{50\times100}$, and $\boldsymbol{W}_N \in \mathbb{R}^{10\times50}$. Each layer in the MLP is connected by a linear or rectified linear unit (Relu) activation function. While our theoretical result is only established for linear networks, we will also test the performance on a nonlinear MLP without bias terms.

We apply the GD and RGD for the CE loss function [36] in combination with the softmax function to train MLP models. The weight matrices are initialized using orthogonal initialization [2], except for $\boldsymbol{W}_1$ in the RGD, which follows a uniform distribution within the range of $(-\frac{1}{\sqrt{784}}, \frac{1}{\sqrt{784}})$ [34]. In addition, we perform a grid search to fine-tune the hyperparameters (μ and γ).

In Fig. 1, it is evident that RGD achieves a quicker convergence compared to GD. Notably, with an increase in the number of layers, the convergence rate decreases in alignment with our theoretical analysis. Moreover, due to the impact of nonlinear activation functions, algorithms that employ ReLU demonstrate a comparatively slower convergence rate compared to those utilizing linear activation functions. However, despite this, the error of the RGD with Relu outperforms that of the RGD using a linear activation function.

V. Conclusion

In this letter, we have provided a convergence analysis of the Riemannian gradient descent for a specific class of loss functions within orthonormal deep linear neural networks. Remarkably, our analysis guarantees a linear convergence rate, provided appropriate initialization. This will serve as a stepping stone for future explorations of training nonlinear orthonormal deep neural networks with adaptive learning rates.

REFERENCES

[1] D. Mishkin and J. Matas, "All you need is a good init," 2015, *arXiv:1511.06422*.

[2] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," 2013, *arXiv:1312.6120*.

[3] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[4] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1310–1318.

[5] Q. V. Le, N. Jaitly, and G. E. Hinton, "A simple way to initialize recurrent networks of rectified linear units," 2015, *arXiv:1504.00941*.

[6] M. Arjovsky, A. Shah, and Y. Bengio, "Unitary evolution recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1120–1128.

[7] B. Hanin, "Which neural net architectures give rise to exploding and vanishing gradients?," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, vol. 31, pp. 580–589.

[8] M. Harandi and B. Fernando, "Generalized backpropagation Etude De Cas: Orthogonality," 2016, *arXiv:1611.05927*.

[9] S. Li, K. Jia, Y. Wen, T. Liu, and D. Tao, "Orthogonal deep neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1352–1368, Apr. 2021.

[10] L. Huang et al., "Controllable orthogonalization in training DNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6429–6438.

[11] J. Li, F. Li, and S. Todorovic, "Efficient Riemannian optimization on the Stiefel manifold via the cayley transform," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–16.

[12] J. Wang, Y. Chen, R. Chakraborty, and S. X. Yu, "Orthogonal convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11505–11515.

[13] F. Malgouyres and F. Mamalet, "Existence, stability and scalability of orthogonal convolutional neural networks," *J. Mach. Learn. Res.*, vol. 23, pp. 15743–15798, 2022.

[14] V. Dorobantu, P. A. Stromhaug, and J. Renteria, "DizzyRNN: Reparameterizing recurrent neural networks for norm-preserving backpropagation," 2016, *arXiv:1612.04035*.

[15] Z. Mhammedi, A. Hellicar, A. Rahman, and J. Bailey, "Efficient orthogonal parametrisation of recurrent neural networks using householder reflections," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2401–2409.

[16] E. Vorontsov, C. Trabelsi, S. Kadoury, and C. Pal, "On orthogonality and learning recurrent networks with long term dependencies," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3570–3578.

[17] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier, "Parseval networks: Improving robustness to adversarial examples," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 854–863.

[18] M. Cogswell, F. Ahmed, R. Girshick, L. Zitnick, and D. Batra, "Reducing overfitting in deep networks by decorrelating representations," 2015, *arXiv:1511.06068*.

[19] N. Bansal, X. Chen, and Z. Wang, "Can we gain more from orthogonality regularizations in training deep networks?," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, vol. 31, pp. 4266–4276.

[20] L. Huang, X. Liu, B. Lang, A. Yu, Y. Wang, and B. Li, "Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 3271–3278.

[21] P. Bartlett, D. Helmbold, and P. Long, "Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 521–530.

[22] S. Arora, N. Cohen, N. Golowich, and W. Hu, "A convergence analysis of gradient descent for deep linear neural networks," 2018, *arXiv:1810.02281*.

[23] D. Zou, P. M. Long, and Q. Gu, "On the global convergence of training deep linear ResNets," 2020, *arXiv:2003.01094*.

[24] O. Shamir, "Exponential convergence time of gradient descent for one-dimensional deep linear neural networks," in *Proc. Conf. Learn. Theory*, 2019, pp. 2691–2713.

[25] Z. Allen-Zhu, Y. Li, and Z. Song, "On the convergence rate of training recurrent neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 6676–6688.

[26] S. Chatterjee, "Convergence of gradient descent for deep neural networks," 2022, *arXiv:2203.16462*.

[27] M. Zhou, R. Ge, and C. Jin, "A local convergence theory for mildly over-parameterized two-layer neural network," in *Proc. Conf. Learn. Theory*, 2021, pp. 4577–4632.

[28] X. Zhang, Y. Yu, L. Wang, and Q. Gu, "Learning one-hidden-layer relu networks via gradient descent," in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, 2019, pp. 1524–1534.

[29] R. Han, R. Willett, and A. R. Zhang, "An optimal statistical and computational framework for generalized tensor estimation," 2020, *arXiv:2002.11255*.

[30] X. Li, S. Chen, Z. Deng, Q. Qu, Z. Zhu, and A. Man-Cho So, "Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods," *SIAM J. Optim.*, vol. 31, no. 3, pp. 1605–1634, 2021.

[31] Z. Zhu, D. Soudry, Y. C. Eldar, and M. B. Wakin, "The global optimization geometry of shallow linear neural networks," *J. Math. Imag. Vis.*, vol. 62, pp. 279–292, 2020.

[32] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 964–973.

[33] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, "The global optimization geometry of low-rank matrix optimization," *IEEE Trans. Inf. Theory*, vol. 67, no. 2, pp. 1308–1331, Feb. 2021.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[35] L. Xiao, Y. Bahri, J. Sohl-Dickstein, S. Schoenholz, and J. Pennington, "Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5393–5402.

[36] J. Zhou et al., "Are all losses created equal: A neural collapse perspective," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 31697–31710.