Generalization Guarantee of Training Graph Convolutional Networks with Graph Topology Sampling

Hongkang Li [1] **Meng Wang** [1] **Sijia Liu** [2 3] **Pin-Yu Chen** [4] **Jinjun Xiong** [5]

Abstract

Graph convolutional networks (GCNs) have recently achieved great empirical success in learning graph-structured data. To address its scalability issue due to the recursive embedding of neighboring features, graph topology sampling has been proposed to reduce the memory and computational cost of training GCNs, and it has achieved comparable test performance to those without topology sampling in many empirical studies. To the best of our knowledge, this paper provides the first theoretical justification of graph topology sampling in training (up to) three-layer GCNs for semi-supervised node classification. We formally characterize some sufficient conditions on graph topology sampling such that GCN training leads to a diminishing generalization error. Moreover, our method tackles the non-convex interaction of weights across layers, which is under-explored in the existing theoretical analyses of GCNs. This paper characterizes the impact of graph structures and topology sampling on the generalization performance and sample complexity explicitly, and the theoretical findings are also justified through numerical experiments.

1. Introduction

Graph convolutional neural networks (GCNs) aggregate the embedding of each node with the embedding of its neighboring nodes in each layer. GCNs can model graph-structured data more accurately and compactly than conventional neural networks and have demonstrated great empirical advantage in text analysis (Hamilton et al., 2017; Kipf & Welling, 2017; Veličković et al., 2018; Peng et al., 2017), computer vision (Satorras & Estrach, 2018; Wang et al., 2018; Hu et al., 2018), recommendation systems (Ying et al., 2018; Van den Berg et al., 2018), physical reasoning (Battaglia et al., 2016; Sanchez-Gonzalez et al., 2018), and biological science (Duvenaud et al., 2015). Such empirical success is often achieved at a cost of higher computational and memory costs, especially for large graphs, because the embedding of one node depends recursively on the neighbors. To alleviate the exponential increase of computational cost in training deep GCNs, various graph topology sampling methods have been proposed to only aggregate the embeddings of a selected subset of neighbors in training GCNs. Node-wise neighbor-sampling methods such as GraphSAGE (Hamilton et al., 2017), VRGCN (Chen et al., 2018b), and Cluster-GCN (Chiang et al., 2019) sample a subset of neighbors for each node. Layer-wise importance sampling methods such as FastGCN (Chen et al., 2018a) and LADIES (Zou et al., 2019) sample a fixed number of nodes for each layer based on the estimate of node importance. Another line of works such as (Zheng et al., 2020; Li et al., 2020; Chen et al., 2021) employ graph sparsification or pruning to reduce the computational and memory cost. Surprisingly, these sampling methods often have comparable or even better testing performance compared to training with the original graph in many empirical studies (Chen et al., 2018a; 2021).

In contrast to the empirical success, the theoretical foundation of training GCNs with graph sampling is much less investigated. Only Cong et al. (2021) analyzes the convergence rate of graph sampling, but no generalization analysis is provided. One fundamental question about training GCNs is still vastly open, which is:

Under what conditions does a GCN learned with graph topology sampling achieve satisfactory generalization?

Our contributions: To the best of our knowledge, this paper provides the first generalization analysis of training GCNs with graph topology sampling. We focus on semi-

[1]Department of Electrical, Computer, and System Engineering, Rensselaer Polytechnic Institute, NY, USA [2]Department of Computer Science and Engineering, Michigan State University, MI, USA [3]MIT-IBM Watson AI Lab, IBM Research, MA, USA [4]IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA [5]Department of Computer Science and Engineering, University at Buffalo, NY, USA. Correspondence to: Hongkang Li <lih35@rpi.edu>, Meng Wang <wangm7@rpi.edu>, Sijia Liu <liusiji5@msu.edu>, Pin-yu Chen <Pin-Yu.Chen@ibm.com>, Jinjun Xiong <jinjun@buffalo.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

supervised node classification problems where, with all node features and partial node labels, the objective is to predict unknown node labels. We summarize our contributions from the following dimensions.

First, this paper proposes a training framework that implements both stochastic gradient descent (SGD) and graph topology sampling, and the learned GCN model with Rectified Linear Unit (ReLU) activation is guaranteed to approach the best generalization performance of a large class of target functions. Moreover, as the number of labeled nodes and the number of neurons increase, the class of target function enlarges, indicating improved generalization.

Second, this paper explicitly characterizes the impact of graph topology sampling on the generalization performance through the proposed *effective adjacency matrix* $\boldsymbol{A}^*$ of a directed graph that models the node correlations. $\boldsymbol{A}^*$ depends on both the given normalized graph adjacency matrix in GCNs and the graph sampling strategy. We provide the general insights that (1) if a node is sampled with a low frequency, its impact on other nodes is reduced in $\boldsymbol{A}^*$ compared with $\boldsymbol{A}$; (2) graph sampling on a highly-unbalanced $\boldsymbol{A}$, where some nodes have a dominating impact in the graph, results in a more balanced $\boldsymbol{A}^*$. Moreover, these insights apply to other graph sampling methods such as FastGCN (Chen et al., 2018a).

We show that learning with topology sampling has the same generalization performance as training GCNs using $\boldsymbol{A}^*$. Therefore, a satisfactory generalization can still be achieved even when the number of sampled nodes is small, provided that the resulting $\boldsymbol{A}^*$ still characterizes the data correlations properly. This is the first theoretical explanation of the empirical success of graph topology sampling.

Third, this paper shows that the required number of labeled nodes, referred to as the sample complexity, is a polynomial of $\|\boldsymbol{A}^*\|_\infty$ and the maximum node degree, where $\|\cdot\|_\infty$ measures the maximum absolute row sum. Moreover, our sample complexity is only logarithmic in the number of neurons m and consistent with the practical over-parameterization of GCNs, in contrast to the loose bound of poly(m) in (Zhang et al., 2020) in the restrictive setting of two-layer (one-hidden-layer) GCNs without graph topology sampling.

1.1. Related Works

Generalization analyses of GCNs without graph sampling. Some recent works analyze GCNs trained on the original graph. Xu et al. (2019); Cong et al. (2021) characterize the expressive power of GCNs. Xu et al. (2021) analyzes the convergence of gradient descent in training linear GCNs. Lv (2021); Liao et al. (2021); Garg et al. (2020); Oono & Suzuki (2020) characterize the generalization gap, which is the difference between the training error and testing error, through Rademacher complexity. Verma & Zhang (2019); Cong et al. (2021); Zhou & Wang (2021) analyze the generalization gap of training GCNs using SGD via the notation of algorithmic stability.

To analyze the training error and generalization performance simultaneously, Du et al. (2019) uses the neural tangent kernel (NTK) approach, where the neural network width is infinite and the step size is infinitesimal, shows that the training error is zero, and characterizes the generalization bound. Zhang et al. (2020) proves that gradient descent can learn a model with zero population risk, provided that all data are generated by an unknown target model. The result in (Zhang et al., 2020) is limited to two-layer GCNs and requires a proper initialization in the local convex region of the optimal solution.

Generalization analyses of feed-forward neural networks. The NTK approach was first developed to analyze fully connected neural networks (FCNNs), see, e.g., (Jacot et al., 2018). The works of Zhong et al. (2017); Fu et al. (2020); Li et al. (2022) analyze one-hidden-layer neural networks with Gaussian input data. Daniely (2017) analyzes multi-layer FCNNs but focuses on training the last layer only, while the changes in the hidden layers are negligible. Allen-Zhu et al. (2019) provides the optimization and generalization of three-layer FCNNs. Our proof framework is built upon (Allen-Zhu et al., 2019) but makes two important technical contributions. First, this paper provides the first generalization analysis of graph topology sampling in training GCNs, while Allen-Zhu et al. (2019) considers FCNNs with neither graph topology nor graph sampling. Second, Allen-Zhu et al. (2019) considers i.i.d. training samples, while this paper considers semi-supervised GCNs where the training data are correlated through graph convolution.

1.2. Notations

Vectors are in bold lowercase, matrices and tensors in are bold uppercase. Scalars are in normal fonts. For instance, $\boldsymbol{Z}$ is a matrix, and $\boldsymbol{z}$ is a vector. z_i denotes the i-th entry of $\boldsymbol{z}$, and $Z_{i,j}$ denotes the (i,j)-th entry of $\boldsymbol{Z}$. $[K]$ $(K>0)$ denotes the set including integers from 1 to K. $\boldsymbol{I}_d \in \mathbb{R}^{d\times d}$ and $\boldsymbol{e}_i$ represent the identity matrix in $\mathbb{R}^{d\times d}$ and the i-th standard basis vector, respectively. We denote the column ℓ_p norm for $\boldsymbol{W} \in \mathbb{R}^{d\times N}$ (for $p \geq 1$) as

$$\|\boldsymbol{W}\|_{2,p} = \Big(\sum_{i\in[m]} \|\boldsymbol{w}_i\|_2^p\Big)^{\frac{1}{p}} \tag{1}$$

Hence, $\|\boldsymbol{W}\|_{2,2} = \|\boldsymbol{W}\|_F$ is the Frobenius norm of $\boldsymbol{W}$. We use $\boldsymbol{w}_i$ ($\tilde{\boldsymbol{w}}_i$) to denote the i-th column (row) vector of $\boldsymbol{W}$. We follow the convention that $f(x) = O(g(x))$ (or $\Omega(g(x))$, $\Theta(g(x))$) means that $f(x)$ increases at most (or at least, or in the same, respectively,) order of $g(x)$. With high probability (w.h.p.) means with probability $1 - e^{-c\log^2(m_1,m_2)}$ for a

sufficient large constant c where m_1 and m_2 are the number of neurons in the two hidden layers.

Function complexity. For any smooth function $\phi(z)$ with its power series representation as $\phi(z) = \sum_{i=0}^{\infty} c_i z^i$, define two useful parameters as follows,

$$\mathcal{C}_\epsilon(\phi, R) = \sum_{i=0}^{\infty} \left((C^* R)^i + (\frac{\sqrt{\log(1/\epsilon)}}{\sqrt{i}} C^* R)^i \right) |c_i| \quad (2)$$

$$\mathcal{C}_s(\phi, R) = C^* \sum_{i=0}^{\infty} (i+1)^{1.75} R^i |c_i| \quad (3)$$

where $R \geq 0$ and C^* is a sufficiently large constant. These two quantities are used in the model complexity and sample complexity, which represent the required number of model parameters and training samples to learn ϕ up to ϵ error, respectively. Many population functions have bounded complexity. For instance, if $\phi(z)$ is $\exp(z)$, $\sin(z)$, $\cos(z)$ or polynomials of z, then $\mathcal{C}_\epsilon(\phi, O(1)) \leq O(\text{poly}(1/\epsilon))$ and $\mathcal{C}_s(\phi, O(1)) \leq O(1)$.

The main notations are summarized in Table 2 in Appendix.

2. Training GCNs with Topology Sampling: Formulation and Main Components

GCN setup. Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ denote an un-directed graph, where $\mathcal{V}$ is the set of nodes with size $|\mathcal{V}| = N$ and $\mathcal{E}$ is the set of edges. Let $\tilde{\boldsymbol{A}} \in \{0,1\}^{N \times N}$ be the adjacency matrix of $\mathcal{G}$ with added self-connections. Let $\boldsymbol{D}$ be the degree matrix with diagonal elements $D_{i,i} = \sum_j \tilde{A}_{i,j}$ and zero entries otherwise. $\boldsymbol{A}$ denotes the normalized adjacency matrix with $\boldsymbol{A} = \boldsymbol{D}^{-\frac{1}{2}} \tilde{\boldsymbol{A}} \boldsymbol{D}^{-\frac{1}{2}}$. Let $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ denote the matrix of the features of N nodes, where the n-th row of $\boldsymbol{X}$, denoted by $\tilde{\boldsymbol{x}}_n \in \mathbb{R}^{1 \times d}$, represents the feature of node n. Assume $\|\tilde{\boldsymbol{x}}_n\| = 1$ for all n without loss of generality. $y_n \in \mathcal{Y}$ represents the label of node n, where $\mathcal{Y}$ is a set of all labels. y_n depends on not only $\boldsymbol{x}_n$ but the neighbors. Let $\Omega \subset \mathcal{V}$ denote the set of labeled nodes. Given $\boldsymbol{X}$ and labels in Ω, the objective of semi-supervised node-classification is to predict the unknown labels in $\mathcal{V}/\Omega$.

Learner network We consider the setting of training a three-layer GCN $F : \mathbb{R}^N \times \mathbb{R}^{N \times d} \to \mathbb{R}^{1 \times K}$ with

$$F_{\boldsymbol{A}}(\boldsymbol{e}_g, \boldsymbol{X}; \boldsymbol{W}, \boldsymbol{V}) = \boldsymbol{e}_{\mathbf{g}}^\top \boldsymbol{A} \sigma(\boldsymbol{r} + \boldsymbol{B_2}) \boldsymbol{C} \text{ and } \boldsymbol{r} = \boldsymbol{A} \sigma(\boldsymbol{A X W} + \boldsymbol{B}_1) \boldsymbol{V} \quad (4)$$

where $\sigma(x) = \max(x, 0)$ is the ReLU activation function, $\boldsymbol{W} \in \mathbb{R}^{d \times m_1}$ and $\boldsymbol{V} \in \mathbb{R}^{m_1 \times m_2}$ represent the weights of m_1 and m_2 hidden nodes in the first and second layer, respectively. $\boldsymbol{B}_1 \in \mathbb{R}^{N \times m_1}$ and $\boldsymbol{B}_2 \in \mathbb{R}^{m_1 \times m_2}$ represent the bias matrices. $\boldsymbol{C} \in \mathbb{R}^{m \times K}$ is the output weight vector. $\boldsymbol{e}_g \in \mathbb{R}^N$ belongs to $\{\boldsymbol{e}_i\}_{i=1}^N$ and selects the index of the node label. We write F as $F_{\boldsymbol{A}}(\boldsymbol{e}_g, \boldsymbol{X}; \boldsymbol{W}, \boldsymbol{V})$, because we only update $\boldsymbol{W}$ and $\boldsymbol{V}$ in training, and $\boldsymbol{A}$ represents the graph topology. Note that in conventional GCNs such as (Kipf & Welling, 2017), $\boldsymbol{C}$ is a learnable parameter, and $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ can be zero. Here for the analytical purpose, we consider a slightly different model where $\boldsymbol{C}$, $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are fixed as randomly selected values.

Consider a loss function $L : \mathbb{R}^{1 \times k} \times \mathcal{Y} \to \mathbb{R}$ such that for every $y \in \mathcal{Y}$, the function $L(\cdot, y)$ is nonnegative, convex, 1-Lipschitz continuous and 1-Lipschitz smooth and $L(0, y) \in [0, 1]$. This includes both the cross-entropy loss and the ℓ_2-regression loss (for bounded $\mathcal{Y}$). The learning problem solves the following empirical risk minimization problem:

$$\min_{\boldsymbol{W}, \boldsymbol{V}} L_\Omega(\boldsymbol{W}, \boldsymbol{V}) = \frac{1}{|\Omega|} \sum_{i \in \Omega} L(F_{\boldsymbol{A}}(\boldsymbol{e}_i, \boldsymbol{X}; \boldsymbol{W}, \boldsymbol{V}), y^i) \quad (5)$$

where L_Ω is the empirical risk of the labeled nodes in Ω. The trained weights are used to estimate the unknown labels on $\mathcal{V}/\Omega$. Note that the results in this paper are *distribution-free*, and no assumption is made on the distributions of $\tilde{x}_n$ and y_n.

Training with SGD. In practice, (5) is often solved by gradient type of methods, where in iteration t, the currently estimations are updated by subtracting the product of a positive step size and the gradient of L_Ω evaluated at the current estimate. To reduce the computational complexity in estimating the gradient, an SGD method is often employed to compute the gradient of the risk of a randomly selected subset of Ω rather than using the whole set Ω.

However, due to the recursive embedding of neighboring features in GCNs, see the concatenations of $\boldsymbol{A}$ in (4), the computation and memory cost of computing the gradient can be high. Thus, graph topology sampling methods have been proposed to further reduce the computational cost.

Graph topology sampling. A node sampling method randomly removes a subset of nodes and the incident edges from $\mathcal{G}$ in each iteration independently, and the embedding aggregation is based on the reduced graph. Mathematically, in iteration s, replace $\boldsymbol{A}$ in (4) with[1] $\boldsymbol{A}^s = \boldsymbol{A}\boldsymbol{P}^s$, where $\boldsymbol{P}^s$ is a diagonal matrix, and the ith diagonal entry is 0, if node i is removed in iteration s. The non-zero diagonal entries of $\boldsymbol{P}^s$ are selected differently based on different sampling methods. Because $\boldsymbol{A}^s$ is much more sparse than $\boldsymbol{A}$, the computation and memory cost of embedding neighboring features is significantly reduced.

[1]Here we use the same sampled matrix $\boldsymbol{A}^s$ in all three layers in (4) to simplify the representation. Our analysis applies to the more general setting that each layer uses a different sampled adjacency matrix, i.e., the three $\boldsymbol{A}$ matrices in (4) are replaced with $\boldsymbol{A}^{s(1)} = \boldsymbol{A}\boldsymbol{P}^{s(1)}$, $\boldsymbol{A}^{s(2)} = \boldsymbol{A}\boldsymbol{P}^{s(2)}$, $\boldsymbol{A}^{s(3)} = \boldsymbol{A}\boldsymbol{P}^{s(3)}$, respectively, as in (Zou et al., 2019; Ramezani et al., 2020), where $\boldsymbol{P}^{s(1)}$, $\boldsymbol{P}^{s(2)}$, and $\boldsymbol{P}^{s(3)}$ are independently sampled following the same sampling strategy.

This paper will analyze the generalization performance, i.e., the prediction accuracy of unknown labels, of our algorithm framework that implements both SGD and graph topology sampling to solve (5). The details of our algorithm are discussed in Section 3.2-3.3, and the generalization performance is presented in Section 3.4.

3. Main Algorithmic and Theoretical Results

3.1. Informal Key Theoretical Findings

We first summarize the main insights of our results before presenting them formally.

1. A provable generalization guarantee of GCNs beyond two layers and with graph topology sampling. The learned GCN by our Algorithm 1 can approach the best performance of label prediction using a large class of target functions. Moreover, the prediction performance improves when the number of labeled nodes and the number of neurons m_1 and m_2 increase. This is the first generalization performance guarantee of training GCNs with graph topology sampling.

2. The explicit characterization of the impact of graph sampling through the effective adjacency matrix $\boldsymbol{A}^*$. We show that training with graph sampling returns a model that has the same label prediction performance as that of a model trained by replacing $\boldsymbol{A}$ with $\boldsymbol{A}^*$ in (4), where $\boldsymbol{A}^*$ depends on both $\boldsymbol{A}$ and the graph sampling strategy. As long as $\boldsymbol{A}^*$ can characterize the correlation among nodes properly, the learned GCN maintains a desirable prediction performance. This explains the empirical success of graph topology sampling in many datasets.

3. The explicit sample complexity bound on graph properties. We provide explicit bounds on the sample complexity and the required number of neurons, both of which grow as the node correlation increase. Moreover, the sample complexity depends on the number of neurons only logarithmically, which is consistent with the practical over-parameterization. To the best of our knowledge, (Zhang et al., 2020) is the only existing work that provides a sample complexity bound based on the graph topology, but in the non-practical and restrictive setting of two-layer GCNs. Moreover, the sample complexity bound by (Zhang et al., 2020) is polynomial in the number of neurons.

4. Tackling the non-convex interaction of weights between different layers. The convexity plays a critical role in many exiting analyses of GCNs. For instance, the analyses in (Zhang et al., 2020) require a special initialization in the local convex region of the global minimum, and the results only apply to two-layer GCNs. The NTK approach in (Du et al., 2019) considers the limiting case that the interactions across layers are negligible. Here, we directly address the non-convex interaction of weights $\boldsymbol{W}$ and $\boldsymbol{V}$ in both algorithmic design and theoretical analyses.

3.2. Graph Topology Sampling Strategy

Here we describe our graph topology sampling strategy using $\boldsymbol{A}^s$, which we randomly generate to replace $\boldsymbol{A}$ in the sth SGD iteration. Although our method is motivated for analysis and different from the existing graph sampling strategies, our insights generalize to other sampling methods like FastGCN (Chen et al., 2018a). The outline of our algorithmic framework of training GCNs with graph sampling is deferred to Section 3.3.

Suppose the node degrees in $\mathcal{G}$ can be divided into L groups with $L \geq 1$, where the degrees of nodes in group l are in the order of d_l, i.e., between cd_l and Cd_l for some constants $c \leq C$, and d_l is order-wise smaller than d_{l+1}, i.e., $d_l = o(d_{l+1})$. Let N_l denote the number of nodes in group l.

Graph sampling strategy[2]**.**. We consider a group-wise uniform sampling strategy, where S_l out of N_l nodes are sampled uniformly from each group l. For all unsampled nodes, we set the corresponding diagonal entries of a diagonal matrix $\boldsymbol{P}^s$ to be zero. If node i is sampled in this iteration and belongs to group l for any i and l, the ith diagonal entry of $\boldsymbol{P}^s$ is set as $p_l^* N_l/S_l$ for some non-negative constant p_l^*. Then $\boldsymbol{A}^s = \boldsymbol{A}\boldsymbol{P}^s$. N_l/S_l can be viewed as the scaling to compensate for the unsampled nodes in group l. p_l^* can be viewed as the scaling to reflect the impact of sampling on nodes with different importance that will be discussed in detail soon.

Effective adjacency matrix $\boldsymbol{A}^*$ by graph sampling. To analyze the impact of graph topology sampling on the learning performance, we define the effective adjacency matrix as follows:

$$\boldsymbol{A}^* = \boldsymbol{A}\boldsymbol{P}^* \tag{6}$$

where $\boldsymbol{P}^*$ is a diagonal matrix defined as

$$\boldsymbol{P}^*_{ii} = p_l^* \quad \text{if node } i \text{ belongs to degree group } l \tag{7}$$

Therefore, compared with $\boldsymbol{A}$, all the columns with indices corresponding to group l are scaled by a factor of p_l^*. We will formally analyze the impact of graph topology sampling on the generalization performance in Section 3.4, but an intuitive understanding is that our graph sampling strategy effectively changes the normalized adjacency matrix $\boldsymbol{A}$ in the GCN network model (4) to $\boldsymbol{A}^*$.

$\boldsymbol{A}^*$ can be viewed as an adjacency matrix of a weighted directed graph $\mathcal{G}'$ that reflects the node correlations, where each un-directed edge in $\mathcal{G}$ corresponds to two directed edges in $\mathcal{G}'$ with possibly different weights. $\boldsymbol{A}^*_{ji}$ measures

[2]Here we discuss asymmetric sampling as a general case. The special case of symmetric sampling is introduced in Section A.1

the impact of the feature of node i on the label of node j. If p_l^* is in the range of $(0,1)$, the corresponding entries of columns with indices in group l in $\boldsymbol{A}^*$ are smaller than those in $\boldsymbol{A}$. That means the impact of a node in group l on all other nodes is reduced from those in $\boldsymbol{A}$. Conversely, if $p_l^* > 1$, then the impact of nodes in group l in $\boldsymbol{A}^*$ is enhanced from that in $\boldsymbol{A}$.

Parameter selection and insights

(1) The scaling factor p_l^* should satisfy

$$0 \leq p_l^* \leq \frac{c_1}{L\psi_l}, \quad \forall l \tag{8}$$

for a positive constant c_1 that can be sufficiently large. ψ_l is defined as follows,

$$\psi_l := \frac{\sqrt{d_L d_l}\boldsymbol{N}_l}{\sum_{i=1}^{L} d_i \boldsymbol{N}_i} \qquad \forall l \in [L] \tag{9}$$

Note that (8) is a minor requirement for most graphs. To see this, suppose L is a constant, and every N_l is in the order of N. Then ψ_l is less than $O(1)$ for all l. Thus, all constant values of p_l^* satisfy (8) with ψ_l from (9). A special example is that p_l^* are all equal, i.e., $\boldsymbol{A}^* = c_2\boldsymbol{A}$ for some constant c_2. Because one can scale $\boldsymbol{W}$ and $\boldsymbol{V}$ by $1/c_2$ in (4) without changing the results, $\boldsymbol{A}^*$ is equivalent to $\boldsymbol{A}$ in this case.

The upper bound in (9) only becomes active in highly unbalanced graphs where there exists a dominating group $\hat{l}$ such that $\sqrt{d_{\hat{l}}}N_{\hat{l}} \gg \sqrt{d_l}N_l$ for all other l. Then the upper bound of $p_{\hat{l}}^*$ is much smaller than those for other p_l^*. Therefore, the columns of $\boldsymbol{A}^*$ that correspond to group $\hat{l}$ are scaled down more significantly than other columns, indicating that the impact of group $\hat{l}$ is reduced more significantly than other groups in $\boldsymbol{A}^*$. Therefore, the takeaway is that **graph topology sampling reduces the impact of dominating nodes more than other nodes, resulting in a more balanced $\boldsymbol{A}^*$ compared with $\boldsymbol{A}$.**

(2) The number of sampled nodes shall satisfy

$$\frac{S_l}{N_l} \geq (1 + \frac{c_1 \text{poly}(\epsilon)}{Lp_l^*\psi_l})^{-1} \qquad \forall l \in [L] \tag{10}$$

where ϵ is a small positive value. The sampling requirement in (10) has two takeaways. **First, the higher-degree groups shall be sampled more frequently than lower-degree groups.** To see this, consider a special case that $p_l^* = 1$, and $N_l = N/L$ for all l. Then (10) indicates that S_l is larger in a group l with a larger d_l. This intuition is the same as FastGCN (Chen et al., 2018a), which also samples high-degree nodes with a higher probability in many cases. Therefore, the insights from our graph sampling method also apply to other sampling methods such as FastGCN. We will show the connection to FastGCN empirically in Section 4.2. **Second, reducing the number of samples in group l corresponds to reducing the impact of group l in $\boldsymbol{A}^*$.** To see this, note that decreasing p_l^* reduces the right-hand side of (10).

Algorithm 1 Training with SGD and graph topology sampling

Input: Normalized adjacency matrix $\boldsymbol{A}$, node features $\boldsymbol{X}$, known node labels in Ω, the step size η, the number of inner iterations T_w, the number of outer iterations T, σ_w, σ_v, λ_w, λ_v.
Initialize $\boldsymbol{W}^{(0)}$, $\boldsymbol{V}^{(0)}$, $\boldsymbol{B}_1$, $\boldsymbol{B}_2$, $\boldsymbol{C}$.
$\boldsymbol{W}_0 = 0$, $\boldsymbol{V}_0 = 0$.
for $t = 0, 1, \cdots, T-1$ **do**
  Apply noisy SGD with step size η on the stochastic objective $\hat{L}_\Omega(\lambda_t; \boldsymbol{W}, \boldsymbol{V})$ in (11) for T_w steps. To generate the stochastic objective in each step s, randomly sample a batch of labeled nodes Ω^s from Ω; generate $\boldsymbol{A}^s$ using graph sampling; randomly generate $\boldsymbol{W}^\rho$, $\boldsymbol{V}^\rho$ and $\boldsymbol{\Sigma}$.
  Let the starting point be $\boldsymbol{W} = \boldsymbol{W}_t$, $\boldsymbol{V} = \boldsymbol{V}_t$ and suppose it reaches $\boldsymbol{W}_{t+1}$ and $\boldsymbol{V}_{t+1}$.
  $\lambda_{t+1} = \lambda_t \cdot (1-\eta)$.
end for
Output:
$\boldsymbol{W}^{(out)} = \sqrt{\lambda_{T-1}}(\boldsymbol{W}^{(0)} + \boldsymbol{W}^\rho + \boldsymbol{W}_T\boldsymbol{\Sigma})$
$\boldsymbol{V}^{(out)} = \sqrt{\lambda_{T-1}}(\boldsymbol{V}^{(0)} + \boldsymbol{V}^\rho + \boldsymbol{\Sigma}\boldsymbol{V}_T)$.

3.3. The Algorithmic Framework of Training GCNs

Because (5) is non-convex, solving it directly using SGD can get stuck at a bad local minimum in theory. The main idea in the theoretical analysis to address this non-convexity is to add weight decay and regularization in the objective of (5) such that with a proper regularization, any second-order critical point is *almost* a global minimum.

Specifically, for initialization, entries of $\boldsymbol{W}^{(0)}$ are i.i.d. from $\mathcal{N}(0, \frac{1}{m_1})$, and entries of $\boldsymbol{V}^{(0)}$ are i.i.d. from $\mathcal{N}(0, \frac{1}{m_2})$. $\boldsymbol{B}_1$ (or $\boldsymbol{B}_2$) is initialized to be an all-one vector multiplying a row vector with i.i.d. samples from $\mathcal{N}(0, \frac{1}{m_1})$ (or $\mathcal{N}(0, \frac{1}{m_2})$). Entries of $\boldsymbol{C}$ are drawn i.i.d. from $\mathcal{N}(0,1)$.

In each outer loop $t = 0, ..., T-1$, we use noisy SGD[3] with step size η for T_w iterations to minimize the stochastic objective function $\hat{L}_\Omega$ in (11) with some fixed λ_{t-1}, where

[3]Noisy SGD is vanilla SGD plus Gaussian perturbation. It is a common trick in the theoretical analyses of non-convex optimization (Ge et al., 2015) and is not needed in practice.

$\lambda_0 = 1$, and the weight decays with $\lambda_{t+1} = (1-\eta)\lambda_t$.

$$\begin{aligned}&\hat{L}_\Omega(\lambda_t; \boldsymbol{W}, \boldsymbol{V})\\ =&L_\Omega(\sqrt{\lambda_t}(\boldsymbol{W}^{(0)} + \boldsymbol{W}^\rho + \boldsymbol{W}\boldsymbol{\Sigma}), \sqrt{\lambda_t}(\boldsymbol{V}^{(0)} + \boldsymbol{V}^\rho + \boldsymbol{\Sigma}\boldsymbol{V}))\\ &+ \lambda_w \|\sqrt{\lambda_t}\boldsymbol{W}\|_{2,4}^4 + \lambda_v \|\sqrt{\lambda_t}\boldsymbol{V}\|_F^2\end{aligned} \tag{11}$$

$\hat{L}_\Omega(\lambda_t; \boldsymbol{W}, \boldsymbol{V})$ is stochastic because in each inner iteration s, (1) we randomly sample a subset Ω^s of labeled nodes; (2) we randomly sample $\boldsymbol{A}^s$ from the graph topology sampling method in Section 3.2; (3) $\boldsymbol{W}^\rho$ and $\boldsymbol{V}^\rho$ are small perturbation matrices with entries i.i.d. drawn from $\mathcal{N}(0, \sigma_w^2)$ and $\mathcal{N}(0, \sigma_v^2)$, respectively; and (4) $\boldsymbol{\Sigma} \in \mathbb{R}^{m_1 \times m_1}$ is a random diagonal matrix with diagonal entries uniformly drawn from $\{1, -1\}$. $\boldsymbol{W}^\rho$ and $\boldsymbol{V}^\rho$ are standard Gaussian smoothing in the literature of theoretical analyses of non-convex optimization, see, e.g. (Ge et al., 2015), and are not needed in practice. $\boldsymbol{\Sigma}$ is similar to the practical Dropout (Srivastava et al., 2014) technique that randomly masks out neurons and is also introduced for the theoretical analysis only.

The last two terms in (11) are additional regularization terms for some positive λ_w and λ_v. As shown in (Allen-Zhu et al., 2019), $\|\cdot\|_{2,4}$ is used for the analysis to drive the weights to be evenly distributed among neurons. The practical regularization $\|\cdot\|_F$ has the same effect in empirical results, while the theoretical justification is open.

Algorithm 1 summarizes the algorithm with the parameter selections in Table 1. Let $\boldsymbol{W}^{out}$ and $\boldsymbol{V}^{out}$ denote the returned weights. We use $F_{\boldsymbol{A}^*}(\boldsymbol{e}_i, \boldsymbol{X}; \boldsymbol{W}^{out}, \boldsymbol{V}^{out})$ to predict the label of node i. This might sound different from the conventional practice which uses $\boldsymbol{A}$ in predicting unknown labels. However, note that $\boldsymbol{A}^*$ only differs from $\boldsymbol{A}$ by a column-wise scaling as from (6). Moreover, $\boldsymbol{A}^*$ can be set as $\boldsymbol{A}$ in many practical datasets based on our discussion after (9). Here we use the general form of $\boldsymbol{A}^*$ for the purpose of analysis.

We remark that our framework of algorithm and analysis can be easily applied to the simplified setup of two-layer GCNs. The resulting algorithm is much simplified to a vanilla SGD plus graph topology sampling. All the additional components above are introduced to address the non-convex interaction of $\boldsymbol{W}$ and $\boldsymbol{V}$ theoretically and may not be needed for practical implementation. We skip the discussion of two-layer GCNs in this paper.

Table 1. Parameter choices for Algorithm 1

λ_v	$2\epsilon_0 m_2 / m_1^{1-0.01}$	σ_v	$1/m_2^{1/2+0.01}$
λ_w	$2\epsilon_0 m_1^{3-0.002} / C_0^4$	σ_w	$1/m_1^{1-0.01}$
C	$\mathcal{C}_\epsilon(\phi, \|\boldsymbol{A}^*\|_\infty)\sqrt{\|\boldsymbol{A}^*\|_\infty^2 + 1}$	C'	$10C\sqrt{p_2}$
C''	$\mathcal{C}_\epsilon(\Phi, C')\sqrt{\|\boldsymbol{A}^*\|_\infty^2 + 1}$	C_0	$\tilde{O}(p_1^2 p_2 K^2 C C'')$

3.4. Generalization Guarantee

Our formal generalization analysis shows that our learning method returns a GCN model that approaches the minimum prediction error that can be achieved by the best function in a large concept class of target functions, which have two important properties: (1) the prediction error decreases as size of the function class increases; and (2) the concept class uses $\boldsymbol{A}^*$ in (6) as the adjacency matrix of the graph topology. Therefore, the result implies that if $\boldsymbol{A}^*$ accurately captures the correlations among node features and labels, the learned GCN model can achieve a small prediction error of unknown labels. Moreover, no other functions in a large concept class can perform better than the learned GCN model. To formalize the results, we first define the target functions as follows.

Concept class and target function F^*. Consider a concept class consisting of target functions $F^* : \mathbb{R}^N \times \mathbb{R}^{N \times d} \to \mathbb{R}^{1 \times K}$:

$$\begin{aligned}F^*_{\boldsymbol{A}^*}(\boldsymbol{e}_g, \boldsymbol{X}) &= \boldsymbol{e}_g^\top \boldsymbol{A}^* \big(\Phi(\boldsymbol{r}_1) \odot \boldsymbol{r}_2\big)\boldsymbol{C}^*\\ \boldsymbol{r}_1 &= \boldsymbol{A}^* \phi_1(\boldsymbol{A}^* \boldsymbol{X} \boldsymbol{W}_1^*)\boldsymbol{V}_1^*\\ \boldsymbol{r}_2 &= \boldsymbol{A}^* \phi_2(\boldsymbol{A}^* \boldsymbol{X} \boldsymbol{W}_2^*)\boldsymbol{V}_2^*\end{aligned} \tag{12}$$

where ϕ_1, ϕ_2, Φ: $\mathbb{R} \to \mathbb{R}$ all infinite-order smooth[4]. The parameters $\boldsymbol{W}_1^*, \boldsymbol{W}_2^* \in \mathbb{R}^{d \times p_2}$, $\boldsymbol{V}_1^*, \boldsymbol{V}_2^* \in \mathbb{R}^{p_2 \times p_1}$, $\boldsymbol{C}^* \in \mathbb{R}^{p_1 \times k}$ satisfy that every column of $\boldsymbol{W}_1^*$, $\boldsymbol{W}_2^*$, $\boldsymbol{V}_1^*$, $\boldsymbol{V}_2^*$ is unit norm, and the maximum absolute value of $\boldsymbol{C}^*$ is at most 1. The effective adjacency matrix $\boldsymbol{A}^*$ is defined in (6). Define

$$\mathcal{C}_\epsilon(\phi, R) = \max\big(\mathcal{C}_\epsilon(\phi_1, R), \mathcal{C}_\epsilon(\phi_2, R)\big), \tag{13}$$

$$\mathcal{C}_s(\phi, R) = \max\big(\mathcal{C}_s(\phi_1, R), \mathcal{C}_2(\phi_1, R)\big). \tag{14}$$

We focus on target functions where the function complexity $\mathcal{C}_\epsilon(\Phi, R), \mathcal{C}_s(\Phi, R), \mathcal{C}_\epsilon(\phi, R), \mathcal{C}_s(\phi, R)$, defined in (2)-(3), (13)-(14), as well as p_1 and p_2, are all bounded.

(12) is more general than GCNs. If $\boldsymbol{r}_2$ is a constant matrix, (12) models a GCN, where $\boldsymbol{W}_1^*$ and $\boldsymbol{V}_1^*$ are weight matrices in the first and second layer, respectively, and ϕ_1 and Φ are the activation functions in each layer.

Modeling the prediction error of unknown labels. We will show that the learned GCN by our method performs almost the same as the best function in the concept class in (12) in predicting unknown labels. Because the practical datasets usually contain noise in features and labels, we employ a probabilistic model to model the data. Note that our

[4]When Φ is operated on a matrix $\boldsymbol{r}_1$, $\Phi(\boldsymbol{r}_1)$ means applying Φ on each entry of $\boldsymbol{r}_1$. In fact, our results still hold for a more general case that a different function Φ_j is applied to every entry of the jth column of $\boldsymbol{r}_1$, $j \in [p_2]$. We keep the simpler model to have a more compact representation. The similar arguments hold for ϕ_1, ϕ_2.

result is distribution-free , and the following distributions are introduced for the presentation of the results.

Specifically, let $\mathcal{D}_{\tilde{x}_n}$ denote the distribution from which the feature $\tilde{x}_n$ of node n is drawn. For example, when the noise level is low, $\mathcal{D}_{\tilde{x}_n}$ can be a distribution centered at the observed feature of node n with a small variance. Similarly, let $\mathcal{D}_{y_n}$ denote the distribution from which the label y_n at node n is drawn. Let $\boldsymbol{e}_g$ be uniformly selected from $\{\boldsymbol{e}_i\}_{i=1}^N \in \mathbb{R}^N$. Let $\mathcal{D}$ denote the concatenation of these distributions of a data point

$$z = (\boldsymbol{e}_g, \boldsymbol{X}, y) \in \mathbb{R}^N \times \mathbb{R}^{N\times d} \times \mathcal{Y}. \tag{15}$$

Then the given feature matrix $\boldsymbol{X}$ and partial labels in Ω can be viewed as $|\Omega|$ identically distributed but *correlated* samples from $\mathcal{D}$. The correlation results from the fact that the label of node i depends on not only the feature of node i but also neighboring features. This model of correlated samples is different from the conventional assumption of i.i.d. samples in supervised learning and makes our analyses more involved.

Let

$$\mathrm{OPT}_{\boldsymbol{A}^*} = \min_{\substack{\boldsymbol{W}_1^*,\, \boldsymbol{W}_2^*,\\ \boldsymbol{V}_1^*,\, \boldsymbol{V}_2^*,\, \boldsymbol{C}^*}} \mathbb{E}_{(\boldsymbol{e}_g,\boldsymbol{X},y)\sim\mathcal{D}} L(F_{\boldsymbol{A}^*}^*(\boldsymbol{e}_g, \boldsymbol{X}), y) \tag{16}$$

be the smallest population risk achieved by the best target function (over the choices of $\boldsymbol{W}_1^*, \boldsymbol{W}_2^*, \boldsymbol{V}_1^*, \boldsymbol{V}_2^*, \boldsymbol{C}^*$) in the concept class $F_{\boldsymbol{A}^*}^*$ in (12). $\mathrm{OPT}_{\boldsymbol{A}^*}$ measures the average loss of predicting the unknown labels if the estimates are computed using the best target function in (12). Clearly, $\mathrm{OPT}_{\boldsymbol{A}^*}$ decreases as the size of the concept increases, i.e., when p_1 and p_2 increase. Moreover, if $\boldsymbol{A}^*$ indeed models the node correlations accurately, $\mathrm{OPT}_{\boldsymbol{A}^*}$ can be very small, indicating a desired generalization performance. We next show that the population risk of the learned GCN model by our method can be arbitrarily close to $\mathrm{OPT}_{\boldsymbol{A}^*}$.

Theorem 3.1. *For every $\gamma \in (0, \frac{1}{4}]$, every $\epsilon_0 \in (0, \frac{1}{100}]$, every $\epsilon \in (0, (Kp_1p_2^2\mathcal{C}_s(\Phi, p_2\mathcal{C}_s(\phi, O(1)))\mathcal{C}_s(\phi, O(1)) \cdot \|\boldsymbol{A}^*\|_\infty^2)^{-1}\epsilon_0)$, as long as*

$$\begin{aligned} &m_1 = m_2 = m \\ \geq &\mathit{poly}\Big(\mathcal{C}_\epsilon\big(\Phi, \mathcal{C}_\epsilon(\phi, O(1))\big), p_2, \|\boldsymbol{A}^*\|_\infty, \frac{1}{\epsilon}\Big) \end{aligned} \tag{17}$$

$$\begin{aligned} |\Omega| \geq &\Theta(\epsilon_0^{-2}\|\boldsymbol{A}^*\|_\infty^8 K^6(1 + p_1^4 p_2^5 \mathcal{C}_\epsilon(\Phi, \sqrt{p_2}\mathcal{C}_\epsilon(\phi, O(1))) \\ &\cdot \mathcal{C}_\epsilon(\phi, O(1))(\|\boldsymbol{A}^*\|_\infty + 1)^4)(1+\delta)^4 \log N \log m), \end{aligned} \tag{18}$$

(8) and (10) hold, there is a choice $\eta = 1/\mathit{poly}(\|\boldsymbol{A}^*\|_\infty, K, m)$ and $T = \mathit{poly}(\|\boldsymbol{A}^*\|_\infty, K, m)$ such that with probability at least 0.99 ,

$$\begin{aligned} &\mathbb{E}_{(\boldsymbol{e}_g,\boldsymbol{X},y)\in\mathcal{D}} L(F_{\boldsymbol{A}^*}(\boldsymbol{e}_g, \boldsymbol{X}; \boldsymbol{W}^{(out)}, \boldsymbol{V}^{(out)}), y) \\ \leq &(1+\gamma)\mathrm{OPT}_{\boldsymbol{A}^*} + \epsilon_0, \end{aligned} \tag{19}$$

where $\boldsymbol{A}^*$ is the effective adjacency matrix in (12).

Theorem 3.1 shows that the required sample complexity is polynomial in $\|\boldsymbol{A}^*\|$ and δ, where δ is the maximum node degree without self-connections in $\boldsymbol{A}$. Note that condition (8) implies that $\|\boldsymbol{A}^*\|_\infty$ is $O(1)$. Then as long as δ is $O(N^\alpha)$ for some small α in $(0, 1)$, say $\alpha = 1/5$, then one can accurately infer the unknown labels from a small percentage of labeled nodes. Moreover, our sample complexity is sufficient but not necessary. It is possible to achieve a desirable generalization performance if the number of labeled nodes is less than the bound in (18).

Graph topology sampling affects the generalization performance through $\boldsymbol{A}^*$. From the discussion in Section 3.2, graph sampling reduces the node correlation in $\boldsymbol{A}^*$, especially for dominating nodes. The generalization performance does not degrade when $\mathrm{OPT}_{\boldsymbol{A}^*}$ is small, i.e., the resulting $\boldsymbol{A}^*$ is sufficient to characterize the node correlation in a given dataset. That explains the empirical success of graph sampling in many datasets.

4. Numerical Results

To unveil how our theoretical results are aligned with GCN's generalization performance in experiments, we will focus on numerical evaluations on synthetic data where we can control target functions and compare with $\boldsymbol{A}^*$ explicitly. We also evaluate both our graph sampling method and FastGCN (Chen et al., 2018a) to validate that insights for our graph sampling method also apply to FastGCN.

We generate a graph $\mathcal{G}$ with $N = 2000$ nodes. $\mathcal{G}$ has two degree groups. Group 1 has N_1 nodes, and every node degree approximately equals d_1. Group 2 has N_2 nodes, and every node degree approximately equals d_2. The edges between nodes are randomly selected. $\boldsymbol{A}$ is the normalized adjacency matrix of $\mathcal{G}$.

The node labels are generated by the target function

$$y = (\sin(\hat{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}^*) \odot \tanh(\hat{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}^*))\boldsymbol{C}^*, \tag{20}$$

where $\hat{\boldsymbol{A}} \in \mathbb{R}^{N\times N}$, $\boldsymbol{X} \in \mathbb{R}^{N\times d}$, $\boldsymbol{W}^* \in \mathbb{R}^{d\times p}$ and $\boldsymbol{C}^* \in \mathbb{R}^{p\times K}$. The feature dimension $d = 10$. $p = 10$, and $K = 2$. $\boldsymbol{X}$, $\boldsymbol{W}^*$ and $\boldsymbol{C}^*$ are all randomly generated with each entry i.i.d. from $\mathcal{N}(0, 1)$.

We consider a regression task with the ℓ_2-regression loss function. A three-layer GCN as defined in (4) with m neurons in each hidden layer is trained on a randomly selected set Ω of labeled nodes. The rest $N - |\Omega|$ labels are used for testing. The learning rate $\eta = 10^{-3}$. The mini-batch size is 5, and the dropout rate as 0.4. The total number of iterations is $TT_w = 4|\Omega|$. Our graph topology sampling method samples $S_1 = 0.9N_1$ and $S_2 = 0.9N_2$ nodes for both groups in each iteration.

4.1. Sample Complexity and Neural Network Width with respect to $\|\boldsymbol{A}^*\|_\infty$

We fix $N_1 = 100$, $N_2 = 1900$ and vary $\boldsymbol{A}$ by changing node degrees d_1 and d_2. In the graph topology sampling method, $p_1^* = 0.7$ and $p_2^* = 0.3$. For every fixed $\boldsymbol{A}$, the effective adjacency matrix $\boldsymbol{A}^*$ is computed based on (6) using p_1^* and p_2^*. Synthetic labels are generated based on (20) using $\boldsymbol{A}^*$ as $\hat{\boldsymbol{A}}$.

Figure 1 shows the testing error decreases as the number of labeled nodes $|\Omega|$ increases, when the number of neurons per layer m is fixed as 500. Moreover, as $\|\boldsymbol{A}^*\|_\infty$ increases, the required number of labeled nodes increases to achieve the same level of testing error. This verifies our sample complexity bound in (18).

Figure 2 shows the testing error decreases as m increases when $|\Omega|$ is fixed as 1500. Moreover, as $\|\boldsymbol{A}^*\|_\infty$ increases, a larger m is needed to achieve the same level of testing error. This verifies our bound on the number of neurons in (17).

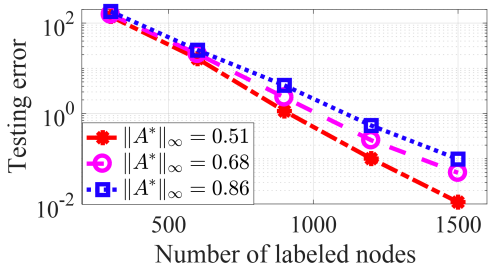

Figure 1. The testing error when $|\Omega|$ and $\|\boldsymbol{A}^*\|_\infty$ change. $m = 500$

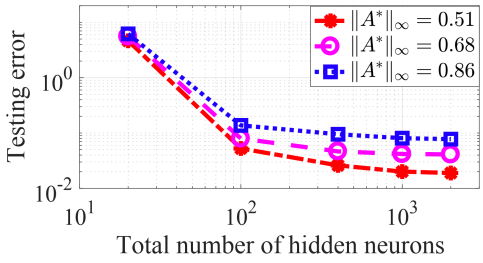

Figure 2. The testing error when m and $\|\boldsymbol{A}^*\|_\infty$ change. $|\Omega| = 1500$.

4.2. Graph Sampling Affects $\boldsymbol{A}^*$

Here we fix $\boldsymbol{A}$ and the graph sampling strategy, and evaluate the prediction performance on datasets generated by (20) using different $\hat{\boldsymbol{A}}$. We generate $\hat{\boldsymbol{A}}$ from $\hat{\boldsymbol{A}} = \boldsymbol{A}\hat{\boldsymbol{P}}$, where $\hat{\boldsymbol{P}}$ is a diagonal matrix with $\hat{\boldsymbol{P}}_{ii} = \hat{p}_1$ for nodes i in group 1 and $\hat{\boldsymbol{P}}_{ii} = \hat{p}_2$ for nodes i in group 2. We vary $\hat{p}_1$ and $\hat{p}_2$ to generate three different datasets from (20). We consider both our graph sampling method in Section 3.2 and FastGCN (Chen et al., 2018a).

In Figure 3, $N_1 = 100$ and $N_2 = 1900$. $d_1 = 10$ and $d_2 = 1$. Figure 3(a) shows the testing performance of a learned GCN by Algorithm 1, where $p_1^* = 0.9$ and $p_2^* = 0.1$. the method indeed performs the best on Dataset 1 when $\hat{\boldsymbol{A}}$ is generated using $\hat{p}_1 = 0.9$ and $\hat{p}_2 = 0.1$, in which case $\boldsymbol{A}^* = \hat{\boldsymbol{A}}$. This verifies our theoretical result that graph sampling affects $\boldsymbol{A}^*$ in the target functions, i.e., it achieves the best performance if $\boldsymbol{A}^*$ is the same as $\hat{\boldsymbol{A}}$ in the target function.

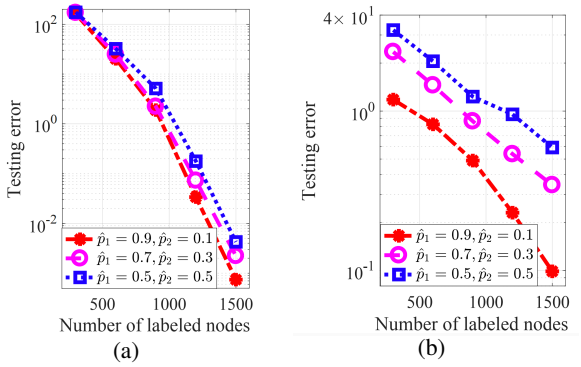

Figure 3. Generalization performance of learned GCNs on datasets generated from different $\hat{\boldsymbol{A}}$ by (a) our graph sampling strategy and (b) FastGCN. $\boldsymbol{A}$ is very unbalanced.

Fig. 3 (b) shows the performance on the same three datasets where in each iteration of Algorithm 1, the graph sampling strategy is replaced with FastGCN (Chen et al., 2018a). The method also performs the best in Dataset 1 when $\boldsymbol{A}^*$ is generated using $\hat{p}_1 = 0.9$ and $\hat{p}_2 = 0.1$. The reason is that the graph topology is highly unbalanced in the sense that $\sqrt{d_2}N_2 \gg \sqrt{d_1}N_1$, which means group 2 has a much higher impact on other nodes in group 1 in $\boldsymbol{A}$. The graph sampling reduces the impact of group 2 nodes more significantly than group 1 nodes, as discussed in Section 3.2.

To further illustrate this, in Figure 4 we change the graph topology by setting $N_1 = 1000$ and $N_2 = 1000$, and all the other settings remain the same. In this case, the graph is balanced because $\sqrt{d_2}N_2$ and $\sqrt{d_1}N_1$ are in the same order. We generate different datasets using the new $\boldsymbol{A}$ following the same method and evaluate the performance of both our graph sampling method and FastGCN. Both methods perform the best in Dataset 3 when $\hat{\boldsymbol{A}}$ is generated using $\hat{p}_1 = 0.5$ and $\hat{p}_2 = 0.5$. That is because on a balanced graph, graph sampling reduces the impact of both groups equally.

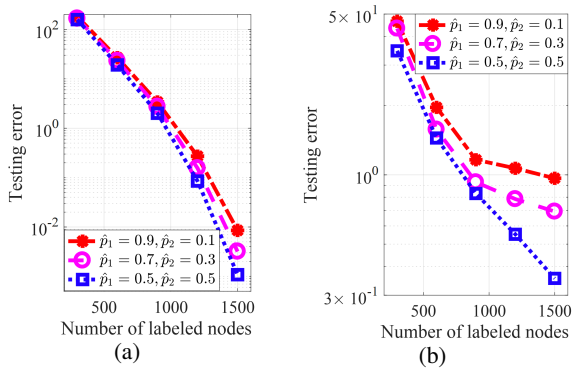

Figure 4. Generalization performance of learned GCNs on datasets generated from different $\boldsymbol{A}^*$ by (a) our graph sampling strategy and (b) FastGCN. $\boldsymbol{A}$ is balanced.

5. Conclusion

This paper provides a new theoretical framework for explaining the empirical success of graph sampling in training GCNs. It quantifies the impact of graph sampling explicitly through the effective adjacency matrix and provides generalization and sample complexity analyses. One future direction is to develop active graph sampling strategies based on the presented insights and analyze its generalization performance. Other potential extension includes the construction of statistical-model-based characterization of $\boldsymbol{A}^*$ and fitness to real-world data, and the generalization analysis of deep GCNs, graph auto-encoders, and jumping knowledge networks.

Acknowledgements

This work was supported by AFOSR FA9550-20-1-0122, ARO W911NF-21-1-0255, NSF 1932196 and the Rensselaer-IBM AI Research Collaboration (http://airc.rpi.edu), part of the IBM AI Horizons Network (http://ibm.biz/AIHorizons). We thank Ruisi Jian, Haolin Xiong at Rensselaer Polytechnic Institute for the help in formulating numerical experiments. We thank all anonymous reviewers for their constructive comments.

References

Allen-Zhu, Z., Li, Y., and Liang, Y. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in neural information processing systems*, pp. 6158–6169, 2019.

Battaglia, P., Pascanu, R., Lai, M., Rezende, D. J., et al. Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems*, pp. 4502–4510, 2016.

Chen, J., Ma, T., and Xiao, C. Fastgcn: Fast learning with graph convolutional networks via importance sampling. In *International Conference on Learning Representations*, 2018a.

Chen, J., Zhu, J., and Song, L. Stochastic training of graph convolutional networks with variance reduction. In *International Conference on Machine Learning*, pp. 942–950. PMLR, 2018b.

Chen, T., Sui, Y., Chen, X., Zhang, A., and Wang, Z. A unified lottery ticket hypothesis for graph neural networks. In *International Conference on Machine Learning*, pp. 1695–1706. PMLR, 2021.

Chiang, W.-L., Liu, X., Si, S., Li, Y., Bengio, S., and Hsieh, C.-J. Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 257–266, 2019.

Cong, W., Ramezani, M., and Mahdavi, M. On provable benefits of depth in training graph convolutional networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Daniely, A. Sgd learns the conjugate kernel class of the network. *Advances in Neural Information Processing Systems*, 30:2422–2430, 2017.

Du, S. S., Hou, K., Salakhutdinov, R. R., Poczos, B., Wang, R., and Xu, K. Graph neural tangent kernel: Fusing graph neural networks with graph kernels. In *Advances in Neural Information Processing Systems*, pp. 5724–5734, 2019.

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. P. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pp. 2224–2232, 2015.

Fu, H., Chi, Y., and Liang, Y. Guaranteed recovery of one-hidden-layer neural networks via cross entropy. *IEEE Transactions on Signal Processing*, 68:3225–3235, 2020.

Garg, V., Jegelka, S., and Jaakkola, T. Generalization and representational limits of graph neural networks. In *International Conference on Machine Learning*, pp. 3419–3430. PMLR, 2020.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pp. 797–842. PMLR, 2015.

Hamilton, W., Ying, Z., and Leskovec, J. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pp. 1024–1034, 2017.

Hu, H., Gu, J., Zhang, Z., Dai, J., and Wei, Y. Relation networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3588–3597, 2018.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *Proc. International Conference on Learning (ICLR)*, 2017.

Li, H., Zhang, S., and Wang, M. Learning and generalization of one-hidden-layer neural networks, going beyond standard gaussian data. In *2022 56th Annual Conference on Information Sciences and Systems (CISS)*, pp. 37–42. IEEE, 2022.

Li, J., Zhang, T., Tian, H., Jin, S., Fardad, M., and Zafarani, R. Sgcn: A graph sparsifier based on graph convolutional networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 275–287. Springer, 2020.

Liao, R., Urtasun, R., and Zemel, R. A pac-bayesian approach to generalization bounds for graph neural networks. In *International Conference on Learning Representations*, 2021.

Lv, S. Generalization bounds for graph convolutional neural networks via rademacher complexity. *arXiv preprint arXiv:2102.10234*, 2021.

Oono, K. and Suzuki, T. Optimization and generalization analysis of transduction through gradient boosting and application to multi-scale graph neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

Peng, N., Poon, H., Quirk, C., Toutanova, K., and Yih, W.-t. Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics*, 5:101–115, 2017.

Ramezani, M., Cong, W., Mahdavi, M., Sivasubramaniam, A., and Kandemir, M. Gcn meets gpu: Decoupling "when to sample" from "how to sample". *Advances in Neural Information Processing Systems*, 33:18482–18492, 2020.

Sanchez-Gonzalez, A., Heess, N., Springenberg, J. T., Merel, J., Riedmiller, M., Hadsell, R., and Battaglia, P. Graph networks as learnable physics engines for inference and control. In *International Conference on Machine Learning*, pp. 4470–4479. PMLR, 2018.

Satorras, V. G. and Estrach, J. B. Few-shot learning with graph neural networks. In *International Conference on Learning Representations*, 2018.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Van den Berg, R., Kipf, T. N., and Welling, M. Graph convolutional matrix completion. In *KDD*, 2018.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. Graph attention networks. *International Conference on Learning Representations (ICLR)*, 2018.

Verma, S. and Zhang, Z.-L. Stability and generalization of graph convolutional neural networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1539–1548, 2019.

Wang, X., Ye, Y., and Gupta, A. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6857–6866, 2018.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? *International Conference on Learning Representations (ICLR)*, 2019.

Xu, K., Zhang, M., Jegelka, S., and Kawaguchi, K. Optimization of graph neural networks: Implicit acceleration by skip connections and more depth. In *International Conference on Machine Learning*. PMLR, 2021.

Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., and Leskovec, J. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 974–983, 2018.

Zhang, S., Wang, M., Liu, S., Chen, P.-Y., and Xiong, J. Fast learning of graph neural networks with guaranteed generalizability: One-hidden-layer case. *arXiv preprint arXiv:2006.14117*, 2020.

Zheng, C., Zong, B., Cheng, W., Song, D., Ni, J., Yu, W., Chen, H., and Wang, W. Robust graph representation learning via neural sparsification. In *International Conference on Machine Learning*, pp. 11458–11468. PMLR, 2020.

Zhong, K., Song, Z., Jain, P., Bartlett, P. L., and Dhillon, I. S. Recovery guarantees for one-hidden-layer neural

networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 4140–4149. JMLR. org, https://arxiv.org/abs/1706.03175, 2017.

Zhou, X. and Wang, H. The generalization error of graph convolutional networks may enlarge with more layers. *Neurocomputing*, 424:97–106, 2021.

Zou, D., Hu, Z., Wang, Y., Jiang, S., Sun, Y., and Gu, Q. Layer-dependent importance sampling for training deep and large graph convolutional networks. *Advances in Neural Information Processing Systems*, 32:11249–11259, 2019.

A. Preliminaries

Lemma A.1. $\|\tilde{\boldsymbol{a}}_n \boldsymbol{X}\| \leq \|\boldsymbol{A}\|_\infty$.

Proof:

$$\begin{aligned}\|\tilde{\boldsymbol{a}}_n \boldsymbol{X}\| &= \|\sum_{k=1}^{N} a_{n,k}\tilde{\boldsymbol{x}}_k\| \\ &= \|\sum_{k=1}^{N} \frac{a_{n,k}}{\sum_{k=1}^{N} a_{n,k}}\tilde{\boldsymbol{x}}_k\| \cdot \sum_{k=1}^{N} a_{n,k} \\ &\leq \sum_{k=1}^{N} \frac{a_{n,k}}{\sum_{k=1}^{N} a_{n,k}}\|\tilde{\boldsymbol{x}}_k\| \cdot \|\boldsymbol{A}\|_\infty \\ &= \|\boldsymbol{A}\|_\infty\end{aligned} \tag{21}$$

where the second to last step is by the convexity of $\|\cdot\|$.

Lemma A.2. *Given a graph $\mathcal{G}$ with $L(\geq 1)$ groups of nodes, where the group i with node degree d_i is denoted as $\mathcal{N}_i$. Suppose that in iteration t, $\boldsymbol{A}^t$ (or any of $\boldsymbol{A}^{t(1)}$, $\boldsymbol{A}^{t(2)}$, $\boldsymbol{A}^{t(3)}$ in the general setting) is generated from the sampling strategy in Section 3.2, if the number of sampled nodes satisfies $l_i \geq |\mathcal{N}_i|/(1 + \frac{c_1 poly(\epsilon)}{L p_i^* \Psi_i})$, we have*

$$\|\boldsymbol{A}^t - \boldsymbol{A}^*\|_\infty \leq poly(\epsilon) \tag{22}$$

Proof:
From Section 3.2, we can rewrite that

$$\tilde{\boldsymbol{a}}_n^t = \begin{cases} \frac{|\mathcal{N}_k|}{l_k} p_k^* A_{n,j}, & \text{if the nodes } n, j \text{ are connected and } j \text{ is selected and } j \in \mathcal{N}_k \\ 0, & \text{else} \end{cases} \tag{23}$$

$$\tilde{\boldsymbol{a}^*}_n = \begin{cases} p_k^* A_{n,j}, & \text{if the nodes } n, j \text{ are connected and } j \in \mathcal{N}_k \\ 0, & \text{else} \end{cases} \tag{24}$$

Let $\boldsymbol{A}^* = (\tilde{\boldsymbol{a}^*}_1^\top, \tilde{\boldsymbol{a}^*}_2^\top, \cdots, \tilde{\boldsymbol{a}^*}_n)^\top$. Since that we need that $\sum_{j=1}^{N} A_{n,j}^* \leq O(1)$, we require

$$p_i^* \sum_{j \in \mathcal{N}_i} A_{n,j} \leq O(1/L), \text{ holds for any } i \in [L], n \in [N] \tag{25}$$

We first roughly compute the ratio of edges that one node is connected to the nodes in another group. For the node with degree $\deg(i)$, it has $\deg(i) - 1$ open edges except the self-connection. Hence, the group with degree $\deg(j)$ has $(\deg(j) - 1)|\mathcal{N}_j|$ open edges except self-connections in total. Therefore, the ratio of the edges connected to the group j to all groups is

$$\frac{(\deg(j) - 1)|\mathcal{N}_j|}{\sum_{l=1}^{L}(\deg(l) - 1)|\mathcal{N}_l|} \approx \frac{d_j|\mathcal{N}_j|}{\sum_{l=1}^{L} d_l|\mathcal{N}_l|} \tag{26}$$

Define

$$\Psi(n, i) = \sqrt{\frac{d_n}{d_i}} \cdot \frac{d_i|\mathcal{N}_i|}{\sum_{l=1}^{L} d_l|\mathcal{N}_l|} \tag{27}$$

Then, as long as

$$p_i^* \sum_{j \in |\mathcal{N}_i|} A_{n,j} \approx p_i^* \frac{1}{\sqrt{d_i d_n}} \cdot \frac{d_i|\mathcal{N}_i|}{\sum_{l=1}^{L} d_l|\mathcal{N}_l|} d_n \lesssim p_i^* \Psi(n, i) \leq O(1/L) \tag{28}$$

i.e.,

$$p_i^* \leq \frac{c_1}{L \cdot \max_{n \in [L]}\{\Psi(n, i)\}} = \frac{c_1}{L \cdot \Psi(L, i)} = \frac{c_1}{L}\sqrt{\frac{d_i}{d_L}}\frac{\sum_{l=1}^{L} d_l|\mathcal{N}_l|}{d_i|\mathcal{N}_i|} \tag{29}$$

for some constant $c_1 > 0$, we can obtain that $\|\boldsymbol{A}^*\|_\infty \leq O(1)$. Since that

$$\sum_{j\in\mathcal{S}_k} A_{n,j} \approx \frac{1}{\sqrt{d_i d_n}} \cdot \frac{d_i|\mathcal{N}_i|}{\sum_{l=1}^L d_l|\mathcal{N}_l|} d_n \frac{l_k}{|\mathcal{N}_k|} \approx \sum_{j\in\mathcal{N}_k} A_{n,j}\frac{l_k}{|\mathcal{N}_k|} \tag{30}$$

$$\sum_{j\notin\mathcal{S}_k} A_{n,j} \approx \frac{1}{\sqrt{d_i d_n}} \cdot \frac{d_i|\mathcal{N}_i|}{\sum_{l=1}^L d_l|\mathcal{N}_l|} d_n (1-\frac{l_k}{|\mathcal{N}_k|}) \approx \sum_{j\in\mathcal{N}_k} A_{n,j}(1-\frac{l_k}{|\mathcal{N}_k|}), \tag{31}$$

the difference between $\tilde{\boldsymbol{a}}_n^t$ and $\tilde{\boldsymbol{a}^*}_n$ can then be derived as

$$\begin{aligned}
&\|\tilde{\boldsymbol{a}}_n^t - \tilde{\boldsymbol{a}^*}_n\|_1 \\
=&\Big|\sum_{k=1}^L \sum_{j\in\mathcal{S}_k} A_{n,j}p_k^*(\frac{|\mathcal{N}_k|}{l_k}-1) + \sum_{k=1}^L \sum_{j\notin\mathcal{S}_k} A_{n,j}p_k^*\Big| \\
\lesssim&\sum_{k=1}^L (p_k^*(\frac{|\mathcal{N}_k|}{l_k}-1)\frac{l_k}{|\mathcal{N}_k|}\sum_{j\in\mathcal{N}_k} A_{n,j} + (1-\frac{l_k}{|\mathcal{N}_k|})p_k^* \sum_{j\in\mathcal{N}_k} A_{n,j}) \\
\lesssim&\text{poly}(\epsilon)\sum_{k=1}^L \frac{1}{L\Psi(L,k)}\sum_{j\in\mathcal{N}_k} A_{n,j} \\
:=&\text{poly}(\epsilon)\Gamma(\boldsymbol{A}^*)
\end{aligned} \tag{32}$$

where the first inequality is by (30, 31) and the second inequality holds as long as $l_i \geq |\mathcal{N}_i|/(1+\frac{c_1\text{poly}(\epsilon)}{Lp_i^*\Psi(L,i)})$. Combining (41), we have

$$\sum_{i=1}^L p_i^* \sum_{j\in\mathcal{N}_i} A_{n,j} \lesssim \sum_{i=1}^L \frac{1}{L\Psi(L,i)}\sum_{j\in\mathcal{N}_i} A_{n,j} = \Gamma(\boldsymbol{A}^*) \leq O(1) \tag{33}$$

Hence, (32) can be bounded by poly(ϵ).

A.1. Symmetric graph sampling method

We provide and discuss a symmetric graph sampling method in this section. The insights behind this version of sampling strategy is the same as in Section 3.2.

Similar to the asymmetric construction in Section 3.2, we consider a group-wise uniform sampling strategy, where S_l nodes are sampled uniformly from N_l nodes. For all unsampled nodes, we set the corresponding diagonal entries of a diagonal matrix $\boldsymbol{P}^s$ to be zero. If node i is sampled in this iteration and belongs to group l for any i and l, the ith diagonal entry of $\boldsymbol{P}^s$ is set as $\sqrt{p_l^* N_l/S_l}$ for some non-negative constant p_l^*. Then $\boldsymbol{A}^s = \boldsymbol{P}^s\boldsymbol{A}\boldsymbol{P}^s$.

Based on this symmetric graph sampling method, we define the effective adjacency matrix as

$$\boldsymbol{A}^* = \boldsymbol{P}^*\boldsymbol{A}\boldsymbol{P}^*, \tag{34}$$

where $\boldsymbol{P}^*$ is a diagonal matrix defined as

$$\boldsymbol{P}_{ii}^* = \sqrt{p_l^*} \quad \text{if node } i \text{ belongs to degree group } l \tag{35}$$

The scaling factor p_l^* should satisfy

$$0 \leq p_l^* \leq \frac{c_2}{L^2\psi_l^2}, \quad \forall l \tag{36}$$

for a positive constant c_2 that can be sufficiently large. ψ_l is defined in (9). The number of sampled nodes shall satisfy

$$\frac{S_l}{N_l} \geq (1+\frac{c_2\text{poly}(\epsilon)}{L\sqrt{p_l^*}\psi_l})^{-2} \qquad \forall l \in [L] \tag{37}$$

where ϵ is a small positive value.

Lemma A.3. *Given a graph $\mathcal{G}$ with $L(\geq 1)$ groups of nodes, where the group i with node degree d_i is denoted as $\mathcal{N}_i$. Suppose $\boldsymbol{A}^t$ (or any of $\boldsymbol{A}^{t(1)}$, $\boldsymbol{A}^{t(2)}$, $\boldsymbol{A}^{t(3)}$ in the general setting) is generated from the sampling strategy in Section A.1, if the number of sampled nodes satisfies $l_i \geq |\mathcal{N}_i|/(1+\frac{c_2 poly(\epsilon)}{L p_i^* \Psi_i})$, then we have*

$$\|\boldsymbol{A}^t - \boldsymbol{A}^*\|_\infty \leq poly(\epsilon) \tag{38}$$

Proof:

From Section A.1, we can rewrite that

$$\tilde{\boldsymbol{a}}_n^t = \begin{cases} \sqrt{\frac{|\mathcal{N}_k||\mathcal{N}_u|}{l_k l_u} p_k^* p_u^*} A_{n,j}, & \text{if the nodes } n,j \text{ are connected and } j \text{ is selected and } n \in \mathcal{N}_u,\ j \in \mathcal{N}_k \\ 0, & \text{else} \end{cases} \tag{39}$$

$$\tilde{\boldsymbol{a}^*}_n = \begin{cases} \sqrt{p_k^* p_u^*} A_{n,j}, & \text{if the nodes } n,j \text{ are connected and } n \in \mathcal{N}_u, j \in \mathcal{N}_k \\ 0, & \text{else} \end{cases} \tag{40}$$

Then $\boldsymbol{A}^* = (\tilde{\boldsymbol{a}^*}_1^\top, \tilde{\boldsymbol{a}^*}_2^\top, \cdots, \tilde{\boldsymbol{a}^*}_n)^\top$. Then, for $n \in \mathcal{N}_u$, as long as

$$\sum_{j \in |\mathcal{N}_i|} \sqrt{p_i^* p_u^*} A_{n,j} \approx \sqrt{p_u^* p_i^*} \frac{1}{\sqrt{d_i d_n}} \cdot \frac{d_i |\mathcal{N}_i|}{\sum_{l=1}^L d_l |\mathcal{N}_l|} d_n \lesssim \sqrt{p_u^* p_i^*} \Psi(n,i) \leq \sqrt{p_i^*} \Psi(n,i) \leq O(1/L) \tag{41}$$

i.e.,

$$\sqrt{p_i^*} \leq \frac{c_2}{L \cdot \max_{n \in [L]} \{\Psi(n,i)\}} = \frac{c_1}{L \cdot \Psi(L,i)} = \frac{c_2}{L} \sqrt{\frac{d_i}{d_L}} \frac{\sum_{l=1}^L d_l |\mathcal{N}_l|}{d_i |\mathcal{N}_i|} \tag{42}$$

for some constant $c_2 > 0$, we can obtain that $\|\boldsymbol{A}^*\|_\infty \leq O(1)$.

The difference between $\tilde{\boldsymbol{a}}_n^t$ and $\tilde{\boldsymbol{a}^*}_n$ can then be derived as

$$\begin{aligned} &\|\tilde{\boldsymbol{a}}_n^t - \tilde{\boldsymbol{a}^*}_n\|_1 \\ =&\Big| \sum_{k=1}^L \sum_{j \in \mathcal{S}_k} A_{n,j} \sqrt{p_u^* p_k^*} \Big(\sqrt{\frac{|\mathcal{N}_k||\mathcal{N}_u|}{l_k l_u}} - 1 \Big) + \sum_{k=1}^L \sum_{j \notin \mathcal{S}_k} A_{n,j} \sqrt{p_u^* p_k^*} \Big| \\ \approx&\Big| \sum_{k=1}^L \sum_{j \in \mathcal{N}_k} A_{n,j} \sqrt{p_u^* p_k^*} \Big(\sqrt{\frac{|\mathcal{N}_k||\mathcal{N}_u|}{l_k l_u}} - 1 \Big) \frac{l_k}{|\mathcal{N}_k|} + \sum_{k=1}^L \sum_{j \in \mathcal{N}_k} A_{n,j} \sqrt{p_u^* p_k^*} \Big(1 - \frac{l_k}{|\mathcal{N}_k|} \Big) \Big| \\ \lesssim&\text{poly}(\epsilon) \end{aligned} \tag{43}$$

as long as $l_i \geq |\mathcal{N}_i|/(1+\frac{c_2 \text{poly}(\epsilon)}{L\sqrt{p_i^*}\Psi(L,i)})^2$.

B. Node classification for three layers

In the whole proof, we consider a more general target function compared to (12). We write $F^* : \mathbb{R}^N \times \mathbb{R}^{N \times d} \to \mathbb{R}^K$:

$$\begin{aligned} &F_{\boldsymbol{A}^*}^* = (f_1^*, f_2^*, \cdots, f_K^*), \\ &f_r^*(\boldsymbol{e}_g, \boldsymbol{X}) = \boldsymbol{e}_g^\top \sum_{k \in [p_1]} c_{k,r}^* \Phi\Big(\boldsymbol{A}^* \sum_{j \in [p_2]} v_{1,k,j}^* \phi_{1,j}(\boldsymbol{A}^* \boldsymbol{X} \boldsymbol{w}_{1,j}^*) \Big) \odot \Big(\boldsymbol{A}^* \sum_{l \in [p_2]} v_{2,k,l}^* \phi_{2,l}(\boldsymbol{A}^* \boldsymbol{X} \boldsymbol{w}_{2,l}^*) \Big),\ \forall r \in [K], \end{aligned} \tag{44}$$

where each $\phi_{1,j}$, $\phi_{2,j}$, Φ_i: $\mathbb{R} \to \mathbb{R}$ is infinite-order smooth.
Table 2 shows some important notations used in our theorem and algorithm. Table 3 gives the full parameter choices for the three-layer GCN. ploy$(\log(m_1 m_2))$ in the following analysis.

Table 2. Summary of notations

$\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$	$\mathcal{G}$ is an un-directed graph consisting of a set of nodes $\mathcal{V}$ and a set of edges $\mathcal{E}$.
N	The total number of nodes in a graph.
$\boldsymbol{A} = \boldsymbol{D}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\boldsymbol{D}^{-\frac{1}{2}}$	$\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is the normalized adjacency matrix computed by the degree matrix $\boldsymbol{D}$ and the initial adjacency matrix $\tilde{\boldsymbol{A}}$.
$\boldsymbol{A}^*$	The effective adjacency matrix.
$\boldsymbol{A}^t$	The sampled adjacency matrix using our sampling strategy in Section 3.2 at the t-th iteration.
$\boldsymbol{e}_g$, $\boldsymbol{X}$ y_n	$\boldsymbol{e}_g$ belongs to $\{\boldsymbol{e}_i\}_{i=1}^N$ and selects the index of the node label. $\boldsymbol{X} \in \mathbb{R}^{N\times d}$ is the feature matrix. y_n is the label of the n-th node.
m_1, m_2	m_1, m_2 are the number of neurons in the first and second hidden layer, respectively.
$\boldsymbol{W}$, $\boldsymbol{V}$, $\boldsymbol{B}_1$, $\boldsymbol{B}_2$	$\boldsymbol{X} \in \mathbb{R}^{N\times d}$ is the data matrix. $\boldsymbol{W}$, $\boldsymbol{V}$ are the weight matrices of the first and second hidden layer, respectively. $\boldsymbol{B}_1$, $\boldsymbol{B}_2$ are the corresponding bias matrices.
$\boldsymbol{W}^{(0)}$, $\boldsymbol{V}^{(0)}$	$\boldsymbol{W}^{(0)}$ and $\boldsymbol{V}^{(0)}$ are random initializations of $\boldsymbol{W}$ and $\boldsymbol{V}$, respectively.
$\boldsymbol{W}^{\rho}$, $\boldsymbol{V}^{\rho}$	$\boldsymbol{W}^{\rho}$ and $\boldsymbol{V}^{\rho}$ are two random matrices used for Gaussian smoothing.
$\boldsymbol{\Sigma}$	The Dropout technique.
Ω, Ω^t	Ω is the set of labeled nodes and Ω_t is the batch of labeled nodes at the t-th iteration.
T, T_w, η, λ_t	In Algorithm 1, T is the number of outer iterations for the weight decay step, while T_w is the number of inner iterations for the SGD steps. η is the step size and λ_t is the weight decay coefficient at the t-th iteration.
L, d_l, S_l, N_l	L is the number of node groups in a graph. d_l is the order-wise degree in the l-th group. N_l is the number of nodes in group l.
S_l	The number of nodes we sample in group l.

Table 3. Full parameter choices for three-layer GCN

τ_v'	$m_1^{1/2-0.005}/(\sqrt{\epsilon_0}m_2^{1/2})$
τ_w'	$C_0/(\epsilon_0^{1/4}m_1^{3/4-0.005})$
τ_v	$m_1^{1/2-0.001}/m_2^{1/2} \gg \tau_v'$
τ_w	$1/m_1^{3/4-0.01} \gg \tau_w'$
λ_v	$2/(\tau_v')^2$
λ_w	$2/(\tau_w')^4$
σ_v	$1/m_2^{1/2+0.01}$
σ_w	$\sigma_w = 1/m_1^{1-0.01}$
C	$\mathcal{C}_\epsilon(\phi, \|\boldsymbol{A}\|_\infty)\sqrt{\|\boldsymbol{A}\|_\infty^2 + 1}$
C'	$10C\sqrt{p_2}$
C''	$\mathcal{C}_\epsilon(\Phi, C')\sqrt{\|\boldsymbol{A}\|_\infty^2 + 1}$
C_0	$\tilde{O}(p_1^2 p_2 K^2 C C'')$
ϵ_c	1

B.1. Lemmas

B.1.1. Function approximation

To show that the target function can be learnt by the learner network with the Relu function, a good approach is to firstly find a function $h(\cdot)$ such that the ϕ functions in the target function can be approximated by $h(\cdot)$ with an indicator function. In this section, Lemma B.1 provides the existence of such $h(\cdot)$ function. Lemma B.2 and B.3 are two supporting lemmas to prove Lemma B.1.

Lemma B.1. *For every smooth function ϕ, every $\epsilon \in (0, \frac{1}{\mathcal{C}(\phi,a)\sqrt{a^2+1}})$, there exists a function $h : \mathbb{R}^2 \to [-\mathcal{C}_\epsilon(\phi, a)\sqrt{a^2+1}, \mathcal{C}_\epsilon(\phi, a)\sqrt{a^2+1}]$ that is also $\mathcal{C}_\epsilon(\phi, a)\sqrt{a^2+1}$-Lipschitz continuous on its first coordinate with the following two (equivalent) properties:*

(a) For every $x_1 \in [-a, a]$ where $a > 0$:

$$\left|\mathbb{E}\left[\mathbb{1}_{\alpha_1 x_1 + \beta_1\sqrt{a^2 - x_1^2} + b_0 \geq 0} h(\alpha_1, b_0)\right] - \phi(x_1)\right| \leq \epsilon$$

where $\alpha_1, \beta_1, b_0 \sim \mathcal{N}(0,1)$ are independent random variables.
(b) For every $\boldsymbol{w}^*, \boldsymbol{x} \in \mathbb{R}^d$ with $\|\boldsymbol{w}^*\|_2 = 1$ and $\|\boldsymbol{x}\| \leq a$:

$$\left|\mathbb{E}\left[\mathbb{1}_{\boldsymbol{w}\boldsymbol{X} + b_0 \geq 0} h(\boldsymbol{w}^\top \boldsymbol{w}^*, b_0)\right] - \phi({\boldsymbol{w}^*}^\top \boldsymbol{x})\right| \leq \epsilon$$

where $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I})$ is an d-dimensional Gaussian, $b_0 \sim \mathcal{N}(0,1)$.
Furthermore, we have $\mathbb{E}_{\alpha_1, b_0 \sim \mathcal{N}(0,1)}[h(\alpha_1, b_0)^2] \leq (\mathcal{C}_s(\phi, a))^2(a^2 + 1)$.
(c) For every $\boldsymbol{w}^*, \boldsymbol{x} \in \mathbb{R}^d$ with $\|\boldsymbol{w}^*\|_2 = 1$, let $\tilde{\boldsymbol{w}} = (\boldsymbol{w}, b_0) \in \mathbb{R}^{d+1}$, $\tilde{\boldsymbol{x}} = (\boldsymbol{x}, 1) \in \mathbb{R}^{d+1}$ with $\|\tilde{\boldsymbol{x}}\| \leq \sqrt{a^2 + 1}$, then we have

$$\left|\mathbb{E}\left[\mathbb{1}_{\tilde{\boldsymbol{w}}^\top \tilde{\boldsymbol{x}} \geq 0} h(\tilde{\boldsymbol{w}}[1:d]^\top \boldsymbol{w}^*, \tilde{\boldsymbol{w}}[d+1])\right] - \phi({\boldsymbol{w}^*}^\top \tilde{\boldsymbol{x}}[1:d])\right| \leq \epsilon$$

where $\tilde{\boldsymbol{w}} \sim \mathcal{N}(0, \boldsymbol{I}_{d+1})$ is an d-dimensional Gaussian.
We also have $\mathbb{E}_{\tilde{\boldsymbol{w}} \in \mathcal{N}(0, \boldsymbol{I}_{d+1})}[h(\tilde{\boldsymbol{w}}[1:d]^\top \boldsymbol{w}^*, \tilde{\boldsymbol{w}}[d+1])^2] \leq (\mathcal{C}_s(\phi, a))^2(a^2 + 1)$.

Proof:
Firstly, since we can assume $\boldsymbol{w}^* = (1, 0, \cdots, 0)$ without loss of generality by rotating $\boldsymbol{x}$ and $\boldsymbol{w}$, it can be derived that $\boldsymbol{x}$, $\boldsymbol{w}$, $\boldsymbol{w}^*$ are equivalent to that they are two-dimensional. Therefore, proving Lemma B.1b suffices in showing Lemma B.1a.
Let $\boldsymbol{w}_0 = (\alpha, \beta)$, $\boldsymbol{x} = (x_1, \sqrt{t^2 - x_1^2})$ where α and β are independent. Following the idea of Lemma 6.3 in (Allen-Zhu et al., 2019), we use another randomness as an alternative, i.e., we write $\boldsymbol{x}^\perp = (\sqrt{t^2 - x_1^2}, -x_1)$, $\boldsymbol{w}_0 = \alpha\frac{\boldsymbol{x}}{t} + \beta\frac{\boldsymbol{x}^\perp}{t} \sim \mathcal{N}(0, \boldsymbol{I})$. Then $\boldsymbol{w}_0\boldsymbol{X} = t\alpha$. Let $\alpha_1 = w_{01} = \alpha\frac{x_1}{t} + \beta\sqrt{1 - \frac{x_1^2}{t^2}}$, where $\alpha, \beta \sim \mathcal{N}(0,1)$. Hence, $\alpha_1 \sim \mathcal{N}(0,1)$.
We first use Lemma B.2 to fit $\phi(x_1)$. By Taylor expansion, we have

$$\begin{aligned}\phi(x_1) &= c_0 + \sum_{i=1,\text{ odd } i}^{\infty} c_i x_1^i + \sum_{i=2,\text{ even } i}^{\infty} c_i x_1^i \\ &= c_0 + \sum_{i=1}^{\infty} c_i' \mathbb{E}_{\alpha,\beta\sim\mathcal{N}(0,1)}[h_i(\alpha_1)\mathbb{1}[q_i(b_0)]\mathbb{1}[\boldsymbol{w}_0\boldsymbol{X} + b_0 \geq 0]]\end{aligned} \tag{45}$$

where $h_i(\cdot)$ is the Hermite polynomial defined in Definition A.5 in (Allen-Zhu et al., 2019), and

$$c_i' = \frac{c_i}{p_i'},\ |c_i'| \leq \frac{200i^2|c_i|}{(i-1)!!}\frac{\sqrt{t^2+1}}{t^{1-i}} \text{ and } q_i(b_0) = \begin{cases} |b_0| \leq t/(2i), & i \text{ is odd} \\ 0 < -b_0 \leq t/(2i), & i \text{ is even} \end{cases} \tag{46}$$

Let $B_i = 100i^{\frac{1}{2}} + 10\sqrt{\log(\frac{1}{\epsilon}\frac{\sqrt{t^2+1}}{t^{1-i}})}$. Define $\hat{h}_i(\alpha_1) = h_i(\alpha_1) \cdot \mathbb{1}[|\alpha_1| \leq B_i] + h_i(\text{sign}(\alpha_1)B_i) \cdot \mathbb{1}[|\alpha_1| > B_i]$ as the truncated version of the Hermite polynomial. Then we have

$$\phi(x_1) = c_0 + R(x_1) + \sum_{i=1}^{\infty} c_i' \mathbb{E}_{\alpha,\beta\sim\mathcal{N}(0,1)}[\hat{h}_i(\alpha_1)\mathbb{1}[q_i(b_0)]\mathbb{1}[\boldsymbol{w}_0\boldsymbol{X} + b_0 \geq 0]],$$

where

$$R(x_1) = \sum_{i=1}^{\infty} c_i' \mathbb{E}_{\alpha,\beta\sim\mathcal{N}(0,1)}\Big[\big(h_i(\alpha_1) \cdot \mathbb{1}[|\alpha_1| > B_i] - h_i(\text{sign}(\alpha_1)B_i \cdot \mathbb{1}[|\alpha| > B_i])\big)\mathbb{1}[q_i(b_0)]\mathbb{1}[\boldsymbol{w}_0\boldsymbol{X} + b_0 \geq 0]\Big]$$

Define

$$h(\alpha_1, b_0) = c_0 + \sum_{i=1}^{\infty} c_i' \cdot \hat{h}_i(\alpha_1) \cdot \mathbb{1}[q_i(b_0)]$$

Then by Lemma B.3, we have

$$|\mathbb{E}_{\alpha,\beta,b_0\sim\mathcal{N}(0,1)}[\mathbb{1}[\boldsymbol{w}_0\boldsymbol{X}+b_0\geq 0]\cdot h(\alpha_1,b_0)-\phi(x_1)|\leq|R(x_1)|\leq\frac{\epsilon}{4}$$

We also have

$$\begin{aligned}\mathbb{E}_{\alpha_1,b_0\sim\mathcal{N}(0,1)}[h(\alpha_1,b_0)^2]&\leq(\epsilon^2+c_0^2)+O(1)\cdot\sum_{i=1}^{\infty}\frac{i!\cdot|c_i|^2i^3}{((i-1)!!)^2}\cdot(\frac{\sqrt{t^2+1}}{t^{1-i}})^2\\&\leq(\epsilon^2+c_0^2)+\sum_{i=1}^{\infty}i^{3.5}\cdot|c_i|^2\cdot(\frac{\sqrt{t^2+1}}{t^{1-i}})^2\\&\leq(\epsilon^2+c_0^2)+\Big(\sum_{i=0}^{\infty}(i+1)^{1.75}\cdot|c_i|\cdot t^i\sqrt{t^2+1}\Big)^2\\&\leq\mathcal{C}_s(\phi,t)^2(t^2+1)\end{aligned}\tag{47}$$

Lemma B.2. *Denote $h_i(x)$ as the degree-i Hermite polynomial as in Definition A.5 in (Allen-Zhu et al., 2019). For every integer $i\geq 1$, there exists constant p_i' with $|p_i'|\geq\frac{t^{1-i}}{\sqrt{t^2+1}}\frac{(i-1)!!}{100i^2}$ such that*

$$\textit{for even } i:\quad x_1^i=\frac{1}{p_i'}\mathbb{E}_{\boldsymbol{w}_0\sim\mathcal{N}(0,\boldsymbol{I}),b_0\sim\mathcal{N}(0,1)}[h_i(\alpha_1)\mathbb{1}[\alpha\geq-\frac{b_0}{t}]\mathbb{1}[0<-b_0\leq\frac{t}{2i}]]\tag{48}$$

$$\textit{for odd } i:\quad x_1^i=\frac{1}{p_i'}\mathbb{E}_{\boldsymbol{w}_0\sim\mathcal{N}(0,\boldsymbol{I}),b_0\sim\mathcal{N}(0,1)}[h(\alpha_1)\mathbb{1}[\alpha\geq-\frac{b_0}{t}]\mathbb{1}[|b_0|\leq\frac{t}{2i}]]\tag{49}$$

for $\|\boldsymbol{x}\|\leq t$.

Proof:
For even i, by Lemma A.6 in (Allen-Zhu et al., 2019), we have

$$\mathbb{E}_{\boldsymbol{w}_0\sim\mathcal{N}(0,\boldsymbol{I}),b_0\sim\mathcal{N}(0,1)}[h_i(\alpha_1)\mathbb{1}[\alpha\geq-\frac{b_0}{t}]\mathbb{1}[0<-b_0\leq\frac{t}{2i}]]=\mathbb{E}_{b_0\sim\mathcal{N}(0,1)}[p_i\cdot\mathbb{1}[0<-b_0\leq\frac{t}{2i}]]\cdot\frac{x_1^i}{t^i}$$

, where

$$p_i=(i-1)!!\frac{\exp(-b_0^2/(2t^2))}{\sqrt{2\pi}}\sum_{r=1,r\text{ odd}}^{i-1}\frac{(-1)^{\frac{i-1-r}{2}}}{r!!}\binom{i/2-1}{(r-1)/2}(-b_0/t)^r$$

Define $c_r=\frac{(-1)^{\frac{i-1-r}{2}}}{r!!}\binom{i/2-1}{(r-1)/2}$. Then $\text{sign}(c_r)=-\text{sign}(c_{r+2})$. We can derive

$$\Big|\frac{c_r(-b_0/t)^r}{c_{r-2}(-b_0/t)^{r-2}}\Big|=\Big|(\frac{b_0}{t})^2\frac{i+1-r}{r(r-1)}\Big|\leq\frac{1}{4i}\leq\frac{1}{4}$$

Therefore,

$$\Big|\sum_{r=1,r\text{ odd}}^{i-1}c_r(-b_0/t)^r\Big|\geq\frac{3}{4}|b_0/t|$$

$$\begin{aligned}&|\mathbb{E}_{b_0\sim\mathcal{N}(0,1)}[p_i\cdot\mathbb{1}[0\leq-b_0/t\leq1/(2i)]]|\cdot t^{-i}\\\geq&|\mathbb{E}_{b_0\sim\mathcal{N}(0,1)}[(i-1)!!\frac{\exp(-b_0^2/2t^2)}{\sqrt{2\pi}}\cdot\frac{3}{4}|b_0/t|\cdot\mathbb{1}[0\leq-b_0/t\leq1/(2i)]]|\cdot t^{-i}\\=&t^{-i}\cdot\int_{-\frac{t}{2i}}^{0}(i-1)!!\frac{\exp(-\frac{b_0^2}{2}(1+\frac{1}{t^2}))}{2\pi}\cdot\frac{3}{4}(-\frac{b_0}{t})db_0\\=&t^{-i}\cdot\frac{t}{t^2+1}\exp(-\frac{b_0^2}{2}(1+\frac{1}{t^2}))(i-1)!!\frac{3}{8\pi}\Big|_{-\frac{t}{2i}}^{0}\\=&t^{-i}\frac{t}{t^2+1}(i-1)!!\frac{3}{8\pi}\Big(1-\exp(-\frac{t^2+1}{8i^2})\Big)\\\geq&t^{1-i}\frac{(i-1)!!}{100i^2}\end{aligned}\tag{50}$$

For odd i, similarly by Lemma A.6 in (Allen-Zhu et al., 2019), we can obtain

$$\mathbb{E}_{\boldsymbol{w}_0\sim\mathcal{N}(0,\boldsymbol{I}),b_0\sim\mathcal{N}(0,1)}[h(\alpha_1)\mathbb{1}[\alpha\geq-\frac{b_0}{t}]\mathbb{1}[|b_0|\leq\frac{t}{2i}]]=\mathbb{E}_{b_0\sim\mathcal{N}(0,1)}[p_i\cdot\mathbb{1}[|b_0|\leq\frac{t}{2i}]]\cdot\frac{x_1^i}{t^i}$$

, where

$$p_i=(i-1)!!\frac{\exp(-b_0^2/(2t^2))}{\sqrt{2\pi}}\sum_{r=1,r\text{ even}}^{i-1}\frac{(-1)^{\frac{i-1-r}{2}}}{r!!}\binom{i/2-1}{(r-1)/2}(-b_0/t)^r$$

Then we also have

$$\Big|\frac{c_r(-b_0/t)^r}{c_{r-2}(-b_0/t)^{r-2}}\Big|=\Big|(\frac{b_0}{t})^2\frac{i+1-r}{r(r-1)}\Big|\leq\frac{1}{4i}\leq\frac{1}{4}$$

Therefore,

$$\Big|\sum_{r=1,r\text{ odd}}^{i-1}c_r(-b_0/t)^r\Big|\geq\frac{3}{4}|c_0|=\frac{3}{4}\frac{(\frac{i}{2}-1)!}{\pi(\frac{1}{2}\cdot\frac{3}{2}\cdots\frac{i-1}{2})}\geq\frac{3}{4}\frac{1}{\pi\frac{i-1}{2}}\geq\frac{3}{2\pi i}$$

$$\begin{aligned}
&|\mathbb{E}_{b_0\sim\mathcal{N}(0,1)}[p_i\cdot\mathbb{1}[|b_0|/t\leq1/(2i)]]|\cdot t^{-i}\\
\geq&t^{-i}\cdot|\mathbb{E}_{b_0\sim\mathcal{N}(0,1)}[(i-1)!!\frac{\exp(-b_0^2/2t^2)}{\sqrt{2\pi}}\cdot\frac{3}{2\pi i}\cdot\mathbb{1}[|b_0|/t\leq1/(2i)]]|\\
=&t^{-i}\cdot\int_{-\frac{t}{2i}}^{\frac{t}{2i}}(i-1)!!\frac{\exp(-\frac{b_0^2}{2}(1+\frac{1}{t^2}))}{2\pi}\cdot\frac{3}{2\pi i}db_0\\
=&t^{-i}\cdot(i-1)!!\frac{3}{4\pi^2i}\cdot\frac{t}{\sqrt{t^2+1}}\cdot\sqrt{2\pi}\cdot\Big(2\Phi(\frac{\sqrt{t^2+1}}{2i})-1\Big)\\
=&t^{-i}\cdot(i-1)!!\frac{3}{4\pi^2i}\cdot\frac{t}{\sqrt{t^2+1}}\cdot\sqrt{2\pi}\cdot\frac{2\Phi(\frac{\sqrt{t^2+1}}{2})-1}{i}\\
\geq&\frac{t^{1-i}}{\sqrt{t^2+1}}\frac{(i-1)!!}{100i^2}
\end{aligned}\tag{51}$$

Lemma B.3. *For $B_i=100i^{1/2}+10\sqrt{\log(t^{i-1}\sqrt{t^2+1}/\epsilon_i^2)}$ where $\epsilon_i^2=t^{i-1}\sqrt{t^2+1}\epsilon^2$, we have*

1. $\sum_{i=1}^{\infty}|c_i'|\cdot|\mathbb{E}_{x\sim\mathcal{N}(0,1)}[|h_i(x)|\cdot\mathbb{1}[|x|\geq b]]|\leq\frac{\epsilon}{8}\sqrt{t^2+1}$
2. $\sum_{i=1}^{\infty}|c_i'|\cdot|\mathbb{E}_{x\sim\mathcal{N}(0,1)}[|h_i(b)|\cdot\mathbb{1}[|x|\geq b]]|\leq\frac{\epsilon}{8}\sqrt{t^2+1}$
3. $\sum_{i=1}^{\infty}|c_i'|\mathbb{E}_{z\in\mathcal{N}(0,1)}[|h_i(z)|\mathbb{1}[|z|\leq B_i]]\leq\mathcal{C}_\epsilon(\phi,t)\sqrt{t^2+1}$
4. $\sum_{i=1}^{\infty}|c_i'|\mathbb{E}_{z\in\mathcal{N}(0,1)}[|\frac{d}{dz}h_i(z)|\mathbb{1}[|z|\leq B_i]]\leq\mathcal{C}_\epsilon(\phi,t)\sqrt{t^2+1}$

Proof:
By the definition of Hermite polynomial in Definition A.5 in (Allen-Zhu et al., 2019), we have

$$h_i(x)\leq\sum_{j=1}^{\lfloor i/2\rfloor}\frac{|x|^{i-2j}i^{2j}}{j!}$$

Combining (46), we can obtain

$$|c_i'h_i(x)|\leq O(1)|c_i|\frac{\sqrt{t^2+1}}{t^{1-i}}\frac{i^4}{i!!}\sum_{j=1}^{\lfloor i/2\rfloor}\frac{|x|^{i-2j}i^{2j}}{j!}\tag{52}$$

(1) Let $b = 100i^{\frac{1}{2}}\theta_i$ and $\theta_i = 1 + \frac{\sqrt{\log(\frac{1}{\epsilon_i^2}\frac{\sqrt{t^2+1}}{t^{1-i}})}}{10\sqrt{i}}$ for $\epsilon_i^2 = \frac{\sqrt{t^2+1}}{t^{1-i}}\epsilon^2$ where $i \geq 1$, then we have

$$\begin{aligned}(\theta_i \cdot e^{-10^2\theta_i^2})^i =& \Big((1 + \frac{\sqrt{\log(\frac{1}{\epsilon_i^2}\frac{\sqrt{t^2+1}}{t^{1-i}})}}{10\sqrt{i}}) \cdot e^{-10^2} e^{-2\cdot 10^2 \frac{\sqrt{\log(\frac{1}{\epsilon_i^2}\frac{\sqrt{t^2+1}}{t^{2-i}})}}{10\sqrt{i}}} \cdot e^{-10^2 \frac{\log(\frac{1}{\epsilon_i^2}\frac{\sqrt{t^2+1}}{t^{2-i}})}{100i}}\Big)^i \\ =& \epsilon_i^2 \frac{t^{1-i}}{\sqrt{t^2+1}} \cdot e^{-10^2 i} \cdot (1 + \frac{\sqrt{\log(\frac{1}{\epsilon_i^2}\frac{\sqrt{t^2+1}}{t^{1-i}})}}{10\sqrt{i}}) e^{-2\cdot 10^2 \frac{\sqrt{\log(\frac{1}{\epsilon_i^2}\frac{\sqrt{t^2+1}}{t^{1-i}})}}{10\sqrt{i}}} \\ \leq& \frac{\epsilon_i^2}{100000^i} \frac{t^{1-i}}{\sqrt{t^2+1}}\end{aligned} \tag{53}$$

where the second step comes from that $(1+s)\cdot e^{-2\cdot 10^4 \cdot s} \leq 1$ for any $s > 0$. Combining the equation C.6, C.7 in (Allen-Zhu et al., 2019) and (53), we can derive

$$\begin{aligned}&\sum_{i=1}^{\infty} |c_i'| \cdot \mathbb{E}_{x\sim\mathcal{N}(0,1)}[|h_i(z)| \cdot \mathbb{1}[|x| \geq b]] \\ \leq& \sum_{i=1}^{\infty} O(1)|c_i| \frac{\sqrt{t^2+1}}{t^{1-i}} \frac{i^4}{i!!} \cdot i^{\frac{i}{2}} \cdot 1200^i \cdot (\theta_i \cdot e^{-10^2\theta_i^2})^i \\ \leq& \frac{\epsilon}{8}\sqrt{t^2+1}\end{aligned} \tag{54}$$

for any $\epsilon > 0$ and $t \leq O(1)$.
(b) Similarly, following (53) and (54), we have

$$\sum_{i=1}^{\infty} |c_i'| \cdot |\mathbb{E}_{x\sim\mathcal{N}(0,1)}[|h_i(b)| \cdot \mathbb{1}[|x| \geq b]]| \leq \sum_{i=1}^{\infty} O(1) \frac{\sqrt{t^2+1}}{t^{1-i}} |c_i| \frac{i^4}{i!!} \cdot e^{-\frac{b^2}{2}} (3b)^i \leq \frac{\epsilon}{8}\sqrt{t^2+1}$$

(c) Similar to (52),

$$\begin{aligned}\sum_{i=1}^{\infty} |c_i'| \mathbb{E}_{z\in\mathcal{N}(0,1)}[|h_i(z)| \mathbb{1}[|z| \leq B_i]] \leq& O(1) \sum_{i=1}^{\infty} |c_i| \frac{i^4}{i!!} \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{B_i^{i-2j} i^{2j}}{j!} t^{i-1}\sqrt{t^2+1} \\ \leq& \sum_{i=1}^{\infty} |c_i| (O(1)\theta_i)^i t^{i-1}\sqrt{t^2+1} \\ \leq& \mathcal{C}_\epsilon(\phi, t)\sqrt{t^2+1},\end{aligned} \tag{55}$$

where the step follows from Claim C.2 (c) in (Allen-Zhu et al., 2019).
(d) Since we have

$$|\frac{d}{dx} h_i(x)| \leq \sum_{j=0}^{\lfloor i/2 \rfloor} |x|^{i-2j} i^{2j} \tag{56}$$

by Definition A.5 in (Allen-Zhu et al., 2019), we can derive

$$\sum_{i=1}^{\infty} |c_i'| \mathbb{E}_{z\in\mathcal{N}(0,1)}[|\frac{d}{dz} h_i(z)| \mathbb{1}[|z| \leq B_i]] \leq \mathcal{C}_\epsilon(\phi, t)\sqrt{t^2+1} \tag{57}$$

B.1.2. Existence of a good pseudo network

We hope to find some good pseudo network that can approximate the target network. In such a pseudo network, the activation $\mathbb{1}_{\boldsymbol{x}\geq 0}$ is replaced by $\mathbb{1}_{\boldsymbol{x}^{(0)}\geq 0}$ where $\boldsymbol{x}^{(0)}$ is the value at the random initialization. We can define a pseudo network without bias as

$$g_r^{(0)}(\boldsymbol{q}, \boldsymbol{A}, \boldsymbol{W}, \boldsymbol{V}, \boldsymbol{B}) = \sum_{n=1}^{N} \boldsymbol{q}^\top \boldsymbol{a}_n \sum_{i\in[m_2]} c_{i,r} \mathbb{1}_{\boldsymbol{r}_{n,i}+B_{2(n,i)}\geq 0} \sum_{j=1}^{N} a_{n,j} \sum_{l\in[m_1]} v_{i,l} \mathbb{1}_{\boldsymbol{a}_j \boldsymbol{X}\boldsymbol{w}_l + B_{1(j,l)}} \boldsymbol{a}_j \boldsymbol{X}\boldsymbol{w}_l \tag{58}$$

Lemma B.4 shows the target function can be approximated by the pseudo network with some parameters. Lemma B.5 to B.8 provides how the existence of such a pseudo network is developed step by step.

Lemma B.4. *For every* $\epsilon \in (0, \frac{1}{K\|\boldsymbol{q}\|_1 p_1 p_2^2 \mathcal{C}_s(\Phi, p_2\mathcal{C}_s(\phi,\|\boldsymbol{A}\|_\infty))\mathcal{C}_s(\phi,\|\boldsymbol{A}\|_\infty)\sqrt{\|\boldsymbol{A}\|_\infty^2+1}})$*, there exists*

$$M = poly(\mathcal{C}_\epsilon(\Phi, \sqrt{p_2}\mathcal{C}_\epsilon(\phi, \|\boldsymbol{A}\|_\infty)\sqrt{\|\boldsymbol{A}\|_\infty^2+1}), 1/\epsilon)$$

$$C = \mathcal{C}_\epsilon(\phi, \|\boldsymbol{A}\|_\infty)\sqrt{\|\boldsymbol{A}\|_\infty^2+1} \tag{59}$$

$$C' = 10C\sqrt{p_2} \tag{60}$$

$$C'' = \mathcal{C}_\epsilon(\Phi, C')\sqrt{\|\boldsymbol{A}\|_\infty^2+1} \tag{61}$$

$$C_0 = \tilde{O}(p_1^2 p_2 K^2 C C'') \tag{62}$$

such that with high probability, there exists $\widehat{\boldsymbol{W}}$, $\widehat{\boldsymbol{V}}$ *with* $m_1, m_2 \geq M$,

$$\|\widehat{\boldsymbol{W}}\|_{2,\infty} \leq \frac{C_0}{m_1}, \quad \|\widehat{\boldsymbol{V}}\|_{2,\infty} \leq \frac{\sqrt{m_1}}{m_2}$$

such that

$$\mathbb{E}_{(\boldsymbol{X},y)\in\mathcal{D}}\Big[\sum_{r=1}^K |f_r^*(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X}) - g_r^{(0)}(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X},\widehat{\boldsymbol{W}}.\widehat{\boldsymbol{V}})|\Big] \leq \epsilon$$

$$\mathbb{E}_{(\boldsymbol{X},y)\in\mathcal{D}}[|L(G^{(0)}(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X},\widehat{\boldsymbol{W}},\widehat{\boldsymbol{V}}))|] \leq OPT + \epsilon$$

Proof:

For each $\phi_{2,j}$, we can construct $h_{\phi,j} : \mathbb{R}^2 \to [-C, C]$ where $C = \mathcal{C}_\epsilon(\phi, \|\boldsymbol{A}\|_\infty)\sqrt{\|\boldsymbol{A}\|_\infty^2+1}$ using Lemma B.1 satisfying

$$\mathbb{E}[h_{\phi,j}({\boldsymbol{w}_{2,j}^*}^\top \boldsymbol{w}_i^{(0)}, B_{1(n,i)}^{(0)})\mathbb{1}_{\tilde{\boldsymbol{a}}_n \boldsymbol{X}\boldsymbol{w}_i^{(0)}+B_{1(n,i)}\geq 0}] = \phi_{2,j}(\tilde{\boldsymbol{a}}_n \boldsymbol{X}\boldsymbol{w}_{2,j}) \pm \epsilon \tag{63}$$

for $i \in [m_1]$. Consider any arbitrary $\boldsymbol{b} \in \mathbb{R}^{m_1}$ with $v_i \in \{-1, 1\}$. Define

$$\widehat{\boldsymbol{W}} = \frac{(C_0 C''/C)^{\frac{1}{2}}}{\epsilon_c^2 m_1}(v_i \sum_{j\in[p_2]} v_{2,j}^* h_{\phi,j}({\boldsymbol{w}_{2,j}^*}^\top \boldsymbol{w}_i^{(0)}, B_{1(i)}^{(0)})\boldsymbol{e}_d)_{i\in[m_1]} \tag{64}$$

$$\widehat{\boldsymbol{V}} = (C_0 C''/C)^{-\frac{1}{2}} \sum_{k\in[p_1]} \frac{c_k^*}{m_2}(\boldsymbol{v} h(\sqrt{m_2} \sum_{j\in[p_2]} v_{1,j}^* \alpha_{i,j}, B_{2(i)}^{(0)}) \sum_{r=1}^K c_{i,r})_{i\in[m_2]} \tag{65}$$

Then,

$$\begin{aligned}
&g_r^{(0)}(\boldsymbol{q},\boldsymbol{A},\widehat{\boldsymbol{W}},\widehat{\boldsymbol{V}},\boldsymbol{B}) \\
&= \sum_{n=1}^N \boldsymbol{q}^\top \boldsymbol{a}_n \sum_{i\in[m_1]} c_{i,r}\mathbb{1}_{\boldsymbol{r}_{n,i}+B_{2(n,i)}\geq 0} \sum_{i'\in[m_2]} \sum_{j=1}^N a_{n,j}\mathbb{1}_{\boldsymbol{a}_j \boldsymbol{X}\boldsymbol{w}_i^{(0)}+B_{1(j,i)}\geq 0} \boldsymbol{a}_j \boldsymbol{X}\widehat{\boldsymbol{W}}_i \widehat{V}_{i,i'} \\
&= \sum_{k\in[p_1]} \frac{c_k^*}{m_2\epsilon_c^2} \sum_{n=1}^N \boldsymbol{q}^\top \boldsymbol{a}_n \sum_{i\in[m_1]} c_{i,r}^2 \mathbb{1}_{\boldsymbol{r}_{n,i}+B_{2(n,i)}\geq 0} h(\sqrt{m_2} \sum_{j\in[p_2]} v_{1,j}^* \alpha_{i,j}, B_{2(i)}^{(0)}) \sum_{j=1}^N a_{n,j} \sum_{l\in[p_2]} v_{2,l}^* \phi_{2,l}(\boldsymbol{a}_j \boldsymbol{X}\boldsymbol{w}_{2,l}^*) \\
&= \sum_{k\in[p_1]} \sum_{n=1}^N \boldsymbol{q}^\top \boldsymbol{a}_n c_k^* \Phi(\sum_{j\in[p_2]} v_{1,j}^* \sum_{m=1}^N a_{m,n}\phi_{1,j}(\boldsymbol{a}_m \boldsymbol{X}\boldsymbol{w}_{1,j}^*)) \sum_{j=1}^N a_{n,j} \sum_{l\in[p_2]} v_{2,l}^* \phi_{2,l}(\boldsymbol{a}_j \boldsymbol{X}\boldsymbol{w}_{2,l}^*) \\
&\quad \pm O(p_1 p_2^2 \mathcal{C}_s(\Phi, p_2\mathcal{C}_s(\phi,\|\boldsymbol{A}\|_\infty))\mathcal{C}_s(\phi,\|\boldsymbol{A}\|_\infty)\sqrt{\|\boldsymbol{A}\|_\infty^2+1}\epsilon) \\
&= \sum_{n=1}^N \boldsymbol{q}^\top \boldsymbol{a}_n \sum_{k\in[p_1]} c_k^* \Phi(\tilde{\boldsymbol{a}}_n \sum_{j\in[p_2]} v_{1,j}^* \phi_{1,j}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{w}_{1,j}^*)) \tilde{\boldsymbol{a}}_n \sum_{l\in[p_2]} v_{2,l}^* \phi_{2,l}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{w}_{2,l}^*) \\
&\quad \pm O(\|\boldsymbol{q}\|_1 p_1 p_2^2 \mathcal{C}_s(\Phi, p_2\mathcal{C}_s(\phi,\|\boldsymbol{A}\|_\infty))\mathcal{C}_s(\phi,\|\boldsymbol{A}\|_\infty)\sqrt{\|\boldsymbol{A}\|_\infty^2+1}\epsilon)
\end{aligned} \tag{66}$$

where the first step comes from definition of $g^{(0)}$, the second step is derived from (64) and (65) and the second to last step is by Lemma B.8.

Lemma B.5. *For every smooth function ϕ, every $\boldsymbol{w}^* \in \mathbb{R}^d$ with $\|\boldsymbol{w}^*\| = 1$, for every $\epsilon \in (0, \frac{1}{\mathcal{C}_s(\phi,\|\boldsymbol{A}\|_\infty)\sqrt{\|\boldsymbol{A}\|_\infty^2+1}})$, there exists real-valued functions $\rho(\boldsymbol{v}_1^{(0)}, \boldsymbol{W}^{(0)}, \boldsymbol{B}_{1(n)}^{(0)})$, $J(\tilde{\boldsymbol{a}}_n\boldsymbol{X}, \boldsymbol{v}_1^{(0)}, \boldsymbol{W}^{(0)}, \boldsymbol{B}_{1(n)}^{(0)})$, $R(\tilde{\boldsymbol{a}}_n\boldsymbol{X}, \boldsymbol{v}_1^{(0)}, \boldsymbol{W}^{(0)}, \boldsymbol{B}_{1(n)}^{(0)})$ and $\phi_\epsilon(\tilde{\boldsymbol{a}}_n\boldsymbol{X})$ such that for every $\boldsymbol{X}$*

$$r_{n,1}(\boldsymbol{X}) = \rho(\boldsymbol{v}_1^{(0)}, \boldsymbol{W}^{(0)}, \boldsymbol{B}_{1(n)}^{(0)}) \sum_{j=1}^{N} a_{j,n}\phi_\epsilon(\boldsymbol{a}_j\boldsymbol{X}\boldsymbol{w}^*) + J(\boldsymbol{X}, \boldsymbol{v}_1^{(0)}, \boldsymbol{W}^{(0)}, \boldsymbol{B}_{1(n)}^{(0)}) + R(\boldsymbol{X}, \boldsymbol{v}_1^{(0)}, \boldsymbol{W}^{(0)}, \boldsymbol{B}_{1(n)}^{(0)})$$

Moreover, letting $C = \mathcal{C}_\epsilon(\phi, \|\boldsymbol{A}\|_\infty)\sqrt{\|\boldsymbol{A}\|_\infty^2 + 1}$ be the complexity of ϕ, and if $v_{1,i} \sim \mathcal{N}(0, \frac{1}{m_2})$ and $w_{i,j}^{(0)}, \boldsymbol{B}_{1(n)}^{(0)} \sim \mathcal{N}(0, \frac{1}{m_1})$ are at random initialization, then we have

1. for every fixed $\tilde{\boldsymbol{a}}_n\boldsymbol{X}$, $\rho(\boldsymbol{v}_1^{(0)}, \boldsymbol{W}^{(0)}, \boldsymbol{B}_{1(n)}^{(0)})$ is independent of $J(\tilde{\boldsymbol{a}}_n\boldsymbol{X}, \boldsymbol{v}_1^{(0)}, \boldsymbol{W}^{(0)}, \boldsymbol{B}_{1(n)}^{(0)})$.

2. $\rho(\boldsymbol{v}_1^{(0)}, \boldsymbol{W}^{(0)}, \boldsymbol{B}_{1(n)}^{(0)}) \sim \mathcal{N}(0, \frac{1}{100C^2m_2})$.

3. $|\phi_\epsilon(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_i^*) - \phi(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_i^*)| \leq \epsilon$

4. with high probability, $|R(\boldsymbol{X}, \boldsymbol{v}_1^{(0)}, \boldsymbol{W}^{(0)}, \boldsymbol{B}_{1(n)}^{(0)})| \leq \tilde{O}(\frac{\|\boldsymbol{A}\|_\infty}{\sqrt{m_1m_2}})$, $|J(\boldsymbol{X}, \boldsymbol{v}_1^{(0)}, \boldsymbol{W}^{(0)}, \boldsymbol{B}_{1(n)}^{(0)})| \leq \tilde{O}(\frac{\|\boldsymbol{A}\|_\infty(1+\|\boldsymbol{A}\|_\infty)}{\sqrt{m_2}})$ and $\mathbb{E}[J(\boldsymbol{X}, \boldsymbol{v}_1^{(0)}, \boldsymbol{W}^{(0)}, \boldsymbol{B}_{1(n)}^{(0)})] = 0$.

With high probability, we also have

$$\tilde{\rho}(v_1^{(0)}) \sim \mathcal{N}(0, \frac{\tau}{C^2m_2})$$

$$\mathcal{W}_2(\rho|_{\boldsymbol{W}^{(0)}, \boldsymbol{B}_{1(n)}^{(0)}}, \tilde{\rho}) \leq \tilde{O}(\frac{1}{C\sqrt{m_2}})$$

Proof:
By Lemma B.1, we have

$$\mathbb{E}_{\boldsymbol{w}_i^{(0)} \sim \mathcal{N}(0, \frac{\boldsymbol{I}}{m_1}), b_{1(n,i)} \sim \mathcal{N}(0, \frac{1}{m_1})}[h(\sqrt{m_1}{\boldsymbol{w}_i^{(0)}}^\top \boldsymbol{w}^*, b_{1(n,i)})\mathbb{1}[\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_i^{(0)} + b_{1(n,i)} \geq 0]] = \frac{\phi_\epsilon(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_i^*)}{C}$$

with

$$|\phi_\epsilon(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}^*) - \phi(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}^*)| \leq \epsilon$$

and $|h(\sqrt{m_1}{\boldsymbol{w}_i^{(0)}}^\top \boldsymbol{w}^*, b_{1(n,i)})| \in [0, 1]$. Note that here the h function is rescaled by $1/C$.
Then, applying Lemma A.4 of (Allen-Zhu et al., 2019), we define

$$I_i = I(h(\sqrt{m_1}{\boldsymbol{w}_i^{(0)}}^\top \boldsymbol{w}^*, B_{1(n,i)})) \subset [-2, 2]$$

$$S = \{i \in [m_1] : \sqrt{m_2}v_{1,i}^{(0)} \in I_i\}$$

$$s_i = s(h(\sqrt{m_1}{\boldsymbol{w}_i^{(0)}}^\top \boldsymbol{w}^*, B_{1(n,i)})), \sqrt{m_2}v_{1,i}^{(0)})$$

$$u_i = \begin{cases} \frac{s_i}{\sqrt{|S|}}, & \text{if } i \in S \\ 0, & \text{if } s \notin S \end{cases}$$

where $u_i,\ i \in [m_1]$ is independent of $\boldsymbol{W}^{(0)}$. We can write

$$\boldsymbol{W}^{(0)} = \alpha\boldsymbol{e}_d\boldsymbol{u}^\top + \boldsymbol{\beta},$$

where $\alpha = \boldsymbol{u}^\top\boldsymbol{e}_d^\top\boldsymbol{W}^{(0)} \sim \mathcal{N}(0, 1/m_1)$ and $\beta \in \mathbb{R}^{d\times m_1}$ are two independent random variables given $\boldsymbol{u}$. We know α is independent of $\boldsymbol{u}$. Since each $i \in S$ with probability τ, we know with high probability,

$$|S| = \tilde{\Theta}(\tau m_1) \tag{67}$$

Since $\alpha = \sum_{i\in S} u_i[\boldsymbol{e}_d^\top \boldsymbol{W}^{(0)}]_i$ and $|u_i[\boldsymbol{e}_d^\top \boldsymbol{W}^{(0)}]_i| \leq \tilde{O}(1/\sqrt{m_1|S|})$, by (67) and the Wasserstein distance bound of central limit theorem we know there exists $g \sim \mathcal{N}(0, \frac{1}{m_1})$ such that

$$\mathcal{W}_2(\alpha|_{\boldsymbol{W}^{(0)},\boldsymbol{B}^{(0)}_{1(n)}}, g) \leq \tilde{O}(\frac{1}{\sqrt{\tau}m_1})$$

Then,

$$\begin{aligned} r_{n,1}(\boldsymbol{X}) &= \sum_{j=1}^N a_{j,n} \sum_{i=1}^{m_1} v_{i,1}^{(0)} \sigma(\boldsymbol{a}_j \boldsymbol{X} \boldsymbol{w}_1^{(0)} + B_{1(n,i)}^{(0)}) \\ &= \sum_{j=1}^N a_{j,n} \sum_{i\notin S} v_{i,1}^{(0)} \sigma(\boldsymbol{a}_j \boldsymbol{X} \boldsymbol{w}_1^{(0)} + B_{1(n,i)}^{(0)}) + \sum_{j=1}^N a_{j,n} \sum_{i\in S} v_{i,1}^{(0)} \sigma(\boldsymbol{a}_j \boldsymbol{X} \boldsymbol{w}_1^{(0)} + B_{1(n,i)}^{(0)}) \\ &= J_1 + \sum_{j=1}^N a_{j,n} \sum_{i\in S} v_{i,1}^{(0)} \sigma(\boldsymbol{a}_j \boldsymbol{X} \boldsymbol{w}_1^{(0)} + B_{1(n,i)}^{(0)}) \end{aligned} \tag{68}$$

$$\begin{aligned} & r_{n,1}(\boldsymbol{X}) - J_1 \\ =& \sum_{j=1}^N a_{j,n} \sum_{i\in S} v_{i,1}^{(0)} \mathbb{1}[\boldsymbol{a}_j \boldsymbol{X} \boldsymbol{w}_1^{(0)} + B_{1(n,i)}^{(0)}] \frac{s_i}{2\sqrt{|S|}} \alpha + \sum_{j=1}^N a_{j,n} \sum_{i\in S} v_{i,1}^{(0)} \mathbb{1}[\boldsymbol{a}_j \boldsymbol{X} \boldsymbol{w}_1^{(0)} + B_{1(n,i)}^{(0)}](\boldsymbol{a}_j \boldsymbol{X} \boldsymbol{\beta}_i + B_{1(n,i)}^{(0)}) \\ =& P_1 + P_2 \end{aligned} \tag{69}$$

Here, we know that since

$$\mathbb{E}[v_{i,1}^{(0)} \sigma(\boldsymbol{a}_j \boldsymbol{X} \boldsymbol{w}_1^{(0)} + B_{1(n,i)}^{(0)})] = \mathbb{E}[v_{i,1}^{(0)}] \cdot \mathbb{E}[\sigma(\boldsymbol{a}_j \boldsymbol{X} \boldsymbol{w}_1^{(0)} + B_{1(n,i)}^{(0)})] = 0 \tag{70}$$

Hence,

$$\mathbb{E}[J_1] = \mathbb{E}[\sum_{j=1}^N a_{j,n} \sum_{i\notin S} v_{i,1}^{(0)} \sigma(\boldsymbol{a}_j \boldsymbol{X} \boldsymbol{w}_1^{(0)} + B_{1(n,i)}^{(0)})] = 0 \tag{71}$$

Then we can derive

$$P_1 = \sum_{j=1}^N a_{j,n} \sum_{i\in S} \mathbb{1}[\boldsymbol{a}_j \boldsymbol{X} \boldsymbol{w}_1^{(0)} + B_{1(n,j)}^{(0)}] \frac{\alpha}{2\sqrt{|S|m_2}} h(\sqrt{m_1} {\boldsymbol{w}_i^{(0)}}^\top \boldsymbol{w}^*, B_{1(n,i)}) + R_1 \tag{72}$$

where $|R_1| \leq \tilde{O}(\sqrt{\frac{|S|}{m_1 m_2}})$. We write $P_3 = \frac{P_1 - R_1}{\alpha}$. Then,

$$|P_3 - \frac{\sqrt{|S|}}{\sqrt{m_2}C} \sum_{j=1}^N a_{j,n} \phi_\epsilon(\boldsymbol{a}_j \boldsymbol{X} \boldsymbol{w}^*)| \leq \tilde{O}(\|\boldsymbol{A}\|_\infty \frac{1}{\sqrt{m_2}})$$

$$|\frac{C\sqrt{m_2}}{\sqrt{\tau m_1}} P_1 - \sum_{j=1}^N a_{j,n} \phi_\epsilon(\boldsymbol{a}_j \boldsymbol{X} \boldsymbol{w}^*)| \leq \tilde{O}(\|\boldsymbol{A}\|_\infty \frac{C}{\sqrt{\tau m_1}})$$

Define

$$\rho(\boldsymbol{v}_1^{(0)}, \boldsymbol{W}^{(0)}, \boldsymbol{B}_{1(n)}^{(0)}) = \frac{\sqrt{\tau m_1}}{C\sqrt{m_2}} \alpha \sim \mathcal{N}(0, \frac{\tau}{C^2 m_2})$$

Then,

$$P_1 = \rho(\boldsymbol{v}_1^{(0)}, \boldsymbol{W}^{(0)}, \boldsymbol{B}_{1(n)}^{(0)}) \cdot \sum_{j=1}^N a_{j,n} \phi_\epsilon(\boldsymbol{a}_j \boldsymbol{X} \boldsymbol{w}^*) + R_1 + R_2(\boldsymbol{X}, \boldsymbol{v}_1^{(0)}, \boldsymbol{W}^{(0)}, \boldsymbol{B}_{1(n)}^{(0)})$$

where $|R_2| \leq \tilde{O}(\frac{1}{\sqrt{m_1 m_2}})$.
We can also define

$$\tilde{\rho}(v_1^{(0)}) = \frac{\sqrt{\tau m_1}}{C\sqrt{m_2}} g \sim \mathcal{N}(0, \frac{\tau}{C^2 m_2})$$

Therefore,

$$\mathcal{W}_2(\rho|_{\boldsymbol{W}^{(0)},\boldsymbol{B}^{(0)}_{1(n)}},\tilde{\rho}) \leq \tilde{O}(\frac{1}{C\sqrt{m_2}})$$

Meanwhile,

$$\boldsymbol{a}_j\boldsymbol{X}\boldsymbol{w}_i^{(0)} = \alpha\frac{s_i}{\sqrt{|S|}}\boldsymbol{a}_j\boldsymbol{X}\boldsymbol{e}_d + \boldsymbol{a}_j\boldsymbol{X}\boldsymbol{\beta}_i + B^{(0)}_{1(n,i)} = \boldsymbol{a}_j\boldsymbol{X}\boldsymbol{\beta}_i + B^{(0)}_{1(n,i)} \pm \tilde{O}(\frac{1}{\sqrt{|S|m_1}})$$

we have

$$P_2 = \sum_{j=1}^{N} a_{j,n}\sum_{i\in S} v_{i,1}^{(0)}\mathbb{1}[\boldsymbol{a}_j\boldsymbol{X}\boldsymbol{\beta}_i + b^{(0)}_{1(n,j)}](\boldsymbol{a}_j\boldsymbol{X}\boldsymbol{\beta}_i + b^{(0)}_{1(n,i)}) + R_3 = J_2 + R_3$$

$$\mathbb{E}[J_2] = 0 \tag{73}$$

with $|R_3| \leq \tilde{O}(\frac{\|\boldsymbol{A}\|_\infty}{\sqrt{m_1m_2}})$.

Let $J = J_1 + J_2$, $R = R_1 + R_2 + R_3$. Then, w.h.p., $\mathbb{E}[J] = 0$, $|J| \leq \tilde{O}(\frac{\|\boldsymbol{A}\|_\infty(1+\|\boldsymbol{A}\|_\infty)}{\sqrt{m_2}})$, $|R| \leq \tilde{O}(\frac{\|\boldsymbol{A}\|_\infty}{\sqrt{m_1m_2}})$.

Lemma B.6. *For every $\epsilon \in (0, \frac{1}{\mathcal{C}_s(\phi,\|\boldsymbol{A}\|_\infty)\sqrt{\|\boldsymbol{A}\|_\infty^2+1}})$, there exists real-valued functions $\phi_{1,j.\epsilon}(\cdot)$ such that*

$$|\phi_{1,j,\epsilon}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}^*_{1,j}) - \phi_{1,j}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}^*_{1,j})| \leq \epsilon$$

for $j \in [p_2]$. Denote by

$$C = \mathcal{C}_\epsilon(\phi,\|\boldsymbol{A}\|_\infty)\sqrt{\|\boldsymbol{A}\|_\infty^2+1},\ C' = 10C\sqrt{p_2},\ \phi_{1,j,\epsilon}(\boldsymbol{a}_j\boldsymbol{X}\boldsymbol{w}^*_{1,i}) = \frac{1}{C'}\phi_{1,j,\epsilon}(\boldsymbol{a}_j\boldsymbol{X}\boldsymbol{w}^*_{1,i})$$

For every $i \in [m_2]$, there exist independent Gaussians

$$\alpha_{i,j} \sim \mathcal{N}(0,\frac{1}{m_2}),\ \beta_i(\boldsymbol{X}) \sim \mathcal{N}(0,\frac{1}{m_2}),$$

satisfying

$$\mathcal{W}_2(r_{n,i}(\boldsymbol{X}),\sum_{j\in[p_2]}\alpha_{i,j}\sum_{m=1}^{N}a_{m,n}\phi_{1,j,\epsilon}(\boldsymbol{a}_m\boldsymbol{X}\boldsymbol{w}^*_{1,i}) + C_i\beta_i(\boldsymbol{X})) \leq \tilde{O}(\frac{p_2^{\frac{2}{3}}}{m_1^{\frac{1}{6}}\sqrt{m_2}})$$

Proof:

Define p_2S many chunks of the first layer with each chunk corresponding to a set $S_{j,l}$, where $|S_{j,l}| = m_1/(p_2S)$ for $j \in [p_2]$ and $l \in [S]$, such that

$$\mathcal{S}_{j,l} = \{(j-1)\frac{m_1}{p_2} + (l-1)\frac{m_1}{p_2}S + k | k \in [\frac{m_1}{p_2S}]\} \subset [m_1]$$

By Lemma B.5, we have

$$\begin{aligned} r_{n,i}(\boldsymbol{X}) = &\sum_{j\in[p_2],l\in[S]}\rho(\boldsymbol{v}_i^{(0)}[j,l],\boldsymbol{W}^{(0)}[j,l],\boldsymbol{B}^{(0)}_{1(n)}[j,l])\sum_{m=1}^{N}a_{m,n}\phi_\epsilon(\boldsymbol{a}_m\boldsymbol{X}\boldsymbol{w}^*_{1,j}) \\ &+ \sum_{j\in[p_2],l\in[S]} J_j(\boldsymbol{X},\boldsymbol{v}_i^{(0)}[j,l],\boldsymbol{W}^{(0)}[j,l],\boldsymbol{B}^{(0)}_{1(n)}[j,l]) + R_j(\boldsymbol{X},\boldsymbol{v}_i^{(0)}[j,l],\boldsymbol{W}^{(0)}[j,l],\boldsymbol{B}^{(0)}_{1(n)}[j,l]), \end{aligned} \tag{74}$$

where $\rho(\boldsymbol{v}_i^{(0)}[j,l],\boldsymbol{W}^{(0)}[j,l],\boldsymbol{B}^{(0)}_{1(n)}[j,l]) \sim \mathcal{N}(0,\frac{1}{100C^2m_2p_2S})$. Then $\rho_j = \sum_{l\in[S]}\rho_{j,l} \sim \mathcal{N}(0,\frac{1}{C'^2m_2})$ for $C' = 10C\sqrt{p_2}$. Define

$$J_j^S(\boldsymbol{X}) = \sum_{l\in[S]} J_j(\boldsymbol{X},\boldsymbol{v}_i^{(0)}[j,l],\boldsymbol{W}^{(0)}[j,l],\boldsymbol{B}^{(0)}_{1(n)}[j,l])$$

$$R_j^S(\boldsymbol{X}) = \sum_{l\in[S]} R_j(\boldsymbol{X},\boldsymbol{v}_i^{(0)}[j,l],\boldsymbol{W}^{(0)}[j,l],\boldsymbol{B}^{(0)}_{1(n)}[j,l])$$

Then there exists Gaussian random variables $\beta_j(\boldsymbol{X})$ and $\beta'(\boldsymbol{X}) = \sum_{i\in[p_2]}\beta_j(\boldsymbol{X})$ such that

$$\mathcal{W}_2(J_j^S(\boldsymbol{X}), \beta_j(\boldsymbol{X})) \leq \frac{\|\boldsymbol{A}\|_\infty(1+\|\boldsymbol{A}\|_\infty)}{\sqrt{m_2 pS}}$$

$$\mathcal{W}_2(r_{n,i}(\boldsymbol{X}), \sum_{j\in[p_2]}\rho_j\sum_{m=1}^N a_{m,n}\phi_{1,j,\epsilon}(\boldsymbol{a}_m\boldsymbol{X}\boldsymbol{w}_{1,j}^*) + \beta'(\boldsymbol{X})) \leq \tilde{O}(\frac{Sp_2}{\sqrt{m_1 m_2}} + \frac{\sqrt{p_2}\|\boldsymbol{A}\|_\infty(1+\|\boldsymbol{A}\|_\infty)}{m_2 S})$$

We know there exists a positive constant C_i such that $\beta'/C_i \sim \mathcal{N}(0, \frac{1}{m_2})$. Let $\alpha_{i,j} = C'\rho_j$, $\beta_i' = \beta'/C_i$. Notice that $\mathbb{E}\big[\sum_{l\in[S],i\in[p_2]}[J_j^2(\boldsymbol{X}, \boldsymbol{v}_i^{(0)}[j,l], \boldsymbol{W}^{(0)}[j,l], b_1^{(0)}[j,l])]\big] = \tilde{O}(\|\boldsymbol{A}\|_\infty^2(1+\|\boldsymbol{A}\|_\infty)^2/m_2)$. Hence, we have

$$C_i \leq \tilde{O}(\|\boldsymbol{A}\|_\infty(1+\|\boldsymbol{A}\|_\infty)$$

Let $S = (m_1/p_2)^{\frac{1}{3}}$, we can obtain

$$\mathcal{W}_2(r_{n,i}(\boldsymbol{X}), \sum_{j\in[p_2]}\alpha_{i,j}\sum_{m=1}^N a_{m,n}\phi_{1,j,\epsilon}(\boldsymbol{a}_m\boldsymbol{X}\boldsymbol{w}_{1,i}^*) + C_i\beta_i(\boldsymbol{X})) \leq \tilde{O}(\frac{p_2^{\frac{2}{3}}}{m_1^{\frac{1}{6}}\sqrt{m_2}})$$

Lemma B.7. *There exists function $h:\mathbb{R}^2\to[-C'',C'']$ for $C'' = \mathcal{C}_\epsilon(\Phi, C')\sqrt{\|\boldsymbol{A}\|_\infty^2+1}$ such that*

$$\begin{aligned}
&\mathbb{E}[\mathbb{1}_{r_{n,i}(\boldsymbol{X})+b_{2(n,i)}^{(0)}\geq 0}h(\sqrt{m_2}\sum_{j\in[p_2]}v_{1,j}^*\alpha_{i,j}, b_{2(n,i)}^{(0)})(\sum_{j\in[p_2]}v_{2,j}^*\phi_{2,j}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_{2,j}^*))]\\
=&\Phi(\sum_{j\in[p_2]}v_{1,j}^*\sum_{m=1}^N a_{m,n}\phi_{1,j})\sum_{j\in[p_2]}v_{2,j}^*\phi_{2,j}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_{2,j}^*)) \pm \tilde{O}(p_2^2\mathcal{C}_s(\Phi, p_2\mathcal{C}_s(\phi,\|\boldsymbol{A}\|_\infty))\mathcal{C}_s(\phi,\|\boldsymbol{A}\|_\infty)\sqrt{\|\boldsymbol{A}\|_\infty^2+1}\epsilon)
\end{aligned} \tag{75}$$

Proof:
Choose $\boldsymbol{w} = (\alpha_{i,1},\cdots,\alpha_{i,p_2},\beta_i)$, $\boldsymbol{x} = (\sum_{m=1}^N a_{m,n}\phi_{1,1,\epsilon},\cdots,\sum_{m=1}^N a_{m,n}\phi_{1,p_2,\epsilon}, C_i)$ and $\boldsymbol{w}^* = (v_{1,1}^*,\cdots,v_{1,p_2}^*,0)$. Then, $\|\boldsymbol{x}\| \leq O(\|\boldsymbol{A}\|_\infty^2 + \|\boldsymbol{A}\|_\infty)$. By Lemma B.1, there exists $h:\mathbb{R}^2\to[-C'',C'']$ for $C'' = \mathcal{C}_s(\Phi, C')\sqrt{\|\boldsymbol{A}\|_\infty^2+1}$ such that

$$\begin{aligned}
&\mathbb{E}[\mathbb{1}_{\boldsymbol{w}\boldsymbol{X}+b_{2(n,i)}^{(0)}\geq 0}h(\sqrt{m_2}\boldsymbol{w}^\top\boldsymbol{w}^*, b_{2(n,i)}^{(0)})(\sum_{j\in[p_2]}v_{2,j}^*\phi_{2,j}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_{2,j}^*))]\\
=&\mathbb{E}_{\alpha_i,\beta_i}[\mathbb{1}_{\sum_{j\in[p_2]}\alpha_{i,j}\sum_{m=1}^N a_{m,n}\phi_{1,j,\epsilon}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_{1,i}^*)+C_i\beta_i'+b_{2(n,i)}^{(0)}\geq 0}h(\sqrt{m_2}\boldsymbol{w}^\top\boldsymbol{w}^*, b_{2(n,i)}^{(0)})(\sum_{j\in[p_2]}v_{2,j}^*\phi_{2,j}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_{2,j}^*))]\\
=&\Phi(C'\sum_{j\in[p_2]}v_{1,j}^*\sum_{m=1}^N a_{m,n}\phi_{1,j,\epsilon})\sum_{j\in[p_2]}v_{2,j}^*\phi_{2,j}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_{2,j}^*)) \pm \epsilon C'''
\end{aligned} \tag{76}$$

where

$$C''' = \sup|\sum_{j\in[p_2]}v_{2,j}^*\phi_{2,j}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_{2,j}^*)| \leq p_2\mathcal{C}_s(\phi,\|\boldsymbol{A}\|_\infty)\sqrt{\|\boldsymbol{A}\|_\infty^2+1}$$

By Lemma B.6, we know

$$\mathcal{W}_2(r_{n,i}(\boldsymbol{X}), \sum_{j\in[p_2]}\alpha_{i,j}\sum_{m=1}^N a_{m,n}\phi_{1,j,\epsilon}(\boldsymbol{a}_m\boldsymbol{X}\boldsymbol{w}_{1,i}^*) + C_i\beta_i(\boldsymbol{X})) \leq \tilde{O}(\frac{p_2^{\frac{2}{3}}}{m_1^{\frac{1}{6}}\sqrt{m_2}})$$

Denote $\mathcal{H} = \{i\in[m_1] : |\sum_{j\in[p_2]}\alpha_{i,j}\sum_{m=1}^N a_{m,n}\phi_{1,j,\epsilon}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_{1,i}^*) + C_i\beta_i'| \geq \tilde{O}(\frac{2p_2^{\frac{2}{3}}}{m_1^{\frac{1}{6}}\sqrt{m_2}})\}$. Then, for every $i\in[\mathcal{H}]$, we have that

$$\mathbb{1}_{r_{n,i}(\boldsymbol{X})+b_{2(n,i)}^{(0)}\geq 0} = \mathbb{1}_{\sum_{j\in[p_2]}\alpha_{i,j}\sum_{m=1}^N a_{m,n}\phi_{1,j,\epsilon}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_{1,i}^*)+C_i\beta_i'+b_{2(n,i)}^{(0)}\geq 0} \tag{77}$$

$$\Pr\Big(\big|\sum_{j\in[p_2]}\alpha_{i,j}\sum_{m=1}^{N}a_{m,n}\phi_{1,j,\epsilon}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_{1,i}^*)+C_i\beta_i'\big|\le\tilde{O}(\frac{2p_2^{\frac{2}{3}}}{m_1^{\frac{1}{6}}\sqrt{m_2}})\Big)\le\tilde{O}(\frac{2p_2^{\frac{2}{3}}}{m_1^{\frac{1}{6}}\sqrt{m_2}})\cdot\sqrt{m_2}=\tilde{O}(\frac{2p_2^{\frac{2}{3}}}{m_1^{\frac{1}{6}}}),\tag{78}$$

which implies with probability at least $1-2p_2^{2/3}/m_1^{1/6}$, (77) holds. Therefore,

$$\begin{aligned}
&\mathbb{E}[\mathbb{1}_{r_{n,i}(\boldsymbol{X})+b_{2(n,i)}^{(0)}\ge0}h(\sqrt{m_2}\sum_{j\in[p_2]}v_{1,j}^*\alpha_{i,j},b_{2(n,i)}^{(0)})(\sum_{j\in[p_2]}v_{2,j}^*\phi_{2,j}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_{2,j}^*))]\\
=&\mathbb{E}[\mathbb{1}_{\sum_{j\in[p_2]}\alpha_{i,j}\sum_{m=1}^{N}a_{m,n}\phi_{1,j,\epsilon}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_{1,i}^*)+C_i\beta_i'+b_{2(n,i)}^{(0)}\ge0}h(\sqrt{m_2}\sum_{j\in[p_2]}v_{1,j}^*\alpha_{i,j},b_{2(n,i)}^{(0)})(\sum_{j\in[p_2]}v_{2,j}^*\phi_{2,j}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_{2,j}^*))]\\
&\pm\mathbb{E}[\mathbb{1}_{r_{n,i}(\boldsymbol{X})+b_{2(n,i)}^{(0)}\ge0}\ne\mathbb{1}_{\sum_{j\in[p_2]}\alpha_{i,j}\sum_{m=1}^{N}a_{m,n}\phi_{1,j,\epsilon}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_{1,i}^*)+C_i\beta_i'+b_{2(n,i)}^{(0)}\ge0}]O(C'''C'')\\
=&\mathbb{E}[\mathbb{1}_{\sum_{j\in[p_2]}\alpha_{i,j}\sum_{m=1}^{N}a_{m,n}\phi_{1,j,\epsilon}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_{1,i}^*)+C_i\beta_i'+b_{2(n,i)}^{(0)}\ge0}h(\sqrt{m_2}\sum_{j\in[p_2]}v_{1,j}^*\alpha_{i,j},b_{2(n,i)}^{(0)})(\sum_{j\in[p_2]}v_{2,j}^*\phi_{2,j}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_{2,j}^*))]\\
&\pm\tilde{O}(\frac{2p_2^{2/3}}{m_1^{1/6}}C'''C'')\\
=&\Phi(\sum_{j\in[p_2]}v_{1,j}^*\sum_{m=1}^{N}a_{m,n}\phi_{1,j})\sum_{j\in[p_2]}v_{2,j}^*\phi_{2,j}(\boldsymbol{x}))\pm\tilde{O}(p_2^2\mathcal{C}_s(\Phi,p_2\mathcal{C}_s(\phi,\|\boldsymbol{A}\|_\infty))\mathcal{C}_s(\phi,\|\boldsymbol{A}\|_\infty)\sqrt{\|\boldsymbol{A}\|_\infty^2+1}\cdot\epsilon),
\end{aligned}\tag{79}$$

where the first step is by Lemma B.6, the second step is by (77) and (78) and the last step comes from (76) and $m_1\ge M$.

Lemma B.8.

$$\begin{aligned}
&\frac{1}{m_2}\mathbb{E}[\sum_{i=1}^{m_2}\frac{c_{i,l}^2}{\epsilon_c^2}\mathbb{1}_{r_{n,i}(\boldsymbol{X})+b_{2(n,i)}^{(0)}\ge0}h(\sqrt{m_2}\sum_{j\in[p_2]}v_{1,j}^*\alpha_{i,j},b_{2(n,i)}^{(0)})(\sum_{j\in[p_2]}v_{2,j}^*\phi_{2,j}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_{2,j}^*))]\\
=&\Phi(\sum_{j\in[p_2]}v_{1,j}^*\sum_{m=1}^{N}a_{m,n}\boldsymbol{a}_m\boldsymbol{X}\boldsymbol{\delta}\phi_{1,j})\sum_{j\in[p_2]}v_{2,j}^*\phi_{2,j}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_{2,j}^*))\\
&\pm\tilde{O}(p_2^2\mathcal{C}_s(\Phi,p_2\mathcal{C}_s(\phi,\|\boldsymbol{A}\|_\infty))\mathcal{C}_s(\phi,\|\boldsymbol{A}\|_\infty)\sqrt{\|\boldsymbol{A}\|_\infty^2+1}\cdot\epsilon)
\end{aligned}\tag{80}$$

Proof:
Recall $\tilde{\rho}(\boldsymbol{v}_1^{(0)})\sim\mathcal{N}(0,\frac{\tau}{C^2m_2})$. Define $\tilde{\rho}_{j,l}=\tilde{\rho}(\boldsymbol{v}_1^{(0)}[j,l])$. Therefore,

$$\mathcal{W}_2(\rho_{j,l}|_{\boldsymbol{W}^{(0)},\boldsymbol{B}_{1(n)}^{(0)}},\tilde{\rho}_{j,l})\le\tilde{O}(\frac{1}{C'\sqrt{m_2}S})\tag{81}$$

$$\mathcal{W}_2(\rho_j|_{\boldsymbol{W}^{(0)},\boldsymbol{B}_{1(n)}^{(0)}},\tilde{\rho}_j)\le\tilde{O}(\frac{1}{C'\sqrt{m_2}})\tag{82}$$

where $\tilde{\rho}_j=\sum_{l\in[S]}\rho_{j,l}$. We then define $\tilde{\alpha}_{i,j}=C'\tilde{\rho}_j$
Next modify $r_{n,i}(\boldsymbol{X})$. Define

$$\tilde{r}_{n,i}(\boldsymbol{X})=\frac{\sum_{m=1}^{N}a_{m,n}\sum_{j\in[m_1]}v_{j,i}^{(0)}\sigma(\boldsymbol{a}_m\boldsymbol{X}\boldsymbol{w}_i^{(0)}+b_{1(n,i)}^{(0)})}{\|\boldsymbol{u}\|}\mathbb{E}[\|\boldsymbol{u}\|]$$

where $u=(\sigma(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_j^{(0)}+b_{1(n,j)}^{(0)}))_{j\in[m_1]}$. By definition, we know

$$\tilde{r}_{n,i}\sim\mathcal{N}(0,\frac{\|\tilde{\boldsymbol{a}}_n\|_\infty^2}{m_2}\mathbb{E}[\|\boldsymbol{u}\|]^2)$$

Then we have

$$\mathcal{W}_2(r_{n,i}(\boldsymbol{X}),\tilde{r}_{n,i}(\boldsymbol{X}))\le\tilde{O}(\frac{\|\boldsymbol{A}\|_\infty\sqrt{\|\boldsymbol{A}\|_\infty^2+1}}{\sqrt{m_2}})\tag{83}$$

Combining (81), (82), (83) and Lemma B.7, we have

$$\begin{aligned}&\frac{1}{m_2}\mathbb{E}[\sum_{i=1}^{m_2}\frac{c_{i,l}^2}{\epsilon_c^2}\mathbb{1}_{r_{n,i}(\boldsymbol{X})+b_{2(n,i)}^{(0)}\geq 0}h(\sqrt{m_2}\sum_{j\in[p_2]}v_{1,j}^*\alpha_{i,j},b_{2(n,i)}^{(0)})(\sum_{j\in[p_2]}v_{2,j}^*\phi_{2,j}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_{2,j}^*))]\\ =&\Phi(\sum_{j\in[p_2]}v_{1,j}^*\sum_{m=1}^{N}a_{m,n}\phi_{1,j})\sum_{j\in[p_2]}v_{2,j}^*\phi_{2,j}(\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_{2,j}^*))\pm\tilde{O}(p_2^2\mathcal{C}_s(\Phi,p_2\mathcal{C}_s(\phi,\|\boldsymbol{A}\|_\infty))\mathcal{C}_s(\phi,\|\boldsymbol{A}\|_\infty)(\sqrt{\|\boldsymbol{A}\|_\infty^2+1}\cdot\epsilon)\end{aligned} \tag{84}$$

B.1.3. COUPLING

This section illustrates the coupling between the real and pseudo networks. We first define diagonal matrices $\boldsymbol{D}_{n,\boldsymbol{w}}$, $\boldsymbol{D}_{n,\boldsymbol{w}}+\boldsymbol{D}''_{n,\boldsymbol{w}}$, $\boldsymbol{D}_{n,\boldsymbol{w}}+\boldsymbol{D}'_{n,\boldsymbol{w}}$ for node n as the sign of Relu's in the first layer at weights $\boldsymbol{W}^{(0)}$, $\boldsymbol{W}^{(0)}+\boldsymbol{W}^{\rho}$ and $\boldsymbol{W}^{(0)}+\boldsymbol{W}^{\rho}+\boldsymbol{W}'$, respectively. We also define diagonal matrices $\boldsymbol{D}_{n,\boldsymbol{v}}$, $\boldsymbol{D}_{n,\boldsymbol{v}}+\boldsymbol{D}''_{n,\boldsymbol{v}}$, $\boldsymbol{D}_{n,\boldsymbol{v}}+\boldsymbol{D}'_{n,\boldsymbol{v}}$ for node n as the sign of Relu's in the second layer at weights $\{\boldsymbol{W}^{(0)},\boldsymbol{V}^{(0)}\}$, $\{\boldsymbol{W}^{(0)}+\boldsymbol{W}^{\rho},\boldsymbol{V}^{(0)}+\boldsymbol{V}^{\rho}\}$ and $\{\boldsymbol{W}^{(0)}+\boldsymbol{W}^{\rho}+\boldsymbol{W}',\boldsymbol{V}^{(0)}+\boldsymbol{V}^{\rho}+\boldsymbol{V}'\}$, respectively. For every $l\in[K]$, we then introduce the pseudo network and its semi-bias, bias-free version as

$$g_l(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X},\boldsymbol{W},\boldsymbol{V})=\boldsymbol{q}^\top\boldsymbol{A}((\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{W}+\boldsymbol{B}_1)\odot(\boldsymbol{D}_{\boldsymbol{w}}+\boldsymbol{D}'_{\boldsymbol{w}})\boldsymbol{V}+\boldsymbol{B}_2)\odot(\boldsymbol{D}_{\boldsymbol{v}}+\boldsymbol{D}'_{\boldsymbol{v}}))\boldsymbol{c}_l \tag{85}$$

$$g_l^{(b)}(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X},\boldsymbol{W},\boldsymbol{V})=\boldsymbol{q}^\top\boldsymbol{A}((\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{W}+\boldsymbol{B}_1)\odot(\boldsymbol{D}_{\boldsymbol{w}}+\boldsymbol{D}'_{\boldsymbol{w}})\boldsymbol{V})\odot(\boldsymbol{D}_{\boldsymbol{v}}+\boldsymbol{D}'_{\boldsymbol{v}}))\boldsymbol{c}_l \tag{86}$$

$$g_l^{(b,b)}(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X},\boldsymbol{W},\boldsymbol{V})=\boldsymbol{q}^\top\boldsymbol{A}((\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{W})\odot(\boldsymbol{D}_{\boldsymbol{w}}+\boldsymbol{D}'_{\boldsymbol{w}})\boldsymbol{V})\odot(\boldsymbol{D}_{\boldsymbol{v}}+\boldsymbol{D}'_{\boldsymbol{v}}))\boldsymbol{c}_l \tag{87}$$

Lemma B.9 gives the final result of coupling with added Drop-out noise. Lemma B.10 states the sparse sign change in Relu and the function value changes of pseudo network by some update. To be more specific, Lemma B.11 shows that the sign pattern can be viewed as fixed for the smoothed objective when a small update is introduced to the current weights. Lemma B.12 proves the bias-free pseudo network can also approximate the target function.

Lemma B.9. *Let $F_{\boldsymbol{A}}=(f_1,f_2,\cdots,f_K)$. With high probability, we have for any $\|\boldsymbol{W}'\|_{2,4}\leq\tau_w$, $\|\boldsymbol{V}'\|_F\leq\tau_v$, such that*

$$\begin{aligned}&f_l(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X},\boldsymbol{W}^{(0)}+\boldsymbol{W}'\boldsymbol{\Sigma},\boldsymbol{V}^{(0)}+\boldsymbol{\Sigma}\boldsymbol{V}')\\ =&\boldsymbol{q}^\top\boldsymbol{A}(\boldsymbol{A}((\boldsymbol{A}\boldsymbol{X}\boldsymbol{W}^{(0)}+\boldsymbol{B}_1^{(0)})\odot\boldsymbol{D}_{\boldsymbol{w},\boldsymbol{x}}^{(0)}\boldsymbol{V}^{(0)}+\boldsymbol{B}_2^{(0)})\odot\boldsymbol{D}_{\boldsymbol{v},\boldsymbol{x}}^{(0)})\boldsymbol{c}+\boldsymbol{q}^\top\boldsymbol{A}(\boldsymbol{A}((\boldsymbol{A}\boldsymbol{X}\boldsymbol{W}')\odot\boldsymbol{D}_{\boldsymbol{w},\boldsymbol{x}}^{(0)}\boldsymbol{V}')\odot\boldsymbol{D}_{\boldsymbol{v},\boldsymbol{x}}^{(0)})\boldsymbol{c}_l\\ &\pm\tilde{O}(\tau_v\frac{\sqrt{m_2}}{\sqrt{m_1}}+m_1^{\frac{9}{5}}\tau_w^{\frac{16}{5}}\sqrt{m_2}+\tau_w^{\frac{8}{5}}m_1^{\frac{9}{10}})\cdot\|\boldsymbol{q}^\top\boldsymbol{A}\|_1\|\boldsymbol{A}\|_\infty^2,\end{aligned} \tag{88}$$

where we use $\boldsymbol{D}_{\boldsymbol{w},\boldsymbol{x}}^{(0)}$ and $\boldsymbol{D}_{\boldsymbol{v},\boldsymbol{x}}^{(0)}$ to denote the sign matrices at random initialization $\boldsymbol{W}^{(0)}$, $\boldsymbol{V}^{(0)}$ and we let $\boldsymbol{D}_{\boldsymbol{w},\boldsymbol{x}}^{(0)}+\boldsymbol{D}'_{\boldsymbol{w},\boldsymbol{x}}$, $\boldsymbol{D}_{\boldsymbol{v},\boldsymbol{x}}^{(0)}+\boldsymbol{D}'_{\boldsymbol{v},\boldsymbol{x}}$ be the sign matrices at $\boldsymbol{W}+\boldsymbol{W}'\boldsymbol{\Sigma}$, $\boldsymbol{V}+\boldsymbol{\Sigma}\boldsymbol{V}'$.

Proof:
Since $\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_i^{(0)}+B_{1(n,i)}^{(0)}=\tilde{\boldsymbol{a}}_n\tilde{\boldsymbol{X}}\tilde{\boldsymbol{w}}_i^{(0)}$ where $\tilde{\boldsymbol{w}}_i^{(0)}=(\boldsymbol{w}_i^{(0)},B_{1(n,i)}^{(0)})\in\mathbb{R}^{d+1}$ and $\tilde{\boldsymbol{X}}=(\boldsymbol{X},\boldsymbol{1})\in\mathbb{R}^{N\times(d+1)}$, we can ignore the bias term for simplicity. Define

$$\boldsymbol{Z}=\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{W}^{(0)})\odot\boldsymbol{D}_{\boldsymbol{w},\boldsymbol{x}}^{(0)}$$

$$\boldsymbol{Z}_1=\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{W}'\boldsymbol{\Sigma})\odot\boldsymbol{D}_{\boldsymbol{w},\boldsymbol{x}}^{(0)}$$

$$\boldsymbol{Z}_2=\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}(\boldsymbol{W}^{(0)}+\boldsymbol{W}')\boldsymbol{\Sigma})\odot\boldsymbol{D}'_{\boldsymbol{w},\boldsymbol{x}}$$

Then by Fact C.9 in (Allen-Zhu et al., 2019) we have

$$\begin{aligned}\|\boldsymbol{Z}_n\boldsymbol{\Sigma}\boldsymbol{V}'\|_2^2\leq&\sum_{i=1}^{m_2}(\boldsymbol{Z}_n\boldsymbol{\Sigma}\boldsymbol{V}'_i)^2\leq\sum_{i=1}^{m_2}\tilde{O}(\|\boldsymbol{Z}_n\|_\infty^2\cdot\|\boldsymbol{V}'_i\|_2^2)\\ \leq&\tilde{O}(\|\boldsymbol{A}\|_\infty^2m_1^{-1}\tau_v^2)\end{aligned} \tag{89}$$

Therefore, we have $\|\boldsymbol{Z}_n\boldsymbol{\Sigma}\boldsymbol{V}'\|_2 \leq \tilde{O}(\|\boldsymbol{A}\|_\infty m_1^{-\frac{1}{2}}\tau_v)$.
Let s be the total number of sign changes in the first layer caused by adding $\boldsymbol{W}'$. Note that the total number of coordinated i such that $|\tilde{\boldsymbol{a}}_n\boldsymbol{X}\boldsymbol{w}_i^{(0)}| \leq s'' = \frac{2\tau_w}{s^{\frac{1}{4}}}$ is at most $s''m_1^{\frac{3}{2}}$ with high probability. Since $\|\boldsymbol{W}'\|_{2,4} \leq \tau_w$, we must have $s \leq \tilde{O}(s''m^{\frac{3}{2}}) = \tilde{O}(\frac{\tau_w}{s^{\frac{1}{4}}}m_1^{\frac{3}{2}})$. Therefore, $\|\boldsymbol{Z}_{2,n}\|_0 \leq s = \tilde{O}(\tau_w^{\frac{4}{5}}m_1^{\frac{6}{5}})$. Then,

$$\begin{aligned}
\|\boldsymbol{Z}_{2,n}\|_2 =& \|(\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}(\boldsymbol{W}^{(0)}+\boldsymbol{W}'\boldsymbol{\Sigma}))\odot \boldsymbol{D}'_{\boldsymbol{w},\boldsymbol{x}})_n\| \\
\leq& \Big(s\cdot \sum_{(\boldsymbol{D}'_{\boldsymbol{w},\boldsymbol{x}})_n\neq 0}(\boldsymbol{A}\boldsymbol{A}\boldsymbol{X}\boldsymbol{W}^{(0)}+\boldsymbol{A}\boldsymbol{A}\boldsymbol{X}\boldsymbol{W}'\boldsymbol{\Sigma})_{n,i}^4\Big)^{\frac{1}{4}} \\
\leq& \Big(s\cdot \sum_{(\boldsymbol{D}'_{\boldsymbol{w},\boldsymbol{x}})_n\neq 0}(\boldsymbol{A}\boldsymbol{A}\boldsymbol{X}\boldsymbol{W}'\boldsymbol{\Sigma})_{n,i}^4\Big)^{\frac{1}{4}} \\
\leq& s^{\frac{1}{4}}\|\boldsymbol{A}\|_\infty\tau_w \\
\leq& \tilde{O}(\tau_w^{\frac{6}{5}}m_1^{\frac{3}{10}}\|\boldsymbol{A}\|_\infty)
\end{aligned} \tag{90}$$

Then we have

$$\|\boldsymbol{Z}_{2,n}\boldsymbol{\Sigma}\boldsymbol{V}'\|_2 \leq \tilde{O}(\tau_v\tau_w^{\frac{6}{5}}m_1^{\frac{3}{10}}\|\boldsymbol{A}\|_\infty)$$

With high probability, we have

$$\sum_{n=1}^N \boldsymbol{q}^\top\boldsymbol{a}_n\sum_{i=1}^{m_2}c_{i,l}(\sigma(r_{n,i}+r'_{n,i})-\sigma(r_{n,i})) \leq \tilde{O}(\|\boldsymbol{q}\|\sqrt{m_2})\|r'_{n,i}\|$$

$$\begin{aligned}
&f_l(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X},\boldsymbol{W}^{(0)}+\boldsymbol{W}'\boldsymbol{\Sigma},\boldsymbol{V}^{(0)}+\boldsymbol{\Sigma}\boldsymbol{V}') \\
=&\sum_{n=1}^N \boldsymbol{q}^\top\boldsymbol{a}_n\sum_{i=1}^{m_2}c_{i,l}\sigma\Big((\boldsymbol{Z}+\boldsymbol{Z}_1+\boldsymbol{Z}_2)_n^\top(\boldsymbol{V}_i+(\boldsymbol{\Sigma})_i\boldsymbol{V}')\Big) \\
=&\sum_{n=1}^N \boldsymbol{q}^\top\boldsymbol{a}_n\sum_{i=1}^{m_2}c_{i,l}\sigma\Big((\boldsymbol{Z}_n+\boldsymbol{Z}_{1,n}+\boldsymbol{Z}_{2,n})^\top\boldsymbol{V}_i+\boldsymbol{Z}_{1,n}^\top(\boldsymbol{\Sigma}\boldsymbol{V}')_i\Big) \pm \tilde{O}(\|\boldsymbol{q}\|\|\boldsymbol{A}\|_\infty\frac{\sqrt{m_2}}{\sqrt{m_1}}\tau_v+\sqrt{m_2}\|\boldsymbol{q}\|\|\boldsymbol{A}\|_\infty\tau_w^{\frac{6}{5}}m_1^{\frac{3}{10}})
\end{aligned} \tag{91}$$

We consider the difference between

$$A_1 = \boldsymbol{q}^\top\boldsymbol{A}(((\boldsymbol{Z}+\boldsymbol{Z}_1+\boldsymbol{Z}_2)\boldsymbol{V}^{(0)}+\boldsymbol{Z}_1\boldsymbol{\Sigma}\boldsymbol{V}')\odot(\boldsymbol{D}^{(0)}_{\boldsymbol{v},\boldsymbol{x}}+\boldsymbol{D}''_{\boldsymbol{v},\boldsymbol{x}}))\boldsymbol{c}_l$$

$$A_2 = \boldsymbol{q}^\top\boldsymbol{A}(((\boldsymbol{Z}+\boldsymbol{Z}_1+\boldsymbol{Z}_2)\boldsymbol{V}^{(0)}+\boldsymbol{Z}_1\boldsymbol{\Sigma}\boldsymbol{V}')\odot\boldsymbol{D}^{(0)}_{\boldsymbol{v},\boldsymbol{x}})\boldsymbol{c}_l$$

where $\boldsymbol{D}''_{\boldsymbol{v},\boldsymbol{x}}$ is the diagonal sign change matrix from $\boldsymbol{Z}\boldsymbol{V}^{(0)}$ to $(\boldsymbol{Z}+\boldsymbol{Z}_1+\boldsymbol{Z}_2)\boldsymbol{V}^{(0)}+\boldsymbol{Z}_1\boldsymbol{\Sigma}\boldsymbol{V}'$. The difference includes three terms.

$$\|\boldsymbol{Z}_{1,n}\boldsymbol{V}^{(0)}\|_\infty \leq \tilde{O}(\|\boldsymbol{A}\|_\infty m_1^{\frac{1}{4}}\tau_w m_2^{-\frac{1}{2}}) \tag{92}$$

$$\|\boldsymbol{Z}_{2,n}\boldsymbol{V}^{(0)}\|_\infty \leq \tilde{O}(\|\boldsymbol{Z}_{2,n}\|m_2^{-\frac{1}{2}}\sqrt{s}) \leq \tilde{O}(\|\boldsymbol{A}\|_\infty m_1^{\frac{9}{10}}\tau_w^{\frac{8}{5}}m_2^{-\frac{1}{2}}) \tag{93}$$

$$\|\boldsymbol{Z}_{1,n}\boldsymbol{\Sigma}\boldsymbol{V}'\| \leq \tilde{O}(\tau_v\|\boldsymbol{A}\|_\infty\tau_w m_1^{\frac{1}{4}}) \tag{94}$$

where (92) is by Fact C.9 in (Allen-Zhu et al., 2019). Then we have

$$|A_1-A_2| \leq \|\boldsymbol{q}^\top\boldsymbol{A}\|_1\cdot\tilde{O}(m_2^{\frac{3}{2}}\|\boldsymbol{A}\|_\infty^2(m_1^{\frac{1}{4}}\tau_w m_2^{-\frac{1}{2}}+m_1^{\frac{9}{10}}\tau_w^{\frac{8}{5}}m_2^{-\frac{1}{2}})^2+m_2^{\frac{1}{2}}\tau_v^{\frac{4}{3}}\|\boldsymbol{A}\|_\infty^{\frac{4}{3}}\tau_w^{\frac{4}{3}}m_1^{\frac{1}{3}})$$

From A_2 to our goal

$$A_3 = \boldsymbol{q}^\top\boldsymbol{A}(\boldsymbol{Z}\boldsymbol{V}^{(0)}\odot\boldsymbol{D}^{(0)}_{\boldsymbol{v},\boldsymbol{x}})\boldsymbol{c}_l+\boldsymbol{q}^\top\boldsymbol{A}(\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{W}'\odot\boldsymbol{D}^{(0)}_{\boldsymbol{w},\boldsymbol{x}}\boldsymbol{V}')\odot\boldsymbol{D}^{(0)}_{\boldsymbol{v},\boldsymbol{x}})\boldsymbol{c}_l$$

There are two more terms.

$$|\boldsymbol{q}^\top \boldsymbol{A}(\boldsymbol{Z}_2\boldsymbol{V}^{(0)} \odot \boldsymbol{D}^{(0)}_{\boldsymbol{v},\boldsymbol{x}})\boldsymbol{c}_l| \le \tilde{O}(\|\boldsymbol{q}^\top \boldsymbol{A}\|_1 \|\boldsymbol{Z}_{2,n}\|\sqrt{s}) \le \tilde{O}(\|\boldsymbol{q}\|_1 \|\boldsymbol{A}\|_\infty \tau_w^{\frac{8}{5}} m_1^{\frac{9}{10}})$$

$$|\boldsymbol{q}^\top \boldsymbol{A}(\boldsymbol{Z}_1\boldsymbol{V}^{(0)} \odot \boldsymbol{D}^{(0)}_{\boldsymbol{v},\boldsymbol{x}})\boldsymbol{c}_l| \le \tilde{O}(\|\boldsymbol{q}^\top \boldsymbol{A}\|_1 \|\boldsymbol{A}\|_\infty \tau_w m_1^{\frac{1}{4}})$$

Therefore, we have

$$|A_2 - A_3| \le \tilde{O}(\|\boldsymbol{q}^\top \boldsymbol{A}\|_1 \|\boldsymbol{A}\|_\infty \tau_w^{\frac{8}{5}} m_1^{\frac{9}{10}} + \|\boldsymbol{q}\|_1 \|\boldsymbol{A}\|_\infty \tau_w m_1^{\frac{1}{4}} + \|\boldsymbol{q}\|_1 \tau_v \|\boldsymbol{A}\|_\infty \tau_w m_1^{\frac{1}{4}})$$

Finally, we have

$$\begin{aligned} & f_l(\boldsymbol{q}, \boldsymbol{A}, \boldsymbol{X}, \boldsymbol{W}^{(0)} + \boldsymbol{W}'\boldsymbol{\Sigma}, \boldsymbol{V}^{(0)} + \boldsymbol{\Sigma}\boldsymbol{V}') \\ =& \boldsymbol{q}^\top \boldsymbol{A}(\boldsymbol{Z}\boldsymbol{V}^{(0)} \odot \boldsymbol{D}^{(0)}_{\boldsymbol{v},\boldsymbol{x}})\boldsymbol{c}_l + \boldsymbol{q}^\top \boldsymbol{A}(\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{W}' \odot \boldsymbol{D}^{(0)}_{\boldsymbol{w},\boldsymbol{x}}\boldsymbol{V}') \odot \boldsymbol{D}^{(0)}_{\boldsymbol{v},\boldsymbol{x}})\boldsymbol{c}_l \\ & \pm \tilde{O}(\tau_v \frac{\sqrt{m_2}}{\sqrt{m_1}} + m_1^{\frac{9}{5}} \tau_w^{\frac{16}{5}} \sqrt{m_2} + \tau_w^{\frac{8}{5}} m_1^{\frac{9}{10}}) \cdot \|\boldsymbol{q}\|_1 \|\boldsymbol{A}\|_\infty \end{aligned} \tag{95}$$

Lemma B.10. *Suppose $\tau_v \in (0,1]$, $\tau_w \in [\frac{1}{m_1^{\frac{3}{2}}}, \frac{1}{m_1^{\frac{1}{2}}}]$, $\sigma_w \in [\frac{1}{m_1^{\frac{3}{2}}}, \frac{\tau_w}{m_1^{\frac{1}{4}}}]$, $\sigma_v \in (0, \frac{1}{m_2^{\frac{1}{2}}})]$. The perturbation matrices satisfies $\|\boldsymbol{W}'\|_{2,4} \le \tau_w$, $\|\boldsymbol{V}'\|_F \le \tau_v$, $\|\boldsymbol{W}''\|_{2,4} \le \tau_w$, $\|\boldsymbol{V}''\|_F \le \tau_v$ and random diagonal matrix $\boldsymbol{\Sigma}$ has each diagonal entry i.i.d. drawn from $\{\pm 1\}$. Then with high probability, we have*
(1) Sparse sign change

$$\|\boldsymbol{D}'_{n,\boldsymbol{w}}\|_0 \le \tilde{O}(\tau_w^{\frac{4}{5}} m_1^{\frac{6}{5}})$$

$$\|\boldsymbol{D}'_{n,\boldsymbol{v}}\|_0 \le \tilde{O}(m_2^{\frac{3}{2}} \sigma_v (\|\boldsymbol{A}\|_\infty + \|\boldsymbol{A}\|_\infty \tau_w m_1^{\frac{1}{4}}) + m_2 \|\boldsymbol{A}\|_\infty^{\frac{2}{3}} (\|\boldsymbol{A}\|_\infty \tau_v + \|\boldsymbol{A}\|_\infty \tau_w m_1^{\frac{1}{4}}(1+\tau_v))^{\frac{2}{3}})$$

(2) Cross term vanish

$$\begin{aligned} & g_r(\boldsymbol{q}, \boldsymbol{A}, \boldsymbol{X}, \boldsymbol{W}^{(0)} + \boldsymbol{W}^\rho + \boldsymbol{W}' + \eta\boldsymbol{W}''\boldsymbol{\Sigma}, \boldsymbol{V}^{(0)} + \boldsymbol{V}^\rho + \boldsymbol{V}' + \eta\boldsymbol{\Sigma}\boldsymbol{V}'') \\ =& g_r(\boldsymbol{q}, \boldsymbol{A}, \boldsymbol{X}, \boldsymbol{W}^{(0)} + \boldsymbol{W}^\rho + \boldsymbol{W}', \boldsymbol{V}^{(0)} + \boldsymbol{V}^\rho + \boldsymbol{V}') + g_r^{(b,b)}(\boldsymbol{q}, \boldsymbol{A}, \boldsymbol{X}, \eta\boldsymbol{W}''\boldsymbol{\Sigma}, \eta\boldsymbol{\Sigma}\boldsymbol{V}'') + g'_r \end{aligned} \tag{96}$$

for every $r \in [K]$, where $\mathbb{E}_{\boldsymbol{\Sigma}}[g'_r] = 0$ and $|g'_r| \le \eta \|\boldsymbol{q}^\top \boldsymbol{A}\|_1 \|\boldsymbol{A}\|_\infty^2 \tau_v$.

Proof:
(1) We first consider the sign changes by $\boldsymbol{W}^\rho$. Since $\tilde{\boldsymbol{a}}_n \boldsymbol{X} \boldsymbol{w}_i^{(0)} + B^{(0)}_{1(n,i)} = \tilde{\boldsymbol{a}}_n \tilde{\boldsymbol{X}} \tilde{\boldsymbol{w}}_i^{(0)}$ where $\tilde{\boldsymbol{w}}_i^{(0)} = (\boldsymbol{w}_i^{(0)}, B^{(0)}_{1(n,i)}) \in \mathbb{R}^{d+1}$ and $\tilde{\boldsymbol{X}} = (\boldsymbol{X}, \boldsymbol{1}) \in \mathbb{R}^{N \times (d+1)}$, we can ignore the bias term for simplicity. We have

$$\tilde{\boldsymbol{a}}_n \tilde{\boldsymbol{X}} \tilde{\boldsymbol{w}}_i^{(0)} \sim \mathcal{N}(0, \frac{\|\tilde{\boldsymbol{a}}_n \tilde{\boldsymbol{X}}\|^2}{m_1})$$

$$\tilde{\boldsymbol{a}}_n \tilde{\boldsymbol{X}} \tilde{\boldsymbol{w}}_i^{\rho} \sim \mathcal{N}(0, \|\tilde{\boldsymbol{a}}_n \tilde{\boldsymbol{X}}\|^2 \sigma_w^2)$$

Therefore,

$$\frac{\tilde{\boldsymbol{a}}_n \tilde{\boldsymbol{X}} \tilde{\boldsymbol{w}}_i^{(0)}}{\tilde{\boldsymbol{a}}_n \tilde{\boldsymbol{X}} \tilde{\boldsymbol{w}}_i^{\rho}} \sim p(z) = \frac{1}{\pi(\sigma_w \sqrt{m_1} z^2 + \frac{1}{\sigma_w \sqrt{m_1}})}$$

$$\begin{aligned} \Pr[|\tilde{\boldsymbol{a}}_n \tilde{\boldsymbol{X}} \tilde{\boldsymbol{w}}_i^{(0)}| \le |\tilde{\boldsymbol{a}}_n \tilde{\boldsymbol{X}} \tilde{\boldsymbol{w}}_i^{\rho}|] &= \Pr[|z| \le 1] \\ &= \int_{-1}^{1} \frac{1}{\pi(\sigma_w \sqrt{m_1} z^2 + \frac{1}{\sigma_w \sqrt{m_1}})} dz \\ &= \int_{-(\sigma_w^2 m_1)^{\frac{1}{2}}}^{(\sigma_w^2 m_1)^{\frac{1}{2}}} \frac{1}{\pi(t^2+1)} dt \\ &= \frac{2}{\pi} \arctan \sigma_w \sqrt{m_1} \\ &\le \tilde{O}(\sigma_w \sqrt{m_1}) \end{aligned} \tag{97}$$

Then, we have

$$\|\boldsymbol{D}''_{n,\boldsymbol{w}}\|_0 \leq \tilde{O}(\sigma_w m_1^{\frac{3}{2}})$$

$$\|\tilde{\boldsymbol{a}}_n\tilde{\boldsymbol{X}}\tilde{\boldsymbol{W}}^{(0)}\boldsymbol{D}''_{n,\boldsymbol{w}}\|_2 \leq \tilde{O}(\|\tilde{\boldsymbol{a}}_n\tilde{\boldsymbol{X}}\|\sigma_w^{\frac{3}{2}} m_1^{\frac{3}{4}})$$

We then consider the sign changes by $\boldsymbol{W}'$. Let $s = \|\boldsymbol{D}'_{n,\boldsymbol{w}} - \boldsymbol{D}''_{n,\boldsymbol{w}}\|_0$ be the total number of sign changes in the first layer caused by adding $\boldsymbol{W}'$. Note that the total number of coordinated i such that $|\tilde{\boldsymbol{a}}_n\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{W}}^{(0)} + \tilde{\boldsymbol{W}}^{\rho})_i| \leq s'' = \frac{2\tau_w}{s^{\frac{1}{4}}}$ is at most $s'' m_1^{\frac{3}{2}}$ with high probability. Since $\|\boldsymbol{W}'\|_{2,4} \leq \tau_w$, we must have

$$s \leq \tilde{O}(s'' m^{\frac{3}{2}}) = \tilde{O}(\frac{\tau_w}{s^{\frac{1}{4}}} m_1^{\frac{3}{2}})$$

$$\|\boldsymbol{D}'_{n,\boldsymbol{w}} - \boldsymbol{D}''_{n,\boldsymbol{w}}\|_0 = s \leq \tilde{O}(\tau_w^{\frac{4}{5}} m_1^{\frac{6}{5}})$$

$$\|\tilde{\boldsymbol{a}}_n\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{W}}^{(0)} + \tilde{\boldsymbol{W}}^{\rho})(\boldsymbol{D}'_{n,\boldsymbol{w}} - \boldsymbol{D}''_{n,\boldsymbol{w}})\|_2 \leq \tilde{O}(s^{\frac{1}{4}}\tau_w) \leq \tilde{O}(\tau_w^{\frac{6}{5}} m_1^{\frac{3}{10}})$$

To sum up, we have

$$\|\boldsymbol{D}'_{n,\boldsymbol{w}}\|_0 \leq \tilde{O}(\sigma_w m_1^{\frac{3}{2}} + \tau_w^{\frac{4}{5}} m_1^{\frac{6}{5}}) \leq \tilde{O}(\tau_w^{\frac{4}{5}} m_1^{\frac{6}{5}})$$

Denote $\boldsymbol{z}_{n,0} = \tilde{\boldsymbol{a}}_n\tilde{\boldsymbol{X}}\tilde{\boldsymbol{W}}^{(0)}\boldsymbol{D}_{n,\boldsymbol{w}}$ and $\boldsymbol{z}_{n,2} = \tilde{\boldsymbol{a}}_n\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{W}}^{(0)} + \tilde{\boldsymbol{W}}^{\rho} + \boldsymbol{W}')(\boldsymbol{D}_{n,\boldsymbol{w}} + \boldsymbol{D}'_{n,\boldsymbol{w}}) - \tilde{\boldsymbol{a}}_n\tilde{\boldsymbol{X}}\tilde{\boldsymbol{W}}^{(0)}\boldsymbol{D}_{n,\boldsymbol{w}}$. With high probability, we know

$$\begin{aligned}\|\boldsymbol{z}_{n,2}\| \leq& \|\tilde{\boldsymbol{a}}_n\tilde{\boldsymbol{X}}\boldsymbol{W}'\| + \|\tilde{\boldsymbol{a}}_n\tilde{\boldsymbol{X}}\tilde{\boldsymbol{W}}^{\rho}\| \\ \leq& \tilde{O}(m_1^{\frac{1}{4}}\tau_w\|\boldsymbol{A}\|_\infty + \|\boldsymbol{A}\|_\infty\sigma_w m_1^{\frac{1}{2}}) \\ \leq& \tilde{O}(\|\boldsymbol{A}\|_\infty\tau_w m_1^{\frac{1}{4}})\end{aligned} \tag{98}$$

Denote $\boldsymbol{Z}_0 = (\boldsymbol{z}_{1,0}^\top, \cdots, \boldsymbol{z}_{N,0}^\top)^\top \in \mathbb{R}^{N\times m_1}$, $\boldsymbol{Z}_2 = (\boldsymbol{z}_{1,2}^\top, \cdots, \boldsymbol{z}_{N,2}^\top)^\top \in \mathbb{R}^{N\times m_1}$. The sign change in the second layer is from $\tilde{\boldsymbol{a}}_n\boldsymbol{Z}_0\boldsymbol{V}^{(0)}$ to $\tilde{\boldsymbol{a}}_n(\boldsymbol{Z}_0 + \boldsymbol{Z}_2)(\boldsymbol{V}^{(0)} + \boldsymbol{V}^{\rho} + \boldsymbol{V}')$. We have

$$\|\tilde{\boldsymbol{a}}_n(\boldsymbol{Z}_0 + \boldsymbol{Z}_2)\boldsymbol{V}^{\rho}\|_\infty \leq \tilde{O}(\sigma_v\|\boldsymbol{A}\|_\infty(\|\boldsymbol{z}_{1,0}\| + \|\boldsymbol{z}_{1,2}\|))$$

$$\|\tilde{\boldsymbol{a}}_n(\boldsymbol{Z}_0 + \boldsymbol{Z}_2)\boldsymbol{V}' + \tilde{\boldsymbol{a}}_n\boldsymbol{Z}_2\boldsymbol{V}^{(0)}\|_2 \leq \tilde{O}(\|\boldsymbol{A}\|_\infty((\|\boldsymbol{z}_{1,0}\| + \|\boldsymbol{z}_{1,2}\|)\tau_v + \|\boldsymbol{z}_{1,2}\|))$$

Combining $\|\boldsymbol{z}_{1,0}\| \leq \tilde{O}(\|\boldsymbol{A}\|_\infty)$, by Claim C.8 in (Allen-Zhu et al., 2019) we have

$$\|\boldsymbol{D}'_{n,\boldsymbol{v}}\|_0 \leq \tilde{O}(m_2^{\frac{3}{2}}\sigma_v(\|\boldsymbol{A}\|_\infty^2 + \|\boldsymbol{A}\|_\infty^2\tau_w m_1^{\frac{1}{4}}) + m_2\|\boldsymbol{A}\|_\infty^{\frac{2}{3}}(\|\boldsymbol{A}\|_\infty\tau_v + \|\boldsymbol{A}\|_\infty\tau_w m_1^{\frac{1}{4}}(1+\tau_v))^{\frac{2}{3}})$$

(2) Diagonal Cross terms.
Denote $\boldsymbol{D}_{\boldsymbol{w}} = (\text{diag}(\boldsymbol{D}_{1,\boldsymbol{w}})^\top, \cdots, \text{diag}(\boldsymbol{D}_{N,m_1})^\top)^\top \in \mathbb{R}^{N\times m_1}$ and define $\boldsymbol{D}'_{\boldsymbol{w}}$, $\boldsymbol{D}''_{\boldsymbol{w}}$, $\boldsymbol{D}_{\boldsymbol{v}}$, $\boldsymbol{D}'_{\boldsymbol{v}}$, $\boldsymbol{D}''_{\boldsymbol{v}}$ accordingly.
Recall

$$\begin{aligned}g_r(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X},\boldsymbol{W},\boldsymbol{V}) &= \boldsymbol{q}^\top\boldsymbol{A}((\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{W} + \boldsymbol{B}_1)\odot(\boldsymbol{D}_{\boldsymbol{w}} + \boldsymbol{D}'_{\boldsymbol{w}})\boldsymbol{V} + \boldsymbol{B}_2)\odot(\boldsymbol{D}_{\boldsymbol{v}} + \boldsymbol{D}'_{\boldsymbol{v}}))\boldsymbol{c}_r \\ g_r^{(b)}(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X},\boldsymbol{W},\boldsymbol{V}) &= \boldsymbol{q}^\top\boldsymbol{A}((\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{W} + \boldsymbol{B}_1)\odot(\boldsymbol{D}_{\boldsymbol{w}} + \boldsymbol{D}'_{\boldsymbol{w}})\boldsymbol{V})\odot(\boldsymbol{D}_{\boldsymbol{v}} + \boldsymbol{D}'_{\boldsymbol{v}}))\boldsymbol{c}_r \\ g_r^{(b,b)}(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X},\boldsymbol{W},\boldsymbol{V}) &= \boldsymbol{q}^\top\boldsymbol{A}((\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{W})\odot(\boldsymbol{D}_{\boldsymbol{w}} + \boldsymbol{D}'_{\boldsymbol{w}})\boldsymbol{V})\odot(\boldsymbol{D}_{\boldsymbol{v}} + \boldsymbol{D}'_{\boldsymbol{v}}))\boldsymbol{c}_r\end{aligned}$$

Then

$$\begin{aligned}&g_r(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X},\boldsymbol{W}^{(0)} + \boldsymbol{W}^{\rho} + \boldsymbol{W}' + \eta\boldsymbol{W}''\boldsymbol{\Sigma},\boldsymbol{V}^{(0)} + \boldsymbol{V}^{\rho} + \boldsymbol{V}' + \eta\boldsymbol{\Sigma}\boldsymbol{V}'') \\ =&g_r(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X},\boldsymbol{W}^{(0)} + \boldsymbol{W}^{\rho} + \boldsymbol{W}',\boldsymbol{V}^{(0)} + \boldsymbol{V}^{\rho} + \boldsymbol{V}') + g_r^{(b,b)}(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X},\eta\boldsymbol{W}''\boldsymbol{\Sigma},\eta\boldsymbol{\Sigma}\boldsymbol{V}'') \\ &+ g_r^{(b)}(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X},\boldsymbol{W}^{(0)} + \boldsymbol{W}^{\rho} + \boldsymbol{W}',\eta\boldsymbol{\Sigma}\boldsymbol{V}'') + g_r^{(b,b)}(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X},\eta\boldsymbol{W}''\boldsymbol{\Sigma},\boldsymbol{V}^{(0)} + \boldsymbol{V}^{\rho} + \boldsymbol{V}')\end{aligned} \tag{99}$$

where the last two terms are the error terms. We know that

$$\|\boldsymbol{W}^{(0)}\| \leq \max_{\|\boldsymbol{a}\|=1}\|\boldsymbol{a}^\top\boldsymbol{W}^{(0)}\| \leq \max_{\|\boldsymbol{a}\|=1}\sqrt{\sum_{i=1}^{m_1}(\boldsymbol{a}^\top\boldsymbol{w}_i^{(0)})^2} \leq \max_{\|\boldsymbol{a}\|=1}\sqrt{\sum_{i=1}^{m_1}(\frac{1}{\sqrt{m_1}})^2} = 1$$

Therefore,

$$\begin{aligned}
&|g_r^{(b)}(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X},\boldsymbol{W}^{(0)}+\boldsymbol{W}^{\rho}+\boldsymbol{W}',\eta\boldsymbol{\Sigma}\boldsymbol{V}'')|\\
=&\eta\sum_{n=1}^{N}\sum_{i=1}^{m_2}\boldsymbol{q}^\top\boldsymbol{a}_n c_{i,r}D_{n,\boldsymbol{v}_i}\sum_{k=1}^{N}a_{n,k}\sum_{l=1}^{m_1}(\boldsymbol{\Sigma}\boldsymbol{V})''_{l,i}D_{k,\boldsymbol{w}_l}(\boldsymbol{a}_k\boldsymbol{X}(\boldsymbol{W}^{(0)}+\boldsymbol{W}^{\rho}+\boldsymbol{W}')_l+\boldsymbol{B}_{1(k,l)})\\
=&\eta\sum_{n=1}^{N}\boldsymbol{q}^\top\boldsymbol{a}_n\sum_{k=1}^{N}a_{n,k}((\boldsymbol{a}_k\boldsymbol{X}(\boldsymbol{W}^{(0)}+\boldsymbol{W}^{\rho}+\boldsymbol{W}')+\boldsymbol{B}_{1k})\odot\boldsymbol{D}_{k,\boldsymbol{w}})\boldsymbol{\Sigma}\boldsymbol{V}''\odot\boldsymbol{D}_{n,\boldsymbol{v}}\boldsymbol{c}_r\\
\leq&\eta\|\boldsymbol{q}^\top\boldsymbol{A}\|_1\|\boldsymbol{A}\|_\infty(\|\boldsymbol{A}\|_\infty(m^{-\frac{1}{2}}+\sigma_w+\tau_w)+m^{-\frac{1}{2}})\tau_v m_2^{\frac{1}{2}}\\
\leq&\tilde{O}(\eta\|\boldsymbol{q}^\top\boldsymbol{A}\|_1\|\boldsymbol{A}\|_\infty^2\tau_v m_2^{\frac{1}{2}}m_1^{-\frac{1}{2}}),
\end{aligned}\tag{100}$$

where the last step is by the value selection of σ_w, τ_w and τ_v.

$$\begin{aligned}
&|g_r^{(b,b)}(\boldsymbol{q},\boldsymbol{A},\eta\boldsymbol{W}''\boldsymbol{\Sigma},\boldsymbol{X},\boldsymbol{V}^{(0)}+\boldsymbol{V}^{\rho}+\boldsymbol{V}')|\\
=&|\eta\sum_{n=1}^{N}\boldsymbol{q}^\top\boldsymbol{a}_n\sum_{k=1}^{N}a_{n,k}(\boldsymbol{a}_k\boldsymbol{X}\boldsymbol{W}''\boldsymbol{\Sigma}\odot(\boldsymbol{D}_{k,\boldsymbol{w}}+\boldsymbol{D}_{k,\boldsymbol{w}}{}'))(\boldsymbol{V}^{(0)}+\boldsymbol{V}^{\rho}+\boldsymbol{V}'')\odot(\boldsymbol{D}_{n,\boldsymbol{v}}+\boldsymbol{D}_{n,\boldsymbol{v}}{}')\boldsymbol{c}_r|\\
\leq&|\eta\sum_{n=1}^{N}\boldsymbol{q}^\top\boldsymbol{a}_n\sum_{k=1}^{N}a_{n,k}(\boldsymbol{a}_k\boldsymbol{X}\boldsymbol{W}''\boldsymbol{\Sigma}\odot(\boldsymbol{D}_{k,\boldsymbol{w}}+\boldsymbol{D}_{k,\boldsymbol{w}}{}'))\boldsymbol{V}'\odot(\boldsymbol{D}_{n,\boldsymbol{v}}+\boldsymbol{D}_{n,\boldsymbol{v}}{}')\boldsymbol{c}_r|\\
&+|\eta\sum_{n=1}^{N}\boldsymbol{q}^\top\boldsymbol{a}_n\sum_{k=1}^{N}a_{n,k}(\boldsymbol{a}_k\boldsymbol{X}\boldsymbol{W}''\boldsymbol{\Sigma}\odot(\boldsymbol{D}_{k,\boldsymbol{w}}+\boldsymbol{D}_{k,\boldsymbol{w}}{}'))(\boldsymbol{V}^{(0)}+\boldsymbol{V}^{\rho})\odot\boldsymbol{D}_{n,\boldsymbol{v}}\boldsymbol{c}_r|\\
&+|\eta\sum_{n=1}^{N}\boldsymbol{q}^\top\boldsymbol{a}_n\sum_{k=1}^{N}a_{n,k}(\boldsymbol{a}_k\boldsymbol{X}\boldsymbol{W}''\boldsymbol{\Sigma}\odot(\boldsymbol{D}_{k,\boldsymbol{w}}+\boldsymbol{D}_{k,\boldsymbol{w}}{}'))(\boldsymbol{V}^{(0)}+\boldsymbol{V}^{\rho})\odot\boldsymbol{D}_{n,\boldsymbol{v}}{}'\boldsymbol{c}_r|\\
\leq&|\eta\|\boldsymbol{q}^\top\boldsymbol{A}\|_1\|\boldsymbol{A}\|_\infty^2\tau_w\tau_v m_2^{\frac{1}{2}}|+2|\eta\|\boldsymbol{q}^\top\boldsymbol{A}\|_1\|\boldsymbol{A}\|_\infty^2\tau_w m_1^{\frac{1}{2}}|\\
\leq&\tilde{O}(|\eta\|\boldsymbol{q}^\top\boldsymbol{A}\|_1\|\boldsymbol{A}\|_\infty^2\tau_w m_1^{\frac{1}{2}}|)
\end{aligned}\tag{101}$$

Lemma B.11. *Denote*

$$\begin{aligned}
P_{\rho,\eta}=&F_{\boldsymbol{A}}(\boldsymbol{q},\boldsymbol{X},\boldsymbol{W}+\boldsymbol{W}^{\rho}+\eta\boldsymbol{W}''\boldsymbol{\Sigma},\boldsymbol{V}+\boldsymbol{V}^{\rho}+\eta\boldsymbol{\Sigma}\boldsymbol{V}'')\\
=&\boldsymbol{q}^\top\boldsymbol{A}(\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}(\boldsymbol{W}+\boldsymbol{W}^{\rho}+\eta\boldsymbol{W}''\boldsymbol{\Sigma})+\boldsymbol{B}_1)\odot\boldsymbol{D}_{\boldsymbol{w},\rho,\eta}(\boldsymbol{V}+\boldsymbol{V}^{\rho}+\eta\boldsymbol{\Sigma}\boldsymbol{V}'')\odot\boldsymbol{D}_{\boldsymbol{v},\rho,\eta})\boldsymbol{c}_r
\end{aligned}\tag{102}$$

$$\begin{aligned}
P'_{\rho,\eta}=&G(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X},\boldsymbol{W}+\boldsymbol{W}^{\rho}+\eta\boldsymbol{W}''\boldsymbol{\Sigma},\boldsymbol{V}+\boldsymbol{V}^{\rho}+\eta\boldsymbol{\Sigma}\boldsymbol{V}'')\\
=&\boldsymbol{q}^\top\boldsymbol{A}(\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}(\boldsymbol{W}+\boldsymbol{W}^{\rho}+\eta\boldsymbol{W}''\boldsymbol{\Sigma})+\boldsymbol{B}_1)\odot\boldsymbol{D}_{\boldsymbol{w},\rho}(\boldsymbol{V}+\boldsymbol{V}^{\rho}+\eta\boldsymbol{\Sigma}\boldsymbol{V}'')\odot\boldsymbol{D}_{\boldsymbol{v},\rho})\boldsymbol{c}_r
\end{aligned}\tag{103}$$

There exists $\eta_0=\frac{1}{poly(m_1,m_2)}$ such that for every $\eta\leq\eta_0$, for every $\boldsymbol{W}''$, $\boldsymbol{V}''$ that satisfies $\|\boldsymbol{W}''\|_{2,\infty}\leq\tau_{\boldsymbol{w},\infty}$, $\|\boldsymbol{V}''\|_{2,\infty}\leq\tau_{\boldsymbol{v},\infty}$, we have

$$\mathbb{E}_{\boldsymbol{W}^{\rho},\boldsymbol{V}^{\rho}}[\frac{|P_{\rho,\eta}-P'_{\rho,\eta}|}{\eta^2}]=\tilde{O}(\boldsymbol{q}^\top\boldsymbol{A}\boldsymbol{1}\|\boldsymbol{A}\|^4(\frac{\tau_{\boldsymbol{w},\infty}^2}{\sigma_w}m_1+\frac{(\tau_{\boldsymbol{w},\infty}^2+\tau_{\boldsymbol{v},\infty}^2m_1^{-1})}{\sigma_v}m_2))+O_p(\eta),$$

where O_p hides polynomial factor of m_1 and m_2.

Proof:

$$\begin{aligned}
P_{\rho,\eta}-P'_{\rho,\eta}=&\boldsymbol{q}^\top\boldsymbol{A}(\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}(\boldsymbol{W}+\boldsymbol{W}^{\rho}+\eta\boldsymbol{W}''\boldsymbol{\Sigma})+\boldsymbol{B}_1)\odot(\boldsymbol{D}_{\boldsymbol{w},\rho,\eta}-\boldsymbol{D}_{\boldsymbol{w},\rho})(\boldsymbol{V}+\boldsymbol{V}^{\rho}+\eta\boldsymbol{\Sigma}\boldsymbol{V}'')\odot\boldsymbol{D}_{\boldsymbol{v},\rho})\boldsymbol{c}_r\\
&+\boldsymbol{q}^\top\boldsymbol{A}(\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}(\boldsymbol{W}+\boldsymbol{W}^{\rho}+\eta\boldsymbol{W}''\boldsymbol{\Sigma})+\boldsymbol{B}_1)\odot\boldsymbol{D}_{\boldsymbol{w},\rho,\eta}(\boldsymbol{V}+\boldsymbol{V}^{\rho}+\eta\boldsymbol{\Sigma}\boldsymbol{V}'')\odot(\boldsymbol{D}_{\boldsymbol{v},\rho,\eta}-\boldsymbol{D}_{\boldsymbol{v},\rho}))\boldsymbol{c}_r
\end{aligned}\tag{104}$$

We write

$$\boldsymbol{Z}=\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}(\boldsymbol{W}+\boldsymbol{W}^{\rho}+\eta\boldsymbol{W}''\boldsymbol{\Sigma})+\boldsymbol{B}_1)\odot\boldsymbol{D}_{\boldsymbol{w},\rho}$$

$$\boldsymbol{Z}+\boldsymbol{Z}'=\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}(\boldsymbol{W}+\boldsymbol{W}^{\rho}+\eta\boldsymbol{W}''\boldsymbol{\Sigma})+\boldsymbol{B}_1)\odot\boldsymbol{D}_{\boldsymbol{w},\rho,\eta}$$

Since for all $n\in[N]$, $\|\eta(\boldsymbol{A}\boldsymbol{A}\boldsymbol{X}\boldsymbol{W}''\boldsymbol{\Sigma})_n\|_\infty\le\eta\|\boldsymbol{A}\|_\infty\tau_{\boldsymbol{w},\infty}$, we have

$$\|\boldsymbol{Z}'_n\|_\infty\le\eta\|\boldsymbol{A}\|_\infty\tau_{\boldsymbol{w},\infty}$$

$$\Pr_{\boldsymbol{W}^\rho}[Z'_{n,i}\neq 0]\le\tilde{O}(\frac{\eta\|\boldsymbol{A}\|_\infty\tau_{\boldsymbol{w},\infty}}{\sigma_w}),\ i\in[m_1]$$

Then we have

$$\Pr[\|\boldsymbol{Z}'_n\|_0\ge 2]\le O_p(\eta^2)$$

Then we only need to consider the case $\|\boldsymbol{Z}'_n\|_0=1$. Let $Z'_{n,n_i}\neq 0$. Then the first term in (104), $\boldsymbol{q}^\top\boldsymbol{A}(\boldsymbol{Z}'(\boldsymbol{V}+\boldsymbol{V}^\rho+\eta\boldsymbol{\Sigma}\boldsymbol{V}'')\odot\boldsymbol{D}_{\boldsymbol{v},\rho})\boldsymbol{c}_r$ should be dealt with separately.
The term $\boldsymbol{q}^\top\boldsymbol{A}(\boldsymbol{Z}'\eta\boldsymbol{\Sigma}\boldsymbol{V}''\odot\boldsymbol{D}_{\boldsymbol{v},\rho})\boldsymbol{c}_r)$ contributes to $O_p(\eta^3)$ to the whole term.
Then we have

$$\|\boldsymbol{q}^\top\boldsymbol{A}(\boldsymbol{Z}'\eta(\boldsymbol{V}+\boldsymbol{V}^\rho)\odot\boldsymbol{D}_{\boldsymbol{v},\rho})\boldsymbol{c}_r\|\le\tilde{O}(\eta\|\|\boldsymbol{q}^\top\boldsymbol{A}\|_1\|\boldsymbol{A}\|_\infty\tau_{\boldsymbol{w},\infty})$$

We also have that

$$\tilde{O}((\frac{\eta\|\boldsymbol{A}\|_\infty\tau_{\boldsymbol{w},\infty}}{\sigma_{\boldsymbol{w}}}m_1)^N)\le\tilde{O}(\frac{\eta\|\boldsymbol{A}\|_\infty\tau_{\boldsymbol{w},\infty}}{\sigma_{\boldsymbol{w}}}m_1)\le 1$$

Therefore, the contribution to the first term is $\tilde{O}(\eta^2\|\boldsymbol{q}^\top\boldsymbol{A}\|_1\|\boldsymbol{A}\|_\infty^2\frac{\tau^2_{\boldsymbol{w},\infty}}{\sigma_w}m_1)+O_p(\eta^3)$.
Denote

$$\begin{aligned}\boldsymbol{\delta}=&\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}(\boldsymbol{W}+\boldsymbol{W}^\rho+\eta\boldsymbol{W}''\boldsymbol{\Sigma})+\boldsymbol{B}_1)\odot\boldsymbol{D}_{\boldsymbol{w},\rho,\eta}(\boldsymbol{V}+\boldsymbol{V}^\rho+\eta\boldsymbol{\Sigma}\boldsymbol{V}'')\\&-\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}(\boldsymbol{W}+\boldsymbol{W}^\rho)+\boldsymbol{B}_1)\odot\boldsymbol{D}_{\boldsymbol{w},\rho}(\boldsymbol{V}+\boldsymbol{V}^\rho)\end{aligned}\tag{105}$$

$\boldsymbol{\delta}\in\mathbb{R}^{m_2}$ has the following terms:
1. $\boldsymbol{Z}'(\boldsymbol{V}+\boldsymbol{V}^\rho+\eta\boldsymbol{\Sigma}\boldsymbol{V}'')$. We have its n-th row norm bounded by $O_p(\eta)$.
2. $\boldsymbol{Z}\eta\boldsymbol{\Sigma}\boldsymbol{V}''$. We have its n-th row infinity norm bounded by $\tilde{O}(\|\boldsymbol{A}\|_\infty\eta\tau_{v,\infty}m_1^{-\frac{1}{2}})$.
3. $\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}\eta\boldsymbol{W}''\boldsymbol{\Sigma}\odot\boldsymbol{D}_{\boldsymbol{w},\rho})(\boldsymbol{V}+\boldsymbol{V}^\rho)$, of which the n-th row infinity is bounded by $\tilde{O}(\|\boldsymbol{A}\|_\infty\eta\tau_{\boldsymbol{w},\infty})$.
4. $\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}\eta^2\boldsymbol{W}''\boldsymbol{\Sigma}\odot\boldsymbol{D}_{\boldsymbol{w},\rho,\eta}\boldsymbol{\Sigma}\boldsymbol{V}'')$. Bounded by $O_p(\eta^2)$.
Therefore,

$$\|\boldsymbol{\delta}_n\|_\infty\le\tilde{O}(\|\boldsymbol{A}\|_\infty\eta(\tau_{\boldsymbol{v},\infty}m_1^{-\frac{1}{2}}+\tau_{\boldsymbol{w},\infty}))+O_p(\eta^2)$$

Similarly, we can derive that the contribution to the second term is $\tilde{O}(\eta^2\|\boldsymbol{q}^\top\boldsymbol{A}\|_1\|\boldsymbol{A}\|_\infty^2\frac{(\tau^2_{\boldsymbol{w},\infty}+\tau^2_{\boldsymbol{v},\infty}m_1^{-1})}{\sigma_v}m_2)+O_p(\eta^3)$.

Lemma B.12. *Let $\boldsymbol{F}^*_{\boldsymbol{A}}=(f_1^*,\cdots,f_K^*)$. Perturbation matrices $\boldsymbol{W}'$, $\boldsymbol{V}'$ satisfy*

$$\|\boldsymbol{W}'\|_{2,4}\le\tau_w,\ \ \|\boldsymbol{V}'\|_F\le\tau_v$$

There exists $\widehat{\boldsymbol{W}}$ and $\widehat{\boldsymbol{V}}$ such that

$$\|\widehat{\boldsymbol{W}}\|_{2,\infty}\le\frac{C_0}{m_1},\ \ \|\widehat{\boldsymbol{V}}\|_{2,\infty}\le\frac{K\sqrt{m_1}}{m_2}$$

$$\mathbb{E}[\sum_{r=1}^K|f_r^*(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X},\widehat{\boldsymbol{W}},\widehat{\boldsymbol{V}})-g_r^{(b,b)}(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X},\widehat{\boldsymbol{W}},\widehat{\boldsymbol{V}})|]\le\epsilon$$

$$\mathbb{E}[G^{(b,b)}(\boldsymbol{q},\boldsymbol{A},\boldsymbol{X},\widehat{\boldsymbol{W}},\widehat{\boldsymbol{V}})]\le OPT+\epsilon$$

Proof:
By Lemma B.10, we have

$$\|\boldsymbol{D}'_{n,\boldsymbol{w}}\|_0\le\tilde{O}(\tau_w^{\frac{4}{5}}m_1^{\frac{6}{5}})\ll\tilde{O}(m_1)$$

$$\|\boldsymbol{D}'_{n,\boldsymbol{v}}\|_0\le\tilde{O}(m_2^{\frac{3}{2}}\sigma_v(\|\boldsymbol{A}\|_\infty^2+\|\boldsymbol{A}\|_\infty^2\tau_w m_1^{\frac{1}{4}})+m_2\|\boldsymbol{A}\|_\infty^{\frac{2}{3}}(\|\boldsymbol{A}\|_\infty\tau_v+\|\boldsymbol{A}\|_\infty\tau_w m_1^{\frac{1}{4}}(1+\tau_v))^{\frac{2}{3}})\le\tilde{O}(m_2\|\boldsymbol{A}\|_\infty^2(\epsilon/C_0)^{\Theta(1)})$$

Applying Lemma B.9, we know

$$\begin{aligned}
&\boldsymbol{q}^\top \boldsymbol{A}((\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}\widehat{\boldsymbol{W}})\odot(\boldsymbol{D}'_{\boldsymbol{w}})\widehat{\boldsymbol{V}})\odot(\boldsymbol{D}_{\boldsymbol{v}}))\boldsymbol{c}_r\\
=&\sum_{n=1}^{N}\boldsymbol{q}^\top\boldsymbol{a}_n\sum_{k=1}^{N}a_{n,k}((\boldsymbol{a}_k\boldsymbol{X}\widehat{\boldsymbol{W}})\odot\boldsymbol{D}_{k,\boldsymbol{w}}{}')\widehat{\boldsymbol{V}}\odot\boldsymbol{D}_{n,\boldsymbol{v}}\boldsymbol{c}_r\\
\leq&\|\boldsymbol{q}^\top\boldsymbol{A}\|_1\|\boldsymbol{A}\|_\infty^2 m_1^{\frac{3}{10}}\frac{C_0}{m_1}\frac{K\sqrt{m_1}}{m_2}\cdot m_2\\
\leq&\epsilon
\end{aligned}\tag{106}$$

$$\begin{aligned}
&\boldsymbol{q}^\top \boldsymbol{A}((\boldsymbol{A}(\boldsymbol{A}\boldsymbol{X}\widehat{\boldsymbol{W}})\odot(\boldsymbol{D}_{\boldsymbol{w}})\widehat{\boldsymbol{V}})\odot(\boldsymbol{D}'_{\boldsymbol{v}}))\boldsymbol{c}_r\\
=&\sum_{n=1}^{N}\boldsymbol{q}^\top\boldsymbol{a}_n\sum_{k=1}^{N}a_{n,k}((\boldsymbol{a}_k\boldsymbol{X}\widehat{\boldsymbol{W}})\odot\boldsymbol{D}_{k,\boldsymbol{w}})\widehat{\boldsymbol{V}}\odot\boldsymbol{D}'_{n,\boldsymbol{v}}\boldsymbol{c}_r\\
\leq&\|\boldsymbol{q}^\top\boldsymbol{A}\|_1\|\boldsymbol{A}\|_\infty^2 m_1^{\frac{1}{2}}\frac{C_0}{m_1}\frac{K\sqrt{m_1}}{m_2}\cdot m_2\cdot\|\boldsymbol{A}\|_\infty^2(\frac{\epsilon}{C_0})^{\Theta(1)}\\
\leq&\epsilon
\end{aligned}\tag{107}$$

Then, the conclusion can be derived.

B.1.4. OPTIMIZATION

This section states the optimization process and convergence performance of the algorithm. Lemma B.13 shows that during the optimization, either there exists an updating direction that decreases the objective, or weight decay decreases the objective. Lemma B.14 provides the convergence result of the algorithm.
Define

$$\begin{aligned}
&L'(\boldsymbol{A}^*,\boldsymbol{A}^*,\boldsymbol{A}^*,\lambda_t,\boldsymbol{W}_t,\boldsymbol{V}_t)\\
=&\frac{1}{\Omega^t}\sum_{i=1}^{|\Omega^t|}\mathbb{E}_{\boldsymbol{W}^\rho,\boldsymbol{V}^\rho,\boldsymbol{\Sigma}'}[L(\lambda_t F_{\boldsymbol{A}^*}(\boldsymbol{e}_g,\boldsymbol{X};\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho+\boldsymbol{W}_t\boldsymbol{\Sigma}',\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho+\boldsymbol{\Sigma}'\boldsymbol{V}_t),y_i)]+R(\sqrt{\lambda_t}\boldsymbol{W}_t,\sqrt{\lambda_t}\boldsymbol{V}_t)
\end{aligned}\tag{108}$$

where

$$R(\sqrt{\lambda}\boldsymbol{W}_t,\sqrt{\lambda}\boldsymbol{V}_t)=\lambda_v\|\sqrt{\lambda}\boldsymbol{V}_t\|_F^2+\lambda_w\|\sqrt{\lambda}\boldsymbol{W}_t\|_{2,4}^2$$

Lemma B.13. *For every $\epsilon_0\in(0,1)$ and $\epsilon\in(0,\frac{\epsilon_0}{K\|\boldsymbol{A}\|_\infty p_1 p_2^2\mathcal{C}_s(\Phi,p_2\mathcal{C}_s(\phi,\|\boldsymbol{A}\|_\infty))\mathcal{C}_s(\phi,\|\boldsymbol{A}\|_\infty)\sqrt{\|\boldsymbol{A}\|_\infty^2+1}})$ and $\gamma\in(0,\frac{1}{4}]$, consider any $\boldsymbol{W}_t,\boldsymbol{V}_t$ with*

$$L'(\boldsymbol{A}^*,\boldsymbol{A}^*,\boldsymbol{A}^*,\lambda_t,\boldsymbol{W}_t,\boldsymbol{V}_t)\in[(1+\gamma)OPT+\Omega(\boldsymbol{q}^\top\boldsymbol{A}^*\mathbf{1}\|\boldsymbol{A}^*\|_\infty^4\epsilon_0/\gamma),\tilde{O}(1)]$$

With high probability on random initialization, there exists $\widehat{\boldsymbol{W}}$, $\widehat{\boldsymbol{V}}$ with $\|\widehat{\boldsymbol{W}}\|_F\leq 1$, $\|\widehat{\boldsymbol{V}}\|_F\leq 1$ such that for every $\eta\in(0,\frac{1}{poly(m_1,m_2)}]$,

$$\begin{aligned}
&\min\{\mathbb{E}_{\boldsymbol{\Sigma}}[L'(\boldsymbol{A}^*,\boldsymbol{A}^*,\boldsymbol{A}^*,\lambda_t,\boldsymbol{W}_t+\sqrt{\eta}\widehat{\boldsymbol{W}}\boldsymbol{\Sigma},\boldsymbol{V}_t+\sqrt{\eta}\boldsymbol{\Sigma}\widehat{\boldsymbol{V}})],L'(\boldsymbol{A}^*,\boldsymbol{A}^*,\boldsymbol{A}^*,(1-\eta)\lambda_t,\boldsymbol{W}_t,\boldsymbol{V}_t)\}\\
\leq&(1-\eta\gamma/4)L'(\boldsymbol{A}^*,\boldsymbol{A}^*,\boldsymbol{A}^*,\lambda_t,\boldsymbol{W}_t,\boldsymbol{V}_t)
\end{aligned}\tag{109}$$

Proof:
Recall the pseudo network and the real network for every $r\in[K]$ as

$$g_r(\boldsymbol{q},\boldsymbol{A}^*,\boldsymbol{X},\boldsymbol{W}',\boldsymbol{V}')=\boldsymbol{q}^\top\boldsymbol{A}^*(\boldsymbol{A}^*(\boldsymbol{A}^*\boldsymbol{X}(\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho+\boldsymbol{W}')+\boldsymbol{B}_1)\odot\boldsymbol{D}_{\boldsymbol{w},\rho,t}(\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho+\boldsymbol{V}')\odot\boldsymbol{D}_{\boldsymbol{v},\rho,t})\boldsymbol{c}_r$$

$$f_r(\boldsymbol{q},\boldsymbol{A}^*,\boldsymbol{X},\boldsymbol{W}',\boldsymbol{V}')=\boldsymbol{q}^\top\boldsymbol{A}^*(\boldsymbol{A}^*(\boldsymbol{A}^*\boldsymbol{X}(\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho+\boldsymbol{W}')+\boldsymbol{B}_1)\odot\boldsymbol{D}_{\boldsymbol{w},\rho,\boldsymbol{W}'}(\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho+\boldsymbol{V}')\odot\boldsymbol{D}_{\boldsymbol{v},\rho,\boldsymbol{V}'})\boldsymbol{c}_r$$

where $\boldsymbol{D}_{\boldsymbol{w},\rho,t}$ and $\boldsymbol{D}_{\boldsymbol{v},\rho,t}$ are the diagonal matrices at weights $\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho+\boldsymbol{W}_t$ and $\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho+\boldsymbol{V}_t$. $\boldsymbol{D}_{\boldsymbol{w},\rho,\boldsymbol{W}'}$ and $\boldsymbol{D}_{\boldsymbol{v},\rho,\boldsymbol{V}'}$ are the diagonal matrices at weights $\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho+\boldsymbol{W}'$ and $\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho+\boldsymbol{V}'$.

Denote $G(\boldsymbol{q}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{W}', \boldsymbol{V}') = (g_1, \cdots, g_K)$, $F_{\boldsymbol{A}^*}(\boldsymbol{q}, \boldsymbol{X}, \boldsymbol{W}', \boldsymbol{V}') = (f_1, \cdots, f_K)$.
As long as $L'(\boldsymbol{A}^*, \boldsymbol{A}^*, \boldsymbol{A}^*, \lambda_t, \boldsymbol{W}_t, \boldsymbol{V}_t) \leq \tilde{O}(1)$, according to C.32 to C.34 in (Allen-Zhu et al., 2019), we have

$$\lambda_w \|\sqrt{\lambda_t}\widehat{\boldsymbol{W}}\|_{2,4}^4 \leq \epsilon_0$$

$$\lambda_v \|\sqrt{\lambda_t}\widehat{\boldsymbol{V}}\|_F^2 \leq \epsilon_0$$

$$\|\widehat{\boldsymbol{W}}\|_F \ll 1$$

$$\|\widehat{\boldsymbol{V}}\|_F \ll 1$$

The we need to study an update direction

$$\widetilde{\boldsymbol{W}} = \boldsymbol{W}_t + \sqrt{\eta}\widehat{\boldsymbol{W}}\boldsymbol{\Sigma}$$

$$\widetilde{\boldsymbol{V}} = \boldsymbol{V}_t + \sqrt{\eta}\boldsymbol{\Sigma}\widehat{\boldsymbol{V}}$$

Changes in Regularizer. Note that here $\boldsymbol{W}_t \in \mathbb{R}^{d\times m_1}$, $\boldsymbol{V}_t \in \mathbb{R}^{m_1\times m_2}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{m_1\times m_1}$. We know that

$$\mathbb{E}_{\boldsymbol{\Sigma}}[\|\boldsymbol{V}_t + \sqrt{\eta}\boldsymbol{\Sigma}\widehat{\boldsymbol{V}}\|_F^2] = \|\boldsymbol{V}_t\|_F^2 + \eta\|\widehat{\boldsymbol{V}}\|_F^2$$

$$\mathbb{E}_{\boldsymbol{\Sigma}}[\|\boldsymbol{W}_t + \sqrt{\eta}\widehat{\boldsymbol{W}}\boldsymbol{\Sigma}\|_{2,4}^4] = \sum_{i\in[m_1]} \mathbb{E}[\|\boldsymbol{w}_{t,i} + \sqrt{\eta}\widehat{\boldsymbol{W}}\boldsymbol{\Sigma}_i\|_2^4]$$

For each term $i \in [m_1]$, we can bound

$$\|\boldsymbol{w}_{t,i} + \sqrt{\eta}\widehat{\boldsymbol{W}}\boldsymbol{\Sigma}_i\|_2^2 = \|\boldsymbol{w}_{t,i}\|_2^2 + \eta\|\widehat{\boldsymbol{W}}\boldsymbol{\Sigma}_i\|_2^2 + 2\sqrt{\eta}{\boldsymbol{w}_{t,i}}^\top\widehat{\boldsymbol{W}}\boldsymbol{\Sigma}_i$$

$$\begin{aligned}\|\boldsymbol{w}_{t,i} + \sqrt{\eta}\widehat{\boldsymbol{W}}\boldsymbol{\Sigma}_i\|_2^4 &= \|\boldsymbol{w}_{t,i}\|_2^4 + \eta^2\|\widehat{\boldsymbol{W}}\boldsymbol{\Sigma}_i\|_2^4 + 4\eta\|{\boldsymbol{w}_{t,i}}^\top\widehat{\boldsymbol{W}}\boldsymbol{\Sigma}_i\|^2 + 2\eta\|\boldsymbol{w}_{t,i}\|_2^2\|\widehat{\boldsymbol{W}}\boldsymbol{\Sigma}_i\|_2^2 \\ &\leq \|\boldsymbol{w}_{t,i}\|_2^4 + 6\eta\|\boldsymbol{w}_{t,i}\|_2^2\|\widehat{\boldsymbol{W}}\boldsymbol{\Sigma}_i\|_2^2 + O_p(\eta^2)\end{aligned} \tag{110}$$

Therefore, by Cauchy-Schwarz inequality, we have

$$\mathbb{E}_{\boldsymbol{\Sigma}}[\|\boldsymbol{W}_t + \sqrt{\eta}\widehat{\boldsymbol{W}}\boldsymbol{\Sigma}\|_{2,4}^4] \leq \|\boldsymbol{W}_t\|_{2,4}^4 + 6\eta\|\boldsymbol{W}_t\|_{2,4}^2\|\widehat{\boldsymbol{W}}\|_{2,4}^2 + O_p(\eta^2)$$

Therefore, by $\lambda_w\|\sqrt{\lambda_t}\boldsymbol{W}_t\|_{2,4}^4 \leq R(\sqrt{\lambda_t}\boldsymbol{W}_t, \sqrt{\lambda_t}\boldsymbol{V}_t)$, we have

$$\begin{aligned}\mathbb{E}[R(\sqrt{\lambda_t}\widetilde{\boldsymbol{W}}, \sqrt{\lambda_t}\widetilde{\boldsymbol{V}})] \leq& R(\sqrt{\lambda_t}\boldsymbol{W}_t, \sqrt{\lambda_t}\boldsymbol{V}_t) + 6\eta\sqrt{\epsilon_0}\sqrt{R(\sqrt{\lambda_t}\boldsymbol{W}_t, \sqrt{\lambda_t}\boldsymbol{V}_t)} + \eta\epsilon_0 \\ \leq& R(\sqrt{\lambda_t}\boldsymbol{W}_t, \sqrt{\lambda_t}\boldsymbol{V}_t) + \frac{1}{4}\eta R(\sqrt{\lambda_t}\boldsymbol{W}_t, \sqrt{\lambda_t}\boldsymbol{V}_t) + 143\eta\epsilon_0\end{aligned} \tag{111}$$

Changes in Objective. Recall that here $\widehat{\boldsymbol{W}}$ and $\widehat{\boldsymbol{V}}$ satisfy $\tau_{\boldsymbol{w},\infty} \leq \frac{1}{m_1^{\frac{999}{1000}}}$ and $\tau_{\boldsymbol{v},\infty} \leq \frac{1}{m_2^{\frac{999}{2000}}}$. By Lemma B.11, we have for every $r \in [K]$

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{W}^\rho,\boldsymbol{V}^\rho}[|f_r(\boldsymbol{q}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{W}+\boldsymbol{W}^\rho+\widetilde{\boldsymbol{W}}\boldsymbol{\Sigma}, \boldsymbol{V}+\boldsymbol{V}^\rho+\boldsymbol{\Sigma}'\widetilde{\boldsymbol{V}}) - g_r(\boldsymbol{q}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{W}+\boldsymbol{W}^\rho+\widetilde{\boldsymbol{W}}\boldsymbol{\Sigma}, \boldsymbol{V}+\boldsymbol{V}^\rho+\boldsymbol{\Sigma}\widetilde{\boldsymbol{V}})|] \\ &\leq \tilde{O}(\|\boldsymbol{q}^\top\boldsymbol{A}^*\|_1\|\boldsymbol{A}^*\|_\infty^2\epsilon_0\eta) + O_p(\eta^{1.5})\end{aligned} \tag{112}$$

By Lemma B.10, we have

$$\begin{aligned}G(\boldsymbol{q}, \boldsymbol{A}^*, \boldsymbol{X}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{V}}) &= G(\boldsymbol{q}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{W}_t, \boldsymbol{V}_t) + \eta G^{(b,b)}(\boldsymbol{q}, \boldsymbol{A}^*, \boldsymbol{X}, \widehat{\boldsymbol{W}}\boldsymbol{\Sigma}, \boldsymbol{\Sigma}\widehat{\boldsymbol{V}}) + \sqrt{\eta}G' \\ &= F_{\boldsymbol{A}^*}(\boldsymbol{q}, \boldsymbol{X}, \boldsymbol{W}_t, \boldsymbol{V}_t) + \eta G^{(b,b)}(\boldsymbol{q}, \boldsymbol{A}^*, \boldsymbol{X}, \widehat{\boldsymbol{W}}\boldsymbol{\Sigma}, \boldsymbol{\Sigma}\widehat{\boldsymbol{V}}) + \sqrt{\eta}G' \\ &= F_{\boldsymbol{A}^*}(\boldsymbol{q}, \boldsymbol{X}, \boldsymbol{W}_t, \boldsymbol{V}_t) + \eta G^{(b,b)}(\boldsymbol{q}, \boldsymbol{A}^*, \boldsymbol{X}, \widehat{\boldsymbol{W}}, \widehat{\boldsymbol{V}}) + \sqrt{\eta}G'\end{aligned} \tag{113}$$

where $\mathbb{E}_{\boldsymbol{\Sigma}}[G'] = 0$ and $|G'| \leq \epsilon$ with high probability. By C.38 in (Allen-Zhu et al., 2019), we have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{W}^\rho,\boldsymbol{V}^\rho,\boldsymbol{\Sigma}}[L(\lambda_t F_{\boldsymbol{A}^*}(\boldsymbol{q}, \boldsymbol{X}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{V}}), y)] \\ \leq&\mathbb{E}_{\boldsymbol{W}^\rho,\boldsymbol{V}^\rho}[L(\lambda_t F_{\boldsymbol{A}^*}(\boldsymbol{q}, \boldsymbol{X}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{V}}) + \eta F^*_{\boldsymbol{A}^*}(\boldsymbol{q}, \boldsymbol{A}^*, \boldsymbol{X}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{V}}), y)] + O(\|\boldsymbol{q}^\top\boldsymbol{A}^*\|_1\|\boldsymbol{A}^*\|_\infty^2\epsilon_0\eta) + O_p(\eta^{1.5})\end{aligned} \tag{114}$$

Following C.40 in (Allen-Zhu et al., 2019), we have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{W}^\rho,\boldsymbol{V}^\rho}[L(\lambda_t F_{\boldsymbol{A}^*}(\boldsymbol{q},\boldsymbol{X},\boldsymbol{W}_t,\boldsymbol{V}_t)+\eta F^*_{\boldsymbol{A}^*}(\boldsymbol{q},\boldsymbol{A}^*,\boldsymbol{X},\boldsymbol{W}_t,\boldsymbol{V}_t),y)]\\ \leq&(1-\eta)(2L(\lambda_t F_{\boldsymbol{A}^*}(\boldsymbol{q},\boldsymbol{X},\boldsymbol{W}_t,\boldsymbol{V}_t),y)-L((1-\eta)\lambda_t F_{\boldsymbol{A}^*}(\boldsymbol{q},\boldsymbol{X},\boldsymbol{W}_t,\boldsymbol{V}_t),y))+\eta L(F^*_{\boldsymbol{A}^*},y)+O_p(\eta^2)\end{aligned} \tag{115}$$

Putting all of them together. Denote

$$c_1=\frac{1}{|\Omega^t|}\sum_{i=1}^{|\Omega|}\mathbb{E}_{\boldsymbol{W}^\rho,\boldsymbol{V}^\rho,\boldsymbol{\Sigma},\boldsymbol{\Sigma}'}[L(\lambda_t F_{\boldsymbol{A}^*}(\boldsymbol{e}_g,\boldsymbol{X},\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho+\widetilde{\boldsymbol{W}}\boldsymbol{\Sigma}',\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho+\boldsymbol{\Sigma}'\widetilde{\boldsymbol{V}}),y_i)] \tag{116}$$

$$c_1'=\mathbb{E}_{\boldsymbol{\Sigma}}[L'(\boldsymbol{A}^*,\boldsymbol{A}^*,\boldsymbol{A}^*,\lambda_t,\widetilde{\boldsymbol{W}},+\widetilde{\boldsymbol{V}})]=c_1+\mathbb{E}_{\boldsymbol{\Sigma}}[R(\sqrt{\lambda_t}\widetilde{\boldsymbol{W}},\sqrt{\lambda}\widetilde{\boldsymbol{V}})] \tag{117}$$

$$c_2=\frac{1}{|\Omega^t|}\sum_{i=1}^{|\Omega|}\mathbb{E}_{\boldsymbol{W}^\rho,\boldsymbol{V}^\rho}[L((1-\eta)\lambda_t F_{\boldsymbol{A}^*}(\boldsymbol{e}_g,\boldsymbol{X},\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho+\boldsymbol{W}_t\boldsymbol{\Sigma}',\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho+\boldsymbol{\Sigma}'\boldsymbol{V}_t),y_i)] \tag{118}$$

$$c_2'=L'(\boldsymbol{A}^*,\boldsymbol{A}^*,\boldsymbol{A}^*,(1-\eta)\lambda_t,\boldsymbol{W}_t,\boldsymbol{V}_t)=c_2+R(\sqrt{(1-\eta)\lambda_t}\boldsymbol{W}_t,\sqrt{(1-\eta)\lambda_t}\boldsymbol{V}_t) \tag{119}$$

$$c_3=\frac{1}{|\Omega^t|}\sum_{i=1}^{|\Omega|}\mathbb{E}_{\boldsymbol{W}^\rho,\boldsymbol{V}^\rho}[L(\lambda_t F_{\boldsymbol{A}^*}(\boldsymbol{e}_g,\boldsymbol{X},\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho+\boldsymbol{W}_t\boldsymbol{\Sigma}',\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho+\boldsymbol{\Sigma}'\boldsymbol{V}_t),y_i)] \tag{120}$$

$$c_3'=L'(\boldsymbol{A}^*,\boldsymbol{A}^*,\boldsymbol{A}^*,\lambda_t,\boldsymbol{W}_t,\boldsymbol{V}_t)=c_3+R(\sqrt{\lambda_t}\boldsymbol{W}_t,\sqrt{\lambda}\boldsymbol{V}_t) \tag{121}$$

Then following from C.38 to C.42 in (Allen-Zhu et al., 2019), we have

$$c_1'\leq(1-\eta)(2c_3'-c_2')+\frac{\eta\gamma}{4}c_3'+\eta(OPT+O(\|\boldsymbol{q}^\top\boldsymbol{A}^*\|_1\|\boldsymbol{A}^*\|_\infty^4\epsilon_0/\gamma))+O_p(\eta^{1.5}), \tag{122}$$

which implies

$$\min\{c_1',c_2'\}\leq(1-\eta\frac{1}{2}+\frac{\eta\gamma}{8})c_3'+\eta\frac{1}{2}OPT+O(\|\boldsymbol{q}^\top\boldsymbol{A}^*\|_1\|\boldsymbol{A}^*\|_\infty^2\epsilon_0\eta/\gamma)+O_p(\eta^{1.5})$$

As long as $c_3'\geq(1+\gamma)OPT+\Omega(\|\boldsymbol{q}^\top\boldsymbol{A}^*\|_1\|\boldsymbol{A}^*\|_\infty^2\epsilon_0/\gamma)$ and $\gamma\in[0,1]$, we have

$$\min\{c_1',c_2'\}\leq(1-\eta\frac{\gamma}{4})c_3'$$

Lemma B.14. *Note that the three sampled aggregation matrices in a three-layer learner network can be be different. We denote them as $\boldsymbol{A}^{t(1)}$, $\boldsymbol{A}^{t(2)}$ and $\boldsymbol{A}^{t(3)}$. Let $\boldsymbol{W}_t$, $\boldsymbol{V}_t$ be the updated weights trained using $\boldsymbol{A}^*$ and let $\boldsymbol{W}_t'$, $\boldsymbol{V}_t'$ be the updated weights trained using $\boldsymbol{A}^{t(i)}$, $i\in[3]$. With probability at least $99/100$, the algorithm converges in $TT_w=poly(m_1,m_2)$ iterations to a point with $\eta\in(0,\frac{1}{poly(m_1,m_2,\|\boldsymbol{A}^*\|_\infty,K)})$*

$$L'(\boldsymbol{A}^*,\boldsymbol{A}^*,\boldsymbol{A}^*,\lambda_t,\boldsymbol{W}_t,\boldsymbol{V}_t)\leq(1+\gamma)OPT+\epsilon_0$$

If

$$\begin{aligned}&L'(\boldsymbol{A}^{t(1)},\boldsymbol{A}^{t(2)},\boldsymbol{A}^{t(3)},\lambda_t,\boldsymbol{W}_t',\boldsymbol{V}_t')\\ =&\frac{1}{|\Omega^t|}\sum_{i=1}^{|\Omega^t|}\mathbb{E}_{\boldsymbol{W}^\rho,\boldsymbol{V}^\rho,\boldsymbol{\Sigma}}[L(\lambda_t F_{\boldsymbol{A}^{(1)},\boldsymbol{A}^{(2)},\boldsymbol{A}^{(3)}}(\boldsymbol{q},\boldsymbol{X}_i,\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho+\boldsymbol{W}_t'\boldsymbol{\Sigma}',\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho+\boldsymbol{\Sigma}'\boldsymbol{V}_t'),y_i)]\\ &+R(\sqrt{\lambda_t}\boldsymbol{W}_t',\sqrt{\lambda_t}\boldsymbol{V}_t'),\end{aligned} \tag{123}$$

where

$$F_{\boldsymbol{A}^{t(1)},\boldsymbol{A}^{t(2)},\boldsymbol{A}^{t(3)}}(\boldsymbol{q},\boldsymbol{X},\boldsymbol{W},\boldsymbol{V})=\boldsymbol{q}^\top\boldsymbol{A}^{t(3)}\sigma(\boldsymbol{A}^{t(2)}\sigma(\boldsymbol{A}^{t(1)}\boldsymbol{X}\boldsymbol{W}+\boldsymbol{B}_1)\boldsymbol{V}+\boldsymbol{B}_2)\boldsymbol{C}, \tag{124}$$

we also have

$$L'(\boldsymbol{A}^{t(1)},\boldsymbol{A}^{t(2)},\boldsymbol{A}^{t(3)},\lambda_{T-1},\boldsymbol{W}_T',\boldsymbol{V}_T')\leq L'(\boldsymbol{A}^*,\boldsymbol{A}^*,\boldsymbol{A}^*,\lambda_T,\boldsymbol{W}_T,\boldsymbol{V}_T)+\lambda_{T-1}\cdot O(poly(\epsilon))\leq(1+\gamma)OPT+\epsilon_0 \tag{125}$$

Proof:
By Lemma B.13, we know that as long as $L'(\boldsymbol{A}^*, \boldsymbol{A}^*, \boldsymbol{A}^*, \lambda_t, \boldsymbol{W}_t, \boldsymbol{V}_t) \in [(1+\gamma)OPT + \Omega(\boldsymbol{q}^\top \boldsymbol{A}^* \boldsymbol{1} \|\boldsymbol{A}^*\|_\infty^4 \epsilon_0/\gamma), \tilde{O}(1)]$, then there exists $\|\widehat{\boldsymbol{W}}\|_F \leq 1$, $\|\widehat{\boldsymbol{V}}\|_F \leq 1$ such that either

$$\mathbb{E}_{\boldsymbol{\Sigma}, \boldsymbol{\Sigma}'}[L'(\boldsymbol{A}^*, \boldsymbol{A}^*, \boldsymbol{A}^*, \lambda_t, \boldsymbol{W}_t\boldsymbol{\Sigma}' + \sqrt{\eta}\widehat{\boldsymbol{W}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}', \boldsymbol{\Sigma}'\boldsymbol{V}_t + \sqrt{\eta}\boldsymbol{\Sigma}'\boldsymbol{\Sigma}\widehat{\boldsymbol{V}})] \leq (1-\eta\gamma/4)L'(\boldsymbol{A}^*, \boldsymbol{A}^*, \boldsymbol{A}^*, \lambda_t, \boldsymbol{W}_t, \boldsymbol{V}_t) \quad (126)$$

or

$$L'(\boldsymbol{A}^*, \boldsymbol{A}^*, \boldsymbol{A}^*, (1-\eta)\lambda_t, \boldsymbol{W}_t, \boldsymbol{V}_t) \leq (1-\eta\gamma/4)L'(\boldsymbol{A}^*, \boldsymbol{A}^*, \boldsymbol{A}^*, \lambda_t, \boldsymbol{W}_t, \boldsymbol{V}_t) \quad (127)$$

Denote $\boldsymbol{W} = \boldsymbol{W}^{(0)} + \boldsymbol{W}^\rho + \boldsymbol{W}_t\boldsymbol{\Sigma}' + \sqrt{\eta}\widehat{\boldsymbol{W}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}'$, $\boldsymbol{V} = \boldsymbol{V}^{(0)} + \boldsymbol{V}^\rho + \boldsymbol{\Sigma}'\boldsymbol{V}_t + \sqrt{\eta}\boldsymbol{\Sigma}'\boldsymbol{\Sigma}\widehat{\boldsymbol{V}}$. Note that

$$\frac{\partial L}{\partial \boldsymbol{w}_j} = \sum_{i=1}^K \frac{\partial L}{\partial f_i}\frac{\partial f_i}{\partial \boldsymbol{w}_j} \quad (128)$$

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{w}_j} f_r(\boldsymbol{q}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{W}^{(0)} + \boldsymbol{W}^\rho + \boldsymbol{W}_t\boldsymbol{\Sigma}' + \sqrt{\eta}\widehat{\boldsymbol{W}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}', \boldsymbol{V}^{(0)} + \boldsymbol{V}^\rho + \boldsymbol{\Sigma}'\boldsymbol{V}_t + \sqrt{\eta}\boldsymbol{\Sigma}'\boldsymbol{\Sigma}\widehat{\boldsymbol{V}}) \\
=&\sum_{n=1}^N \boldsymbol{q}^\top \boldsymbol{a}_n \sum_{i=1}^{m_2} c_{i,r}\mathbb{1}_{r_{n,i}+B_{2(n,i)}\geq 0}\sum_{k=1}^N a_{n,k}v_{j,i}\mathbb{1}_{\tilde{\boldsymbol{a}^*}_n\boldsymbol{X}\boldsymbol{w}_k+B_{1(n,k)}\geq 0}(\tilde{\boldsymbol{a}^*}_n\boldsymbol{X})^\top,
\end{aligned} \quad (129)$$

which implies $\frac{\partial F}{\partial \boldsymbol{w}_t}$, $\frac{\partial^2 F}{\partial \boldsymbol{w}_t^2}$, $\frac{\partial^3 F}{\partial \boldsymbol{w}_t^3}$ are summations of $\mathbb{1}$, δ, δ' functions and their multiplications. It can be found that no $\delta(x)\delta'(x)$, $\delta(x)^2$ or $\delta'^2(x)$ exist in these terms. Therefore, by $\int_{-\infty}^{\infty}\delta(t)f(t)dt = f(0)$ and $\int_{-\infty}^{\infty}\delta'(t)f(t)dt = -f'(0)$, we can obtain that the value of the third-order derivative w.r.t. $\boldsymbol{W}^\rho$ of $\mathbb{E}_{\boldsymbol{W}^\rho, \boldsymbol{V}^\rho, \boldsymbol{\Sigma}}[L(\lambda_t F_{\boldsymbol{A}^*}(\boldsymbol{e}_g, \boldsymbol{X}, \boldsymbol{W}^{(0)} + \boldsymbol{W}^\rho + \boldsymbol{W}_t\boldsymbol{\Sigma}, \boldsymbol{V}^{(0)} + \boldsymbol{V}^\rho + \boldsymbol{\Sigma}\boldsymbol{V}_t), y)]$ is proportional to $\text{poly}(\|\boldsymbol{A}^*\|_\infty, K)$, some certain value of the probability density function of $\boldsymbol{W}^\rho$ and its derivative, i.e., $\text{poly}(\sigma_w^{-1})$. Similarly, the value of the third-order derivative w.r.t. $\boldsymbol{W}^\rho$ of $\mathbb{E}_{\boldsymbol{W}^\rho, \boldsymbol{V}^\rho, \boldsymbol{\Sigma}}[L(\lambda_t F_{\boldsymbol{A}^*}(\boldsymbol{e}_g, \boldsymbol{X}, \boldsymbol{W}^{(0)} + \boldsymbol{W}^\rho + \boldsymbol{W}_t\boldsymbol{\Sigma}, \boldsymbol{V}^{(0)} + \boldsymbol{V}^\rho + \boldsymbol{\Sigma}\boldsymbol{V}_t), y)]$ is polynomially depend on σ_v^{-1} and $\|\boldsymbol{A}^*\|_\infty$. By the value selection of σ_w and σ_v, we can conclude that L' is $B = \text{poly}(m_1, m_2, \|\boldsymbol{A}^*\|_\infty, K)$ second-order smooth. By Fact A.8 in (Allen-Zhu et al., 2019), it satisfies with $\eta \in (0, \frac{1}{\text{poly}(m_1, m_2, \|\boldsymbol{A}^*\|_\infty, K)})$

$$\lambda_{\min}(\nabla^2 L'(\boldsymbol{A}^*, \boldsymbol{A}^*, \boldsymbol{A}^*, \lambda_{t-1}, \boldsymbol{W}_t, \boldsymbol{V}_t)) < -\frac{1}{(m_1 m_2)^8} \quad (130)$$

Meanwhile, for $t \geq 1$, by the escape saddle point theorem of Lemma A.9 in (Allen-Zhu et al., 2019), we know with probability at least $1-p$, $\lambda_{\min}(\nabla^2 L'(\boldsymbol{A}^*, \boldsymbol{A}^*, \boldsymbol{A}^*, \lambda_{t-1}, \boldsymbol{W}_t, \boldsymbol{V}_t)) > -\frac{1}{(m_1 m_2)^8}$ holds. Choosing $p = \frac{1}{100T}$, then this holds for $t = 1, 2, \cdots, T$ with probability at least 0.999.Therefore, for $t = 1, 2, \cdots, T$, the first case cannot happen, i.e., as long as $L'(\boldsymbol{A}^*, \boldsymbol{A}^*, \boldsymbol{A}^*, \lambda_t, \boldsymbol{W}_t, \boldsymbol{V}_t) \geq (1+\gamma)OPT + \Omega(\boldsymbol{q}^\top \boldsymbol{A}^* \boldsymbol{1} \|\boldsymbol{A}^*\|_\infty^4 \epsilon_0/\gamma)$,

$$L'(\boldsymbol{A}^*, \boldsymbol{A}^*, \boldsymbol{A}^*, (1-\eta)\lambda_t, \boldsymbol{W}_t, \boldsymbol{V}_t) \leq (1-\eta\gamma/4)L'(\boldsymbol{A}^*, \boldsymbol{A}^*, \boldsymbol{A}^*, \lambda_t, \boldsymbol{W}_t, \boldsymbol{V}_t) \quad (131)$$

On the other hand, for $t = 1, 2, \cdots, T-1$, as long as $L' \leq \tilde{O}(1)$, by Lemma A.9 in (Allen-Zhu et al., 2019), we have

$$L'(\boldsymbol{A}^*, \boldsymbol{A}^*, \boldsymbol{A}^*, \lambda_t, \boldsymbol{W}_{t+1}, \boldsymbol{V}_{t+1}) \leq L'(\boldsymbol{A}^*, \boldsymbol{A}^*, \boldsymbol{A}^*, \lambda_t, \boldsymbol{W}_t, \boldsymbol{V}_t) + (m_1 m_2)^{-1} \quad (132)$$

By $L'(\boldsymbol{A}^*, \boldsymbol{A}^*, \boldsymbol{A}^*, \lambda_1, \boldsymbol{W}_0, \boldsymbol{V}_0) \leq \tilde{O}(1)$ with high probability, we have $L'(\boldsymbol{A}^*, \boldsymbol{A}^*, \boldsymbol{A}^*, \lambda_t, \boldsymbol{W}_t, , \boldsymbol{V}_t) \leq \tilde{O}(1)$ with high probability for $t = 1, 2, \cdots, T$. Therefore, after $T = \tilde{\Theta}(\eta^{-1}\log\frac{\log m}{\epsilon_0})$ rounds of weight decay, we have $L'(\boldsymbol{A}^*, \boldsymbol{A}^*, \boldsymbol{A}^*, \lambda_t, \boldsymbol{W}_t, \boldsymbol{V}_t) \leq (1+\gamma)OPT + \Omega(\boldsymbol{q}^\top \boldsymbol{A}^* \boldsymbol{1} \|\boldsymbol{A}^*\|_\infty^4 \epsilon_0/\gamma)$. Rescale down ϵ_0 and we can obtain our final result.
Consider $L'(\boldsymbol{A}^{t(1)}, \boldsymbol{A}^{t(2)}, \boldsymbol{A}^{t(3)}, \lambda_t, \boldsymbol{W}'_t, \boldsymbol{V}'_t)$. Let $\boldsymbol{w}_i$, $\boldsymbol{v}_i$ be the output weights updated with all the aggregation matrices equal to $\boldsymbol{A}^*$, and let $\boldsymbol{w}'_i$, $\boldsymbol{v}'_i$ be the output weights updated with our sampling strategy in Section 3.2. We know that

$$\begin{aligned}
\|\boldsymbol{w}_i - \boldsymbol{w}_i{}'\| &\lesssim \sum_{t=0}^{T-1}\|\eta\sum_{l=0}^{T_w-1}\sum_{n=1}^N \boldsymbol{q}^\top \boldsymbol{a}_n{}^*\sum_{i=1}^{m_2} c_{i,r}\mathbb{1}[\boldsymbol{a}^*{}_n\sigma(\boldsymbol{A}^*\boldsymbol{X}\boldsymbol{W})\boldsymbol{v}_i \geq 0]\sum_{k=1}^N a_{n,k}{}^* v_{j,i}\mathbb{1}[\boldsymbol{a}^*{}_k\boldsymbol{X}\boldsymbol{w}_j \geq 0](\boldsymbol{a}^*{}_k - \boldsymbol{a}_k{}^{t(1)})\boldsymbol{X}\| \\
&\leq \frac{1}{\text{poly}(m_1, m_2)}\cdot\frac{1}{\text{poly}(\epsilon)}\text{poly}(m_1, m_2)\epsilon_c\|\boldsymbol{A}^*\|_\infty \cdot \text{poly}(\epsilon) = O(\epsilon)
\end{aligned} \quad (133)$$

$$\begin{aligned}\|\boldsymbol{v}_i-\boldsymbol{v}_i'\| &\lesssim \sum_{t=0}^{T-1}\|\eta\sum_{l=0}^{T_w-1}\sum_{n=1}^{N}\boldsymbol{q}^\top\boldsymbol{a}_n^*\sum_{i=1}^{m_2}c_{i,r}\mathbb{1}[\boldsymbol{a}^*{}_n\sigma(\boldsymbol{A}^*\boldsymbol{X}\boldsymbol{W})\boldsymbol{v}_i\geq 0](\boldsymbol{a}^*{}_n\sigma(\boldsymbol{A}^*\boldsymbol{X}\boldsymbol{W})-\boldsymbol{a}_n^{t(2)}\sigma(\boldsymbol{A}^{t(1)}\boldsymbol{X}\boldsymbol{W}'))\| \\ &\leq \frac{1}{\text{poly}(m_1,m_2)}\cdot\frac{1}{\text{poly}(\epsilon)}\text{poly}(m_1,m_2)\epsilon_c\|\boldsymbol{A}^*\|_\infty\cdot\text{poly}(\epsilon)=O(\epsilon)\end{aligned} \tag{134}$$

With a slight abuse of notation, for $r\in[K]$, we denote

$$f_r(\boldsymbol{q},\boldsymbol{A}^{t(1)},\boldsymbol{A}^{t(2)},\boldsymbol{A}^{t(3)},\boldsymbol{X},\boldsymbol{W}_t',\boldsymbol{V}_t')=\boldsymbol{q}^\top\boldsymbol{A}^{t(3)}\sigma(\boldsymbol{A}^{t(2)}\sigma(\boldsymbol{A}^{t(1)}\boldsymbol{X}\boldsymbol{W}+\boldsymbol{B}_1)\boldsymbol{V}+\boldsymbol{B}_2)\boldsymbol{c}_r \tag{135}$$

The difference between $f_r(\boldsymbol{q},\boldsymbol{A}^*,\boldsymbol{X},\boldsymbol{W}_t,\boldsymbol{V}_t)$ and $f_r(\boldsymbol{q},\boldsymbol{A}^{t(1)},\boldsymbol{A}^{t(2)},\boldsymbol{A}^{t(3)},\boldsymbol{X},\boldsymbol{W}_t',\boldsymbol{V}_t')$ is caused by $\|\boldsymbol{A}^*-\boldsymbol{A}^{t(1)}\|_\infty$, $\|\boldsymbol{A}^*-\boldsymbol{A}^{t(2)}\|_\infty$, $\|\boldsymbol{A}^*-\boldsymbol{A}^{t(3)}\|_\infty$, $\boldsymbol{w}_i^{(t)}-\boldsymbol{w}_i^{(t)'}$ and $\boldsymbol{v}_i^{(t)}-\boldsymbol{v}_i^{(t)'}$. Following the proof in Lemma A.2, we can easily obtain that if $|p_l-p_l^*|\leq p_l^*\cdot O(\text{poly}(\epsilon))$ and $l_i\geq|\mathcal{N}_i|/(1+\frac{c_1\cdot\text{poly}(\epsilon)}{p_l^*L\Phi(L,i)})$, it can be derived that $\|\boldsymbol{A}^*-\boldsymbol{A}^{(1)}\|_\infty\leq O(\text{poly}(\epsilon))$, $\|\boldsymbol{A}^*-\boldsymbol{A}^{(2)}\|_\infty\leq O(\text{poly}(\epsilon))$ and $\|\boldsymbol{A}^*-\boldsymbol{A}^{(3)}\|_\infty\leq O(\text{poly}(\epsilon))$. Then, by (133) and (134), we have

$$\begin{aligned}&|\boldsymbol{q}^\top\boldsymbol{A}^*\sigma(\boldsymbol{A}^*\sigma(\boldsymbol{A}^*\boldsymbol{X}\boldsymbol{W})\boldsymbol{V})\boldsymbol{c}_r-\boldsymbol{q}^\top\boldsymbol{A}^*\sigma(\boldsymbol{A}^{(2)}\sigma(\boldsymbol{A}^{(1)}\boldsymbol{X}\boldsymbol{W}')\boldsymbol{V}')\boldsymbol{c}_r| \\ \leq&\Big|\sum_{n=1}^{N}\boldsymbol{q}^\top\boldsymbol{a}^*{}_n\sum_{i=1}^{m_2}c_{i,r}|\sigma(\boldsymbol{a}^*{}_n\sigma(\boldsymbol{A}^*\boldsymbol{X}\boldsymbol{W})\boldsymbol{v}_i)-\sigma(\boldsymbol{a}_n^{(2)}\sigma(\boldsymbol{A}^{(1)}\boldsymbol{X}\boldsymbol{W}')\boldsymbol{v}_i')|\Big| \\ \leq&\Big|\sum_{n=1}^{N}\boldsymbol{q}^\top\boldsymbol{a}^*{}_n\sum_{i=1}^{m_2}c_{i,r}|\boldsymbol{a}^*{}_n\sigma(\boldsymbol{A}^*\boldsymbol{X}\boldsymbol{W})\boldsymbol{v}_i-\boldsymbol{a}_n^{(2)}\sigma(\boldsymbol{A}^{(1)}\boldsymbol{X}\boldsymbol{W}')\boldsymbol{v}_i'|\Big| \\ \leq&\Big|\sum_{n=1}^{N}\boldsymbol{q}^\top\boldsymbol{a}^*{}_n\sum_{i=1}^{m_2}c_{i,r}|(\boldsymbol{a}^*{}_n-\boldsymbol{a}_n^{(2)})\sigma(\boldsymbol{A}^*\boldsymbol{X}\boldsymbol{W})\boldsymbol{v}_i|\Big|+\Big|\sum_{n=1}^{N}\boldsymbol{q}^\top\boldsymbol{a}^*{}_n\sum_{i=1}^{m_2}c_{i,r}|\boldsymbol{a}_n^{(2)}(\sigma(\boldsymbol{A}^*\boldsymbol{X}\boldsymbol{W})\boldsymbol{v}_i-\sigma(\boldsymbol{A}^{(1)}\boldsymbol{X}\boldsymbol{W}')\boldsymbol{v}_i')|)\Big| \\ \leq&\Big|\sum_{n=1}^{N}\boldsymbol{q}^\top\boldsymbol{a}^*{}_n\sum_{i=1}^{m_2}c_{i,r}|(\boldsymbol{a}^*{}_n-\boldsymbol{a}_n^{(2)})\sigma(\boldsymbol{A}^*\boldsymbol{X}\boldsymbol{W})\boldsymbol{v}_i|\Big| \\ &+\Big|\sum_{n=1}^{N}\boldsymbol{q}^\top\boldsymbol{a}^*{}_n\sum_{i=1}^{m_2}c_{i,r}|\boldsymbol{a}_n^{(2)}((\sigma(\boldsymbol{A}^*\boldsymbol{X}\boldsymbol{W})-\sigma(\boldsymbol{A}^{(1)}\boldsymbol{X}\boldsymbol{W}'))\boldsymbol{v}_i+\sigma(\boldsymbol{A}^{(1)}\boldsymbol{X}\boldsymbol{W}')(\boldsymbol{v}_i-\boldsymbol{v}_i'))|\Big| \\ \leq&\Big|\sum_{n=1}^{N}\boldsymbol{q}^\top\boldsymbol{a}^*{}_n\sum_{i=1}^{m_2}c_{i,r}|(\boldsymbol{a}^*{}_n-\boldsymbol{a}_n^{(2)})\sigma(\boldsymbol{A}^*\boldsymbol{X}\boldsymbol{W})\boldsymbol{v}_i|\Big|+\Big|\sum_{n=1}^{N}\boldsymbol{q}^\top\boldsymbol{a}^*{}_n\sum_{i=1}^{m_2}c_{i,r}|\boldsymbol{a}_n^{(2)}\sigma(\boldsymbol{A}^{(1)}\boldsymbol{X}\boldsymbol{W}')(\boldsymbol{v}_i-\boldsymbol{v}_i')|\Big| \\ &+\Big|\sum_{n=1}^{N}\boldsymbol{q}^\top\boldsymbol{a}^*{}_n\sum_{i=1}^{m_2}c_{i,r}|\sum_{k=1}^{N}a''_{n,k}\sum_{l=1}^{m_1}v_{i,l}|(\boldsymbol{a}^*{}_k-\boldsymbol{a}_k')\boldsymbol{X}\boldsymbol{w}_l+\boldsymbol{a}_k'\boldsymbol{X}(\boldsymbol{w}_l-\boldsymbol{w}_l')||\Big| \\ \leq&O(\text{poly}(\epsilon)).\end{aligned} \tag{136}$$

Hence,

$$\begin{aligned}&|\boldsymbol{e}_g^\top\boldsymbol{A}^*\sigma(\boldsymbol{A}^*\sigma(\boldsymbol{A}^*\boldsymbol{X}\boldsymbol{W})\boldsymbol{V})\boldsymbol{c}_r-\boldsymbol{e}_g^\top\boldsymbol{A}^{(3)}\sigma(\boldsymbol{A}^{(2)}\sigma(\boldsymbol{A}^{(1)}\boldsymbol{X}\boldsymbol{W}')\boldsymbol{V}')\boldsymbol{c}_r| \\ \leq&|\boldsymbol{e}_g^\top\boldsymbol{A}^*\sigma(\boldsymbol{A}^*\sigma(\boldsymbol{A}^*\boldsymbol{X}\boldsymbol{W})\boldsymbol{V})\boldsymbol{c}_r-\boldsymbol{e}_g^\top\boldsymbol{A}^*\sigma(\boldsymbol{A}^{(2)}\sigma(\boldsymbol{A}^{(1)}\boldsymbol{X}\boldsymbol{W}')\boldsymbol{V}')\boldsymbol{c}_r|+|\boldsymbol{e}_g^\top(\boldsymbol{A}^*-\boldsymbol{A}^{(3)})\sigma(\boldsymbol{A}^{(2)}\sigma(\boldsymbol{A}^{(1)}\boldsymbol{X}\boldsymbol{W})\boldsymbol{V})\boldsymbol{c}_r| \\ \leq&O(\text{poly}(\epsilon)).\end{aligned} \tag{137}$$

which implies

$$L'(\boldsymbol{A}^{t(1)},\boldsymbol{A}^{t(2)},\boldsymbol{A}^{t(3)},\lambda_{T-1},\boldsymbol{W}_T',\boldsymbol{V}_T')\leq L'(\boldsymbol{A}^*,\boldsymbol{A}^*,\boldsymbol{A}^*,\lambda_T,\boldsymbol{W}_T,\boldsymbol{V}_T)+\lambda_{T-1}\cdot O(\text{poly}(\epsilon))\leq(1+\gamma)OPT+\epsilon_0 \tag{138}$$

Proof of Theorem 3.1:
By Lemma B.14, we have that the algorithm converges in TT_w iterations to a point

$$L'(\boldsymbol{A}^{t(1)},\boldsymbol{A}^{t(2)},\boldsymbol{A}^{t(3)},\lambda_t,\boldsymbol{W}_t,\boldsymbol{V}_t)\leq(1+\gamma)OPT+\epsilon_0$$

We know w.h.p., among $\tilde{O}(1/\epsilon_0^2)$ choices of j,

$$\min_j\{\mathbb{E}_{\boldsymbol{W}^\rho,\boldsymbol{V}^\rho,\boldsymbol{\Sigma},\, \boldsymbol{z}\in\Omega} L(\lambda_{T-1}F_{\boldsymbol{A}^*}(\boldsymbol{e}_g,\boldsymbol{X},\boldsymbol{W}^{(0)}+\boldsymbol{W}^{\rho,j}+\boldsymbol{W}_T\boldsymbol{\Sigma},\boldsymbol{V}^{(0)}+\boldsymbol{V}^{\rho,j}+\boldsymbol{\Sigma}\boldsymbol{V}_T)\} \le (1+\gamma)OPT+\epsilon_0 \tag{139}$$

Then we have

$$\|\boldsymbol{W}_T\|_{2,4} \le \epsilon_0^{\frac{1}{4}}\tau_w' \tag{140}$$

$$\|\boldsymbol{V}_T\|_F \le \epsilon_0^{\frac{1}{2}}\tau_v' \tag{141}$$

By Lemma B.9, we know that

$$\begin{aligned}&f_r(\boldsymbol{e}_g,\boldsymbol{A}^*,\boldsymbol{X}_i,\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho+\boldsymbol{W}_T\boldsymbol{\Sigma},\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho+\boldsymbol{\Sigma}\boldsymbol{V}_T,\boldsymbol{B})\\ =&f_r(\boldsymbol{e}_g,\boldsymbol{A}^*,\boldsymbol{X}_i,\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho,\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho,\boldsymbol{B})+g_r^{(b,b)}(\boldsymbol{e}_g,\boldsymbol{A}^*,\boldsymbol{X}_i,\boldsymbol{W}_T,\boldsymbol{V}_T,\boldsymbol{B})\pm\frac{\epsilon}{K}\end{aligned} \tag{142}$$

Denote $\boldsymbol{r}'=\boldsymbol{A}^*\sigma(\boldsymbol{A}^*\boldsymbol{X}(\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho)+\boldsymbol{B}_1)(\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho)$. Then,

$$\|\boldsymbol{r}'\| \le \|\boldsymbol{A}^*\|(\|\boldsymbol{A}^*\|_\infty\cdot\tilde{O}(1))\cdot\tilde{O}(1) \le \|\boldsymbol{A}^*\|_\infty \tag{143}$$

Therefore,

$$\begin{aligned}&|f_r(\boldsymbol{e}_g,\boldsymbol{A}^*,\boldsymbol{X}_i,\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho,\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho,\boldsymbol{B})|\\ =&|\boldsymbol{e}_g^\top\boldsymbol{A}^*\sigma(\boldsymbol{r}'+\boldsymbol{B}_2)\boldsymbol{c}_r|\\ \le&\tilde{O}(\|\boldsymbol{A}^*\|_\infty(\|\boldsymbol{A}^*\|_\infty+1)\epsilon_c)\end{aligned} \tag{144}$$

We also have

$$\begin{aligned}&|g_r^{(b,b)}(\boldsymbol{e}_g,\boldsymbol{A}^*,\boldsymbol{X}_i,\boldsymbol{W}_T,\boldsymbol{V}_T,\boldsymbol{B})|\\ \le&|\boldsymbol{e}_g^\top\boldsymbol{A}^*\boldsymbol{A}^*(\boldsymbol{A}^*\boldsymbol{X}\boldsymbol{W}_T\odot\boldsymbol{D}^{(0)}_{\boldsymbol{w},\boldsymbol{x}}\boldsymbol{V}_T)\odot\boldsymbol{D}^{(0)}_{\boldsymbol{v},\boldsymbol{x}}\boldsymbol{c}_r|\\ \le&\|\boldsymbol{A}^*\|_\infty^2\tau_v'\tau_w'm_1^{\frac{1}{4}}\sqrt{m_2}\epsilon_c\\ \le&C_0\|\boldsymbol{A}^*\|_\infty^2\end{aligned} \tag{145}$$

Hence,

$$f_r(\boldsymbol{e}_g,\boldsymbol{A}^*,\boldsymbol{X}_i,\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho+\boldsymbol{W}_T\boldsymbol{\Sigma},\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho+\boldsymbol{\Sigma}\boldsymbol{V}_T,\boldsymbol{B}) \le \tilde{O}(\|\boldsymbol{A}^*\|_\infty^2(\epsilon_c+C_0)) \tag{146}$$

Combining (135, 137), we can obtain

$$f_r(\boldsymbol{e}_g,\boldsymbol{A}^{t(1)},\boldsymbol{A}^{t(2)},\boldsymbol{A}^{t(3)},\boldsymbol{X}_i,\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho+\boldsymbol{W}_T\boldsymbol{\Sigma},\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho+\boldsymbol{\Sigma}\boldsymbol{V}_T,\boldsymbol{B}) \le \tilde{O}(\|\boldsymbol{A}^*\|_\infty^2(\epsilon_c+C_0)) \tag{147}$$

as long as $\|\boldsymbol{A}^*-\boldsymbol{A}^{t(1)}\|_\infty \le \text{poly}(\epsilon)$, $\|\boldsymbol{A}^*-\boldsymbol{A}^{t(2)}\|_\infty \le \text{poly}(\epsilon)$ and $\|\boldsymbol{A}^*-\boldsymbol{A}^{t(3)}\|_\infty \le \text{poly}(\epsilon)$.

For any given $\{\boldsymbol{X}_i,y_i\}_{i=1}^{|\Omega|}$, the dependency between y_i, y_j, where $i,j\in|\Omega|$ can be considered in two steps. Figure 5(a) shows $\boldsymbol{a}_i\boldsymbol{X}$ is dependent with at most $(1+\delta)^2$ $\boldsymbol{a}_j\boldsymbol{X}'s$. This is because each $\boldsymbol{a}_i\boldsymbol{X}$ is determined by at most $(1+\delta)$ row vector $\tilde{\boldsymbol{x}}_l's$, while each $\tilde{\boldsymbol{x}}_l$ is contained by at most $(1+\delta)$ $\boldsymbol{a}_p\boldsymbol{X}'s$. Similarly, y_i is determined by at most $(1+\delta)$ $\boldsymbol{a}_p\boldsymbol{X}'s$ and by Figure 5(b) we can find y_i is dependent with at most $(1+\delta)^4$ y_j (including y_i). Since the matrix $\boldsymbol{A}^*$ shares the same non-zero entries with $\boldsymbol{A}$, the output with $\boldsymbol{A}^*$ indicates the same dependence.

Denote $u_i = 1/|\Omega^t|\sum_{i=1}^{|\Omega^t|}|L(\lambda_{T-1}F_{\boldsymbol{A}^*}(\boldsymbol{e}_g,\boldsymbol{X},\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho+\boldsymbol{W}_T\boldsymbol{\Sigma},\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho+\boldsymbol{\Sigma}\boldsymbol{V}_T),y_i) - \mathbb{E}_{(\boldsymbol{e}_g,\boldsymbol{X},y)\in\mathcal{D}}[L(\lambda_{T-1}F_{\boldsymbol{A}^*}(\boldsymbol{e}_g,\boldsymbol{X},\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho+\boldsymbol{W}_T\boldsymbol{\Sigma},\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho+\boldsymbol{\Sigma}\boldsymbol{V}_T),y_i)]$. Then, $\mathbb{E}[u_i]=0$. Since that L is 1-lipschitz smooth and $L(\boldsymbol{0}^K,y)\in[0,1]$, we have

$$\begin{aligned}&|L(\lambda_{T-1}F_{\boldsymbol{A}^*}(\boldsymbol{e}_g,\boldsymbol{X},\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho+\boldsymbol{W}_T\boldsymbol{\Sigma},\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho+\boldsymbol{\Sigma}\boldsymbol{V}_T),y_i)-L(\boldsymbol{0}^K,y_i)|\\ \le&\|(\lambda_{T-1}F_{\boldsymbol{A}^*}(\boldsymbol{e}_g,\boldsymbol{X},\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho+\boldsymbol{W}_T\boldsymbol{\Sigma},\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho+\boldsymbol{\Sigma}\boldsymbol{V}_T),y_i)-(\boldsymbol{0}^K,y_i)\|\\ \le&\tilde{O}(\sqrt{K}\|\boldsymbol{A}^*\|_\infty^2(\epsilon_c+C_0))\end{aligned} \tag{148}$$

Then,

$$|u_i| \le 2\sqrt{K}\|\boldsymbol{A}^*\|_\infty^2(\epsilon_c+C_0)$$

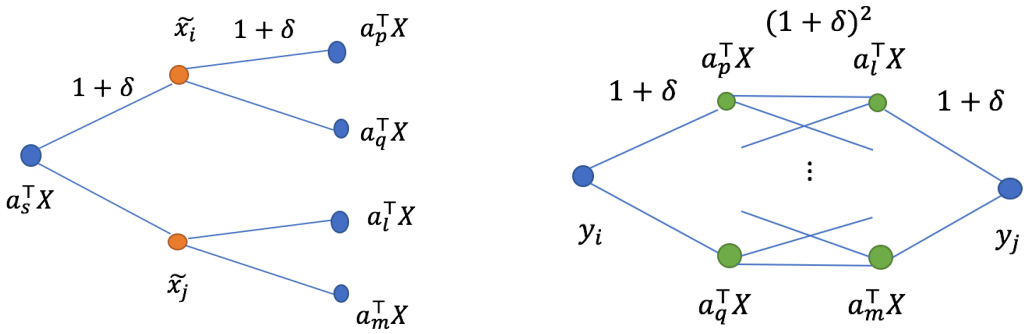

Figure 5. (a) Dependency between $\boldsymbol{a}_s\boldsymbol{X}$ and $\boldsymbol{a}_p\boldsymbol{X}$ (b) Dependency between y_i and y_j

$$\mathbb{P}(|u_i| \geq t) \leq 1 \leq \exp(1 - \frac{t^2}{4K\|\boldsymbol{A}^*\|_\infty^4(\epsilon_c + C_0)^2}) \tag{149}$$

Then, u_i is a sub-Gaussian random variable. We have $\mathbb{E}e^{su_i} \leq e^{\|\boldsymbol{A}^*\|_\infty^4(\epsilon_c+C_0)^2 s^2}$. By Lemma 7 in (Zhang et al., 2020), we have

$$\mathbb{E}e^{s\sum_{i=1}^{|\Omega|} u_i} \leq e^{(1+\delta)^4 K\|\boldsymbol{A}^*\|_\infty^4(\epsilon_c+C_0)^2|\Omega|s^2}$$

Therefore,

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{|\Omega|}\frac{1}{|\Omega|}u_i\Big| \geq k\Big) \leq \exp(\|\boldsymbol{A}^*\|_\infty^4(\epsilon_c + C_0)^2 K(1+\delta)^4|\Omega|s^2 - |\Omega|ks) \tag{150}$$

for any $s > 0$. Let $s = \frac{k}{2\|\boldsymbol{A}^*\|_\infty^4(\epsilon_c+C_0)^2K(1+\delta)^4}$, $k = \|\boldsymbol{A}^*\|_\infty^4(\epsilon_c + C_0)^2 K\sqrt{\frac{(1+\delta)^4\log N}{|\Omega|}}$, we can obtain

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{|\Omega|}\frac{1}{|\Omega|}u_i\Big| \geq k\Big) \leq \exp(-\|\boldsymbol{A}^*\|_\infty^4(\epsilon_c + C_0)^2 K\log N) \leq N^{-K} \tag{151}$$

Therefore, with probability at least $1 - N^{-K}$, we have

$$\begin{aligned}&\Big|\mathbb{E}_{(\boldsymbol{e}_g,\boldsymbol{X},y)\sim\mathcal{D}}[L(\lambda_{T-1}F_{\boldsymbol{A}^*}(\boldsymbol{e}_g,\boldsymbol{X},\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho+\boldsymbol{W}_T\boldsymbol{\Sigma},\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho+\boldsymbol{\Sigma}\boldsymbol{V}_T),y_i)]\\ &-\frac{1}{|\Omega^t|}\sum_{i=1}^{|\Omega^t|}L(\lambda_{T-1}F_{\boldsymbol{A}^*}(\boldsymbol{e}_g,\boldsymbol{X},\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho+\boldsymbol{W}_T\boldsymbol{\Sigma},\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho+\boldsymbol{\Sigma}\boldsymbol{V}_T),y_i)\Big|\\ \leq&\epsilon_0\end{aligned} \tag{152}$$

as long as $|\Omega| \geq \tilde{\Theta}(\epsilon_0^{-2}\|\boldsymbol{A}^*\|_\infty^8(1+p_1^4p_2^5\mathcal{C}_\epsilon(\phi,\|\boldsymbol{A}^*\|_\infty)\mathcal{C}_\epsilon(\Phi,\sqrt{p_2}\mathcal{C}_\epsilon(\phi,\|\boldsymbol{A}^*\|_\infty))(\|\boldsymbol{A}^*\|_\infty+1)^4K^6(1+\delta)^4\log N)$, i.e.,

$$\mathbb{E}_{(\boldsymbol{e}_g,\boldsymbol{X},y)\sim\mathcal{D}}[L(\lambda_{T-1}F_{\boldsymbol{A}^*}(\boldsymbol{e}_g,\boldsymbol{X},\boldsymbol{W}^{(0)}+\boldsymbol{W}^\rho+\boldsymbol{W}_T\boldsymbol{\Sigma},\boldsymbol{V}^{(0)}+\boldsymbol{V}^\rho+\boldsymbol{\Sigma}\boldsymbol{V}_T),y_i)] \leq OPT+\epsilon \tag{153}$$