

# BLOCK COORDINATE DESCENT FOR NEURAL NETWORKS PROVABLY FINDS GLOBAL MINIMA

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

In this paper, we consider a block coordinate descent (BCD) algorithm for training deep neural networks and provide a new global convergence guarantee under strictly monotonically increasing activation functions. While existing works demonstrate convergence to stationary points for BCD in neural networks, our contribution is the first to prove convergence to global minima, ensuring arbitrarily small loss. We show that the loss with respect to the output layer decreases exponentially while the loss with respect to the hidden layers remains well-controlled. Additionally, we derive generalization bounds using the Rademacher complexity framework, demonstrating that BCD not only achieves strong optimization guarantees but also provides favorable generalization performance. Moreover, we propose a modified BCD algorithm with skip connections and non-negative projection, extending our convergence guarantees to ReLU activation, which are not strictly monotonic. Empirical experiments confirm our theoretical findings, showing that the BCD algorithm achieves a small loss for strictly monotonic and ReLU activations.

## 1 INTRODUCTION

Deep learning has led to significant advances across various domains, such as computer vision, natural language processing, and reinforcement learning, achieving unprecedented performance in numerous tasks. However, understanding the training dynamics and optimization behavior of deep neural networks remains an ongoing challenge due to the highly non-convex nature of their loss functions (Li et al., 2018). Proving convergence to global minima of gradient descent via backpropagation, particularly for deep networks with multiple layers, remains an open problem in the field. While the neural tangent kernel (NTK) regime (Jacot et al., 2018) addresses this problem by reducing the non-convex loss to the convex one in RKHS, it fails to fully explain the empirical success of deep learning because it often outperforms kernel methods, even if we employ NTK as the kernel.

Contrary to the backpropagation-based training, the block coordinate descent (BCD), which originated from the mathematical optimization field (see Tseng (2001), for example), is an optimization framework where we divide a variable into several blocks and optimize them alternately. BCD offers computational advantages by updating subsets of parameters iteratively, allowing for tractable optimization of complex systems. The objective function appearing in the neural network training is also highly non-convex, and to overcome this issue, BCD-based neural network optimization methods have been proposed (Carreira-Perpinan & Wang, 2014; Askari et al., 2018; Lau et al., 2018; Zhang & Brand, 2017; Patel et al., 2020; Zeng et al., 2019; Nakamura et al., 2021; Qiao et al., 2021; Zhang et al., 2022; Xu et al., 2024). When we apply BCD to neural network training, the most natural way is that we regard the weights of each layer as a block, and existing works adopt this way. By the formulation of BCD, the loss function of the neural network can be divided into several components, one of which coincides with a loss with respect to a layer. Compared to the original loss, these divided ones have more accessible landscapes to optimize.

Based on such an advantage of BCD for neural networks, its theoretical perspective, mainly about its convergence guarantee, has been explored in recent years. However, existing theoretical works on BCD for neural networks (Zhang & Brand, 2017; Zeng et al., 2019; Zhang et al., 2022; Xu et al., 2024) have only focused on the convergence to stationary points, points with zero gradients. Convergence to stationary points does not imply convergence to global minima, especially when

the objective function is highly non-convex, such as the loss that appears in the training of neural networks (Li et al., 2018; Safran & Shamir, 2018).

How neural network training finds global minima has been one of the most significant topics in deep learning theory literature. However, existing guarantees on BCD remain in convergence to the stationary points. To bridge this gap, we aim to provide the convergence guarantee to the global minima of BCD for neural networks. To this end, we consider multi-layer neural networks and employ a BCD-type algorithm, updating the parameters using vanilla gradient descent. Our contribution can be summarized as follows:

- We prove the global convergence of a block coordinate descent (BCD) algorithm, where we train deep neural network models with strictly monotonically increasing activation. We ensure that the parameters attain arbitrarily small loss by proving that (i) the loss with respect to the output layer will decrease exponentially to zero and (ii) the loss with respect to the hidden layers remains small in every iteration. Through the analysis, we carefully evaluate the difference propagated from the output layer to the input layer. To the best of our knowledge, this is the first result that guarantees convergence to the global minima of neural networks with any number of layers beyond the NTK regime.
- We derive a generalization error bound of deep neural networks trained by BCD under settings with i.i.d. data. In the convergence analysis, we show that the norm of weight matrices of each layer can be bounded by a constant. Combining this and the Rademacher complexity argument from Bartlett et al. (2017) gives an upper bound on generalization error. Compared to the existing works on gradient descent, BCD enables us to provide the generalization gap bound for multi-layer neural networks with an optimization guarantee.
- A notable challenge in applying our approach to commonly used activation functions like ReLU is their non-monotonic nature. Since ReLU is not strictly monotonically increasing, our initial convergence result does not directly apply. To address this issue, we propose a modified BCD algorithm incorporating skip connections (He et al., 2016) and non-negative projection updates. This modification ensures that convergence guarantees extend to ReLU networks, thereby broadening the applicability of our method to real-world architectures that predominantly use ReLU activations.
- We validate our theoretical findings through numerical experiments, showing that BCD for both strictly monotonic and ReLU activations achieves arbitrarily small loss values. These empirical results confirm the practical viability of our proposed methods, demonstrating their effectiveness in optimizing deep neural networks beyond theoretical guarantees.

## 1.1 OTHER RELATED WORKS

**Convergence guarantee of GD/SGD for neural networks** In recent years, theoretical works on the convergence guarantee of (stochastic) gradient descent for neural networks have been intensively investigated. In the neural tangent kernel (NTK) regime (Jacot et al., 2018; Allen-Zhu et al., 2019b; Arora et al., 2019; Du et al., 2019; Zou et al., 2020), to name a few, the training dynamics of deep neural networks can be approximated by the gradient descent in RKHS. While we can ensure its global convergence by exploiting the convexity, the *feature learning ability* of neural networks, which is considered one of the critical ingredients of the practical success of deep learning, is not reflected since the training dynamics are reduced to the kernel method. For example, the parameters of networks trained by the NTK regime hardly move from their initial points as the number of parameters increases. On the other hand, our analysis does not fall into such a situation. Moreover, our analysis does not require any *overparameterization* on hidden layers to ensure global convergence.

The mean-field (MF) regime (Nitanda & Suzuki, 2017; Chizat & Bach, 2018; Mei et al., 2019; Tzen & Raginsky, 2020; Pham & Nguyen, 2021; Nguyen & Pham, 2023) is another promising approach of investigating neural network training. It regards the training of parameters as that of (probability) measure over the parameters, by which we can convert the non-convex optimization with respect to the parameters to the convex one where the distribution of parameters itself is a variable to be optimized. While several studies ensure its global convergence by employing this convexity without loss of feature learning ability, most of their models only focus on two or three-layer networks, our analysis admits the any number of hidden layers.

More recently, Banerjee et al. (2023) proposed restricted strong convexity (RSC) to analyze neural network training, which derives the global convergence guarantee by assuming that the gradient and output of neural networks correlate with each other during the training. However, Banerjee et al. (2023) still requires an analysis of this correlation assumption and does not fully explain the nature of global convergence in the training.

**Generalization error bound of multi-layer neural networks** Investigation of generalization error analysis for multi-layer neural networks has been explored in recent years (Neyshabur et al., 2015; Wei & Ma, 2019; Bartlett et al., 2017; Neyshabur et al., 2017; Golowich et al., 2018; Bartlett et al., 2019; Arora et al., 2018; Suzuki et al., 2020). These works give a generalization error by evaluating the complexity of neural networks from various perspective, such as the VC-dimension, the norm of parameters of networks, and so on. On the other hand, most of these results do not consider the optimization, but we also demonstrate the global convergence guarantee. Moreover, several works on generalization error analysis go beyond two-layer networks. However, most focus only on three-layer networks (Allen-Zhu & Li, 2019; Allen-Zhu et al., 2019a).

## 2 PRELIMINARIES

### 2.1 NOTATIONS

For an integer  $n$ , we define  $[n] := \{1, \dots, n\}$ . For  $x \in \mathbb{R}^d$ ,  $\|x\|$  denotes its Euclidean norm. We denote the  $d$ -dimensional identity matrix by  $I_d$ . For  $A \in \mathbb{R}^{n \times m}$ ,  $\|A\|_F := \sqrt{\sum_{i,j} A_{ij}^2}$  denotes its Frobenius norm, and  $\|A\|_{op} := \max_{\|x\| \leq 1} \|Ax\|$  denotes its operator norm. For two symmetric matrices  $A$  and  $B$ , we denote  $A \preceq B$  ( $A \preceq B$ ) if and only if the matrix  $B - A$  is positive (non-negative) definite. For  $x = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$ ,  $\text{diag}(x) \in \mathbb{R}^{d \times d}$  denotes a diagonal matrix whose  $j$ -th diagonal component is  $x_j$ .

### 2.2 PROBLEM SETTINGS

Here, we introduce problem settings we consider in this paper. We observe  $n$  training examples  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ , where  $x_i \in \mathbb{R}^{d_{in}}$  is a feature vector and  $y_i \in \mathbb{R}^{d_{out}}$  is a label. Let  $X = (x_1 \dots x_n)^\top \in \mathbb{R}^{n \times d_{in}}$ . Throughout the analysis, we consider high-dimensional settings  $n \leq d_{in}$ . Moreover, we make an assumption about the matrix  $X$  as follows:

**Assumption 1** (Data matrix is full row rank).  $\text{rank}(X) = n$ .

This assumption is required to show the global convergence. As we will see in the proof of the main result, we cannot ensure the existence of global minima without Assumption 1.

A multi-layer neural network is defined by

$$f_{NN}(x) := W_L \sigma(W_{L-1} \sigma(\dots W_2 \sigma(W_1 x)) \dots),$$

where  $\sigma$  is element-wise activation and  $W_1 \in \mathbb{R}^{r \times d_{in}}$ ,  $W_j \in \mathbb{R}^{r \times r}$  for  $j \in \{2, \dots, L-1\}$ , and  $W_L \in \mathbb{R}^{d_{out} \times r}$ . We consider that all the hidden layers have the same width  $r$ .

Then, we make the following assumption on the activation function.

**Assumption 2** (Activation).  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  is monotonically increasing and satisfies  $\sigma(0) = 0$ . Especially, there exists a constant  $0 < \alpha < 2$  such that  $\inf_{x \in \mathbb{R}} \sigma'(x) \geq \alpha$  holds<sup>1</sup>. Moreover,  $\sigma$  is  $\ell$ -Lipschitz, i.e., for any  $u_1, u_2 \in \mathbb{R}$ ,  $|\sigma(u_1) - \sigma(u_2)| \leq \ell |u_1 - u_2|$  holds.

A typical example of activation function satisfying Assumption 2 is LeakyReLU activation  $x \mapsto \max\{x, ax\}$  ( $a < 1$ ): which satisfies Assumption 2 with  $\alpha = a$  and  $\ell = 1$ . We note that other activation, such as ReLU  $x \mapsto \max\{x, 0\}$ , does not satisfy Assumption 2. We also provide the global convergence algorithm when we use the ReLU activation in Section 5.

Under this formulation of neural networks, we formalize the regression problem

$$\min_W \sum_{i=1}^n (f_{NN}(x_i) - y_i)^2, \quad (1)$$

<sup>1</sup>If  $\sigma$  is not differentiable, we assume that  $\sigma(x_1) - \sigma(x_2) \geq \alpha(x_1 - x_2)$  for any  $x_1, x_2 \in \mathbb{R}$ .

where  $\mathbf{W} = (W_1, \dots, W_L)$ . One of the most straightforward approaches to solve (1) is (stochastic) gradient method, in which the parameters are updated using the loss gradient. Conversely, we employ a layer-wise optimization method called *block coordinate descent*, as we introduce in the following section.

### 3 BLOCK COORDINATE DESCENT

In this section, after we introduce the basic notion of the *block coordinate descent* (BCD), we provide the algorithm we consider in this paper. BCD, which originated from the mathematical optimization field (see Tseng (2001), for example), is an optimization framework where we divide a variable into several blocks and optimize them alternately.

In BCD, instead of directly utilizing the loss (1), we introduce auxiliary parameters  $V_{1,i} \dots V_{L,i}$ .  $V_{j,i}$  aims to approximate the output of  $j$ -th layer for the  $i$ -th sample  $x_i$ . By construction, we have  $V_{j,i} \in \mathbb{R}^r$  for  $j = 1, \dots, L-1$ . By using these auxiliary parameters, we reformulate (1) as follows:

$$\min_{\mathbf{W}, \mathbf{V}} F(\mathbf{W}, \mathbf{V}) := \sum_{i=1}^n \left[ \|W_L V_{L-1,i} - y_i\|^2 + \gamma \sum_{j=1}^{L-1} \|\sigma(W_j V_{j-1,i}) - V_{j,i}\|^2 \right], \quad (2)$$

where  $\gamma > 0$  is a hyperparameter and we denote  $V_{0,i} := x_i$ ,  $\mathbf{W} = (W_1, \dots, W_L)$ , and  $\mathbf{V} = (V_{1,1}, \dots, V_{L-1,n})$ . In the reformulated problem (2), the second term represents the loss at the  $j$ -th layer, indicating how  $V_{j,i}$  approximates the output of the layer given the input  $x_i$ . The first term represents the loss at the output layer, showing how close the outputs of the network with the approximated (hidden layer) output  $V_{j,i}$  are to the training labels  $y_1, \dots, y_n$ . By the construction, if  $(\mathbf{W}^*, \mathbf{V}^*)$  satisfies  $F = 0$  in (2),  $\mathbf{W}^*$  is the optimal solution of (1).

One of the benefits of the reformulation (2) is that we can treat the objective function with respect to the weights of each layer  $(W_1, \dots, W_L)$  separately. Such a simplification results not only in a faster implementation (e.g., parallelization) but also a favorable loss landscape, including theoretical tractability. While various methods for optimizing (2) have been explored, we consider a relatively simple one, updating weights  $W_j$  and auxiliary variables  $V_{j,i}$  sequentially from the output layer. Specifically, we update the variables in order  $W_L \rightarrow V_{L-1,i} \rightarrow W_{L-1} \rightarrow \dots V_{1,i} \rightarrow W_1$  by using the objective function (2). We summarize the algorithm considered in this paper in Algorithm 1. From now on, we explain its detailed procedure.

---

#### Algorithm 1: Block coordinate descent

---

**input :**  $(W_1)_{ab} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1/d_{in})$ ,  $(W_j)_{ab} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1/r)$  for all  $j = 2, \dots, L$ ,  $V_{0,i} = x_i$ .

1  $K$ : max outer iteration,  $K_V$ ,  $K_W$ : max inner iteration,  $\eta_V$ ,  $\eta_W^{(1)}$ ,  $\eta_W^{(2)}$ : step size;

2  $W_j \leftarrow$  output of Algorithm 2 with inputs  $s_1, s_2$ , and  $W_j$  for  $j = 2, \dots, L$ ;

3  $V_{j,i} \leftarrow \sigma(W_j V_{j-1,i})$  for all  $j = 1, \dots, L-1$  and  $i = 1, \dots, n$ ;

4 **for**  $k \leftarrow 1$  **to**  $K$  **do**

5      $W_L \leftarrow W_L - \eta_W^{(1)} \nabla_{W_L} \sum_{i=1}^n \|W_L V_{L-1,i} - y_i\|^2$ ;

6     **for**  $i \leftarrow 1$  **to**  $n$  **do**

7          $V_{L-1,i} \leftarrow V_{L-1,i} - \eta_V \nabla_{V_{L-1,i}} \|W_L V_{L-1,i} - y_i\|^2$ ;

8     **for**  $j \leftarrow L-1$  **to** 2 **do**

9          $W_j \leftarrow W_j - \gamma \eta_W^{(1)} \sum_{i=1}^n \nabla_{W_j} \|\sigma(W_j V_{j,i}) - V_{j+1,i}\|^2$ ;

10        **for**  $i \leftarrow 1$  **to**  $n$  **do**

11            **for**  $k_{inner} \leftarrow 1$  **to**  $K_V$  **do**

12                 $V_{j-1,i} \leftarrow V_{j-1,i} - \gamma \eta_V \nabla_{V_{j-1,i}} \|\sigma(W_j V_{j-1,i}) - V_{j,i}\|^2$ ;

13        **for**  $k_{inner} \leftarrow 1$  **to**  $K_W$  **do**

14             $W_1 \leftarrow W_1 - \gamma \eta_W^{(2)} \sum_{i=1}^n \nabla_{W_1} \|\sigma(W_1 V_{0,i}) - V_{1,i}\|^2$ ;

---

**Initialization** We consider Gaussian initialization for  $W_j$ ; that is, each element of  $W_1$  is sampled from  $\mathcal{N}(0, d_{in}^{-1})$ , and each element of  $W_j$  ( $j = 2, \dots, L$ ) is sampled from  $\mathcal{N}(0, r^{-1})$ . After that, we apply *singular value bounding* (SVB) (Jia et al., 2017) to  $W_j$  ( $j = 2, \dots, L$ ). In SVB, we

**Algorithm 2:** Singular Value Bounding

---

**input :**  $W_j$ : matrix,  $(s_1, s_2)$ : lower and upper bounds on the singular values  
1  $(U, \Sigma, V) \leftarrow$ : Singular value decomposition of  $W_j = U\Sigma V$ ;  
2 **for**  $s \leftarrow$  diagonal components of  $\Sigma$  **do**  
3    $s \leftarrow \max\{s_1, \min\{s_2, s\}\}$ ;  
**output:**  $U\Sigma V$

---

conduct the singular value decomposition of  $W_j$  as  $W_j = U\Sigma V^\top$ , where  $U$  and  $V$  are orthogonal matrices, and  $\Sigma$  is a non-negative diagonal matrix. Since  $W_j$  is full-rank with probability 1 over the initialization, we also have  $\Sigma \in \mathbb{R}^{r \times r}$  with probability 1. After SVB, we adjust each diagonal component of  $\Sigma$  to be within the interval  $[s_1, s_2]$ . Then, we utilize  $W_j = U\Sigma'V^\top$  as the initial parameter of  $W_j$ , where  $\Sigma'$  be the matrix obtained by the adjustment. We summarize this procedure in Algorithm 2.

In Jia et al. (2017), SVB is conducted at every epoch to enhance the stability of the training and prediction performance of stochastic gradient descent. The upper and lower bounding of the singular value prevents the amplifying or vanishing of a gradient in the backpropagation. Applying SVB also has several advantages in BCD, not only for practical reasons but also from a theoretical perspective. First, the regularity of  $W_j$  results in a preferable condition number of the objective function  $\|\sigma(W_j V_{j-1,i}) - V_{j,i}\|^2$  in  $F$ , the loss at the  $j$ -th layer. Moreover, the upper bound on the singular value prevents  $V_j$  from becoming extremely large at the initialization.

**Remark 3.1.** While Jia et al. (2017) applies SVB at every epoch, we use it only at the initialization. By setting the step size not too large, we can ensure that all the singular values of  $W_j$  remain in a bounded interval, as we show in the proof, with which we enjoy the same benefit throughout the training.

After initializing  $W_j$ , we initialize  $V_j$  in an exact manner, i.e.,  $V_{j,i} = \sigma(W_j V_{j-1,i})$  for all  $j = 1, \dots, L-1$  and  $i = 1, \dots, n$ . While we can employ any initialization scheme for  $V_j$ , the exact manner results in  $\|\sigma(W_j V_{j-1,i}) - V_{j,i}\|^2 = 0$  at the initialization, leading to faster convergence.

**Update of  $V$**  For optimizing  $W$  and  $V$ , we utilize vanilla gradient descent. We employ a common step size  $\eta_V$  for each  $V_{j,i}$  and perform multiple updates using the loss  $\|\sigma(W_j V_{j-1,i}) - V_{j,i}\|^2$  (line 6, 12), given by

$$V_{L-1,i} \leftarrow V_{L-1,i} - \eta_V \nabla_{V_{L-1,i}} \|W_L V_{L-1,i} - y_i\|^2 \quad (3)$$

and

$$V_{j-1,i} \leftarrow V_{j-1,i} - \gamma \eta_V \nabla_{V_{j-1,i}} \|\sigma(W_j V_{j-1,i}) - V_{j,i}\|^2.$$

The first update (3) can be interpreted as solving the linear equation  $W_L V_{L-1,i} = y_i$ , which has a solution if the matrix  $W_L$  is full row rank. We assume that the activation satisfies Assumption 2. In this case, since the mapping  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  is a bijection, there exists an inverse map  $\sigma^{-1}$ , and training  $V_j$  can be viewed as equivalent to solving the linear equation  $W_{j+1} V_{j,i} = \sigma^{-1}(V_{j+1,i})$ . Therefore, it is expected that  $V_{j,i}$  converges to the solution via gradient descent with a suitable choice of  $\eta_V$  as long as the matrix  $W_j \in \mathbb{R}^{r \times r}$  is regular.

**Update of  $W$**  For the update of  $W_j$  ( $j = 1, \dots, L$ ), we use the loss function at  $j$ -th layer, that is,  $\sum_{i=1}^n \|W_L V_{L-1,i} - y_i\|^2$  for  $W_L$  and  $\sum_{i=1}^n \|\sigma(W_j V_{j-1,i}) - V_{j,i}\|^2$  for  $W_j$  ( $j = 1, \dots, L-1$ ).

For  $W_2, \dots, W_L$ , we use a common step size  $\eta_W^{(1)}$  and conduct the gradient descent update:

$$W_L \leftarrow W_L - \eta_W^{(1)} \nabla_{W_L} \sum_{i=1}^n \|W_L V_{L-1,i} - y_i\|^2,$$

and

$$W_j \leftarrow W_j - \gamma \eta_W^{(1)} \sum_{i=1}^n \nabla_{W_j} \|\sigma(W_j V_{j,i}) - V_{j+1,i}\|^2$$

for each iteration (line 5, 9). For  $W_1$ , we employ a different step size  $\eta_W^{(2)}$  and apply

$$W_1 \leftarrow W_1 - \gamma \eta_W^{(2)} \sum_{i=1}^n \nabla_{W_1} \|\sigma(W_1 V_{0,i}) - V_{1,i}\|^2,$$

multiple ( $K_W$ ) times for each iteration (line 14). These update manners are required to attain the global convergence. With respect to the loss of the second to  $L$ -th layer, we update both  $W_j$  and  $V_{j-1,i}$ . In particular, by applying multiple updates to  $V_{j-1,i}$ , we can ensure linear convergence of the loss  $\sum_{i=1}^n \|\sigma(W_j V_{j-1,i}) - V_{j,i}\|^2$  for each iteration while the singular values of matrix  $W_j$  are upper and lower bounded. On the other hand, the existence of  $W^*$  satisfying  $\sum_{i=1}^n \|\sigma(W^* V_{j-1,i}) - V_{j,i}\|^2 = 0$  is not ensured, particularly in the case where  $n > r$ . Hence, it is not necessary to update  $W_j$  for multiple times. Furthermore, as the number of iterations increases, it becomes less likely to maintain the regularity of the matrix  $W_j$ . This is why we only apply gradient descent once to  $W_j$  ( $j = 2, \dots, L$ ). On the other hand, in the first layer, the input  $V_{0,i} = x_i$  is fixed, and we need to demonstrate linear convergence of the loss  $\sum_{i=1}^n \|\sigma(W_1 V_{0,i}) - V_{1,i}\|^2$  through the update of  $W_1$ . In the overparameterized setting  $d_{in} \geq n$ , if the data matrix satisfies  $\text{rank}(U) = n$ , we can ensure the existence of a global minima  $W^*$  satisfying  $\sum_{i=1}^n \|\sigma(W^* V_{0,i}) - V_{1,i}\|^2 = 0$ , and hence linear convergence under a suitable choice of  $\eta_W^{(2)}$ .

**Remark 3.2.** Concerning the recent progress of the block coordinate descent algorithms applied to deep learning, as represented by (Jia et al., 2017; Zhang & Brand, 2017; Lau et al., 2018; Patel et al., 2020), among others, we employ a relatively simple approach using vanilla gradient descent without any regularization, focusing on devising the loss function and the order in which the parameters are updated. While our convergence proof is based on this specific setup, our analysis can be extended to encompass more complex scenarios. Our algorithm is adaptable to different settings, including potential applications to other loss functions and problems, such as classification problems, and the inclusion of regularization terms. We discuss possible extensions in Appendix A.

## 4 GLOBAL CONVERGENCE OF BLOCK COORDINATE DESCENT

In this section, we show that BCD for neural networks with an activation satisfying Assumption 2 finds global minima, in other words, the objective value  $F$  converges to an arbitrarily small value. In this section, we consider the case with single output ( $d_{out} = 1$ ). We discuss its extension to the multi-output case in Appendix B. Moreover, for the single output case, we provide a bound on the generalization error under the i.i.d. setting by utilizing the Rademacher complexity argument.

### 4.1 GLOBAL CONVERGENCE WITH MONOTONICALLY INCREASING ACTIVATION

Here, we consider the case of single outputs  $d_{out} = 1$ . In this case, the objective function is described by

$$\min_{\mathbf{W}, \mathbf{V}} F(\mathbf{W}, \mathbf{V}) := \sum_{i=1}^n \left[ (W_L V_{L-1,i} - y_i)^2 + \gamma \sum_{j=1}^{L-1} \|\sigma(W_j V_{j-1,i}) - V_{j,i}\|^2 \right]. \quad (4)$$

We now state the first main result, the global convergence of BCD with activation satisfying Assumption 2.

**Theorem 4.1** (BCD finds global minima of neural networks). *We assume that activation  $\sigma$  satisfies Assumption 2 and there exists a constant  $C_V > 0$  such that  $\lambda_{\max}(V_j V_j^\top) \leq C_V$  for  $j = 1, \dots, L-1$  during training. We denote  $s := \sigma_{\min}(X) > 0$ . Let  $R_i = |W_L V_{L-1,i} - y_i|$  at the initial value of the objective function with respect to the output layer, and define  $R := \sum_{i=1}^n R_i^2$ ,  $R_{\max} := \max_i R_i$ , and  $C_K := \left(\frac{2}{\alpha}\right)^L \left(4R_{\max}\eta_V + \frac{2}{2-\alpha}\sqrt{\epsilon}\right)$ .*

*Then, under  $(s_1, s_2) = (\frac{3}{4}, \frac{5}{4})$ ,  $\eta_V \leq \frac{1}{8\alpha\ell^2}$ ,  $\eta_W^{(1)} \leq \frac{\eta_V^{-1}}{8\sqrt{r}C_V K} \left(\frac{\alpha}{2}\right)^L$ ,  $\eta_W^{(2)} \leq \frac{1}{2\ell^2 \cdot \max_i \|x_i\|}$ , and*

$$K = \left\lceil \frac{2}{\eta_V} \log\left(\frac{3R}{\epsilon}\right) \right\rceil, K_V = \left\lceil \frac{1}{\gamma\alpha\ell\eta_V} \log\left(\frac{3\gamma(L-2)rnC_K^2}{\epsilon}\right) \right\rceil, K_W = \left\lceil \frac{1}{4\gamma s\alpha^2\eta_W^{(2)}} \log\left(\frac{3rnC_K^2}{\epsilon}\right) \right\rceil,$$

*it holds  $F(\mathbf{W}, \mathbf{V}) \leq \epsilon$ , where  $\mathbf{W} = (W_1, \dots, W_L)$  and  $\mathbf{V} = (V_{1,1}, \dots, V_{L-1,n})$  are the parameters obtained by the output of Algorithm 1.*

The proof can be seen in Appendix C. Theorem 4.1 exhibits that BCD provably finds a global minimum under a suitable choice of hyperparameters. While the definitions of  $K$ ,  $K_V$  and  $K_W$  are

somewhat complex, the total number of gradient computation to achieve  $\epsilon$  error is bounded by  $\tilde{O}(K(LK_V + K_W)) = \tilde{O}(\log^2(\frac{1}{\epsilon}))$ .

The proof consists of two parts: (i) the loss with respect to the output layer is monotonically decreasing in the outer loop, and (ii) the loss with respect to the hidden layer remains sufficiently small at the end of each iteration. We provide more detail to Appendix C due to page limitations.

We should note that the claims presented in Theorem 4.1 lie outside the framework of the so-called NTK regime (Jacot et al., 2018), among others. Specifically, while the NTK regime assumes that the parameters of neural networks remain almost unchanged during training, our analysis allows for scenarios where the parameters undergo changes of  $\Omega(1)$ .

**Remark 4.2.** *The assumption in Theorem 4.1,  $\lambda_{\max}(V_j V_j^\top) \leq C_V$ , ensures that the auxiliary parameters  $V_{j,i}$  are bounded during training. While we assume the existence of  $C_V$  in Theorem 4.1, we can provide a quantitative bound on the  $C_V$  as  $C_V = \mathcal{O}((\gamma\eta_V \ell n K K_V)^2)$  (note that this bound may not be tight). We provide a detailed derivation of this bound in Appendix D.*

## 4.2 GENERALIZATION ERROR BOUND

The objective of this subsection is to show that BCD Algorithm 1 does not only have a strong convergence guarantee, but also attains favorable generalization performance. To this end, we need to make an assumption about the data distribution.

**Assumption 3.** *The training sample  $\{(x_i, y_i)\}_{i=1}^n$  is independently sampled from a distribution  $(x, y) \sim P$ . Under the distribution  $P$ , it holds that  $\|x\| \leq B_X$  and  $|y| \leq B_Y$  almost surely.*

The first statement defines the data distribution, which is essential and standard requirement for describing the generalization error. The one requires that inputs and labels should be bounded with probability one, which is also standard.

We then provide the following result on the generalization error bound.

**Theorem 4.3** (Generalization error bound). *Let  $\hat{f}_{NN}$  be the output of Algorithm 1 under the same condition as Theorem 4.1. Then, if Assumption 3 holds,*

$$\begin{aligned} \mathbb{E}_{(x,y) \sim P} \left[ \left( \hat{f}_{NN}(x) - y \right)^2 \right] &\leq \frac{1}{n} \sum_{i=1}^n \left( \hat{f}_{NN}(x_i) - y_i \right)^2 \\ &+ \tilde{O} \left( \frac{\|X\|}{n} (B_Y + 2^L \ell^{L-1} B_X) d_{in}^{\frac{1}{2}} L^{\frac{3}{2}} (2r)^{\frac{L}{2}} \log r + (B_Y + 2^L \ell^{L-1} B_X)^2 \sqrt{\frac{\log(1/\delta)}{n}} \right). \end{aligned}$$

*with probability at least  $1 - \delta$  over the training sample  $\{(x_i, y_i)\}_{i=1}^n$ .*

The proof can be seen in Appendix E. Notably, Theorem 4.3 provides a bound on the generalization error for multi-layer neural networks with optimization guarantees, beyond the NTK regime. To obtain Theorem 4.3, we utilize a result from Bartlett et al. (2017), which evaluates the generalization gap using the spectral norms of the weight matrix of each layer. As mentioned in the previous section, we can show that the spectral norm (equal to the maximum singular value) of  $W_j$  is upper bounded. Combining this with the result from Bartlett et al. (2017), we can derive the generalization gap of BCD (see Appendix E for details).

## 5 RELU ACTIVATION

In this section, we propose a BCD algorithm specifically for the ReLU activation  $\sigma(x) := \max\{x, 0\}$ , which has been excluded in Theorem 4.1 due to Assumption 2. The difficulty in handling the ReLU activation is that it only takes non-negative values. For attaining zero loss for a hidden layer  $\|\sigma(W_j V_{j-1}) - V_j\|^2$ , we need to prevent  $V_j$  from taking negative value due to this non-negativity. Therefore, we must exclude such situations by modifying Algorithm 1.

### 5.1 BCD FOR NEURAL NETWORKS WITH SKIP CONNECTION

As a solution to overcome the difficulty of ReLU activation, we consider ResNet (He et al., 2016) type networks, where the neural networks includes skip connection. With skip connection, the

objective function treated in BCD is given by

$$\min_{\mathbf{W}, \mathbf{V}} F(\mathbf{W}, \mathbf{V}) := \sum_{i=1}^n \left[ (W_L V_{L-1,i} - y_i)^2 + \gamma \sum_{j=1}^{L-1} \|\sigma(W_j V_{j-1,i}) + V_{j-1,i} - V_{j,i}\|^2 \right],$$

where the loss of the hidden layer,  $\gamma \sum_{j=1}^{L-1} \|\sigma(W_j V_{j-1,i}) + V_{j-1,i} - V_{j,i}\|^2$  differs from (4). We describe the modified algorithm in Algorithm 3. We use the notation  $V_{j,i}^+ = \max\{V_{j,i}, 0\}$ .

---

**Algorithm 3:** Block coordinate descent: ReLU

---

**Input:**  $(W_1)_{ab} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1/d_{in})$ ,  $(W_j)_{ab} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1/r)$  for all  $j = 2, \dots, L$ ,  $V_{0,i} = x_i$

```

1  $K$ : max iteration,  $K_{in}$ : max inner iteration,  $\eta_V, \eta_W^{(1)}, \eta_W^{(2)}$ : step size;
2  $W_j \leftarrow \text{SVB}(W_j)$  with inputs  $s_1, s_2$ , and  $W_j$  for  $j = 2, \dots, L-1$ ;
3  $V_{j,i} = \sigma(W_j V_{j-1,i}) + V_{j-1,i}$  for all  $j = 1, \dots, L-1$  and  $i = 1, \dots, n$ ;
4 for  $k \leftarrow 1$  to  $K$  do
5   for  $i \leftarrow 1$  to  $n$  do
6      $V_{L-1,i} \leftarrow (V_{L-1,i} - \eta_V \nabla_{V_{L-1,i}} \|W_L V_{L-1,i} - y_i\|^2)^+$ ;
7      $W_{L-1} \leftarrow W_{L-1} - \gamma \eta_W^{(1)} \sum_{i=1}^n \nabla_{W_{L-1}} \|\sigma(W_{L-1} V_{L-2,i}) + V_{L-2,i} - V_{L-1,i}\|^2$ ;
8     for  $j \leftarrow L-1$  to  $2$  do
9        $W_j \leftarrow W_j - \gamma \eta_W^{(1)} \sum_{i=1}^n \nabla_{W_j} \|\sigma(W_j V_{j-1,i}) + V_{j-1,i} - V_{j,i}\|^2$ ;
10      for  $i \leftarrow 1$  to  $n$  do
11        for  $k_{inner} \leftarrow 1$  to  $K_V$  do
12           $V_{j-1,i} \leftarrow V_{j-1,i} - \gamma \eta_V \nabla_{V_{j-1,i}} \|\sigma(W_j V_{j-1,i}) + V_{j-1,i} - V_{j,i}\|^2$ ;
13           $V_{j-1,i} \leftarrow (V_{j-1,i})^+$ ;
14      for  $k_{inner} \leftarrow 1$  to  $K_W$  do
15         $W_1 \leftarrow W_1 - \gamma \eta_W^{(2)} \sum_{i=1}^n \nabla_{W_1} \|W_1 V_{0,i} - V_{1,i}\|^2$ ;

```

---

The initialization and update of  $W_1, \dots, W_{L-1}$  are common in Algorithm 1 and Algorithm 3. However, there are several differences between the two algorithms in their update procedures. First, in Algorithm 3, we apply the non-negative projection  $V \mapsto V^+$  for each  $V_{j,i}$  after the inner loop finishes. This is required for the non-negativity of ReLU: to ensure the solvability of the equation  $\|\sigma(W_j V_{j-1}) - V_j\|^2 = 0$ . Next, we do not update  $W_L$  in Algorithm 3. This is required to ensure the existence of  $V_{L-1,i}$  satisfying  $W_L V_{L-1,i} = y_i$  under the condition  $V_{L-1,i} \geq \mathbf{0}$ . To verify this, we first provide the following lemma.

**Lemma 5.1.** *Suppose that the vector  $W_L$  has both positive and negative entries. Then, for any  $y_i$ , there exists a non-negative vector  $V_{L-1,i}$  satisfying  $W_L V_{L-1,i} = y_i$ .*

This lemma implies that, to ensure the global convergence for arbitrary training label  $y_i$ , it is sufficient to check that  $W_L$  has both positive and negative components. Clearly, such a situation will occur frequently as the width of the hidden layer  $r$  increases. Indeed, by the symmetry of the Gaussian distribution, this probability is calculated as  $1 - 2 \cdot (\frac{1}{2})^r = 1 - 2^{-r+1}$ . Additionally, we provide a high probability bound on the norm of the positive and negative components of  $W_L$ , which determines the convergence speed of the gradient descent.

**Lemma 5.2.** *Let  $W_L^\top \sim \mathcal{N}(0, r^{-1} I_r)$ ,  $w_+ := \max\{W_L, \mathbf{0}^\top\}$ , and  $w_- := \min\{W_L, \mathbf{0}^\top\}$ . Then, for any  $\delta > 0$ , with probability at least  $1 - 2\delta$ ,  $\min\{\|w_+\|^2, \|w_-\|^2\} \geq \frac{1}{2} - \sqrt{\frac{8 \log(2/\delta)}{r}}$  holds.*

Since it is not trivial that the similar inequality holds for each iteration when considering the update of  $W_L$ , we assume that  $W_L$  is fixed during training for simplicity.

Similarly to the problem (4) considered in the previous section, we consider 1-dimensional outputs here. We then formally state the convergence result of Algorithm 3 applied to networks with ReLU activation and skip connections.

**Theorem 5.3** (Global convergence of BCD with ReLU activation). *We assume that there exists a constant  $C_V > 0$  such that  $\lambda_{\max}(V_j V_j^\top) \leq C_V$  for  $j = 1, \dots, L-1$  during training. We denote  $s := \sigma_{\min}(X)$ . Let  $R_i = |W_L V_{L-1,i} - y_i|$  at the initial value of the objective function with*



respect to the output layer, and define  $R := \sum_{i=1}^n R_i^2$ ,  $R_{\max} := \max_i R_i$ , and  $C_K := (4R_{\max}\eta_V + 5\sqrt{\epsilon})(\frac{3}{2})^L$ . Then, under  $(s_1, s_2) = (0, \frac{1}{4})$ ,  $\eta_V \leq \frac{1}{2 \min\{\|w_+\|^2, \|w_-\|^2\}}$ ,  $\eta_W^{(1)} \leq \frac{\eta_V^{-1}}{24\sqrt{\epsilon}C_V K} (\frac{2}{3})^L$ ,  $\eta_W^{(2)} \leq \frac{1}{2 \cdot \max_i \|x_i\|}$ , and

$$K = \left\lceil \frac{1}{4\eta_V \min\{\|w_+\|^2, \|w_-\|^2\}} \log\left(\frac{3R}{\epsilon}\right) \right\rceil, K_V = \left\lceil \frac{3}{4\gamma\eta_V} \log\left(\frac{49(L-2)rnC_K^2}{3\epsilon}\right) \right\rceil, K_W = \left\lceil \frac{1}{4\gamma s \eta_W^{(2)}} \log\left(\frac{C_K^2}{\epsilon}\right) \right\rceil,$$

it holds  $F(\mathbf{W}, \mathbf{V}) \leq \epsilon$ , where  $\mathbf{W} = (W_1, \dots, W_L)$  and  $\mathbf{V} = (V_{1,1}, \dots, V_{L-1,n})$  are the parameters obtained by the output of Algorithm 3.

The proof can be seen in Appendix F. Thus, we obtain a global convergence guarantee of BCD for networks with ReLU activation.

## 6 NUMERICAL EXPERIMENT

In this section, we conduct numerical experiments to verify our theoretical findings. Particularly, we numerically confirm that BCD converges to a global minimum for monotonically increasing activation (Algorithm 1) and ReLU (Algorithm 3) using an artificial dataset.

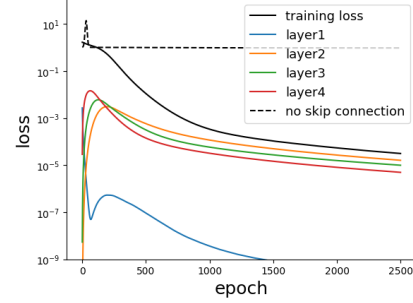
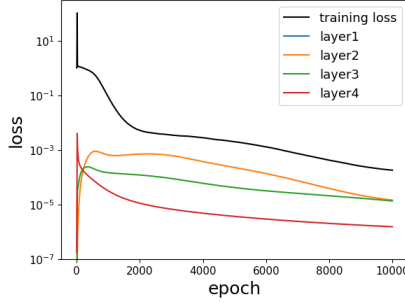


Figure 1: Loss of Algorithm 1 with LeakyReLU      Figure 2: Loss of Algorithm 3 with ReLU

### 6.1 MONOTONICALLY INCREASING ACTIVATION

First, we conduct a numerical experiment for a monotonically increasing activation. We apply Algorithm 1 to a neural network with four hidden layers, each with  $r = 30$  nodes, and LeakyReLU activation  $\sigma(x) = \max\{x, 0.5x\}$ , which satisfies Assumption 2 with  $\alpha = 0.5$  and  $\ell = 1$ . We prepare  $n = 500$  training samples from a teacher network with a single hidden layer and the same activation. We set  $d_{in} = 600$ , sample  $x_i$  from the normal distribution, and define  $y_i$  as the output of the teacher network. For hyperparameters, we employ  $K_V = K_W = 100$  and  $\eta_V = \eta_W^{(1)} = \eta_W^{(2)} = 1$ .

Figure 1 shows the result. The black line means the training error, i.e.,  $\frac{1}{n} \sum_{i=1}^n (f_{NN}(x_i) - y_i)^2$ . Other lines represent the loss of  $j$ -th layer, i.e.,  $\sum_{i=1}^n \|\sigma(W_j V_{j-1,i}) - V_{j,i}\|^2$  for  $j \in \{1, 2, 3, 4\}$ . We can observe that the training error monotonically decreases while the losses for each layer remain small, which reflects our theoretical findings.

### 6.2 RELU ACTIVATION

Next, we experimentally examine BCD for ReLU activation using Algorithm 3. We apply Algorithm 3 to a neural network with four hidden layers,  $r = 30$ , ReLU activation and skip connection. Similarly to the monotonically increasing activation, we prepare a dataset with  $n = 500$  and  $d_{in} = 600$  using a teacher network. For hyperparameters, we employ  $K_V = K_W = 100$  and  $\eta_V = \eta_W^{(1)} = \eta_W^{(2)} = 1$ .

Figure 2 shows the result. Like Figure 1, the black line means the training error. Other lines represent the loss of  $j$ -th layer, i.e.,  $\sum_{i=1}^n \|\sigma(W_j V_{j-1,i}) + V_{j-1,i} - V_{j,i}\|^2$  for  $j \in \{1, 2, 3, 4\}$ . We can observe the same convergence procedure here: the training error monotonically decreases and the losses for each layer remain small.

Additionally, we plot the training loss without using the skip connection as the dashed black line. While the training loss for BCD without skip connections does not decrease due to the difficulty of maintaining non-negativity, the skip connection drastically improves BCD training.

## 7 CONCLUSION

In this paper, we proposed a block coordinate descent (BCD) algorithm for training deep neural networks and ensured the convergence to global minima for networks with strictly monotonically increasing activation functions. We also derived a generalization bound using Rademacher complexity, ensuring both strong optimization and generalization performance. For ReLU activations, we introduced a modified BCD algorithm with skip connections and non-negative projection updates to ensure convergence. Empirical validation demonstrated the practical effectiveness of our algorithms for both monotonic and ReLU activations. Overall, this work advances the understanding of BCD in neural networks, offering provable convergence and generalization guarantees.

## REFERENCES

- Zeyuan Allen-Zhu and Yuanzhi Li. What can resnet learn efficiently, going beyond kernels? *Advances in Neural Information Processing Systems*, 32, 2019.
- Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32, 2019a.
- Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via overparameterization. In *International conference on machine learning*, pp. 242–252. PMLR, 2019b.
- Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International conference on machine learning*, pp. 254–263. PMLR, 2018.
- 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*, pp. 322–332. PMLR, 2019.
- Armin Askari, Geoffrey Negiar, Rajiv Sambharya, and Laurent El Ghaoui. Lifted neural networks. *arXiv preprint arXiv:1805.01532*, May 2018.
- Arindam Banerjee, Pedro Cisneros-Velarde, Libin Zhu, and Misha Belkin. Restricted strong convexity of deep learning models with smooth activations. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=PINRbk7h01>.
- Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.
- Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.
- Miguel Carreira-Perpinan and Weiran Wang. Distributed optimization of deeply nested systems. In *Artificial Intelligence and Statistics*, pp. 10–19. PMLR, 2014.
- Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for overparameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.
- Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pp. 1675–1685. PMLR, 2019.

- Spencer Frei, Yuan Cao, and Quanquan Gu. Agnostic learning of a single neuron with gradient descent. *Advances in Neural Information Processing Systems*, 33:5417–5428, 2020.
- Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pp. 297–299. PMLR, 2018.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Arthur Jacot, Franck Gabriel, and Cl  ment Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.
- Kui Jia, Dacheng Tao, Shenghua Gao, and Xiangmin Xu. Improving training of deep neural networks via singular value bounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4344–4352, 2017.
- Tim Tsz-Kit Lau, Jinshan Zeng, Baoyuan Wu, and Yuan Yao. A proximal block coordinate descent algorithm for deep neural network training. *arXiv preprint arXiv:1803.09082*, 2018.
- Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.
- Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on learning theory*, pp. 2388–2464. PMLR, 2019.
- Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.
- Kensuke Nakamura, Stefano Soatto, and Byung-Woo Hong. Block-cyclic stochastic coordinate descent for deep neural networks. *Neural Networks*, 139:348–357, 2021.
- Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on learning theory*, pp. 1376–1401. PMLR, 2015.
- Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.
- Phan-Minh Nguyen and Huy Tuan Pham. A rigorous framework for the mean field limit of multi-layer neural networks. *Mathematical Statistics and Learning*, 6(3):201–357, 2023.
- Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.
- Ravi G Patel, Nathaniel A Trask, Mamikon A Gulian, and Eric C Cyr. A block coordinate descent optimizer for classification problems exploiting convexity. *arXiv preprint arXiv:2006.10123*, 2020.
- Huy Tuan Pham and Phan-Minh Nguyen. Global convergence of three-layer neural networks in the mean field regime. *arXiv preprint arXiv:2105.05228*, 2021.
- Linbo Qiao, Tao Sun, Hengyue Pan, and Dongsheng Li. Inertial proximal deep learning alternating minimization for efficient neutral network training. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3895–3899. IEEE, 2021.
- Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *International conference on machine learning*, pp. 4433–4441. PMLR, 2018.
- Taiji Suzuki, Hiroshi Abe, and Tomoaki Nishimura. Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=ByeGzlrKwH>.

- Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109:475–494, 2001.
- Belinda Tzen and Maxim Raginsky. A mean-field theory of lazy training in two-layer neural nets: entropic regularization and controlled mckean-vlasov dynamics. *arXiv preprint arXiv:2002.01987*, 2020.
- Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.
- Colin Wei and Tengyu Ma. Data-dependent sample complexity of deep neural networks via lipschitz augmentation. *Advances in Neural Information Processing Systems*, 32, 2019.
- Jintao Xu, Chenglong Bao, and Wenxun Xing. Convergence rates of training deep neural networks via alternating minimization methods. *Optimization Letters*, 18(4):909–923, 2024.
- Tian Ye and Simon S Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. *Advances in Neural Information Processing Systems*, 34:1429–1439, 2021.
- Gilad Yehudai and Shamir Ohad. Learning a single neuron with gradient methods. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3756–3786. PMLR, 2020.
- Jinshan Zeng, Tim Tsz-Kit Lau, Shaobo Lin, and Yuan Yao. Global convergence of block coordinate descent in deep learning. In *International conference on machine learning*, pp. 7313–7323. PMLR, 2019.
- Hui Zhang, Shenglong Zhou, Geoffrey Ye Li, and Naihua Xiu. 0/1 deep neural networks via block coordinate descent. *arXiv preprint arXiv:2206.09379*, 2022.
- Ziming Zhang and Matthew Brand. Convergent block coordinate descent for training tikhonov regularized deep neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.
- Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109:467–492, 2020.

## A DISCUSSION OF EXTENSION

As we mentioned in Remark 3.2, we consider a simple form of BCD for the convergence guarantee, where we just apply the gradient descent update using the  $\ell_2$  distance. This section discusses the possible extension of the BCD algorithms Algorithm 1 and Algorithm 3.

**General loss function** One possible and somewhat straightforward extension is to employ general loss function  $\ell(\cdot, \cdot)$  instead of  $\ell_2$  distance we consider in this paper. This implies that our results are not restricted to regression. In this case, we can consider the total loss

$$\min_{\mathbf{W}, \mathbf{V}} F(\mathbf{W}, \mathbf{V}) := \sum_{i=1}^n \left[ \ell(W_L V_{L-1,i}, y_i) + \gamma \sum_{j=1}^{L-1} \|\sigma(W_j V_{j-1,i}) - V_{j,i}\|^2 \right],$$

where the loss of the output layer is replaced by  $\ell(\cdot, \cdot)$  compared to (4). Since the term  $\gamma \sum_{j=1}^{L-1} \|\sigma(W_j V_{j-1,i}) - V_{j,i}\|^2$  remains the same as (4), we can employ the same argument as Theorem 4.1 for its convergence proof. Therefore, we only need to ensure the global convergence of the output layer ( $W_L$  and  $V_{L-1}$ ) to obtain a similar result to Theorem 4.1. Indeed, we can consider strongly convex losses and replace the bound on  $F_L$  with an ordinal convergence guarantee for the convex function. One example of such a loss is cross-entropy loss, defined by

$$\ell(W_L V_{L-1}, y_i) = - \sum_{c=1}^{d_{out}} y_{ij} \log \frac{\exp(W_L V_{L-1})_c}{\sum_{c=1}^{d_{out}} \exp(W_L V_{L-1})_c},$$

which is typically used for the  $d_{out}$ -class classification problem. Thus, while we focus on the regression problem, our analysis can be extended to classification problems as well.

**Different activation between layers** While we consider a model that uses the same activation  $\sigma$  for all layers, we can employ different activation  $\sigma_j$  for  $j$ -th layer, provided they satisfy Assumption 2. We can follow the exactly same proof we show in Theorem 4.1, by replacing  $\sigma$  in the convergence argument with respect to the loss of  $j$ -th layer by  $\sigma_j$ .

**Other initialization schemes** In Algorithm 1, we initialize the weights  $W_j$  using the Gaussian initialization and apply singular value bounding to them, and then initialize  $V_{j,i}$  in the exact manner, i.e.,  $V_{j,i} = \sigma(W_{j-1} V_{j-1,i})$ . However, to ensure the global convergence, we only need to preserve the condition in Lemma C.4 during the training. Therefore, several variants of the initialization scheme can be considered. In particular, we do not need to initialize the weights using a Gaussian distribution. Xavier’s initialization, which employs a uniform distribution instead of a Gaussian, is one possible choice.

**Activation violating Assumption 2** Here, we discuss the possibility of employing activation other than those that satisfy Assumption 2 or ReLU. First, we note that our analysis relies on the monotonicity property of the activation function, including ReLU. Without this assumption, we cannot rule out the possibility that the parameters may be trapped in local minima due to the existence of points where  $\sigma' = 0$ . Thus, the monotonicity of the activation function is crucial in our proof.

For monotonically increasing activation that violate Assumption 2, one can employ sigmoid, tanh, and similar activation. As discussed in Section 5 for ReLU, one of the difficulties in handling such activations is that they cannot take any value in  $\mathbb{R}$ . For example, the sigmoid activation  $x \mapsto \frac{1}{1+\exp(-x)}$  only takes values in  $[0, 1]$ , and the tanh activation  $x \mapsto \frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}$  only takes values in  $[-1, 1]$ . In these cases, we need to take care of  $V_{j,i}$  not to go out of these ranges. For ReLU activation, we employ skip connection to overcome this problem.

**Training loss with the regularization term** A line of works investigates BCD methods usually considers the regularization term, meaning the loss function is given by

$$\min_{\mathbf{W}, \mathbf{V}} F(\mathbf{W}, \mathbf{V}) := \sum_{i=1}^n \left[ (W_L V_{L-1,i} - y_i)^2 + r_W(W_L)^2 + \gamma \sum_{j=1}^{L-1} \left[ \|\sigma(W_j V_{j-1,i}) - V_{j,i}\|^2 + r_W(W_j) + r_V(V_j) \right] \right],$$

where  $r_W$  and  $r_V$  denotes the regularization terms with respect to  $W$  and  $V$ , respectively. In this case, additional term appears in gradient descent updates. While these make convergence guarantees more challenging, convergence to a global minimum is still ensured as long as  $r_W$  and  $r_V$  are strongly convex, such as Tikhonov regularization. However, we need to carefully evaluate the gap in training loss caused by the regularization to derive the generalization error bound in Theorem 4.3.

## B EXTENSION TO MULTI DIMENSIONAL OUTPUT

Here, we consider the case with multi-output, where the loss is given by

$$\min_{\mathbf{W}, \mathbf{V}} F(\mathbf{W}, \mathbf{V}) := \sum_{i=1}^n \left[ \|W_L V_{L-1,i} - y_i\|^2 + \gamma \sum_{j=1}^{L-1} \|\sigma(W_j V_{j-1,i}) - V_{j,i}\|^2 \right].$$

with  $y_i \in \mathbb{R}^{d_{out}}$  ( $d_{out} > 1$ ). Comparing to the above result, a difficulty emerges in the optimization of the output layer: convergence analysis of  $W_L$  and  $V_{L-1,i}$ . If  $\text{rank}(W_L) \geq d_{out}$ , we can obtain the same result as Theorem 4.1 since the linear equation

$$W_L V_{L-1,i} - y_i = 0 \quad (5)$$

have a solution, and we can prove the convergence of gradient descent using the same argument as the proof of Theorem 4.1. In the case  $d_{out} > \text{rank}(W_L)$ , the linear equation may not have solutions, which means we cannot ensure the global convergence, and furthermore, nor can we verify the existence of global minima. To overcome such a situation, we introduce the following assumption: the labels have a low-rank representation.

**Assumption 4.** *There exists an integer  $r < d_{out}$  such that there exists a matrix  $U_1 \in \mathbb{R}^{d_{out} \times r}$  satisfying  $y_i = U_1 z_i$  with  $z_i \in \mathbb{R}^r$  for any  $i \in [n]$ .*

In this case, the equation (5) has solutions, including  $W_L = U_1$  and  $V_{L-1,i} = y_i$ . On the other hand, whether the parameters can converge to one of such solutions is not trivial. As an attempt to investigate this problem, we first write down the update of  $W_L$  by line 5 in Algorithm 1.

**Update of  $W_L$**  With general  $d_{out}$ , the straight-forward calculation gives

$$W_L^{(k)} = W_L^{(k-1)} \left( 1 - \eta_W \sum_{i=1}^n V_{L-1,i} V_{L-1,i}^\top \right) + \eta_W U_1 \sum_{i=1}^n z_i V_{L-1,i}^\top.$$

Let us discuss this update. The first term represents that with a sufficient small  $\eta_W$  so that the maximum eigenvalue of the (symmetric) matrix  $\eta_W \sum_{i=1}^n V_{L-1,i} V_{L-1,i}^\top$  is smaller than 1,  $W_L$  shrinks to zero exponentially. In the second term, a matrix spanned by  $U_1$  is added to  $W_L$ . Especially when the matrix  $\sum_{i=1}^n z_i V_{L-1,i}^\top \in \mathbb{R}^{r \times r}$  is full rank,  $W_L = \eta_W U_1 \sum_{i=1}^n z_i V_{L-1,i}^\top$  and  $y_i = (\eta_W \sum_{i=1}^n z_i V_{L-1,i}^\top)^{-1} y_i$  can be a solution of (5). Thus, throughout the gradient descent update, only matrices aligning with  $U_1$  are added to  $W_L$  and  $W_L$  loses other components exponentially. On the other hand, ensuring this procedure rigorously is still not easy, for example, the evaluation of the minimum eigenvalue of the matrix  $\sum_{i=1}^n z_i V_{L-1,i}^\top$  is complicated. Recently, Ye & Du (2021) shows the global convergence of this update as follows:

**Theorem B.1** (Theorem 1.1 in Ye & Du (2021)). *Suppose  $Y = (y_1, \dots, y_n) \in \mathbb{R}^{d_{out} \times n}$  satisfies Assumption 4. Let  $s_1$  and  $s_r$  be the minimum and maximum singular value of  $Y$ . Assume that each entry of  $W_L$  and  $V_{L-1,i}$  are initialized from Gaussian distribution with mean 0 and variance  $\delta^2$ , where  $\epsilon = \tilde{O}\left(\frac{s_r}{\sqrt{r^3 s_1(n+d_{out})}}\right)$ . Then, with setting  $\eta = O\left(\frac{s_r \delta^2}{r s_1}\right)$ , the output of the gradient descent achieves  $\sum_{i=1}^n \|W_L V_{L-1,i} - y_i\|^2 \leq \epsilon$  after  $O\left(\frac{1}{\eta s_1} \log\left(\frac{r s_r}{\epsilon}\right) + \frac{1}{\eta s_r} \log\left(\frac{s_r}{\epsilon}\right)\right)$  iterations.*

To apply this to BCD in the multi-output case, we only need to modify the proof of Theorem 4.1 in two points: (i) convergence analysis of  $W_L$  and  $V_{L-1,i}$  and (ii) adjust the initialization scheme. For (i), while we use a simple convergence analysis in Appendix C, we directly can apply Theorem B.1 instead. Then, the number of the iteration  $K$  in Theorem 4.1 is replaced by  $O\left(\frac{1}{\eta s_1} \log\left(\frac{rs_r}{\epsilon}\right) + \frac{1}{\eta s_r} \log\left(\frac{s_r}{\epsilon}\right)\right)$  as shown in Theorem B.1. For (ii), Theorem B.1 requires Gaussian initialization for each component of  $W_L$ ,  $V_{L-1,i}$ , which does not align with the exact manner initialization  $V_{j,i} = \sigma(V_{j-1,i})$  considered in Theorem 4.1. We need to bridge this gap to attain a convergence guarantee. However, as we discussed in Appendix A, the exact manner initialization is only required for small objective value at initialization, and we can extend our analysis to any initialization scheme. Thus, our analysis can be extended to the multi-dimensional output case.

## C PROOF OF THEOREM 4.1

In this section, we provide the proof to Theorem 4.1. The key notion is the block-wise analysis. First, we provide the preliminary lemmas for the proof. After that, we prepare the block-wise analysis and combine them.

Throughout this section, we suppose that the conditions in Theorem 4.1 are satisfied.

### C.1 PRELIMINARY RESULTS

The following lemma immediately follows from the smoothness of the activation.

**Lemma C.1.** *Let  $d \geq 1$  an integer. For any  $x_1, x_2 \in \mathbb{R}^d$ , it holds that  $\|\sigma(x_1) - \sigma(x_2)\|^2 \leq \ell^2 \|x_1 - x_2\|^2$ .*

Next, by utilizing Assumption 2, we derive the following lemma.

**Lemma C.2.** *For activation function satisfying Assumption 2, for any  $x, y \in \mathbb{R}$ , there exists  $\xi$  such that  $\alpha \leq \xi \leq \ell$  and  $\sigma(x + y) = \sigma(x) + \xi y$  hold.*

*Proof.* We first consider the case  $y > 0$ . Then, we have

$$\sigma(x + y) - \sigma(x) = \int_0^y \sigma'(x + t) dt \geq \alpha y.$$

The Lipschitz continuity of  $\sigma$  gives  $\sigma(x + y) \leq \sigma(x) + \ell y$ . Thus we get

$$\alpha \leq \xi := \frac{\sigma(x + y) - \sigma(x)}{y} \leq \ell,$$

which gives the conclusion.

The case  $y < 0$  can be proven by substituting  $x$  and  $y$  in above discussion by  $x + y$  and  $-y$ .

In the case  $y = 0$  we can take arbitrary  $\xi$  with  $\alpha \leq \xi \leq \ell$  to satisfy the assertion.  $\square$

This lemma gives the following proposition, which we utilize throughout the convergence analysis.

**Proposition C.3.** *For activation functions satisfying Assumption 2 and an integer  $d > 1$ , for any  $x, y \in \mathbb{R}^d$ , there exists a diagonal matrix  $\Xi$  such that each diagonal entry  $\Xi_{jj}$  of  $\Xi$  satisfies  $\alpha < \Xi_{jj} < \ell$  and  $\sigma(x + y) = \sigma(x) + \Xi y$ .*

*Proof.* Note that by Lemma C.2, for each  $j = 1, \dots, d$ , there exists a  $\Xi_{jj}$  satisfying  $\sigma(x + y)_j = \sigma(x)_j + \Xi_{jj} y_j$ . Then,  $\Xi = \text{diag}(\Xi_{11}, \dots, \Xi_{dd})$  satisfies the desired condition.  $\square$

Next, we prove that the singular values of  $W_j$  ( $j = 2, \dots, L$ ) are upper and lower bounded during the training.

**Lemma C.4** (Regularity of weight matrix  $W_j$  during training). *For  $j = 2, \dots, L$ ,  $\frac{1}{2} \leq \lambda_{\min}^{1/2}(W_j W_j^T) \leq \lambda_{\max}^{1/2}(W_j W_j^T) \leq 2$  always holds during the training.*

*Proof.* By Lemma C.5, it suffices to show that every row  $w$  of  $W_j$  satisfies  $\|\Delta w\| \leq \frac{1}{4\sqrt{r}}$ , where  $\Delta w$  denotes the difference between  $w$  at the start and end of the training. Indeed, this implies

$$\begin{aligned}\sigma_{\max}(\Delta W) &= \lambda_{\max}^{\frac{1}{2}}(\Delta W \Delta W^\top) \leq \sqrt{\sum_{p=1}^r \lambda_p(\Delta W \Delta W^\top)} \\ &\leq \sqrt{\text{Tr}(\Delta W \Delta W^\top)} = \|\Delta W\|_F \leq \frac{1}{4}.\end{aligned}$$

Combining this with  $\frac{3}{4} \leq \lambda_{\min} \leq \lambda_{\max} \leq \frac{5}{4}$  gives the conclusion.

To this end, we prove  $\|\Delta w\| \leq \frac{1}{4\sqrt{r}}$ . This follows from

$$\begin{aligned}\eta_W^{(1)} \gamma \nabla_w \|\sigma(wV) - V'\|^2 &= 2\eta_W^{(1)} \gamma \cdot \|\text{diag}(\sigma'(wV))V^\top(\sigma(wV) - V')\| \\ &\leq 2\eta_W^{(1)} \gamma \ell \lambda_{\max}^{1/2}(VV^\top) \cdot \|\sigma(wV) - V'\| \\ &\leq 2\eta_W^{(1)} \gamma \ell C_V \cdot \eta_V \left(\frac{2}{\alpha}\right)^L \leq \frac{1}{4K\sqrt{r}},\end{aligned}$$

where the last inequality follows from the definition of  $\eta_W^{(1)}$ .  $\square$

**Lemma C.5** (Weyl's inequality for singular values). *Let  $A \in \mathbb{R}^{d_1 \times d_2}$  be a real-valued matrix, then, for every matrix  $\Delta \in \mathbb{R}^{d_1 \times d_2}$ , it holds that*

$$\max_k |\sigma_k(A + \Delta) - \sigma_k(A)| \leq \sigma_{\max}(\Delta),$$

where  $\sigma_k(A)$  denotes the  $k$ -th largest singular value of  $A$  and  $\sigma_{\max}(A)$  denotes its maximum singular value.

## C.2 ANALYSIS OF GRADIENT DESCENT IN A GENERAL FORM

First, we introduce the key idea of analysis with general notations<sup>2</sup>. Let us consider the regression problem with an objective

$$\sum_{a=1}^b (\sigma(w^\top x_a) - y_a)^2,$$

where  $w \in \mathbb{R}^d$  is a trainable parameter. Let  $w' := w - \eta \nabla \sum_{a=1}^b (\sigma(w^\top x_a) - y_a)^2$ , where  $w'$  denotes the parameter obtained by a single update of gradient descent with a step-size  $\eta > 0$ . Denote  $X := (x_1, \dots, x_b)^\top \in \mathbb{R}^{b \times d}$  and  $Y := (y_1, \dots, y_b)^\top \in \mathbb{R}^b$ . Then,  $\sum_{a=1}^b (\sigma(w^\top x_a) - y_a)^2 = \|\sigma(Xw) - Y\|^2$  holds and a straightforward calculation shows  $w' = w - 2\eta X^\top D(\sigma(Xw) - Y)$ , where  $D = \text{diag}((\sigma'(w^\top x_1), \dots, \sigma'(w^\top x_b)))$ . Then, we have

$$\begin{aligned}\|\sigma(Xw') - Y\|^2 &= \|\sigma(Xw - 2\eta X X^\top D(\sigma(Xw) - Y)) - Y\|^2 \\ &= \|\sigma(Xw) - 2\eta \Xi X X^\top D(\sigma(Xw) - Y) - Y\|^2,\end{aligned}$$

where  $\Xi$  is a diagonal matrix which is determined by Proposition C.3. Thus, we obtain

$$\|\sigma(Xw') - Y\|^2 = \|(I - 2\eta \Xi X X^\top D)(\sigma(Xw) - Y)\|^2. \quad (6)$$

The obtained relationship (6) implies that with a sufficiently small choice of  $\eta$  satisfying

$$\|I - 2\eta \Xi X X^\top D\|_{op} < 1, \quad (7)$$

the loss  $\|\sigma(Xw) - Y\|^2$  will linearly decrease to zero.

<sup>2</sup>Our analysis is similar to that in Yehudai & Ohad (2020); Frei et al. (2020)



Here, we explain how (7) holds even when the matrix  $2\eta\Xi XX^\top D$  is not symmetric, i.e., positive-semidefinite. Let  $M := XX^\top$ , which is positive-semidefinite. Since diagonal matrices commute, we have

$$\Xi XX^\top D = (\Xi^{-1}D)^{-\frac{1}{2}} D^{\frac{1}{2}} \Xi^{\frac{1}{2}} M \Xi^{\frac{1}{2}} D^{\frac{1}{2}} (\Xi^{-1}D)^{\frac{1}{2}}.$$

Let  $N := (\Xi^{-1}D)^{-\frac{1}{2}}$ . Then, we obtain

$$I - 2\eta\Xi XX^\top D = N \left( I - 2\eta D^{\frac{1}{2}} \Xi^{\frac{1}{2}} M \Xi^{\frac{1}{2}} D^{\frac{1}{2}} \right) N^{-1}.$$

Since  $2\eta D^{\frac{1}{2}} \Xi^{\frac{1}{2}} M \Xi^{\frac{1}{2}} D^{\frac{1}{2}}$  is positive-semidefinite,  $I - 2\eta D^{\frac{1}{2}} \Xi^{\frac{1}{2}} M \Xi^{\frac{1}{2}} D^{\frac{1}{2}}$  only has positive eigenvalue for a sufficiently small  $\eta$ , and its largest eigenvalue is smaller than one if  $2\eta D^{\frac{1}{2}} \Xi^{\frac{1}{2}} M \Xi^{\frac{1}{2}} D^{\frac{1}{2}}$  is positive-definite. Since the eigenvalues are invariant with respect to the change of basis, we obtain (7).

### C.3 BLOCK-WISE CONVERGENCE ANALYSIS

According to the relationship (6), we provide the block-wise convergence analysis, that is, the convergence analysis of the  $W_j$  and  $V_j$  of the each layer.

**Update of  $V_{j,i}$  ( $j = 1, \dots, L-2$ )** According to Algorithm 1, the update of  $V_{j,i}$  ( $j = 1, \dots, L-1$ ) is written by

$$V_{j,i} \leftarrow V_{j,i} - \gamma \eta_V \sum_{p=1}^r \nabla_{V_{j,i}} \left( \sigma(w_{j,p} V_{j,i}) - (V_{j+1,i})_p \right)^2, \quad (8)$$

where  $w_{j,p}$  denotes the  $p$ -th row of the weight matrix of the  $j$ -th layer  $W_j$  and  $(V_{j+1,i})_p$  denotes the  $p$ -th component of  $V_{j+1,i}$ .

Despite the abuse of notation, we omit the layer index  $j$  and the sample index  $i$  for notational simplicity. We note that the analysis here can be independently applied to each layer and sample, as shown in the proof of the main theorem; hence, this abbreviation does not matter in the proof of Theorem 4.1. Then, (8) can be rewritten by

$$V \leftarrow V - \gamma \eta_V \sum_{p=1}^r \nabla_V \left( \sigma(w_p V) - V'_p \right)^2, \quad (9)$$

where we denote  $V'_p := (V_{j+1,i})_p$ .

Let  $F_V(v) := \sum_{p=1}^r (\sigma(w_p v) - V'_p)^2 (= \|\sigma(Wv) - V'\|^2)$  and  $V^{(0)}$  be the initial point of  $V_j$  of the inner loop for each outer iteration (we also use abuse of notation here), and  $V^{(k)}$  be the parameter obtained by  $k$  iterations of the inner loop.

Under these settings, we first show the existence of global minima of  $F_V$  as follows:

**Lemma C.6** (Existence of  $v^*$ ). *Suppose that  $\frac{1}{2} \leq \sigma_{\min}(W)$  and  $\sigma_{\max}(W) \leq 2$  hold. Let  $\Delta v := \sigma(WV^{(0)}) - V'$ . Then, there exists a unique  $v$  satisfying  $F_V(v^*) = 0$  and*

$$\|V^{(0)} - v^*\| \leq \frac{2}{\alpha} \|\Delta v\|.$$

*Proof.* Let  $\Delta v := \sigma^{-1}(\sigma(WV^{(0)}) + \Delta v) - WV^{(0)}$ . Then, it follows that  $v^* = V^{(0)} + W^{-1}\Delta v$  since

$$\sigma(V^{(0)} + W^{-1}\Delta v) = \sigma\left(\sigma^{-1}(\sigma(WV^{(0)}) + \Delta v)\right) = \sigma(WV^{(0)}) + \Delta v.$$

Now,  $\sigma^{(-1)}(\cdot)$  is  $\frac{1}{\alpha}$ -Lipschitz and satisfies  $\sigma(0) = 0$ . Then, we have  $\|\Delta v\| \leq \frac{1}{\alpha} \|\Delta v\|$  and consequently

$$\|V^{(0)} - v^*\| = \|W^{-1}\Delta v\| \leq \|W^{-1}\|_{op} \cdot \|\Delta v\| \leq \frac{2}{\alpha} \|\Delta v\|.$$

This gives the assertion.  $\square$

Next, by using the observation in (6), we provide the convergence analysis to the update (9).

**Lemma C.7** (Convergence analysis of  $V$ ). *Under the same condition as Theorem 4.1, it holds that*

$$\left\| \sigma(WV^{(k)}) - V' \right\|^2 \leq \exp(-\gamma\alpha^2\ell\eta_V k) \left\| \sigma(WV^{(0)}) - V' \right\|^2.$$

*Proof.* Let  $\sigma(WV^{(k)}) := \sigma(WV^{(k-1)}) - \Delta\sigma$ . Then, by letting  $a \rightarrow l$ ,  $b \rightarrow r$ ,  $x_a \rightarrow w_j$ ,  $y_a \rightarrow V'$  in (6), there exists a diagonal matrix  $\Xi$  such that

$$\Delta\sigma = \Xi W \left( 2\gamma\eta_V W^\top D^{(k-1)} \left( \sigma(WV^{(k-1)}) - V' \right) \right)$$

and  $\alpha I \prec \Xi \prec \ell I$ , where

$$D^{(k-1)} = \text{diag} \left( \left( \sigma'(w_1 V^{(k-1)}), \dots, \sigma'(w_r V^{(k-1)}) \right) \right) \in \mathbb{R}^{r \times r}.$$

Then, it holds that

$$\begin{aligned} \left\| \sigma(WV^{(k)}) - V' \right\|^2 &= \left\| \left( I - 2\gamma\eta_V \Xi W W^\top D^{(k-1)} \right) \left( \sigma(WV^{(k-1)}) - V' \right) \right\|^2 \\ &\leq \left\| I - 2\gamma\eta_V \Xi W W^\top D^{(k-1)} \right\|_{op}^2 \left\| \sigma(WV^{(k-1)}) - V' \right\|^2, \end{aligned}$$

By using the fact that  $\frac{1}{4}I \prec WW^\top \prec 4I$  and  $\alpha I \prec D^{(k-1)} \prec \ell I$ , we have  $\frac{\gamma\alpha^2\ell}{2}\eta_V \leq \lambda_{\min}(A) \leq \lambda_{\max}(A) \leq 8\gamma\ell^2\eta_V$  for  $A = 2\gamma\eta_V \Xi W W^\top D^{(k-1)}$ . Hence, by taking  $\eta_V \leq \frac{1}{8\gamma\ell^2}$ , we have

$$0 \leq \left\| I - 2\gamma\eta_V \Xi W W^\top D^{(k-1)} \right\|_{op} \leq 1 - \frac{\gamma\alpha^2\ell}{2}\eta_V$$

and therefore,

$$\left\| \sigma(WV^{(k)}) - V' \right\|^2 \leq \left( 1 - \frac{\gamma\alpha^2\ell}{2}\eta_V \right)^2 \left\| \sigma(WV^{(k-1)}) - V' \right\|^2.$$

This results in

$$\begin{aligned} \left\| \sigma(WV^{(k)}) - V' \right\|^2 &\leq \left( 1 - \frac{\gamma\alpha^2\ell}{2}\eta_V \right)^{2k} \left\| \sigma(WV^{(0)}) - V' \right\|^2 \\ &\leq \exp(-\gamma\alpha^2\ell\eta_V k) \left\| \sigma(WV^{(k-1)}) - V' \right\|^2, \end{aligned}$$

where the last inequality follows from  $1 - x \leq e^{-x}$ . Thus we obtain the conclusion.  $\square$

Finally, we provide a lemma evaluating distance to global minima based on the objective value:

**Lemma C.8.** *Suppose that  $F_V(v) \leq \epsilon$  holds. Then,*

$$\|v - v^*\| \leq \frac{2}{\alpha} \sqrt{\epsilon}.$$

*Proof.* Since

$$\epsilon \geq F_V(v) = \|\sigma(Wv) - \sigma(Wv^*)\|^2 \geq \alpha^2 \|Wv - Wv^*\|^2 \geq \frac{1}{4} \alpha^2 \|v - v^*\|,$$

we obtain  $\|v - v^*\| \leq \frac{2}{\alpha} \sqrt{\epsilon}$ .  $\square$

**Update of  $W_j$  ( $j = 2, \dots, L-1$ )** Let  $w_{j,p} \in \mathbb{R}^r$  be the  $p$ -th row of the weight matrix  $W_j$ . Then, update of each  $w_{j,p}$  is given by

$$w_{j,p} \leftarrow w_{j,p} - \gamma \eta_W^{(1)} \sum_{i=1}^n \nabla_{w_{j,p}} \left( \sigma(w_{j,p} V_{j,i}) - (V_{j+1,i})_p \right)^2, \quad (10)$$

For notational simplicity, we omit the layer index  $j$  and the node index  $p$ . Namely, the update (10) is simply rewritten by

$$w \leftarrow w - \gamma \eta_W^{(1)} \sum_{i=1}^n \nabla_w (\sigma(w V_i) - V'_i)^2,$$

where we denote  $V_i := V_{j,i}$  and  $V'_i := (V_{j+1,i})_p$ . Let  $F_W(w) := \sum_{i=1}^n (\sigma(w V_i) - V'_i)^2 (= \|\sigma(wV) - V'\|^2)$  and  $w^{(0)}$  be the initial point of  $w_{j,p}$  of the inner loop for each outer iteration (we also use abuse of notation here), and  $w^{(k)}$  be the parameter obtained by  $k$  iterations of the inner loop. Against to the argument of  $F_V$  in the above paragraph,  $F_W$  have not a solution  $w^*$  satisfying  $F_W(w^*) = 0$  especially when  $n > r$ . However, we can still ensure that the objective value remains small during the update of  $W_j$  as follows:

**Lemma C.9** (Convergence analysis of  $W_j$ ). *Under the same condition as Theorem 4.1, it holds that*

$$\|\sigma(w'V) - V'\|^2 \leq \|\sigma(wV) - V'\|^2.$$

*Proof.* By letting  $a \rightarrow i$ ,  $b \rightarrow i$ ,  $x_a \rightarrow V_i$  and  $y_a \rightarrow V'$  in (6), there exists a diagonal matrix  $\Xi$  satisfying  $\alpha I \prec \Xi \prec \ell I$  and

$$\begin{aligned} \|\sigma(w'V) - V'\|^2 &= \left\| \left( I - 2\gamma \eta_W^{(1)} \Xi V V^\top D^{(k-1)} \right) (\sigma(w^{(k-1)}V) - V') \right\|^2 \\ &\leq \left\| I - 2\gamma \eta_W^{(1)} \Xi V V^\top D^{(k-1)} \right\|_{op}^2 \left\| \sigma(w^{(k-1)}V) - V' \right\|^2. \end{aligned}$$

By using the fact that  $O \preceq V V^\top \prec C_V I$  and  $\alpha I \prec D^{(k-1)} \prec \ell I$ , we have  $\lambda_{\max} \left( 2\gamma \eta_W^{(1)} \Xi V V^\top D^{(k-1)} \right) \leq 2\gamma C_V \ell^2 \eta_W^{(1)}$ . Hence, by taking  $\eta_W^{(1)} \leq \frac{1}{2\gamma C_V \ell^2}$ , we obtain the conclusion.  $\square$

**Update of  $W_1$**  Let  $w_p \in \mathbb{R}^{d_{in}}$  the  $p$ -th row of the weight matrix  $W_1$ . Then, the update of each  $w_p$  is given by

$$w_p \leftarrow w_p - \gamma \eta_W^{(2)} \sum_{i=1}^n \nabla_{w_i} \left( \sigma(w_p x_i) - (V_{1,i})_p \right)^2. \quad (11)$$

Namely, the update (11) is simply rewritten by

$$w^{(k)} \leftarrow w^{(k-1)} - \gamma \eta_W^{(2)} \sum_{i=1}^n \nabla_w \left( \sigma(w^{(k-1)} x_i) - V_i \right)^2, \quad (12)$$

where we denote  $V_i := (V_{1,i})_l$ .

Let  $F_W(w) := \sum_{i=1}^n (\sigma(w x_i) - V_i)^2 (= \|\sigma(wX) - V\|^2)$  and  $w^{(0)}$  be the initial point of  $w_j$  of the inner loop for each outer iteration (we also use abuse of notation here), and  $w^{(k)}$  be the parameter obtained by  $k$  iterations of the inner loop.

**Lemma C.10** (Existence of  $W^*$ ). *Let  $\Delta v := \sigma(w^{(0)} x_i) - V_i$ . Then, there exists a  $w^*$  such that  $F_w(w^*) = 0$  and*

$$\|w^{(0)} - w^*\| \leq \frac{1}{\alpha \sigma_{\min}(X)} \|\Delta v\|$$

*Proof.* The proof is essentially same as that of Lemma C.6.  $\square$

We then provide the convergence analysis to the update (12) by using the observation in (6), we.

**Lemma C.11** (Convergence analysis of  $W_1$ ). *Under the same condition as Theorem 4.1,*

$$\left\| \sigma(w^{(k)} X) - V_1 \right\|^2 \leq \exp(-4\gamma s \alpha^2 \eta_W^{(2)} k) \left\| \sigma(w^{(0)} X) - V_1 \right\|^2.$$

*Proof.* By letting  $a \rightarrow i$ ,  $b \rightarrow i$ ,  $x_a \rightarrow x_i$  and  $y_a \rightarrow V_i$  in (6), we obtain

$$\begin{aligned} \left\| \sigma(w^{(k)} X) - V_1 \right\|^2 &= \left\| \sigma(w^{(k-1)} X) - V_1 - 2\gamma \eta_W^{(2)} A X X^\top D^{(k-1)} (\sigma(w^{(k-1)} X) - V_1) \right\|^2 \\ &= \left\| \left( I - 2\gamma \eta_W^{(2)} \Xi X X^\top D^{(k-1)} \right) (\sigma(w^{(k-1)} X) - V_1) \right\|^2 \\ &\leq \left\| I - 2\gamma \eta_W^{(2)} \Xi X X^\top D^{(k-1)} \right\|_{op}^2 \left\| \sigma(w^{(k-1)} X) - V_1 \right\|^2. \end{aligned}$$

By using the fact that  $0 < s := \lambda_{\min}(X X^\top)$  and  $\lambda_{\max}(X X^\top) \leq \max_i \|x_i\|^2$ , we have

$$2\gamma s \alpha^2 \eta_W^{(2)} \leq \lambda_{\min}(A) \leq \lambda_{\max}(A) \leq 2\gamma \ell^2 \max_i \|x_i\|^2 \eta_W^{(2)}$$

for  $A = 2\gamma \eta_W^{(2)} \Xi X X^\top D^{(k-1)}$ . Then, by taking  $\eta_W^{(2)} \leq \frac{1}{2\gamma \ell^2 \max_i \|x_i\|^2}$ , we obtain

$$\left\| \sigma(w^{(k)} X) - V_1 \right\|^2 \leq \left( 1 - 2\gamma s \alpha^2 \eta_W^{(2)} \right)^2 \left\| \sigma(w^{(k-1)} X) - V_1 \right\|^2.$$

By using  $1 - x \leq e^{-x}$ , this concludes

$$\begin{aligned} \left\| \sigma(w^{(k)} X) - V_1 \right\|^2 &\leq \left( 1 - 2\gamma s \alpha^2 \eta_W^{(2)} \right)^{2k} \left\| \sigma(w^{(0)} X) - V_1 \right\|^2 \\ &\leq \exp(-4\gamma s \alpha^2 \eta_W^{(2)} k) \left\| \sigma(w^{(0)} X) - V_1 \right\|^2, \end{aligned}$$

which is the desired bound.  $\square$

### C.3.1 PROOF OF THEOREM 4.1

Before providing the proof of Theorem 4.1, we introduce the following lemma:

**Lemma C.12** (Bound on  $\Delta v$  at the output layer). *Let  $R_i := \left\| W_j^{(0)} V_{L-1,i}^{(0)} - y_i \right\|$ . Then, we have*

$$\left\| V_{L-1,i}^{(k)} - V_{L-1,i}^{(k-1)} \right\| \leq 4R_i \eta_V.$$

*Proof.* By the construction of the algorithm, we have

$$\begin{aligned} \left\| V_{L-1,i}^{(k)} - V_{L-1,i}^{(k-1)} \right\| &= \left\| 2\eta_V (W_L^{(k)} V_{L-1,i}^{(k-1)} - y_i) W_L^{(k)} \right\| \\ &\leq 2\eta_V \left\| W_L^{(k)} \right\|_{op} \cdot \left\| W_L^{(k)} V_{L-1,i}^{(k-1)} - y_i \right\| \\ &\leq 4\eta_V \left\| W_L^{(0)} V_{L-1,i}^{(0)} - y_i \right\| = 4\eta_V R_i, \end{aligned}$$

which gives the conclusion.  $\square$

Then, we move to the proof to Theorem 4.1.

*Proof of Theorem 4.1.* Let us consider the decomposition of  $F$  as

$$F = F_L + \sum_{j=1}^{L-1} F_j = \sum_{i=1}^n \left[ F_{L,i} + \sum_{j=1}^{L-1} \sum_{p=1}^r F_{j,i,p} \right],$$

where

$$F_{L,i} := (W_L V_{L-1,i} - y_i)^2, \quad F_L = \sum_{i=1}^n F_{L,i}$$

and

$$F_{j,i,p} := \gamma \left( \sigma(W_j V_{j-1,i})_p - (V_{j,i})_p \right)^2, \quad F_j = \sum_{i=1}^n \sum_{p=1}^r F_{j,i,p}$$

for  $j = 1, \dots, L-1$ . The proof consists of two parts: (I)  $F_L$  is monotonically decreasing in the outer loop and (II)  $F_{j,i,p}$  ( $j = 1, \dots, L-1, i = 1, \dots, n, p = 1, \dots, r$ ) is sufficiently small at the end of each inner iteration.

**(I) Bound on  $F_L$**  The update of  $V_{L-1,i}$  is described by

$$V_{L-1,i}^{(k)} = V_{L-1,i}^{(k-1)} - 2\eta_V \left( W_L^{(k)} V_{L-1,i}^{(k-1)} - y_i \right) W_L^{(k)}.$$

Then, we have

$$W_L^{(k)} V_{L-1,i}^{(k-1)} - y_i = \left( 1 - 2\eta_V \|W_L^{(k)}\|^2 \right) \left( W_L^{(k)} V_{L-1,i}^{(k-1)} - y_i \right).$$

This results in

$$F_{L,i}^{(k)} \leq \left( 1 - 2\eta_V \|W_L^{(k)}\|^2 \right)^2 F_{L,i}^{(k-1)} \leq \exp \left( -4\eta_V \|W_L^{(k)}\|^2 \right) F_{L,i}^{(k-1)} \leq \exp(-\eta_V) F_{L,i}^{(k-1)},$$

where the second inequality follows from  $1 - x \leq e^{-x}$  and the last inequality from  $\|W_L^{(k)}\| \geq \frac{1}{2}$ .

This concludes

$$F_L^{(k)} \leq \exp(-\eta_V k) F_L^{(0)}.$$

Since  $F_L^{(0)} = R$  by the definition of  $R$ , after  $k = \frac{1}{\eta_V} \log \left( \frac{3R}{\epsilon} \right)$  iterations,  $F_L^{(k)} \leq \frac{\epsilon}{3}$  holds.

**(II)-(i) Bound on  $F_j$**  ( $j = 2, \dots, L-1$ ) Let us define  $\Delta v_{j,i} := \sigma(W_{j+1} V_{j,i}) - V_{j+1,i}$  for  $j = 1, \dots, L-1$ , where we denote  $V_{L,i} := y_i$ . Then, by Lemma C.6 and Lemma C.8, we have

$$\|\Delta v_{j,i}\| \leq \frac{2}{\alpha} (\|\Delta v_{j+1,i}\| + \sqrt{\epsilon})$$

for any  $j = 1, \dots, L-2$  and  $i = 1, \dots, n$ . We have  $\|\Delta v_{L-1,i}^{(k)}\| \leq 4R_{\max} \eta_V$  by Lemma C.12. By using this bound, we can derive

$$\|\Delta v_{j,i}^{(k)}\| \leq \left( 4R_{\max} \eta_V + \frac{2}{2-\alpha} \sqrt{\epsilon} \right) \left( \frac{2}{\alpha} \right)^{L-1-j} - \frac{2}{2-\alpha} \sqrt{\epsilon} \quad (13)$$

$$\leq \left( 4R_{\max} \eta_V + \frac{2}{2-\alpha} \sqrt{\epsilon} \right) \left( \frac{2}{\alpha} \right)^L \quad (14)$$

by induction. Indeed, (13) holds for  $j = L-1$  with equality. Moreover, under the induction hypothesis, it holds that

$$\begin{aligned} \|\Delta v_{j-1,i}\| &\leq \frac{2}{\alpha} (\|\Delta v_{j,i}\| + \sqrt{\epsilon}) \leq \frac{2}{\alpha} \left( \left( 4R_{\max} \eta_V + \frac{2}{2-\alpha} \sqrt{\epsilon} \right) \left( \frac{2}{\alpha} \right)^{L-j} - \frac{2}{2-\alpha} \sqrt{\epsilon} + \sqrt{\epsilon} \right) \\ &= \left( 4R_{\max} \eta_V + \frac{2}{2-\alpha} \sqrt{\epsilon} \right) \left( \frac{2}{\alpha} \right)^{L-(j-1)} - \frac{2}{2-\alpha} \sqrt{\epsilon}. \end{aligned}$$

This concludes (13) for  $j = 1, \dots, L-1$ . Then, by using Lemma C.7, we have

$$F_{j,i,p} \leq \gamma \exp(-\gamma \alpha^2 \ell \eta_V k_{\text{inner}}) \cdot \left( \frac{2}{\alpha} \right)^L \left( 4R_{\max} \eta_V + \frac{2}{2-\alpha} \sqrt{\epsilon} \right)^2.$$

Thus,  $k_{\text{inner}} = \frac{2}{\gamma \alpha^2 \ell \eta_V} \log \left( \left( \frac{2}{\alpha} \right)^L \left( 4R_{\max} \eta_V + \frac{2}{2-\alpha} \sqrt{\epsilon} \right)^2 \frac{3(L-2)rn\gamma}{\epsilon} \right)$  gives  $F_{j,i,p} \leq \frac{\epsilon}{3(L-2)rn}$  and hence,  $F_j \leq \frac{\epsilon}{3(L-2)}$  by summing up  $F_{j,i,p}$ .

**(II)-(ii) Bound on  $F_1$**  By using Lemma C.11, we have

$$\sum_{i=1}^n F_{1,i,p} \leq \exp\left(-4\gamma s \alpha^2 \eta_W^{(2)} k_{inner}\right) \left\| \sigma(W^{(0)}U) - V_1 \right\|^2$$

Since  $\Delta v_{1,i}^{(k)} \leq \left(4R_{\max}\eta_V + \frac{2}{2-\alpha}\sqrt{\epsilon}\right) \left(\frac{2}{\alpha}\right)^{L-1}$ , we have

$$\left\| \sigma(W^{(0)}U) - V_1 \right\|^2 \leq \sum_{i=1}^n \left( \frac{2}{\alpha} (\|\Delta v_{1,i}\| + \epsilon) \right)^2 = n \left( 4R_{\max}\eta_V + \frac{2}{2-\alpha}\sqrt{\epsilon} \right)^2 \cdot \left( \frac{2}{\alpha} \right)^{2L}$$

Thus,  $k_{inner} = \frac{1}{4\gamma s \alpha^2 \eta_W^{(2)}} \log \left( n \left( 4R_{\max}\eta_V + \frac{2}{2-\alpha}\sqrt{\epsilon} \right)^2 \cdot \left( \frac{2}{\alpha} \right)^{2L} \frac{3r}{\epsilon} \right)$  gives  $\sum_{i=1}^n F_{1,i,p} \leq \epsilon$  for  $p = 1, \dots, r$ . This results in  $F_1 = \sum_{i=1}^n \sum_{p=1}^r F_{1,i,p} \leq \frac{\epsilon}{3}$ .

**(III) Summing up all** By combining all, after  $K$  outer iterations and  $K_V$  and  $K_W$  inner iterations, we have

$$F = F_L + \sum_{j=1}^{L-1} F_j \leq \underbrace{\frac{\epsilon}{3}}_{F_L} + \sum_{j=2}^{L-1} \underbrace{\frac{\epsilon}{3(L-2)}}_{F_2, \dots, F_{L-1}} + \underbrace{\frac{\epsilon}{3}}_{F_1} = \epsilon,$$

which gives the conclusion.  $\square$

## D QUANTITATIVE EVALUATION OF $C_V$

Here, we provide the quantitative bound on  $C_V$  satisfying  $\lambda_{\max}(V_j V_j^\top) \leq C_V$  for  $j = 1, \dots, L-1$  during the training, which we introduced in Theorem 4.1 and Remark 4.2.

**Proposition D.1.** Let  $c_V := 2 \max_j \sum_{i=1}^n \|V_{j,i}\|^2$ , where  $V_j$ s are the parameters at the initialization. Under the same settings as Theorem 4.1, We can take

$$C_V = c_V + \mathcal{O}((\gamma \eta_V \ell n K K_V)^2).$$

*Proof.* First, we have

$$\begin{aligned} \lambda_{\max}(V_j V_j^\top) &\leq \sum_{j=1}^r \lambda_j(V_j V_j^\top) = \text{tr}(V_j V_j^\top) \\ &= \text{tr} \left( \sum_{i=1}^n V_{j,i} V_{j,i}^\top \right) = \sum_{i=1}^n \text{tr}(V_{j,i} V_{j,i}^\top) = \sum_{i=1}^n \|V_{j,i}\|^2. \end{aligned} \quad (15)$$

This implies that we only need to evaluate the norm of  $V_{j,i}$ s during the training. Remind that the update of  $V_j$  is given by

$$V_{j,i} \leftarrow V_{j,i} - 2\gamma \eta_V W_j^\top D(\sigma(W_j V_{j,i}) - V_{j+1,i}),$$

where  $D = \text{diag}(\sigma'(W_j V_{j,i}))$ . Let  $\Delta V_{j,i} := 2\gamma \eta_V W_j^\top D(\sigma(W_j V_{j,i}) - V_{j+1,i})$ . Then, we have

$$\begin{aligned} \|\Delta V_{j,i}\| &= 2\gamma \eta_V \left\| W_j^\top D(\sigma(W_j V_{j,i}) - V_{j+1,i}) \right\| \leq 2\gamma \eta_V \cdot \|W_j\|_{op} \|D\|_{op} \|\sigma(W_j V_{j,i}) - V_{j+1,i}\| \\ &\leq 4\gamma \ell \eta_V C_K, \end{aligned}$$

where in the second inequality, we use  $\|W_j\|_{op} = \lambda_{\max}^{1/2}(W_j W_j^\top) \leq 2$  from Lemma C.4,  $\|D\|_{op} \leq \ell$ , and  $\|\sigma(W_j V_{j,i}) - V_{j+1,i}\| \leq C_K$  from (14) (Note that the objective function  $\|\sigma(W_j V_{j,i}) - V_{j+1,i}\|^2$  is monotonically decreasing from Lemma C.7, (14) always holds). Since the total number of updating  $V_{j,i}$  is  $K \cdot K_V$ , by using the triangle inequality we have

$$\|V_{j,i}\| \leq \|V_{j,i}^{init}\| + K \cdot K_V \cdot 4\gamma \ell \eta_V C_K,$$

where  $\|V_{j,i}^{init}\|$  is the initial value of  $V_{j,i}$ .

Substituting this bound to (15), we obtain

$$\begin{aligned}\lambda_{\max}(V_j V_j^\top) &\leq \sum_{i=1}^n (\|V_{j,i}^{init}\| + K \cdot K_V \cdot 4\gamma\ell\eta_V C_K)^2 \\ &\leq \sum_{i=1}^n 2(\|V_{j,i}^{init}\|^2 + (K \cdot K_V \cdot 4\gamma\ell\eta_V C_K)^2) \\ &\leq c_V + \mathcal{O}((\gamma\eta_V \ell n K K_V)^2),\end{aligned}$$

where we use the inequality  $(a + b)^2 \leq 2a^2 + 2b^2$  in the second inequality. Thus, we obtain the conclusion.  $\square$

## E PROOF OF THEOREM 4.3

Here, we provide the proof of Theorem 4.3, the generalization error bound of neural networks trained by Algorithm 1.

*Proof of Theorem 4.3.* By using the bound on  $u$  and  $y$  supposed in Assumption 3, we have

$$\begin{aligned}|f(u) - y| &\leq B_Y + |W_L \sigma(W_{L-1} \dots \sigma(W_1 u) \dots)| \\ &\leq B_Y + \ell \|W_L\|_{op} \|W_{L-1} \sigma(W_{L-2} \dots \sigma(W_1 u) \dots)\| \\ &\leq \dots \\ &\leq B_Y + \ell^{L-1} \left( \prod_{j=2}^L \|W_j\|_{op} \right) \|W_1 u\| \\ &\leq B_Y + 2^L \ell^{L-1} \|u\| \\ &\leq B_Y + 2^L \ell^{L-1} B_X.\end{aligned}$$

Hence, by taking  $M = B_Y + 2^L \ell^{L-1} B_X$  and  $\mathcal{R}(\mathcal{F})$  as what derived by Lemma E.2 in Lemma E.1, we obtain the conclusion.  $\square$

**Lemma E.1** (Theorem 11.3 in Mohri et al. (2018)). *For a hypothesis class  $\mathcal{F}$  and a training data  $\{(x_i, y_i)\}_{i=1}^n$ , let us define its (empirical) Rademacher complexity by*

$$\mathcal{R}(\mathcal{F}) := \mathbb{E}_{\sigma} \left[ \sup_{f \in \mathcal{F}} \frac{\sigma^\top f(\mathbf{u})}{n} \right],$$

where  $f(\mathbf{x}) = (f(x_1), \dots, f(x_n))^\top$  and  $\sigma$  is a random vector whose each component independently takes value  $\pm 1$  with probability  $\frac{1}{2}$ . Suppose that  $|h(x) - y| \leq M$  a.s. for any  $h \in \mathcal{F}$ . Then, for any  $0 < \delta < 1$ , with probability at least  $1 - \delta$  over a sample, we have

$$\mathbb{E}_{(x,y) \sim P} [(h(x) - y)^2] \leq \frac{1}{n} \sum_{i=1}^n (h(x_i) - y_i)^2 + 2M\mathcal{R}(\mathcal{F}) + 3M^2 \sqrt{\frac{\log(2/\delta)}{2n}}.$$

**Lemma E.2** (Rademacher complexity bound). *Let  $\mathcal{F}$  be the class of neural network predictors obtained by Algorithm 1. Then, the Rademacher complexity of  $\mathcal{F}$  can be bounded by*

$$\mathcal{R}(\mathcal{F}) \leq \frac{4}{n\sqrt{n}} + \log\left(\frac{1}{\sqrt{n}}\right) \frac{12\sqrt{R_{\mathcal{F}}}}{n}$$

with  $R_{\mathcal{F}} = d_{in}(2r)^L L^3 \|U\|^2 \log(2r^2)(\log n)$ .

To obtain this result, we apply the obtained bound on the spectral of  $W$  to the Rademacher complexity bound shown in Bartlett et al. (2017) as follows:

**Lemma E.3** (Lemma A.8 in Bartlett et al. (2017)). Assume activation functions  $\{\sigma_j(\cdot)\}_{j=1}^L$  such that each  $\sigma_j$  is  $\rho_j$ -Lipschitz continuous and  $\sigma_j(0) = 0$ . Let us define

$$\mathcal{F} := \left\{ \sigma_L(W_L \sigma_{L-1}(\dots \sigma_1(W_1 \cdot) \dots)) \mid \|W_j\|_{op} \leq B_j, \|W_j\|_{2,1} \leq b_j \ (1 \leq j \leq L) \right\}.$$

Then, it holds that

$$\mathcal{R}(\mathcal{F}) \leq \frac{4}{n\sqrt{n}} + \log \left( \frac{1}{\sqrt{n}} \right) \frac{12\sqrt{R_{\mathcal{F}}}}{n},$$

where  $R_{\mathcal{F}} > 0$  is a constant defined by

$$R_{\mathcal{F}} := \|X\|^2 \log(2r^2)(\log n) \left( \prod_{j=1}^L B_j \rho_j \right) \left( \sum_{j=1}^L \left( \frac{b_j}{B_j} \right)^{\frac{2}{3}} \right)^3.$$

*Proof of Lemma E.2.* By applying Lemma E.3 with  $\rho_1 = \dots = \rho_L = 1$ ,  $B_j = 2$ ,  $b_1 = 2d_{in}$  and  $b_j = 2r$  for  $j = 1, \dots, L-2$  and  $b_L = 2$ , we obtain

$$\begin{aligned} R_{\mathcal{F}} &= \|X\|^2 \log(2r^2)(\log n) \left( 4d_{in} \prod_{j=2}^{L-1} (2r) \right) \left( \sum_{j=1}^L \left( \frac{2r}{2} \right)^{\frac{2}{3}} \right)^3 \\ &= \|X\|^2 \log(2r^2)(\log n) \cdot 4d_{in}(2r)^{L-2} L^3 r^2 = d_{in}(2r)^L L^3 \|U\|^2 \log(2r^2)(\log n), \end{aligned}$$

which gives the conclusion.  $\square$

## F PROOF OF THEOREM 5.3

### F.1 PROOF OF LEMMA 5.2

*Proof of Lemma 5.2.* First, we have  $\mathbb{E}[\|w_+\|^2] = \mathbb{E}[\|w_-\|^2] = \frac{1}{2}$ . The first equality follows from the symmetricity, and the second equality follows from

$$\frac{1}{2} = \frac{1}{2} \mathbb{E}[\|W_L\|^2] = \frac{1}{2} \mathbb{E}[\|w_+\|^2 + \|w_-\|^2] = \frac{1}{2} (\mathbb{E}[\|w_+\|^2] + \mathbb{E}[\|w_-\|^2]) = \mathbb{E}[\|w_+\|^2],$$

where we use  $\mathbb{E}[\|w_+\|^2] = \mathbb{E}[\|w_-\|^2]$  in the last equality. Then, by using the concentration inequality argument (see Example 2.11 in Wainwright (2019) for example), we have

$$\mathbb{P} \left( \left| \|w_+\|^2 - \frac{1}{2} \right| \geq t \right) \leq 2 \exp \left( -\frac{rt^2}{8} \right)$$

for any  $t \in (0, 1)$ . By letting  $t = \sqrt{\frac{8 \log(2/\delta)}{r}}$ , we obtain

$$\mathbb{P} \left( \|w_+\|^2 < \frac{1}{2} - \sqrt{\frac{8 \log(2/\delta)}{r}} \right) \leq \delta$$

Since the same argument holds with  $w_-$ , taking a union bound concludes the assertion.  $\square$

### F.2 ANALYSIS OF GRADIENT DESCENT WITH SKIP CONNECTION

We introduce the key idea of analysis with general notations similarly to Appendix C, while there exists a skip connection. Let us consider the regression problem with an objective

$$\sum_{a=1}^b (\sigma(w^\top x_a) + w_a - y_a)^2,$$

where  $w \in \mathbb{R}^d$  is a trainable parameter. Let  $w' := w - \eta \nabla_w \sum_{a=1}^b (\sigma(w^\top x_a) + w_a - y_a)^2$ , where  $w'$  denotes the parameter obtained by a single update of gradient descent with a step-size  $\eta > 0$ . Denote  $X := (x_1, \dots, x_b)^\top \in \mathbb{R}^{b \times d}$  and  $Y := (y_1, \dots, y_b)^\top \in \mathbb{R}^b$ . Then,



$\sum_{a=1}^b (\sigma(w^\top x_a) + w_a - y_a)^2 = \|\sigma(Xw) + w - Y\|^2$  holds and a straightforward calculation shows  $w' = w - 2\eta(X^\top D + I)(\sigma(Xw) + w - Y)$ , where  $D = \text{diag}((\sigma'(w^\top x_1), \dots, \sigma'(w^\top x_b)))$ . Then, we have

$$\begin{aligned} & \sigma(Xw') + w' - Y \\ &= \sigma(Xw - 2\eta X(X^\top D + I)(\sigma(Xw) + w - Y)) + w - 2\eta(X^\top D + I)(\sigma(Xw) + w - Y) - Y \\ &= \sigma(Xw) - 2\eta \Xi X(X^\top D + I)(\sigma(Xw) + w - Y) + w - 2\eta(X^\top D + I)(\sigma(Xw) + w - Y) - Y \\ &= [I - 2\eta(I + \Xi X X^\top D + \Xi X + X^\top D)](\sigma(Xw) + w - Y), \end{aligned}$$

where  $\Xi$  is a diagonal matrix which is determined by Proposition C.3. Thus, we obtain

$$\|\sigma(Xw') + w' - Y\|^2 = \|[I - 2\eta(I + \Xi X X^\top D + \Xi X + X^\top D)](\sigma(Xw) + w - Y)\|^2. \quad (16)$$

The obtained relationship (16) implies that with a sufficiently small choice of  $\eta$  satisfying  $\|I - 2\eta \Xi X X^\top D\|_{op} < 1$ , the loss  $\|\sigma(Xw) - Y\|^2$  will linearly decrease to zero.

**Lemma F.1.**  $0 \preceq D \preceq I$  holds.

*Proof.* The assertion directly follows from  $\sigma'(u) \in \{0, 1\}$  for arbitrary  $u \in \mathbb{R}$ .  $\square$

**Lemma F.2.** Suppose  $\|X\|_{op} \leq \frac{1}{3}$ . Then, the inequality

$$\|I - 2\eta(I + \Xi X X^\top D + \Xi X + X^\top D)\|_{op} \leq 1 - \frac{2}{3}\eta$$

holds.

*Proof.* Since  $\lambda_{\min}(\Xi X X^\top D) \geq 0$ , we have  $\|I - 2\eta(I + \Xi X X^\top D)\|_{op} \leq 1 - 2\eta$ . Moreover, we have

$$\|\Xi X + X^\top D\|_{op} \leq \|\Xi\|_{op}\|X\| + \|X\|_{op}\|D\|_{op} \leq \frac{2}{3}.$$

Then, the triangle inequality gives

$$\begin{aligned} \|I - 2\eta(I + \Xi X X^\top D + \Xi X + X^\top D)\|_{op} &\leq \|I - 2\eta(I + \Xi X X^\top D)\|_{op} + 2\eta\|\Xi X + X^\top D\|_{op} \\ &\leq 1 - 2\eta + 2\eta \cdot \frac{2}{3} = 1 - \frac{2}{3}\eta, \end{aligned}$$

which is the conclusion.  $\square$

**Lemma F.3.** Suppose  $\|W\|_{op} \leq \frac{1}{3}$  and  $V' \geq \mathbf{0}$ . Then, if  $\|\sigma(WV) + V - V'\|^2 \leq \epsilon$ , then

$$\begin{aligned} \|V - (V)^+\|^2 &\leq \epsilon, \\ \|\sigma(W(V)^+) + (V)^+ - V'\|^2 &\leq \frac{49}{9}\epsilon \end{aligned}$$

*Proof.* Since  $\sigma(WV) \geq \mathbf{0}$  and  $V \geq \mathbf{0}$ , we have

$$\begin{aligned} \epsilon &\geq \|\sigma(WV) + V - V'\|^2 \geq \sum_{V_j < 0} [\sigma(WV)_j + V_j - V'_j]^2 \\ &\geq \sum_{V_j < 0} (V_j)^2 = \|V - (V)^+\|^2, \end{aligned}$$

which gives the first conclusion. The second follows from

$$\begin{aligned} \|\sigma(W(V)^+) + (V)^+ - V'\| &\leq \|\sigma(W(V)^+) - \sigma(WV) + (V)^+ - V\| + \|\sigma(WV) + V - V'\| \\ &\leq \|\sigma(W(V)^+) - \sigma(WV)\| + \|(V)^+ - V\| + \|\sigma(WV) + V - V'\| \\ &\leq \|W((V)^+ - V)\| + \epsilon^{\frac{1}{2}} + \epsilon^{\frac{1}{2}} \\ &\leq \frac{1}{3}\epsilon^{\frac{1}{2}} + 2\epsilon^{\frac{1}{2}} = \frac{7}{3}\epsilon^{\frac{1}{2}}, \end{aligned}$$

where we use the triangle inequality in the first and second inequalities, and 1-Lipschitzness of the ReLU activation in the third inequality.  $\square$

**Lemma F.4.** Suppose that  $V^{(0)}$  satisfies  $\sigma(WV^{(0)}) + V^{(0)} - V' =: \Delta v$  and  $V^*$  satisfies  $\sigma(WV^*) + V^* = V'$ . If  $\|W\|_{op} < 1$ , it holds that

$$\|V^{(0)} - V^*\| \leq \frac{1}{1 - \|W\|_{op}} \|\Delta v\|.$$

*Proof.* We have

$$\begin{aligned} \|\Delta v\| &= \|\sigma(WV^{(0)}) + V^{(0)} - V'\| \\ &= \|\sigma(WV^{(0)}) + V^{(0)} - \sigma(WV^*) - V^*\| \\ &\geq \|V^{(0)} - V^*\| - \|\sigma(WV^{(0)}) - \sigma(WV^*)\| \\ &\geq \|V^{(0)} - V^*\| - \|W(V^{(0)} - V^*)\| \\ &\geq \|V^{(0)} - V^*\| - \|W\|_{op} \|V^{(0)} - V^*\| = (1 - \|W\|_{op}) \|V^{(0)} - V^*\|, \end{aligned}$$

where we use the triangle inequality in the first inequality, the 1-Lipschitzness of ReLU activation in the second inequality. Dividing each term by  $1 - \|W\|_{op}$  gives the conclusion.  $\square$

### F.3 PRELIMINARY RESULTS

**Lemma F.5** (Regularity of weight matrix  $W_j$  during training). For  $j = 2, \dots, L - 1$ ,  $\|W_j\|_{op} \leq \frac{1}{3}$  always holds during the training.

*Proof.* By Lemma C.5, it suffices to show that every of  $W_j$  satisfies  $\|\Delta w\| \leq \frac{1}{12\sqrt{r}}$ , where  $\Delta w$  denotes the difference between  $w$  at the start and end of the training by the same as the proof of Lemma C.4.

To this end, we prove  $\|\Delta w\| \leq \frac{1}{12\sqrt{r}}$ . This follows from

$$\begin{aligned} \eta_W^{(1)} \gamma \nabla_w \|\sigma(wV) + V - V'\|^2 &= 2\eta_W^{(1)} \gamma \cdot \|\text{diag}(\sigma'(wV))V^\top (\sigma(wV) + V - V')\| \\ &\leq 2\eta_W^{(1)} \gamma \ell \lambda_{\max}^{1/2}(VV^\top) \cdot \|\sigma(wV) + V - V'\| \\ &\leq 2\eta_W^{(1)} \gamma \ell C_V \cdot \eta_V \left(\frac{3}{2}\right)^L \leq \frac{1}{12K\sqrt{r}}, \end{aligned}$$

where the last inequality follows from the definition of  $\eta_W^{(1)}$ .  $\square$

**Lemma F.6** (Convergence analysis of  $W_j$ ). Under the same condition as Theorem 4.1, it holds that

$$\|\sigma(w^{(k)}V) - V'\|^2 \leq \|\sigma(w^{(0)}V) - V_1\|^2.$$

*Proof.* The proof is essentially same as that of Lemma C.9.  $\square$

**Lemma F.7** (Bound on  $\Delta v$  at the output layer). Let  $R_i := |W_j^{(0)}V_{L-1,i}^{(0)} - y_i|$ . Then, we have

$$\|V_{L-1,i}^{(k)} - V_{L-1,i}^{(k-1)}\| \leq 4R_i \eta_V.$$

*Proof.* Since  $V_{L-1}^{(k)} \geq \mathbf{0}$ , we have

$$\begin{aligned}
\|V_{L-1,i}^{(k)} - V_{L-1,i}^{(k-1)}\| &= \left\| \left( V_{L-1,i}^{(k-1)} - 2\eta_V (W_L V_{L-1,i}^{(k-1)} - y_i) W_L \right)^+ - V_{L-1,i}^{(k-1)} \right\| \\
&\leq \left\| \left( V_{L-1,i}^{(k-1)} - 2\eta_V (W_L V_{L-1,i}^{(k-1)} - y_i) W_L \right) - V_{L-1,i}^{(k-1)} \right\| \\
&= \left\| 2\eta_V (W_L V_{L-1,i}^{(k-1)} - y_i) W_L \right\| \\
&\leq 2\eta_V \|W_L^{(k)}\|_{op} \cdot \|W_L^{(k)} V_{L-1,i}^{(k-1)} - y_i\| \\
&\leq 4\eta_V \|W_L^{(0)} V_{L-1,i}^{(0)} - y_i\| = 4\eta_V R_i,
\end{aligned}$$

which gives the conclusion.  $\square$

#### F.4 PROOF OF THEOREM 5.3

*Proof of Theorem 5.3.* We follow the similar argument as that of Theorem 4.1. Let us consider the decomposition of  $F$  as

$$F = F_L + \gamma \sum_{j=1}^{L-1} F_j = \sum_{i=1}^n \left[ F_{L,i} + \gamma \sum_{j=1}^{L-1} \sum_{p=1}^r F_{j,i,p} \right],$$

where

$$F_{L,i} := (W_L V_{L-1,i} - y_i)^2, \quad F_L = \sum_{i=1}^n F_{L,i}$$

and

$$F_{j,i,p} := \left( \sigma(W_j V_{j-1,i})_p + (V_{j-1,i})_p - (V_{j,i})_p \right)^2, \quad F_j = \sum_{i=1}^n \sum_{p=1}^r F_{j,i,p}$$

for  $j = 1, \dots, L-1$ .

**(I) Bound on  $F_L$**  We only need to consider the case  $W_L V_{L-1,i}^{(k-1)} - y_i \neq 0$ . The update of  $V_{L-1,i}$  is described by

$$V_{L-1,i}^{(k)} = \left( V_{L-1,i}^{(k-1)} - 2\eta_V (W_L V_{L-1,i}^{(k-1)} - y_i) W_L \right)^+ = V_{L-1,i}^{(k-1)} - 2\eta_V (W_L V_{L-1,i}^{(k-1)} - y_i) \tilde{w},$$

where we define  $\tilde{w} := (2\eta_V (W_L V_{L-1,i}^{(k-1)} - y_i))^{-1} (V_{L-1,i}^{(k-1)} - V_{L-1,i}^{(k)})$ . Then, we have

$$W_L^{(k)} V_{L-1,i}^{(k-1)} - y_i = (1 - 2\eta_V \tilde{w}^\top W_L) (W_L V_{L-1,i}^{(k-1)} - y_i).$$

Then, we show an inequality

$$\tilde{w}^\top W_L \geq \min\{\|w_+\|^2, \|w_-\|^2\}. \quad (17)$$

First we consider a case  $W_L V_{L-1,i}^{(k-1)} - y_i > 0$ . In this case, we have

$$\begin{aligned}
&\left( 2\eta_V (W_L V_{L-1,i}^{(k-1)} - y_i) \tilde{w} \right)_j \\
&= \left( \left( V_{L-1,i}^{(k-1)} - 2\eta_V (W_L V_{L-1,i}^{(k-1)} - y_i) W_L \right)^+ - V_{L-1,i}^{(k-1)} \right)_j \\
&= \begin{cases} 2\eta_V (W_L V_{L-1,i}^{(k-1)} - y_i) (W_L)_j & \text{if } j \in J_1 := \{j \mid (W_L)_j \leq 0\}, \\ 2\eta_V (W_L V_{L-1,i}^{(k-1)} - y_i) (W_L)_j & \text{if } j \in J_2 := \{j \mid (W_L)_j > 0 \text{ and } V_{L-1,i}^{(k-1)} > 2\eta_V (W_L V_{L-1,i}^{(k-1)} - y_i) (W_L)_j\}, \\ (V_{L-1,i}^{(k-1)})_j & \text{otherwise.} \end{cases}
\end{aligned}$$

This gives

$$\begin{aligned}
\tilde{w}^\top W_L &= \sum_{j=1}^r (\tilde{w})_j (W_L)_j \\
&= \sum_{j \in J_1 \cup J_2} (W_L)_j^2 + \sum_{j \in (J_1 \cup J_2)^c} 2\eta_V \left( W_L V_{L-1,i}^{(k-1)} - y_i \right)^{-1} \left( V_{L-1,i}^{(k-1)} \right)_j (W_L)_j \\
&\geq \sum_{j \in J_1} (W_L)_j^2 = \|w_-\|^2,
\end{aligned} \tag{18}$$

where in the inequality we use  $\left( V_{L-1,i}^{(k-1)} \right)_j > 0$  and  $(W_L)_j > 0$  for  $j \in (J_1 \cup J_2)^c$ .

If  $W_L V_{L-1,i}^{(k-1)} - y_i < 0$ , it holds that

$$\begin{aligned}
&\left( 2\eta_V \left( W_L V_{L-1,i}^{(k-1)} - y_i \right) \tilde{w} \right)_j \\
&= \begin{cases} 2\eta_V \left( W_L V_{L-1,i}^{(k-1)} - y_i \right) (W_L)_j & \text{if } j \in J_1 := \{j \mid (W_L)_j \geq 0\}, \\ 2\eta_V \left( W_L V_{L-1,i}^{(k-1)} - y_i \right) (W_L)_j & \text{if } j \in J_2 := \{j \mid (W_L)_j < 0 \text{ and } V_{L-1,i}^{(k-1)} > 2\eta_V \left( W_L V_{L-1,i}^{(k-1)} - y_i \right) (W_L)_j\}, \\ \left( V_{L-1,i}^{(k-1)} \right)_j & \text{otherwise.} \end{cases}
\end{aligned}$$

This gives

$$\begin{aligned}
\tilde{w}^\top W_L &= \sum_{j=1}^r (\tilde{w})_j (W_L)_j \\
&= \sum_{j \in J_1 \cup J_2} (W_L)_j^2 + \sum_{j \in (J_1 \cup J_2)^c} 2\eta_V \left( W_L V_{L-1,i}^{(k-1)} - y_i \right)^{-1} \left( V_{L-1,i}^{(k-1)} \right)_j (W_L)_j \\
&\geq \sum_{j \in J_1} (W_L)_j^2 = \|w_+\|^2,
\end{aligned} \tag{19}$$

where in the inequality we use  $\left( V_{L-1,i}^{(k-1)} \right)_j > 0$  and  $(W_L)_j < 0$  for  $j \in (J_1 \cup J_2)^c$ . The two bounds (18) and (19) conclude (17).

This results in

$$\begin{aligned}
F_{L,i}^{(k)} &\leq (1 - 2\eta_V \tilde{w}^\top W_L)^2 F_{L,i}^{(k-1)} \leq \exp(-4\eta_V \tilde{w}^\top W_L) F_{L,i}^{(k-1)} \\
&\leq \exp\left(-4\eta_V \min\{\|w_+\|^2, \|w_-\|^2\}\right) F_{L,i}^{(k-1)},
\end{aligned}$$

where the second inequality follows from  $1 - x \leq e^{-x}$  and the last inequality from (17). This concludes

$$F_L^{(k)} \leq \exp\left(-4\eta_V \min\{\|w_+\|^2, \|w_-\|^2\}k\right) F_L^{(0)}.$$

Since  $F_L^{(0)} = R$  by the definition of  $R$ , as long as we set  $\eta_V \leq \frac{1}{2 \min\{\|w_+\|^2, \|w_-\|^2\}}$  after  $k = \frac{1}{4\eta_V \min\{\|w_+\|^2, \|w_-\|^2\}} \log\left(\frac{3R}{\epsilon}\right)$  iterations,  $F_L^{(k)} \leq \frac{\epsilon}{3}$  holds.

**(II)-(i) Bound on  $F_j$  ( $j = 2, \dots, L-1$ )** Let us define  $\Delta v_{j,i} := \sigma(W_{j+1} V_{j,i}) + V_{j,i} - V_{j+1,i}$  for  $j = 1, \dots, L-1$ , where we denote  $V_{L,i} := y_i$ . Then,  $\|\Delta v_{j,i}^{(k)}\| \leq 2\eta_V R_i$  holds and Lemma F.4 gives

$$\|\Delta v_{j,i}\| \leq \frac{1}{1 - \|W_{j+1}\|} (\|\Delta v_{j+1,i}\| + \sqrt{\epsilon}) + \epsilon \leq \frac{3}{2} \left( \|\Delta v_{j+1,i}\| + \frac{5}{3} \sqrt{\epsilon} \right).$$

By these inequalities, we can ensure

$$\|\Delta v_{j,i}\| \leq (4R_{\max}\eta_V + 5\sqrt{\epsilon}) \left(\frac{3}{2}\right)^L$$

by taking  $\alpha = \frac{4}{3}$  in (13).

By for each  $j$ , it holds that

$$\begin{aligned} & \|\sigma(WV^{(k_{\text{inner}})}) + V^{(k_{\text{inner}})} - V'\|^2 \\ &= \left\| \left[ I - 2\gamma\eta_V \left( I + \Xi W W^\top D^{(k_{\text{inner}}-1)} + \Xi W + W^\top D^{(k_{\text{inner}}-1)} \right) \right] \left( \sigma(WV^{(k_{\text{inner}}-1)}) + V^{(k_{\text{inner}}-1)} - V' \right) \right\|^2 \\ &\leq \left\| I - 2\gamma\eta_V \left( I + \Xi W W^\top D^{(k_{\text{inner}}-1)} + \Xi W + W^\top D^{(k_{\text{inner}}-1)} \right) \right\|_{op}^2 \left\| \sigma(WV^{(k_{\text{inner}}-1)}) + V^{(k_{\text{inner}}-1)} - V' \right\|^2 \\ &\leq \left( 1 - \frac{2}{3}\gamma\eta_V \right)^2 \left\| \sigma(WV^{(k_{\text{inner}}-1)}) + V^{(k_{\text{inner}}-1)} - V' \right\|^2, \end{aligned}$$

which gives

$$\begin{aligned} \left\| \sigma(WV^{(k_{\text{inner}})}) + V^{(k_{\text{inner}})} - V' \right\|^2 &\leq \left( 1 - \frac{2}{3}\gamma\eta_V \right)^{2k_{\text{inner}}} \left\| \sigma(WV^{(0)}) + V^{(0)} - V' \right\|^2 \\ &\leq \exp \left( -\frac{2}{3}\gamma\eta_V \right)^{2k_{\text{inner}}} \left\| \sigma(WV^{(0)}) + V^{(0)} - V' \right\|^2 \\ &= \exp \left( -\frac{4}{3}\gamma\eta_V k_{\text{inner}} \right) \left\| \sigma(WV^{(0)}) + V^{(0)} - V' \right\|^2. \end{aligned}$$

Hence, by taking  $k_{\text{inner}} = \frac{3}{4\gamma\eta_V} \log \left( (4R_{\max}\eta_V + 5\sqrt{\epsilon})^2 \left(\frac{3}{2}\right)^{2L} \frac{49(L-2)rn}{3\epsilon} \right)$ , we obtain  $F_{j,i} \leq \frac{3\epsilon}{49(L-2)rn}$ . Then, Lemma F.3 gives  $F_{j,i} \leq \frac{\epsilon}{3(L-2)rn}$  after the non-negative projection (line 13) is applied.

**(II)-(ii) Bound on  $F_1$**  The update of  $W_1$  is same as what we considered in Theorem 4.1 (Algorithm 1). Therefore, by using Lemma C.11, we have

$$F_1 \leq \exp \left( -s\eta_W^{(2)} k \right) \left\| \sigma(W_1^{(0)} X) - V_1 \right\|^2 \leq \exp \left( -s\eta_W^{(2)} k \right) \cdot (4R_{\max}\eta_V + 5\sqrt{\epsilon})^2 \left(\frac{3}{2}\right)^{2L}.$$

Thus,  $k = \frac{1}{s\eta_W^{(2)}} \log \left( (4R_{\max}\eta_V + 5\sqrt{\epsilon})^2 \left(\frac{3}{2}\right)^{2L} \frac{3}{\epsilon} \right)$  gives  $F_1 \leq \frac{\epsilon}{3}$ .

**(III) Summing up all** By combining all, after  $K$  iterations and  $K_V$  and  $K_W$  iterations, we have

$$F = F_L + \sum_{j=1}^{L-1} F_j \leq \underbrace{\frac{\epsilon}{3}}_{F_L} + \sum_{j=2}^{L-1} \underbrace{\frac{\epsilon}{3\gamma(L-2)rn} rn\epsilon}_{F_2, \dots, F_{L-1}} + \underbrace{\frac{\epsilon}{3}}_{F_1} = \epsilon,$$

which gives the conclusion.  $\square$