Fast Learning of Graph Neural Networks with Guaranteed Generalizability: One-hidden-layer Case

Shuai Zhang [1] **Meng Wang** [1] **Sijia Liu** [2] **Pin-Yu Chen** [3] **Jinjun Xiong** [3]

Abstract

Although graph neural networks (GNNs) have made great progress recently on learning from graph-structured data in practice, their theoretical guarantee on generalizability remains elusive in the literature. In this paper, we provide a theoretically-grounded generalizability analysis of GNNs with one hidden layer for both regression and binary classification problems. Under the assumption that there exists a ground-truth GNN model (with zero generalization error), the objective of GNN learning is to estimate the ground-truth GNN parameters from the training data. To achieve this objective, we propose a learning algorithm that is built on tensor initialization and accelerated gradient descent. We then show that the proposed learning algorithm converges to the ground-truth GNN model for the regression problem, and to a model sufficiently close to the ground-truth for the binary classification problem. Moreover, for both cases, the convergence rate of the proposed learning algorithm is proven to be linear and faster than the vanilla gradient descent algorithm. We further explore the relationship between the sample complexity of GNNs and their underlying graph properties. Lastly, we provide numerical experiments to demonstrate the validity of our analysis and the effectiveness of the proposed learning algorithm for GNNs.

[1]Dept. of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, NY, USA [2]MIT-IBM Waston AI Lab, Cambridge, MA, USA [3]IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA. Correspondence to: Shuai Zhang <zhangs21@rpi.edu>, Meng Wang <wangm7@rpi.edu>, Sijia Liu <Sijia.Liu@ibm.com>, Pin-Yu Chen <Pin-Yu.Chen@ibm.com>, Jinjun Xiong <jinjun@us.ibm.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

1. Introduction

Graph neural networks (GNNs) (Gilbert et al., 2005; Scarselli et al., 2008) have demonstrated great practical performance in learning with graph-structured data. Compared with traditional (feed-forward) neural networks, GNNs introduce an additional neighborhood aggregation layer, where the features of each node are aggregated with the features of the neighboring nodes (Gilmer et al., 2017; Xu et al., 2018). GNNs have a better learning performance in applications including physical reasoning (Battaglia et al., 2016), recommendation systems (Ying et al., 2018), biological analysis (Duvenaud et al., 2015), and compute vision (Monfardini et al., 2006). Many variations of GNNs, such as Gated Graph Neural Networks (GG-NNs) (Li et al., 2016), Graph Convolutional Networks (GCNs) (Kipf & Welling, 2017) and others (Hamilton et al., 2017; Veličković et al., 2018) have recently been developed to enhance the learning performance on graph-structured data.

Despite the numerical success, the theoretical understanding of the generalizability of the learned GNN models to the testing data is very limited. Some works (Xu et al., 2018; 2019; Wu et al., 2019; Morris et al., 2019) analyze the expressive power of GNNs but do not provide learning algorithms that are guaranteed to return the desired GNN model with proper parameters. Only few works (Du et al., 2019; Verma & Zhang, 2019) explore the generalizabilty of GNNs, under the one-hidden-layer setting, as even with one hidden layer the models are already complex to analyze, not to mention the multi-layer setting. Both works show that for regression problems, the generalization gap of the training error and the testing error decays with respect to the number of training samples at a sub-linear rate. The analysis in Ref. (Du et al., 2019) analyzes GNNs through Graph Neural Tangent Kernels (GNTK) which is an extension of Neural Tangent kernel (NTK) model (Jacot et al., 2018; Chizat & Bach, 2018; Nitanda & Suzuki, 2019; Cao & Gu, 2020). When over-parameterized, this line of works shows sub-linear convergence to the global optima of the learning problem with assuming enough filters in the hidden layer (Jacot et al., 2018; Chizat & Bach, 2018). Ref. (Verma & Zhang, 2019) only applies to the case of one single filter in the hidden layer, and the activation function needs to be

smooth, excluding the popular ReLU activation function. Moreover, refs. (Du et al., 2019; Verma & Zhang, 2019) do not consider classification and do not discuss if a small training error and a small generalization error can be achieved simultaneously.

One recent line of research analyzes the generalizability of neural networks (NNs) from the perspective of model estimation (Brutzkus & Globerson, 2017; Du et al., 2018; 2017; Fu et al., 2018; Ge et al., 2018; Safran & Shamir, 2018; Zhong et al., 2017b;a). These works assume the existence of a ground-truth NN model with some unknown parameters that maps the input features to the output labels for both training and testing samples. Then the learning objective is to estimate the ground-truth model parameters from the training data, and this ground-truth model is guaranteed to have a zero generalization error on the testing data. The analyses are focused on one-hidden-layer NNs, assuming the input features following the Gaussian distribution (Shamir, 2018). If one-hidden-layer NNs only have one filter in the hidden layer, gradient descent (GD) methods can learn the ground-truth parameters with a high probability (Du et al., 2018; 2017; Brutzkus & Globerson, 2017). When there are multiple filters in the hidden layer, the learning problem is much more challenging to solve because of the common spurious local minima (Safran & Shamir, 2018). (Ge et al., 2018) revises the learning objective and shows the global convergence of GD to the global optimum of the new learning problem. The required number for training samples, referred to as the sample complexity in this paper, is a high-order polynomial function of the model size. A few works (Zhong et al., 2017b;a; Fu et al., 2018) study a learning algorithm that initializes using the tensor initialization method (Zhong et al., 2017b) and iterates using GD. This algorithm is proved to converge to the ground-truth model parameters with a zero generalization error for the one-hidden-layer NNs with multiple filters, and the sample complexity is shown to be linear in the model size. All these works only consider NNs rather than GNNs.

Contributions. This paper provides the first algorithmic design and theoretical analysis to learn a GNN model with a zero generalization error, assuming the existence of such a ground-truth model. We study GNNs in semi-supervised learning, and the results apply to both regression and binary classification problems. Different from NNs, each output label on the graph depends on multiple neighboring features in GNNs, and such dependence significantly complicates the analysis of the learning problem. Our proposed algorithm uses the tensor initialization (Zhong et al., 2017b) and updates by accelerated gradient descent (AGD). We prove that with a sufficient number of training samples, our algorithm returns the ground-truth model with the zero generalization error for regression problems. For binary classification problems, our algorithm returns a model sufficiently close to the ground-truth model, and its distance to the ground-truth model decays to zero as the number of samples increases. Our algorithm converges linearly, with a rate that is proved to be faster than that of vanilla GD. We quantifies the dependence of the sample complexity on the model size and the underlying graph structural properties. The required number of samples is linear in the model size. It is also a polynomial function of the graph degree and the largest singular value of the normalized adjacency matrix. Such dependence of the sample complexity on graph parameters is exclusive to GNNs and does not exist in NNs.

The rest of the paper is organized as follows. Section 2 introduces the problem formulation. The algorithm is presented in Section 3, and Section 4 summarizes the major theoretical results. Section 5 shows the numerical results, and Section 6 concludes the paper. All the proofs are in the supplementary materials.

Notation: Vectors are bold lowercase, matrices and tensors are bold uppercase. Also, scalars are in normal font, and sets are in calligraphy and blackboard bold font. 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_{ij} denotes the (i,j)-th entry of $\boldsymbol{Z}$. $\mathcal{Z}$ stands for a regular set. Special sets $\mathbb{N}$ (or $\mathbb{N}^+$), $\mathbb{Z}$ and $\mathbb{R}$ denote the sets of all natural numbers (or positive natural numbers), all integers and all real numbers, respectively. Typically, $[Z]$ stands for the set of $\{1, 2, \cdots, Z\}$ for any number $\mathbb{N}^+$. $\boldsymbol{I}$ and $\boldsymbol{e}_i$ denote the identity matrix and the i-th standard basis vector. $\boldsymbol{Z}^T$ denotes the transpose of $\boldsymbol{Z}$, similarly for $\boldsymbol{z}^T$. $\|\boldsymbol{z}\|$ denotes the ℓ_2-norm of a vector $\boldsymbol{z}$, and $\|\boldsymbol{Z}\|_2$ and $\|\boldsymbol{Z}\|_F$ denote the spectral norm and Frobenius norm of matrix $\boldsymbol{Z}$, respectively. We use $\sigma_i(\boldsymbol{Z})$ to denote the i-th largest singular value of $\boldsymbol{Z}$. Moreover, the outer product of a group of vectors $\boldsymbol{z}_i \in \mathbb{R}^{n_i}, i \in [l]$, is defined as $\boldsymbol{T} = \boldsymbol{z}_1 \otimes \cdots \otimes \boldsymbol{z}_l \in \mathbb{R}^{n_1 \times \cdots \times n_l}$ with $T_{j_1,\cdots,j_l} = (\boldsymbol{z}_1)_{j_1} \cdots (\boldsymbol{z}_l)_{j_l}$.

2. Problem Formulation

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 δ and δ_{ave} denote the maximum and average node degree of $\mathcal{G}$, respectively. Let $\tilde{\boldsymbol{A}} \in \{0,1\}^{N \times N}$ be the adjacency matrix of $\mathcal{G}$ with added self-connections. Then, $\tilde{A}_{i,j} = 1$ if and only if there exists an edge between node v_i and node v_j, $i, j \in [N]$, and $\tilde{A}_{i,i} = 1$ for all $i \in [N]$. 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 $\boldsymbol{A} = \boldsymbol{D}^{-1/2}\tilde{\boldsymbol{A}}\boldsymbol{D}^{-1/2}$, and $\sigma_1(\boldsymbol{A})$ is the largest singular value of $\boldsymbol{A}$.

Each node v_n in $\mathcal{V}$ $(n = 1, 2, \cdots, N)$ corresponds to an input feature vector, denoted by $\boldsymbol{x}_n \in \mathbb{R}^d$, and a label $y_n \in \mathbb{R}$. y_n depends on not only $\boldsymbol{x}_n$ but also all $\boldsymbol{x}_j$ where v_j is a neighbor of v_n. Let $\boldsymbol{X} = \left[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\right]^T \in \mathbb{R}^{N \times d}$

denote the feature matrix. Following the analyses of NNs (Shamir, 2018), we assume $\boldsymbol{x}_n$'s are i.i.d. samples from the standard Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. For GNNs, we consider the typical semi-supervised learning problem setup. Let $\Omega \subset [N]$ denote the set of node indices with known labels, and let Ω^c be its complementary set. The objective of the GNN is to predict y_i for every i in Ω^c.

Suppose there exists a one-hidden-layer GNN that maps node features to labels, as shown in Figure 1. There are K filters[1] in the hidden layer, and the weight matrix is denoted by $\boldsymbol{W}^* = \begin{bmatrix}\boldsymbol{w}_1^* & \boldsymbol{w}_2^* & \cdots & \boldsymbol{w}_K^*\end{bmatrix} \in \mathbb{R}^{d\times K}$. The hidden layer is followed by a pooling layer. Different from NNs, GNNs have an additional aggregation layer with $\boldsymbol{A}$ as the aggregation factor matrix (Kipf & Welling, 2017). For every node $v_n \in \mathcal{V}$, the input to the hidden layer is $\boldsymbol{a}_n^T\boldsymbol{X}$, where $\boldsymbol{a}_n^T$ denotes the n-th row of $\boldsymbol{A}$. When there is no edge in $\mathcal{V}$, $\boldsymbol{A}$ is reduced to the identify matrix, and a GNN model is reduced to an NN model.

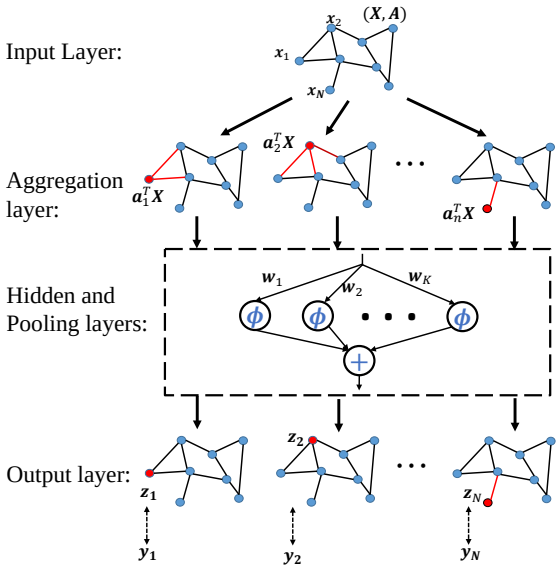


Figure 1. Structure of the graph neural network

The output z_n of the node v_n of the GNN is

$$z_n = g(\boldsymbol{W}^*; \boldsymbol{a}_n^T\boldsymbol{X}) = \frac{1}{K}\sum_{j=1}^{K}\phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j^*), \forall n \in [N], \tag{1}$$

where $\phi(\cdot)$ is the activation function. We consider both regression and binary classification in this paper. For regression, $\phi(\cdot)$ is the ReLU function[2] $\phi(x) = \max\{x, 0\}$, and $y_n = z_n$. For binary classification, we consider the sigmoid activation function where $\phi(x) = 1/(1+e^{-x})$. Then y_n is a binary variable generated from z_n by $\text{Prob}\{y_n = 1\} = z_n$, and $\text{Prob}\{y_n = 0\} = 1 - z_n$.

Given $\boldsymbol{X}$, $\boldsymbol{A}$, and y_i for all $i \in \Omega$, the learning objective is to estimate $\boldsymbol{W}^*$, which is assumed to have a zero generalization error. The training objective is to minimize the empirical risk function,

$$\min_{\boldsymbol{W}\in\mathbb{R}^{d\times K}} \hat{f}_\Omega(\boldsymbol{W}) := \frac{1}{|\Omega|}\sum_{n\in\Omega}\ell(\boldsymbol{W}; \boldsymbol{a}_n^T\boldsymbol{X}), \tag{2}$$

where ℓ is the loss function. For regression, we use the squared loss function , and (2) is written as

$$\min_{\boldsymbol{W}}: \quad \hat{f}_\Omega(\boldsymbol{W}) = \frac{1}{2|\Omega|}\sum_{n\in\Omega}\left|y_n - g(\boldsymbol{W}; \boldsymbol{a}_n^T\boldsymbol{X})\right|^2. \tag{3}$$

For classification, we use the cross entropy loss function, and (2) is written as

$$\begin{aligned}\min_{\boldsymbol{W}}: \quad \hat{f}_\Omega(\boldsymbol{W}) = &\frac{1}{|\Omega|}\sum_{n\in\Omega} -y_n \log\left(g(\boldsymbol{W}; \boldsymbol{a}_n^T\boldsymbol{X})\right) \\ &-(1-y_n)\log\left(1 - g(\boldsymbol{W}; \boldsymbol{a}_n^T\boldsymbol{X})\right).\end{aligned} \tag{4}$$

Both (3) and (4) are nonconvex due to the nonlinear function ϕ. Moreover, while $\boldsymbol{W}^*$ is a global minimum of (3), $\boldsymbol{W}^*$ is not necessarily a global minimum of (4)[3]. Furthermore, compared with NNs, the additional difficulty of analyzing the generalization performance of GNNs lies in the fact that each label y_n is correlated with all the input features that are connected to node v_n, as shown in the risk functions in (3) and (4).

Note that our model with $K = 1$ is equivalent to the one-hidden-layer convolutional network (GCN) (Kipf & Welling, 2017) for binary classification. To study the multi-class classification problem, the GCN model in (Kipf & Welling, 2017) has M nodes for M classes in the second layer and employs the softmax activation function at the output. Here, our model has a pooling layer and uses the sigmoid function for binary classification. Moreover, we consider both regression and binary classification problems using the same model architecture with different activation functions. We consider one-hidden-layer networks following the state-of-art works in NNs (Du et al., 2018; 2017; Brutzkus & Globerson, 2017; Zhong et al., 2017b;a; Fu et al., 2018) and GNNs (Du et al., 2019; Verma & Zhang, 2019) because the theoretical analyses are extremely complex and still being developed for multiple hidden layers.

3. Proposed Learning Algorithm

In what follows, we illustrate the algorithm used for solving problems (3) and (4), summarized in Algorithm 1. Algorithm 1 has two components: a) accelerated gradient descent

[1] We assume $K \le d$ to simplify the representation of the analysis, while the result still holds for $K > d$ with minor changes.

[2] Our result can be extended to the sigmoid activation function with minor changes.

[3] $\boldsymbol{W}^*$ is a global minimum if replacing all y_n with z_n in (4), but z_n's are unknown in practice.

and b) tensor initialization. We initialize $\boldsymbol{W}$ using the tensor initialization method (Zhong et al., 2017b) with minor modification for GNNs and update iterates by the Heavy Ball method (Polyak, 1987).

Accelerated gradient descent. Compared with the vanilla GD method, each iterate in the Heavy Ball method is updated along the combined directions of both the gradient and the moving direction of the previous iterates. Specifically, one computes the difference of the estimates in the previous two iterations, and the difference is scaled by a constant β. This additional momentum term is added to the gradient descent update. When β is 0, AGD reduces to GD.

During each iteration, a fresh subset of data is applied to estimate the gradient. The assumption of disjoint subsets is standard to simplify the analysis (Zhong et al., 2017a;b) but not necessary in numerical experiments.

Algorithm 1 Accelerated Gradient Descent Algorithm with Tensor Initialization

1: **Input:** $\boldsymbol{X}$, $\{y_n\}_{n\in\Omega}$, $\boldsymbol{A}$, the step size η, the momentum constant β, and the error tolerance ε;
2: **Initialization:** Tensor Initialization via Subroutine 1;
3: Partition Ω into $T = \log(1/\varepsilon)$ disjoint subsets, denoted as $\{\Omega_t\}_{t=1}^T$;
4: **for** $t = 1, 2, \cdots, T$ **do**
5: $\quad \boldsymbol{W}^{(t+1)} = \boldsymbol{W}^{(t)} - \eta\nabla\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)}) + \beta(\boldsymbol{W}^{(t)} - \boldsymbol{W}^{(t-1)})$
6: **end for**

Tensor initialization. The main idea of the tensor initialization method (Zhong et al., 2017b) is to utilize the homogeneous property of an activation function such as ReLU to estimate the magnitude and direction separately for each $\boldsymbol{w}_j^*$ with $j \in [K]$. A non-homogeneous function can be approximated by piece-wise linear functions, if the function is strictly monotone with lower-bounded derivatives (Fu et al., 2018), like the sigmoid function. Our initialization is similar to those in (Zhong et al., 2017b; Fu et al., 2018) for NNs with some definitions are changed to handle the graph structure, and the initialization process is summarized in Subroutine 1.

Specifically, following (Zhong et al., 2017b), we define a special outer product, denoted by $\widetilde{\otimes}$, such that for any vector $\boldsymbol{v} \in \mathbb{R}^{d_1}$ and $\boldsymbol{Z} \in \mathbb{R}^{d_1\times d_2}$,

$$\boldsymbol{v}\widetilde{\otimes}\boldsymbol{Z} = \sum_{i=1}^{d_2}(\boldsymbol{v}\otimes\boldsymbol{z}_i\otimes\boldsymbol{z}_i + \boldsymbol{z}_i\otimes\boldsymbol{v}\otimes\boldsymbol{z}_i + \boldsymbol{z}_i\otimes\boldsymbol{z}_i\otimes\boldsymbol{v}), \quad (5)$$

where $\otimes$ is the outer product and $\boldsymbol{z}_i$ is the i-th column of $\boldsymbol{Z}$.

Subroutine 1 Tensor Initialization Method

1: **Input:** $\boldsymbol{X}$, $\{y_n\}_{n\in\Omega}$ and $\boldsymbol{A}$;
2: Partition Ω into three disjoint subsets Ω_1, Ω_2, Ω_3;
3: Calculate $\widehat{\boldsymbol{M}}_1$, $\widehat{\boldsymbol{M}}_2$ following (6), (7) using Ω_1, Ω_2, respectively;
4: Estimate $\widehat{\boldsymbol{V}}$ by orthogonalizing the eigenvectors with respect to the K largest eigenvalues of $\widehat{\boldsymbol{M}}_2$;
5: Calculate $\widehat{\boldsymbol{M}}_3(\widehat{\boldsymbol{V}}, \widehat{\boldsymbol{V}}, \widehat{\boldsymbol{V}})$ using (9) through Ω_3;
6: Obtain $\{\widehat{\boldsymbol{u}}_j\}_{j=1}^K$ via tensor decomposition method (Kuleshov et al., 2015);
7: Obtain $\widehat{\boldsymbol{\alpha}}$ by solving optimization problem (11);
8: **Return:** $\boldsymbol{w}_j^{(0)} = \widehat{\alpha}_j\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j$, $j = 1, ..., K$.

Next, we define [4]

$$\boldsymbol{M}_1 = \mathbb{E}_{\boldsymbol{X}}\{y\boldsymbol{x}\} \in \mathbb{R}^d, \quad (6)$$

$$\boldsymbol{M}_2 = \mathbb{E}_{\boldsymbol{X}}\Big\{y\big[(\boldsymbol{a}_n^T\boldsymbol{X})\otimes(\boldsymbol{a}_n^T\boldsymbol{X}) - \boldsymbol{I}\big]\Big\} \in \mathbb{R}^{d\times d}, \quad (7)$$

$$\boldsymbol{M}_3 = \mathbb{E}_{\boldsymbol{X}}\Big\{y\big[(\boldsymbol{a}_n^T\boldsymbol{X})^{\otimes 3} - (\boldsymbol{a}_n^T\boldsymbol{X})\widetilde{\otimes}\boldsymbol{I}\big]\Big\} \in \mathbb{R}^{d\times d\times d}, \quad (8)$$

where $\boldsymbol{z}^{\otimes 3} := \boldsymbol{z}\otimes\boldsymbol{z}\otimes\boldsymbol{z}$. The tensor M_3 is used to identify the directions of $\{w_j^*\}_{j=1}^K$. M_1 depends on both the magnitudes and directions of $\{w_j^*\}_{j=1}^K$. We will sequentially estimate the directions and magnitudes of $\{w_j^*\}_{j=1}^K$ from M_3 and M_1. The matrix M_2 is used to identify the subspace spanned by w_j^*. We will project to this subspace to reduce the computational complexity of decomposing M_3.

Specifically, the values of $\boldsymbol{M}_1$, $\boldsymbol{M}_2$ and $\boldsymbol{M}_3$ are all estimated through samples, and let $\widehat{\boldsymbol{M}}_1$, $\widehat{\boldsymbol{M}}_2$, $\widehat{\boldsymbol{M}}_3$ denote the corresponding estimates of these high-order momentum. Tensor decomposition method (Kuleshov et al., 2015) provides the estimates of the vectors $\boldsymbol{w}_j^*/\|\boldsymbol{w}_j^*\|_2$ from $\widehat{\boldsymbol{M}}_3$, and the estimates are denoted as $\widehat{\widetilde{\boldsymbol{w}}}_j^*$.

However, the computation complexity of estimate through $\widehat{\boldsymbol{M}}_3$ depends on poly(d). To reduce the computational complexity of tensor decomposition, $\widehat{\boldsymbol{M}}_3$ is in fact first projected to a lower-dimensional tensor (Zhong et al., 2017b) through a matrix $\widehat{\boldsymbol{V}} \in \mathbb{R}^{d\times K}$. $\widehat{\boldsymbol{V}}$ is the estimation of matrix $\boldsymbol{V}$ and can be computed from the right singular vectors of $\widehat{\boldsymbol{M}}_2$. The column vectors of $\boldsymbol{V}$ form a basis for the subspace spanned by $\{\boldsymbol{w}_j^*\}_{j=1}^K$, which indicates that $\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{w}_j^* = \boldsymbol{w}_j^*$ for any $j \in [K]$. Then, from (8), $\boldsymbol{M}_3(\widehat{\boldsymbol{V}}, \widehat{\boldsymbol{V}}, \widehat{\boldsymbol{V}}) \in \mathbb{R}^{K\times K\times K}$ is defined as

$$\boldsymbol{M}_3(\widehat{\boldsymbol{V}}, \widehat{\boldsymbol{V}}, \widehat{\boldsymbol{V}}) := \mathbb{E}_{\boldsymbol{X}}\Big\{y\big[(\boldsymbol{a}_n^T\boldsymbol{X}\widehat{\boldsymbol{V}})^{\otimes 3} - (\boldsymbol{a}_n^T\boldsymbol{X}\widehat{\boldsymbol{V}})\widetilde{\otimes}\boldsymbol{I}\big]\Big\}. \quad (9)$$

[4] $\mathbb{E}_{\boldsymbol{X}}$ stands for the expectation over the distribution of random variable $\boldsymbol{X}$.

Similar to the case of $\widehat{\boldsymbol{M}}_3$, by applying the tensor decomposition method in $\widehat{\boldsymbol{M}}_3(\widehat{\boldsymbol{V}}, \widehat{\boldsymbol{V}}, \widehat{\boldsymbol{V}})$, one can obtain a series of normalized vectors, denoted as $\{\widehat{\boldsymbol{u}}_j\}_{j=1}^K \in \mathbb{R}^K$, which are the estimates of $\{\boldsymbol{V}^T\overline{\boldsymbol{w}}_j^*\}_{j=1}^K$. Then, $\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j$ is an estimate of $\overline{\boldsymbol{w}}_j^*$ since $\overline{\boldsymbol{w}}_j^*$ lies in the column space of $\boldsymbol{V}$ with $\boldsymbol{V}\boldsymbol{V}^T\overline{\boldsymbol{w}}_j^* = \overline{\boldsymbol{w}}_j^*$.

From (Zhong et al., 2017b), (6) can be written as

$$\boldsymbol{M}_1 = \sum_{j=1}^{K} \psi_1(\overline{\boldsymbol{w}}_j^*)\|\boldsymbol{w}_j^*\|_2\overline{\boldsymbol{w}}_j^*, \tag{10}$$

where ψ_1 depends on the distribution of $\boldsymbol{X}$. Since the distribution of $\boldsymbol{X}$ is known, the values of $\psi(\widehat{\overline{\boldsymbol{w}}}_j^*)$ can be calculated exactly. Then, the magnitudes of $\boldsymbol{w}_j^*$'s are estimated through solving the following optimization problem:

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^K} : \quad \left|\widehat{\boldsymbol{M}}_1 - \sum_{j=1}^{K} \psi(\widehat{\overline{\boldsymbol{w}}}_j^*)\alpha_j\widehat{\overline{\boldsymbol{w}}}_j^*\right|. \tag{11}$$

Thus, $\boldsymbol{W}^{(0)}$ is given as $\left[\widehat{\alpha}_1\widehat{\overline{\boldsymbol{w}}}_1^*, \quad \cdots, \quad \widehat{\alpha}_K\widehat{\overline{\boldsymbol{w}}}_K^*\right]$.

4. Main Theoretical Results

Theorems 1 and 2 state our major results about the GNN model for regression and binary classification, respectively. Before formally presenting the results, we first summarize the key findings as follows.

1. Zero generalization error of the learned model. Algorithm 1 can return $\boldsymbol{W}^*$ exactly for regression (see (14)) and approximately for binary classification (see (19)). Specifically, since $\boldsymbol{W}^*$ is often not a solution to (4), Algorithm 1 returns a critical point $\widehat{\boldsymbol{W}}$ that is sufficiently close to $\boldsymbol{W}^*$, and the distance decreases with respect to the number of samples in the order of $\sqrt{1/|\Omega|}$. Thus, with a sufficient number of samples, $\widehat{\boldsymbol{W}}$ will be close to $\boldsymbol{W}^*$ and achieves a zero generalization error approximately for binary classification. Algorithm 1 always returns $\boldsymbol{W}^*$ exactly for regression, a zero generalization error is thus achieved.

2. Fast linear convergence of Algorithm 1. Algorithm 1 is proved to converge linearly to $\boldsymbol{W}^*$ for regression and $\widehat{\boldsymbol{W}}$ for classification, as shown in (14) and (18). That means the distance of the estimate during the iterations to $\boldsymbol{W}^*$ (or $\widehat{\boldsymbol{W}}$) decays exponentially. Moreover, Algorithm 1 converges faster than the vanilla GD. The rate of convergence is $1 - \Theta\big(\frac{1}{\sqrt{K}}\big)$ for regression [5] and $1 - \Theta\big(\frac{1}{K}\big)$ for classification, where K is the number of filters in the hidden layer. In comparison, the convergence rates of GD are $1 - \Theta\big(\frac{1}{K}\big)$ and $1 - \Theta\big(\frac{1}{K^2}\big)$, respectively. Note that a smaller value of the rate of convergence corresponds to faster convergence. We remark that this is the first theoretical guarantee of AGD methods for learning GNNs.

3. Sample complexity analysis. $\boldsymbol{W}^*$ can be estimated exactly for regression and approximately for classification, provided that the number of samples is in the order of $(1+\delta^2)\text{poly}(\sigma_1(\boldsymbol{A}), K)d\log N\log(1/\varepsilon)$, as shown in (13) and (17), where ε is the desired estimation error tolerance. $\boldsymbol{W}^*$ has Kd parameters, where K is the number of nodes in the hidden layer, and d is the feature dimension. Our sample complexity is order-wise optimal with respect to d and only logarithmic with respect to the total number of features N. We further show that the sample complexity is also positively associated with $\sigma_1(\boldsymbol{A})$ and δ. That characterizes the relationship between the sample complexity and graph structural properties. From Lemma 1, we know that given δ, $\sigma_1(\boldsymbol{A})$ is positively correlated with the average node degree δ_{ave}. Thus, the required number of samples increases when the maximum and average degrees of the graph increase. That coincides with the intuition that more edges in the graph corresponds to the stronger dependence of the labels on neighboring features, thus requiring more samples to learn these dependencies. Our sample complexity quantifies this intuition explicitly.

Note that the graph structure affects this bound only through $\sigma_1(\boldsymbol{A})$ and δ. Different graph structures may require a similar number of samples to estimate $\boldsymbol{W}^*$, as long as they have similar $\sigma_1(\boldsymbol{A})$ and δ. We will verify this property on different graphs numerically in Figure 7.

Lemma 1. *Give an un-directed graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ and the normalized adjacency matrix $\boldsymbol{A}$ as defined in Section 2, the largest singular value $\sigma_1(\boldsymbol{A})$ of $\boldsymbol{A}$ satisfies*

$$\frac{1+\delta_{ave}}{1+\delta_{\max}} \le \sigma_1(\boldsymbol{A}) \le 1, \tag{12}$$

where δ_{ave} and δ are the average and maximum node degree, respectively.

4.1. Formal theoretical guarantees

To formally present the results, some parameters in the results are defined as follows. $\sigma_j(\boldsymbol{W}^*)$ $(j \in [N])$ is the j-th singular value of $\boldsymbol{W}^*$. $\kappa = \sigma_1(\boldsymbol{W}^*)/\sigma_K(\boldsymbol{W}^*)$ is the conditional number of $\boldsymbol{W}^*$. γ is defined as $\prod_{j=1}^K \sigma_j(\boldsymbol{W}^*)/\sigma_K(\boldsymbol{W}^*)$. For a fixed $\boldsymbol{W}^*$, both γ and κ can be viewed as constants and do not affect the order-wise analysis.

Theorem 1. *(Regression) Let $\{\boldsymbol{W}^{(t)}\}_{t=1}^T$ be the sequence generated by Algorithm 1 to solve (3) with $\eta = K/(8\sigma_1^2(\boldsymbol{A}))$. Suppose the number of samples satisfies*

$$|\Omega| \ge C_1\varepsilon_0^{-2}\kappa^9\gamma^2(1+\delta^2)\sigma_1^4(\boldsymbol{A})K^8 d\log N\log(1/\varepsilon) \tag{13}$$

[5] $f(d) = O(g(d))$ means that if for some constant $C > 0$, $f(d) \le Cg(d)$ holds when d is sufficiently large. $f(d) = \Theta(g(d))$ means that for some constants $c > 0$ and $C > 0$, $cg(d) \le f(d) \le Cg(d)$ holds when d is sufficiently large.

for some constants $C_1 > 0$ *and* $\varepsilon_0 \in (0, \frac{1}{2})$. *Then* $\{\boldsymbol{W}^{(t)}\}_{t=1}^T$ *converges linearly to* $\boldsymbol{W}^*$ *with probability at least* $1 - K^2T \cdot N^{-10}$ *as*

$$\begin{aligned} \|\boldsymbol{W}^{(t)} - \boldsymbol{W}^*\|_2 \le \nu(\beta)^t \|\boldsymbol{W}^{(0)} - \boldsymbol{W}^*\|_2, \text{and} \\ \|\boldsymbol{W}^{(T)} - \boldsymbol{W}^*\|_2 \le \varepsilon \|\boldsymbol{W}^*\|_2 \end{aligned} \tag{14}$$

where $\nu(\beta)$ *is the rate of convergence that depends on* β. *Moreover, we have*

$$\nu(\beta) < \nu(0) \quad \text{for some small nonzero } \beta. \tag{15}$$

Specifically, let $\beta^* = \left(1 - \sqrt{\frac{1-\varepsilon_0}{88\kappa^2\gamma}}\right)^2$, *we have*

$$\nu(0) \ge 1 - \frac{1-\varepsilon_0}{88\kappa^2\gamma K}, \nu(\beta^*) = 1 - \frac{1-\varepsilon_0}{\sqrt{88\kappa^2\gamma K}}. \tag{16}$$

Theorem 2. *(Classification) Let* $\{\boldsymbol{W}^{(t)}\}_{t=1}^T$ *be the sequence generated by Algorithm 1 to solve (4) with* $\eta = 1/(2\sigma_1^2(\boldsymbol{A}))$. *Suppose the number of samples satisfies*

$$|\Omega| \ge C_2 \varepsilon_0^{-2}(1+\delta^2)\kappa^8\gamma^2\sigma_1^4(\boldsymbol{A})K^8 d \log N \log(1/\varepsilon) \tag{17}$$

for some positive constants C_2 *and* $\varepsilon_0 \in (0,1)$. *Then, let* $\widehat{\boldsymbol{W}}$ *be the nearest critical point of (4) to* $\boldsymbol{W}^*$, *we have that* $\{\boldsymbol{W}^{(t)}\}_{t=1}^T$ *converges linearly to* $\widehat{\boldsymbol{W}}$ *with probability at least* $1 - K^2T \cdot N^{-10}$ *as*

$$\begin{aligned} \|\boldsymbol{W}^{(t)} - \widehat{\boldsymbol{W}}\|_2 \le \nu(\beta)^t \|\boldsymbol{W}^{(0)} - \widehat{\boldsymbol{W}}\|_2, \text{ and} \\ \|\boldsymbol{W}^{(T)} - \widehat{\boldsymbol{W}}\|_2 \le \varepsilon \|\boldsymbol{W}^{(0)} - \widehat{\boldsymbol{W}}\|_2. \end{aligned} \tag{18}$$

The distance between $\widehat{\boldsymbol{W}}$ *and* $\boldsymbol{W}^*$ *is bounded by*

$$\|\widehat{\boldsymbol{W}} - \boldsymbol{W}^*\|_2 \le C_3(1-\varepsilon_0)^{-1}\kappa^2\gamma K\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega|}}, \tag{19}$$

where $\nu(\beta)$ *is the rate of convergence that depends on* β, *and* C_3 *is some positive constant. Moreover, we have*

$$\nu(\beta) < \nu(0) \quad \text{for some small nonzero } \beta, \tag{20}$$

Specifically, let $\beta^* = \left(1 - \sqrt{\frac{1-\varepsilon_0}{11\kappa^2\gamma K^2}}\right)^2$, *we have*

$$\nu(0) = 1 - \frac{1-\varepsilon_0}{11\kappa^2\gamma K^2}, \nu(\beta^*) = 1 - \sqrt{\frac{1-\varepsilon_0}{11\kappa^2\gamma K^2}}. \tag{21}$$

4.2. Comparison with existing works

Only (Verma & Zhang, 2019; Du et al., 2019) analyze the generalization error of one-hidden-layer GNNs in regression problems, while there is no existing work about the generalization error in classification problems. (Verma & Zhang, 2019; Du et al., 2019) show that the difference between the risks in the testing data and the training data decreases in the order of $1/\sqrt{|\Omega|}$ as the sample size increases. The GNN model in (Verma & Zhang, 2019) only has one filter in the hidden layer, i.e., $K = 1$, and the loss function is required to be a smooth function, excluding ReLU. Ref. (Du et al., 2019) only considers infinitely wide GNNs. In contrast, $\boldsymbol{W}^*$ returned by Algorithm1 can achieve zero risks for both training data and testing data in regression problems. Our results apply to an arbitrary number of filters and the ReLU activation function. Moreover, this paper is the first work that characterizes the generalization error of GNNs for binary classification.

When δ is zero, our model reduces to one-hidden-layer NNs, and the corresponding sample complexity is $O\Big(\text{poly}(K)d\log N\log(1/\varepsilon)\Big)$. Our results are at least comparable to, if not better than, the state-of-art theoretical guarantees that from the prespective of model estimation for NNs. For example, (Zhong et al., 2017b) considers one-hidden-layer NNs for regression and proves the linear convergence of their algorithm to the ground-truth model parameters. The sample complexity of (Zhong et al., 2017b) is also linear in d, but the activation function must be smooth. (Zhang et al., 2019) considers one-hidden-layer NNs with the ReLU activation function for regression, but the algorithm cannot converge to the ground-truth parameters exactly but up to a statistical error. Our result in Theorem 1 applies to the nonsmooth ReLU function and can recover $\boldsymbol{W}^*$ exactly. (Fu et al., 2018) considers one-hidden-layer NNs for classification and proves linear convergence of their algorithm to a critical point sufficiently close to $\boldsymbol{W}^*$ with the distance bounded by $O(\sqrt{1/|\Omega|})$. The convergence rate in (Fu et al., 2018) is $1 - \Theta(1/K^2)$, while Algorithm 1 has a faster convergence rate of $1 - \Theta(1/K)$.

5. Numerical Results

We verify our results on synthetic graph-structured data. We consider four types of graph structures as shown in Figure 2: (a) a connected-cycle graph having each node connecting to its δ closet neighbors; (b) a two-dimensional grid having each node connecting to its nearest neighbors in axis-aligned directions; (c) a random δ-regular graph having each node connecting to δ other nodes randomly; (d) a random graph with bounded degree having each node degree selected from 0 with probability $1-p$ and δ with probability p for some $p \in [0,1]$. The feature vectors $\{\boldsymbol{x}_n\}_{n=1}^N$ are randomly generated from the standard Gaussian distribution $\mathcal{N}(0, \boldsymbol{I}_{d\times d})$. Each entry of $\boldsymbol{W}^*$ is generated from $\mathcal{N}(0, 5^2)$ independently. $\{z_n\}_{n=1}^N$ are computed based on (1). The labels $\{y_n\}_{n=1}^N$ are generated by $y_n = z_n$ and $\text{Prob}\{y_n = 1\} = z_n$ for regression and classification problems, respectively.

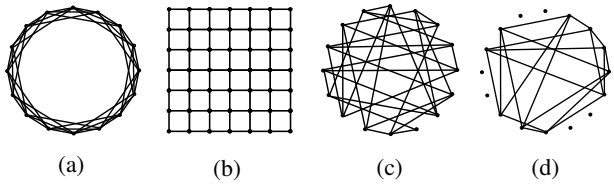
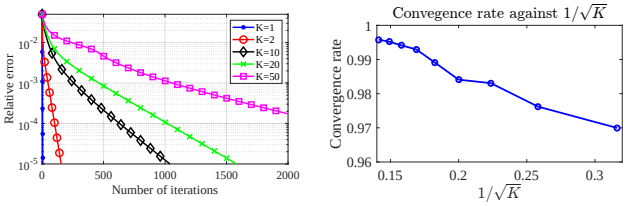


Figure 2. Different graph structures: (a) a connected-cycle graph, (b) a two-dimensional grid, (c) a random regular graph, (d) a random graph with a bounded degree.

During each iteration of Algorithm 1, we use the whole training data to calculate the gradient. The initialization is randomly selected from $\left\{\boldsymbol{W}^{(0)} \middle| \|\boldsymbol{W}^{(0)} - \boldsymbol{W}^*\|_F / \|\boldsymbol{W}^*\|_F < 0.5\right\}$ to reduce the computation. As shown in (Fu et al., 2018; Zhang et al., 2019), the random initialization and the tensor initialization have very similar numerical performance. We consider the learning algorithm to be successful in estimation if the relative error, defined as $\|\boldsymbol{W}^{(t)} - \boldsymbol{W}^*\|_F / \|\boldsymbol{W}^*\|_F$, is less than 10^{-3}, where $\boldsymbol{W}^{(t)}$ is the weight matrix returned by Algorithm 1 when it terminates.

5.1. Convergence rate

We first verify the linear convergence of Algorithm 1, as shown in (14) and (18). Figure 3 (a) and (b) show the convergence rate of Algorithm 1 when varying the number of nodes in the hidden layer K. The dimension d of the feature vectors is chosen as 10, and the sample size $|\Omega|$ is chosen as 2000. We consider the connected-cycle graph in Figure 2 (a) with $\delta = 4$. All cases converge to $\boldsymbol{W}^*$ with the exponential decay. Moreover, from Figure 3, we can also see that the rate of convergence is almost a linear function of $1/\sqrt{K}$. That verifies our theoretical result of the convergence rate of $1 - O(1/\sqrt{K})$ in (16).



(a) (b)

Figure 3. Convergence rate with different K for a connected-circle graph

Figure 4 compares the rates of convergence of AGD and GD in regression problems. We consider a connected-cycle graph with $\delta = 4$. The number of samples $|\Omega| = 500$, $d = 10$, and $K = 5$. Starting with the same initialization, we show the smallest number of iterations needed to reach a certain estimation error, and the results are averaged over 100 independent trials. Both AGD and GD converge linearly. AGD requires a smaller number of the iterations than GD to achieve the same relative error.

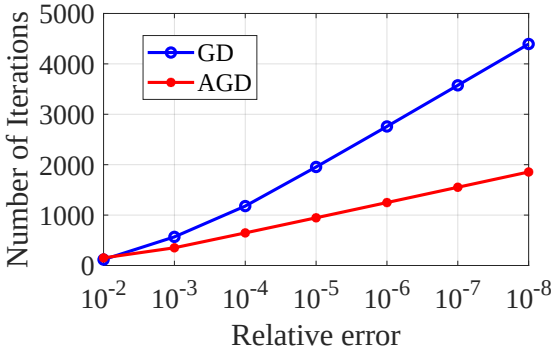
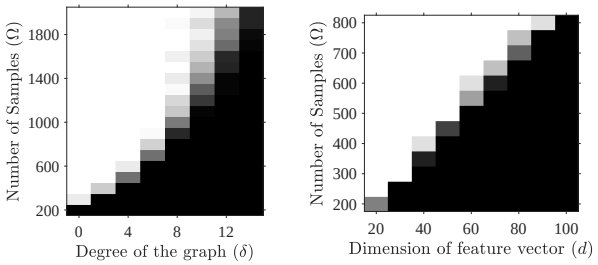


Figure 4. Convergence rates of AGD and GD

5.2. Sample complexity

We next study the influence of d, δ, δ_{ave}, and different graph structures on the estimation performance of Algorithm 1. These relationships are summarized in the sample complexity analyses in (13) and (17) of section 4.1. We have similar numerical results for both regression and classification, and here we only present the regression case.

Figures 5 (a) and (b) show the successful estimation rates when the degree of graph δ and the feature dimension d changes. We consider the connected-cycle graph in Figure 2 (a), and K is kept as 5. d is 40 in Figure 5 (a), and δ is 4 in Figure 5 (b). The results are averaged over 100 independent trials. White block means all trials are successful while black block means all trials fail. We can see that the required number of samples for successful estimation increases as d and δ increases.



(a) (b)

Figure 5. Successful estimation rate for varying the required number of samples, δ, and d in a connected-circle graphs

Figure 6 shows the success rate against the sample size $|\Omega|$ for the random graph in Figure 2(d) with different average node degrees. We vary p to change the average node degree

δ_{ave}. K and d are fixed as 5 and 40, respectively. The successful rate is calculated based on 100 independent trials. We can see that more samples are needed for successful recovery for a larger δ_{ave} when the maximum degree δ is fixed.

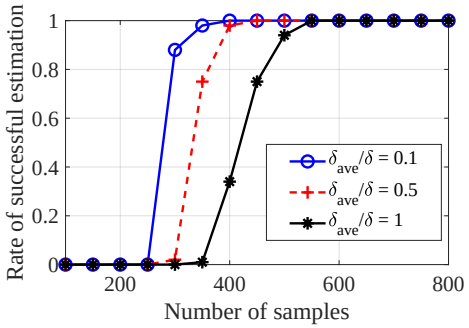


Figure 6. The success rate against the number of samples for different $\delta_{\text{ave}}/\delta$

Figure 7 shows the success rate against $|\Omega|$ for three different graph structures, including a connected cycle, a two-dimensional grid, and a random regular graph in Figure 2 (a), (b), and (c). The maximum degrees of these graphs are all fixed with $\delta = 4$. The average degrees of the connected-circle and the random δ-regular graphs are also $\delta_{\text{ave}} = 4$. δ_{ave} is very close to 4 for the two-dimensional grid when the graph size is large enough, because only the boundary nodes have smaller degrees, and the percentage of boundary nodes decays as the graph size increases. Then from Lemma 1, we have $\sigma_1(\boldsymbol{A})$ is 1 for all these graphs. Although these graphs have different structures, the required numbers of samples to estimate $\boldsymbol{W}^*$ accurately are the same, because both δ and $\sigma_1(\boldsymbol{A})$ are the same. One can verify this property from Figure 7 where all three curves almost coincide.

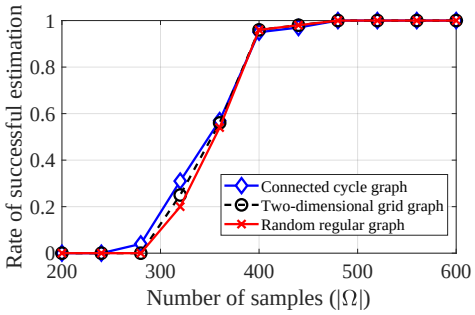


Figure 7. The success rate with respect to sample complexity for various graph structures

5.3. Accuracy in learning $\boldsymbol{W}^*$

We study the learning accuracy of $\boldsymbol{W}^*$, characterized in (14) for regression and (19) for classification. For regression problems, we simulate the general cases when the labels are noisy, i.e., $y_n = z_n + \xi_n$. The noise $\{\xi_n\}_{n=1}^N$ are i.i.d. from $\mathcal{N}(0, \sigma^2)$, and the noise level is measured by σ/E_z, where E_z is the average energy of the noiseless labels $\{z_n\}_{n=1}^N$, calculated as $E_z = \sqrt{\frac{1}{N}\sum_{n=1}^N |z_n|^2}$. The number of hidden nodes K is 5, and the dimension of each feature d is as 60. We consider a connected-circle graph with $\delta = 2$. Figure 8 shows the performance of Algorithm 1 in the noisy case. We can see that when the number of samples exceeds $Kd = 300$, which is the degree of freedom of $\boldsymbol{W}^*$, the relative error decreases dramatically. Also, as N increases, the relative error converges to the noise level. When there is no noise, the estimation of $\boldsymbol{W}^*$ is accurate.

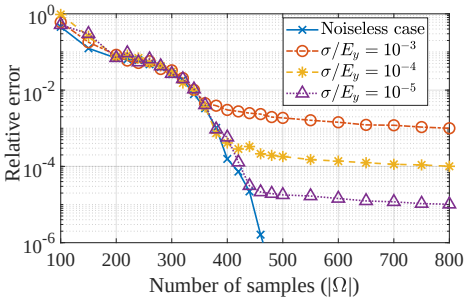


Figure 8. Learning accuracy of Algorithm 1 with noisy measurements for regression

For binary classification problems, Algorithm 1 returns the nearest critical point $\widehat{\boldsymbol{W}}$ to $\boldsymbol{W}^*$. We show the distance between the returned model and the ground-truth model $\boldsymbol{W}^*$ against the number of samples in Figure 9. We consider a connected-cycle graph with the degree $\delta = 2$. $K = 3$ and $d = 20$. The relative error $\|\widehat{\boldsymbol{W}} - \boldsymbol{W}^*\|_F / \|\boldsymbol{W}^*\|_F$ is averaged over 100 independent trials. We can see that the distance between the returned model and the ground-truth model indeed decreases as the number of samples increases.

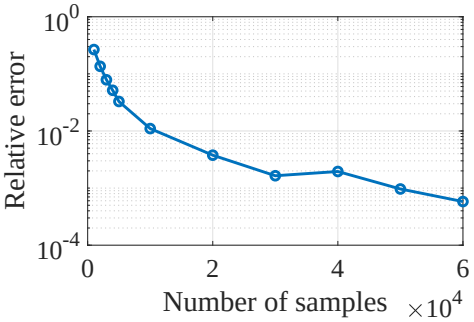


Figure 9. Distance between the returned model by Algorithm1 and the ground-truth model for binary classification

6. Conclusion

Despite the practical success of graph neural networks in learning graph-structured data, the theoretical guarantee of the generalizability of graph neural networks is still elusive. Assuming the existence of a ground-truth model, this paper shows theoretically, for the first time, learning a one-hidden-layer graph neural network with a generation error that is zero for regression or approximately zero for binary classification. With the tensor initialization, we prove that the accelerated gradient descent method converges to the ground-truth model exactly for regression or approximately for binary classification at a linear rate. We also characterize the required number of training samples as a function of the feature dimension, the model size, and the graph structural properties. One future direction is to extend the analysis to multi-hidden-layer neural networks.

7. Acknowledgement

This work was supported by Air Force Office of Scientific Research (AFOSR) FA9550-20-1-0122, National Science Foundation (NSF) 1932196 and the Rensselaer-IBM AI Research Collaboration (http://airc.rpi.edu), part of the IBM AI Horizons Network (http://ibm.biz/AIHorizons).

References

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.

Brutzkus, A. and Globerson, A. Globally optimal gradient descent for a convnet with gaussian inputs. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 605–614. JMLR. org, 2017.

Cao, Y. and Gu, Q. Generalization error bounds of gradient descent for learning overparameterized deep relu networks. 2020.

Chizat, L. and Bach, F. A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*, 8, 2018.

Du, S. S., Lee, J. D., and Tian, Y. When is a convolutional filter easy to learn? *arXiv preprint, http://arxiv.org/abs/1709.06129*, 2017.

Du, S. S., Lee, J. D., Tian, Y., Singh, A., and Poczos, B. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. In *International Conference on Machine Learning*, pp. 1338–1347, 2018.

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. *arXiv preprint arXiv:1802.06463*, 2018.

Ge, R., Lee, J. D., and Ma, T. Learning one-hidden-layer neural networks with landscape design. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BkwHObbRZ.

Gilbert, A. C., Muthukrishnan, S., and Strauss, M. Improved time bounds for near-optimal sparse fourier representation via sampling. In *Proc. SPIE*, 2005.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1263–1272. JMLR. org, 2017.

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

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.

Janson, S. Large deviations for sums of partly dependent random variables. *Random Structures & Algorithms*, 24 (3):234–248, 2004.

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

Kuleshov, V., Chaganty, A., and Liang, P. Tensor factorization via matrix factorization. In *Artificial Intelligence and Statistics*, pp. 507–516, 2015.

Li, Y., Tarlow, D., Brockschmidt, M., and Zemel, R. Gated graph sequence neural networks. *International Conference on Learning Representations (ICLR)*, 2016.

Monfardini, G., Di Massa, V., Scarselli, F., and Gori, M. Graph neural networks for object localization. *Frontiers in Artificial Intelligence and Applications*, 141:665, 2006.

Morris, C., Ritzert, M., Fey, M., Hamilton, W. L., Lenssen, J. E., Rattan, G., and Grohe, M. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4602–4609, 2019.

Nitanda, A. and Suzuki, T. Refined generalization analysis of gradient descent for over-parameterized two-layer neural networks with smooth activations on classification problems. *arXiv preprint arXiv:1905.09870*, 2019.

Polyak, B. T. Introduction to optimization. *New York: Optimization Software, Inc*, 1987.

Safran, I. and Shamir, O. Spurious local minima are common in two-layer relu neural networks. In *International Conference on Machine Learning*, pp. 4430–4438, 2018.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2008.

Shamir, O. Distribution-specific hardness of learning neural networks. *The Journal of Machine Learning Research*, 19(1):1135–1163, 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.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., and Weinberger, K. Simplifying graph convolutional networks. In *International Conference on Machine Learning*, pp. 6861–6871, 2019.

Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K.-i., and Jegelka, S. Representation learning on graphs with jumping knowledge networks. In *International Conference on Machine Learning*, pp. 5453–5462, 2018.

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

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, X., Yu, Y., Wang, L., and Gu, Q. Learning one-hidden-layer relu networks via gradient descent. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1524–1534, 2019.

Zhong, K., Song, Z., and Dhillon, I. S. Learning non-overlapping convolutional neural networks with multiple kernels. *arXiv preprint arXiv:1711.03440*, 2017a.

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, 2017b.

Zhong, K., Song, Z., Jain, P., Bartlett, P. L., and Dhillon, I. S. Recovery guarantees for one-hidden-layer neural networks. *arXiv preprint, http://arxiv.org/abs/1706.03175*, 2017c.

A. Proof of Theorem 1

In this section, before presenting the proof of Theorem 1, we start with defining some useful notations. Recall that in (3), the empirical risk function for linear regression problem is defined as

$$\min_{\boldsymbol{W}}: \quad \hat{f}_{\Omega}(\boldsymbol{W}) = \frac{1}{2|\Omega|}\sum_{n\in\Omega}\Big|y_n - g(\boldsymbol{W};\boldsymbol{a}_n^T\boldsymbol{X})\Big|^2. \tag{22}$$

Population risk function, which is the expectation of the empirical risk function, is defined as

$$\min_{\boldsymbol{W}}: \quad f_{\Omega}(\boldsymbol{W}) = \mathbb{E}_{\boldsymbol{X}}\frac{1}{2|\Omega|}\sum_{n\in\Omega}\Big|y_n - g(\boldsymbol{W};\boldsymbol{a}_n^T\boldsymbol{X})\Big|^2. \tag{23}$$

Then, the road-map of the proof can be summarized in the following three steps.

First, we show the Hessian matrix of the population risk function f_{Ω_t} is positive-definite at ground-truth parameters $\boldsymbol{W}^*$ and then characterize the local convexity region of f_{Ω_t} near $\boldsymbol{W}^*$, which is summarized in Lemma 2.

Second, $\hat{f}_{\Omega_t}$ is non-smooth because of ReLU activation, but f_{Ω_t} is smooth. Hence, we characterize the gradient descent term as $\nabla\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)}) = \langle\nabla^2 f_{\Omega_t}(\widehat{\boldsymbol{W}}^{(t)}), \boldsymbol{W}^{(t)} - \boldsymbol{W}^*\rangle + \big(\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)}) - f_{\Omega_t}(\boldsymbol{W}^{(t)})\big)$. During this step, we need to apply concentration theorem to bound $\nabla\hat{f}_{\Omega_t}$ to its expectation ∇f_{Ω_t} , which is summarized in Lemma 3.

Third, we take the momentum term of $\beta(\boldsymbol{W}^{(t)} - \boldsymbol{W}^{(t-1)})$ into consideration and obtain the following recursive rule:

$$\begin{bmatrix}\boldsymbol{W}^{(t+1)} - \boldsymbol{W}^* \\ \boldsymbol{W}^{(t)} - \boldsymbol{W}^*\end{bmatrix} = \boldsymbol{L}(\beta)\begin{bmatrix}\boldsymbol{W}^{(t)} - \boldsymbol{W}^* \\ \boldsymbol{W}^{(t-1)} - \boldsymbol{W}^*\end{bmatrix}. \tag{24}$$

Then, we know iterates $\boldsymbol{W}^{(t)}$ converge to the ground-truth with a linear rate which is the largest singlar value of matrix $\boldsymbol{L}(\beta)$. Recall that AGD reduces to GD with $\beta = 0$, so our analysis applies to GD method as well. We are able to show the convergence rate of AGD is faster than GD by proving the largest singluar value of $\boldsymbol{L}(\beta)$ is smaller than $\boldsymbol{L}(0)$ for some $\beta > 0$. Lemma 4 provides the estimation error of $\boldsymbol{W}^{(0)}$ and sample complexity to guarantee $\|\boldsymbol{L}(\beta)\|_2$ is less than 1 for $t = 0$.

Lemma 2. *Let f_{Ω_t} be the population risk function in (23) for regression problems, then for any $\boldsymbol{W}$ that satisfies*

$$\|\boldsymbol{W}^* - \boldsymbol{W}\|_2 \le \frac{\varepsilon_0\sigma_K}{44\kappa^2\gamma K^2}, \tag{25}$$

the second-order derivative of f_{Ω_t} is bounded as

$$\frac{(1-\varepsilon_0)\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2}\boldsymbol{I} \preceq \nabla^2 f_{\Omega_t}(\boldsymbol{W}) \preceq \frac{4\sigma_1^2(\boldsymbol{A})}{K}\boldsymbol{I}. \tag{26}$$

Lemma 3. *Let $\hat{f}_{\Omega_t}$ and f_{Ω_t} be the empirical and population risk functions in (22) and (23) for regression problems, respectively. Then, for any fixed point $\boldsymbol{W}$ satisfies (25), we have* [6]

$$\Big\|\nabla f_{\Omega_t}(\boldsymbol{W}) - \nabla\hat{f}_{\Omega_t}(\boldsymbol{W})\Big\|_2 \lesssim \sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}}\|\boldsymbol{W} - \boldsymbol{W}^*\|_2, \tag{27}$$

with probability at least $1 - K^2\cdot N^{-10}$.

Lemma 4. *Assume the number of samples $|\Omega_t| \gtrsim \kappa^3(1+\delta^2)\sigma_1^4(\boldsymbol{A})Kd\log^4 N$, the tensor initialization method via Subroutine 1 outputs $\boldsymbol{W}^{(0)}$ such that*

$$\|\boldsymbol{W}^{(0)} - \boldsymbol{W}^*\|_2 \lesssim \kappa^6\sigma_1^2(\boldsymbol{A})\sqrt{\frac{K^4(1+\delta^2)d\log N}{|\Omega_t|}}\|\boldsymbol{W}^*\|_2 \tag{28}$$

with probability at least $1 - N^{-10}$.

[6] We use $f(d) \gtrsim (\text{ or } \lesssim, \eqsim) g(d)$ to denote there exists some positive constant C such that $f(d) \ge (\text{ or } \le, =) C\cdot g(d)$ when d is sufficiently large.

The proofs of Lemmas 2 and 3 are included in Appendix A.1 and A.2, respectively, while the proof of Lemma 4 can be found in Appendix D. With these three preliminary lemmas on hand, the proof of Theorem 1 is formally summarized in the following contents.

Proof of Theorem 1. The update rule of $\boldsymbol{W}^{(t)}$ is

$$\begin{aligned}\boldsymbol{W}^{(t+1)} =&\boldsymbol{W}^{(t)} - \eta\nabla\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)}) + \beta(\boldsymbol{W}^{(t)} - \boldsymbol{W}^{(t-1)})\\ =&\boldsymbol{W}^{(t)} - \eta\nabla f_{\Omega_t}(\boldsymbol{W}^{(t)}) + \beta(\boldsymbol{W}^{(t)} - \boldsymbol{W}^{(t-1)}) + \eta(\nabla f_{\Omega_t}(\boldsymbol{W}^{(t)}) - \nabla\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)})).\end{aligned} \tag{29}$$

Since $\nabla^2_{\Omega_t}$ is a smooth function, by the intermediate value theorem, we have

$$\begin{aligned}\boldsymbol{W}^{(t+1)} = \boldsymbol{W}^{(t)} &- \eta\nabla^2 f_{\Omega_t}(\widehat{\boldsymbol{W}}^{(t)})(\boldsymbol{W}^{(t)} - \boldsymbol{W}^*)\\ &+ \beta(\boldsymbol{W}^{(t)} - \boldsymbol{W}^{(t-1)})\\ &+ \eta\big(\nabla f_{\Omega_t}(\boldsymbol{W}^{(t)}) - \nabla\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)})\big),\end{aligned} \tag{30}$$

where $\widehat{\boldsymbol{W}}^{(t)}$ lies in the convex hull of $\boldsymbol{W}^{(t)}$ and $\boldsymbol{W}^*$.

Next, we have

$$\begin{bmatrix}\boldsymbol{W}^{(t+1)} - \boldsymbol{W}^*\\ \boldsymbol{W}^{(t)} - \boldsymbol{W}^*\end{bmatrix} = \begin{bmatrix}\boldsymbol{I} - \eta\nabla^2 f_{\Omega_t}(\widehat{\boldsymbol{W}}^{(t)}) + \beta\boldsymbol{I} & \beta\boldsymbol{I}\\ \boldsymbol{I} & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{W}^{(t)} - \boldsymbol{W}^*\\ \boldsymbol{W}^{(t-1)} - \boldsymbol{W}^*\end{bmatrix} + \eta\begin{bmatrix}\nabla f_{\Omega_t}(\boldsymbol{W}^{(t)}) - \nabla\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)})\\ 0\end{bmatrix}. \tag{31}$$

Let $\boldsymbol{L}(\beta) = \begin{bmatrix}\boldsymbol{I} - \eta\nabla^2 f_{\Omega_t}(\widehat{\boldsymbol{W}}^{(t)}) + \beta\boldsymbol{I} & \beta\boldsymbol{I}\\ \boldsymbol{I} & 0\end{bmatrix}$, so we have

$$\left\|\begin{bmatrix}\boldsymbol{W}^{(t+1)} - \boldsymbol{W}^*\\ \boldsymbol{W}^{(t)} - \boldsymbol{W}^*\end{bmatrix}\right\|_2 = \|\boldsymbol{L}(\beta)\|_2\left\|\begin{bmatrix}\boldsymbol{W}^{(t)} - \boldsymbol{W}^*\\ \boldsymbol{W}^{(t-1)} - \boldsymbol{W}^*\end{bmatrix}\right\|_2 + \eta\left\|\begin{bmatrix}\nabla f_{\Omega_t}(\boldsymbol{W}^{(t)}) - \nabla\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)})\\ 0\end{bmatrix}\right\|_2.$$

From Lemma 3, we know that

$$\eta\left\|\nabla f_{\Omega_t}(\boldsymbol{W}^{(t)}) - \nabla\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)})\right\|_2 \lesssim \eta\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}}\|\boldsymbol{W} - \boldsymbol{W}^*\|_2. \tag{32}$$

Then, we have

$$\begin{aligned}\|\boldsymbol{W}^{(t+1)} - \boldsymbol{W}^*\|_2 \lesssim&\left(\|\boldsymbol{L}(\beta)\|_2 + \eta\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}}\right)\|\boldsymbol{W}^{(t)} - \boldsymbol{W}^*\|_2\\ :\eqsim&\nu(\beta)\|\boldsymbol{W}^{(t)} - \boldsymbol{W}^*\|_2.\end{aligned} \tag{33}$$

Let $\nabla^2 f(\widehat{\boldsymbol{W}}^{(t)}) = \boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{S}^T$ be the eigen-decomposition of $\nabla^2 f(\widehat{\boldsymbol{W}}^{(t)})$. Then, we define

$$\widetilde{\boldsymbol{L}}(\beta) := \begin{bmatrix}\boldsymbol{S}^T & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{S}^T\end{bmatrix}\boldsymbol{L}(\beta)\begin{bmatrix}\boldsymbol{S} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{S}\end{bmatrix} = \begin{bmatrix}\boldsymbol{I} - \eta\boldsymbol{\Lambda} + \beta\boldsymbol{I} & \beta\boldsymbol{I}\\ \boldsymbol{I} & 0\end{bmatrix}. \tag{34}$$

Since $\begin{bmatrix}\boldsymbol{S} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{S}\end{bmatrix}\begin{bmatrix}\boldsymbol{S}^T & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{S}^T\end{bmatrix} = \begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix}$, we know $\boldsymbol{L}(\beta)$ and $\widetilde{\boldsymbol{L}}(\beta)$ share the same eigenvalues. Let λ_i be the i-th eigenvalue of $\nabla^2 f_{\Omega_t}(\widehat{\boldsymbol{W}}^{(t)})$, then the corresponding i-th eigenvalue of $\boldsymbol{L}(\beta)$, denoted by $\delta_i(\beta)$, satisfies

$$\delta_i^2 - (1 - \eta\lambda_i + \beta)\delta_i + \beta = 0. \tag{35}$$

Then, we have

$$\delta_i(\beta) = \frac{(1 - \eta\lambda_i + \beta) + \sqrt{(1 - \eta\lambda_i + \beta)^2 - 4\beta}}{2}, \tag{36}$$

and

$$|\delta_i(\beta)| = \begin{cases} \sqrt{\beta}, & \text{if} \quad \beta \geq \left(1-\sqrt{\eta\lambda_i}\right)^2, \\ \frac{1}{2}\left|(1-\eta\lambda_i+\beta)+\sqrt{(1-\eta\lambda_i+\beta)^2-4\beta}\right|, & \text{otherwise.} \end{cases} \tag{37}$$

Note that the other root of (35) is abandoned because the root in (36) is always no less than the other root with $|1-\eta\lambda_i|<1$. By simple calculations, we have

$$\delta_i(0) > \delta_i(\beta), \quad \text{for} \quad \forall \beta \in \left(0,(1-\eta\lambda_i)^2\right). \tag{38}$$

Moreover, δ_i achieves the minimum $\delta_i^* = |1-\sqrt{\eta\lambda_i}|$ when $\beta = \left(1-\sqrt{\eta\lambda_i}\right)^2$.

Let us first assume $\boldsymbol{W}^{(t)}$ satisfies (25), then from Lemma 2, we know that

$$0 < \frac{(1-\varepsilon_0)\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2} \leq \lambda_i \leq \frac{4\sigma_1^2(\boldsymbol{A})}{K}.$$

Let $\gamma_1 = \frac{(1-\varepsilon_0)\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2}$ and $\gamma_2 = \frac{4\sigma_1^2(\boldsymbol{A})}{K}$. If we choose β such that

$$\beta^* = \max\left\{(1-\sqrt{\eta\gamma_1})^2, (1-\sqrt{\eta\gamma_2})^2\right\}, \tag{39}$$

then we have $\beta \geq (1-\sqrt{\eta\lambda_i})^2$ and $\delta_i = \max\left\{|1-\sqrt{\eta\gamma_1}|, |1-\sqrt{\eta\gamma_2}|\right\}$ for any i.

Let $\eta = \frac{1}{2\gamma_2}$, then β^* equals to $\left(1-\sqrt{\frac{\gamma_1}{2\gamma_2}}\right)^2$. Then, for any $\varepsilon_0 \in (0,1/2)$, we have

$$\|\boldsymbol{L}(\beta^*)\|_2 = \max_i \delta_i(\beta^*) = 1-\sqrt{\frac{\gamma_1}{2\gamma_2}} = 1-\sqrt{\frac{1-\varepsilon_0}{88\kappa^2\gamma K}} \leq 1-\frac{1-(3/4)\cdot\varepsilon_0}{\sqrt{88\kappa^2\gamma K}}. \tag{40}$$

Then, let

$$\eta\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}} \lesssim \frac{\varepsilon_0}{4\sqrt{88\kappa^2\gamma K}}, \tag{41}$$

we need $|\Omega_t| \gtrsim \varepsilon_0^{-2}\kappa^2\gamma M(1+\delta^2)\sigma_1^2(\boldsymbol{A})K^3 d\log N$. Combining (40) and (41), we have

$$\nu(\beta^*) \leq 1-\frac{1-\varepsilon_0}{\sqrt{88\kappa^2\gamma K}}. \tag{42}$$

Let $\beta = 0$, we have

$$\nu(0) \geq \|\boldsymbol{A}(0)\|_2 = 1-\frac{1-\varepsilon_0}{88\kappa^2\gamma K},$$

$$\nu(0) \lesssim \|\boldsymbol{A}(0)\|_2 + \eta\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}} \leq 1-\frac{1-2\varepsilon_0}{88\kappa^2\gamma K}$$

if $|\Omega_t| \gtrsim \varepsilon_0^{-2}\kappa^2\gamma M(1+\delta^2)\sigma_1^2(\boldsymbol{A})K^3 d\log N$.

Hence, with $\eta = \frac{1}{2\gamma_2}$ and $\beta = \left(1-\frac{\gamma_1}{2\gamma_2}\right)^2$, we have

$$\|\boldsymbol{W}^{(t+1)} - \boldsymbol{W}^*\|_2 \leq \left(1-\frac{1-\varepsilon_0}{\sqrt{88\kappa^2\gamma K}}\right)\|\boldsymbol{W}^{(t)} - \boldsymbol{W}^*\|_2, \tag{43}$$

provided $\boldsymbol{W}^{(t)}$ satisfies (25), and

$$|\Omega_t| \gtrsim \varepsilon_0^{-2}\kappa^2\gamma(1+\delta^2)\sigma_1^4(\boldsymbol{A})K^3 d\log N. \tag{44}$$

Then, we can start mathematical induction of (43) over t.

Base case: According to Lemma 4, we know that (25) holds for $\boldsymbol{W}^{(0)}$ if

$$|\Omega_t| \gtrsim \varepsilon_0^{-2}\kappa^9\gamma^2(1+\delta^2)\sigma_1^4(\boldsymbol{A})K^8 d\log N. \tag{45}$$

According to Theorem 1, it is clear that the number of samples $|\Omega_t|$ satisfies (45), then (25) indeed holds for $t = 0$. Since (25) holds for $t = 0$ and $|\Omega_t|$ in Theorem 1 satisfies (44) as well, we have (43) holds for $t = 0$.

Induction step: Assuming (43) holds for $\boldsymbol{W}^{(t)}$, we need to show that (43) holds for $\boldsymbol{W}^{(t+1)}$. That is to say, we need $|\Omega_t|$ satisfies (44), which holds naturally from Theorem 1.

Therefore, when $|\Omega_t| \gtrsim \varepsilon_0^{-2}\kappa^9\gamma^2(1+\delta^2)\sigma_1^4(\boldsymbol{A})K^8 d\log N$, we know that (43) holds for all $0 \le t \le T-1$ with probability at least $1 - K^2T \cdot N^{-10}$. By simple calculations, we can obtain

$$\|\boldsymbol{W}^{(T)} - \boldsymbol{W}^*\|_2 \le \Big(1 - \frac{1-\varepsilon_0}{\sqrt{88\kappa^2\gamma K}}\Big)^T \|\boldsymbol{W}^{(0)} - \boldsymbol{W}^*\|_2 \tag{46}$$

□

A.1. Proof of Lemma 2

In this section, we provide the proof of Lemma 2 which shows the local convexity of f_{Ω_t} in a small neighborhood of $\boldsymbol{W}^*$. The roadmap is to first bound the smallest eigenvalue of $\nabla^2 f_{\Omega_t}$ in the ground truth as shown in Lemma 5, then show that the difference of $\nabla^2 f_{\Omega_t}$ between any fixed point $\boldsymbol{W}$ in this region and the ground truth $\boldsymbol{W}^*$ is bounded in terms of $\|\boldsymbol{W} - \boldsymbol{W}^*\|_2$ by Lemma 6.

Lemma 5. *The second-order derivative of f_{Ω_t} at the ground truth $\boldsymbol{W}^*$ satisfies*

$$\frac{\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2} \preceq \nabla^2 f_{\Omega_t}(\boldsymbol{W}^*) \preceq \frac{3\sigma_1^2(\boldsymbol{A})}{K}. \tag{47}$$

Lemma 6. *Suppose $\boldsymbol{W}$ satisfies (25), we have*

$$\big\|\nabla^2 f_{\Omega_t}(\boldsymbol{W}) - \nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\big\|_2 \le 4\sigma_1^2(\boldsymbol{A})\frac{\|\boldsymbol{W}^* - \boldsymbol{W}\|_2}{\sigma_K}. \tag{48}$$

The proofs of Lemmas 5 and 6 can be found in Sec. A.3. With these two preliminary lemmas on hand, the proof of Lemma 2 is formally summarized in the following contents.

Proof of Lemma 2. By the triangle inequality, we have

$$\Big|\big\|\nabla^2 f_{\Omega_t}(\boldsymbol{W})\big\|_2 - \big\|\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\big\|_2\Big| \le \|\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*) - \nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2,$$

and

$$\begin{aligned}\big\|\nabla^2 f_{\Omega_t}(\boldsymbol{W})\big\|_2 &\le \big\|\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\big\|_2 + \|\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*) - \nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2,\\ \big\|\nabla^2 f_{\Omega_t}(\boldsymbol{W})\big\|_2 &\ge \big\|\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\big\|_2 - \|\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*) - \nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2.\end{aligned}$$

The error bound of $\|\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*) - \nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2$ can be derived from Lemma 6, and the error bound of $\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)$ is provided in Lemma 5.

Therefore, for any $\boldsymbol{W}$ satisfies (25), we have

$$\frac{(1-\varepsilon_0)\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2} \le \big\|\nabla^2 f_{\Omega_t}(\boldsymbol{W})\big\|_2 \le \frac{4\sigma_1^2(\boldsymbol{A})}{K}. \tag{49}$$

□

A.2. Proof of Lemma 3

The proof of Lemma 3 is mainly to bound the concentration error of random variables $\boldsymbol{z}_n(j,k)$ as shown in (60). We first show that $\boldsymbol{z}_n(j,k)$ is a sub-exponential random variable, and the definitions of sub-Gaussian and sub-exponential random variables are provided in Definitions 1 and 2. Though Hoeffding's inequality provides the concentration error for sum of independent random variables, random variables $\boldsymbol{z}_n(j,k)$ with different j,k are not independent. Hence, we introduce Lemma 7 to provide the upper bound for the moment generation function of the sum of partly dependent random variables and then apply standard Chernoff inequality. Lemmas 8 and 9 are standard tools in analyzing spectral norms of high-demensional random matrices.

Definition 1 (Definition 5.7, (Vershynin, 2010)). *A random variable X is called a sub-Gaussian random variable if it satisfies*

$$(\mathbb{E}|X|^p)^{1/p} \leq c_1\sqrt{p} \tag{50}$$

for all $p \geq 1$ and some constant $c_1 > 0$. In addition, we have

$$\mathbb{E}e^{s(X-\mathbb{E}X)} \leq e^{c_2\|X\|_{\psi_2}^2 s^2} \tag{51}$$

for all $s \in \mathbb{R}$ and some constant $c_2 > 0$, where $\|X\|_{\psi_2}$ is the sub-Gaussian norm of X defined as $\|X\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2}(\mathbb{E}|X|^p)^{1/p}$.

Moreover, a random vector $\boldsymbol{X} \in \mathbb{R}^d$ belongs to the sub-Gaussian distribution if one-dimensional marginal $\boldsymbol{\alpha}^T\boldsymbol{X}$ is sub-Gaussian for any $\boldsymbol{\alpha} \in \mathbb{R}^d$, and the sub-Gaussian norm of $\boldsymbol{X}$ is defined as $\|\boldsymbol{X}\|_{\psi_2} = \sup_{\|\boldsymbol{\alpha}\|_2=1} \|\boldsymbol{\alpha}^T\boldsymbol{X}\|_{\psi_2}$.

Definition 2 (Definition 5.13, (Vershynin, 2010)). *A random variable X is called a sub-exponential random variable if it satisfies*

$$(\mathbb{E}|X|^p)^{1/p} \leq c_3 p \tag{52}$$

for all $p \geq 1$ and some constant $c_3 > 0$. In addition, we have

$$\mathbb{E}e^{s(X-\mathbb{E}X)} \leq e^{c_4\|X\|_{\psi_1}^2 s^2} \tag{53}$$

for $s \leq 1/\|X\|_{\psi_1}$ and some constant $c_4 > 0$, where $\|X\|_{\psi_1}$ is the sub-exponential norm of X defined as $\|X\|_{\psi_1} = \sup_{p\geq 1} p^{-1}(\mathbb{E}|X|^p)^{1/p}$.

Lemma 7. *Given a sampling set $\mathcal{X} = \{x_n\}_{n=1}^N$ that contains N partly dependent random variables, for each $n \in [N]$, suppose x_n is dependent with at most $d_{\mathcal{X}}$ random variables in $\mathcal{X}$ (including x_n itself), and the moment generate function of x_n satisfies $\mathbb{E}_{x_n} e^{sx_n} \leq e^{Cs^2}$ for some constant C that may depend on x_n. Then, the moment generation function of $\sum_{n=1}^N x_n$ is bounded as*

$$\mathbb{E}_{\mathcal{X}} e^{s\sum_{n=1}^N x_n} \leq e^{Cd_{\mathcal{X}}Ns^2}. \tag{54}$$

Lemma 8 (Lemma 5.2, (Vershynin, 2010)). *Let $\mathcal{B}(0,1) \in \{\boldsymbol{\alpha}\big|\|\boldsymbol{\alpha}\|_2 = 1, \boldsymbol{\alpha} \in \mathbb{R}^d\}$ denote a unit ball in $\mathbb{R}^d$. Then, a subset $\mathcal{S}_\xi$ is called a ξ-net of $\mathcal{B}(0,1)$ if every point $\boldsymbol{z} \in \mathcal{B}(0,1)$ can be approximated to within ξ by some point $\boldsymbol{\alpha} \in \mathcal{B}(0,1)$, i.e. $\|\boldsymbol{z} - \boldsymbol{\alpha}\|_2 \leq \xi$. Then the minimal cardinality of a ξ-net $\mathcal{S}_\xi$ satisfies*

$$|\mathcal{S}_\xi| \leq (1 + 2/\xi)^d. \tag{55}$$

Lemma 9 (Lemma 5.3, (Vershynin, 2010)). *Let $\boldsymbol{A}$ be an $N \times d$ matrix, and let $\mathcal{S}_\xi$ be a ξ-net of $\mathcal{B}(0,1)$ in $\mathbb{R}^d$ for some $\xi \in (0,1)$. Then*

$$\|\boldsymbol{A}\|_2 \leq (1-\xi)^{-1} \max_{\boldsymbol{\alpha}\in\mathcal{S}_\xi} |\boldsymbol{\alpha}^T\boldsymbol{A}\boldsymbol{\alpha}|. \tag{56}$$

The proof of Lemma 7 can be found in Appendex A.3. With these preliminary Lemmas and definition on hand, the proof of Lemma 3 is formally summarized in the following contents.

Proof of Lemma 3 . We have

$$\hat{f}_{\Omega_t}(\boldsymbol{W}) = \frac{1}{2|\Omega_t|}\sum_{n\in\Omega_t}\Big|y_n - g(\boldsymbol{W};\boldsymbol{a}_n^T\boldsymbol{X})\Big|^2 = \frac{1}{2|\Omega_t|}\sum_{n\in\Omega_t}\Big|y_n - \sum_{j=1}^K \phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j)\Big|^2, \tag{57}$$

and

$$f_{\Omega_t}(\boldsymbol{W}) = \mathbb{E}_{\boldsymbol{X}}\hat{f}_{\Omega_t}(\boldsymbol{W}) = \frac{1}{2|\Omega_t|}\sum_{n\in\Omega_t}\mathbb{E}_{\boldsymbol{x}}\Big|y_n - \sum_{j=1}^K \phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j)\Big|^2. \tag{58}$$

The gradients of $\hat{f}_{\Omega_t}$ are

$$\begin{aligned}\frac{\partial \hat{f}_{\Omega_t}}{\partial \boldsymbol{w}_k}(\boldsymbol{W}) =& \frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\Big(y_n - \sum_{j=1}^{K}\phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j)\Big)\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_k)\\ =& \frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\Big(\sum_{j=1}^{K}\phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j^*) - \sum_{j=1}^{K}\phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j)\Big)\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_k)\\ =& \sum_{j=1}^{K}\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\big(\phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j^*) - \phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j)\big)\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_k).\end{aligned} \tag{59}$$

Let us define

$$\boldsymbol{z}_n(k,j) = \boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_k)\big(\phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j^*) - \phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j)\big), \tag{60}$$

then for any normalized $\boldsymbol{\alpha}\in\mathbb{R}^d$, we have

$$\begin{aligned}&p^{-1}\Big(\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_k)\big(\phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j^*) - \phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j)\big)\big|^p\Big)^{1/p}\\ \le& p^{-1}\Big(\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\big|^{2p}\cdot\mathbb{E}_{\boldsymbol{X}}\big|\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_k)\big(\phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j^*) - \phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j)\big)\big|^{2p}\Big)^{1/2p}\\ \le& p^{-1}\Big(\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\big|^{2p}\Big)^{1/2p}\cdot\Big(\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{a}_n^T\boldsymbol{X}(\boldsymbol{w}_j^* - \boldsymbol{w}_j)\big|^{2p}\Big)^{1/2p}\end{aligned} \tag{61}$$

where the first inequality comes from the Cauchy-Schwarz inequality. Furthermore, $\boldsymbol{a}_n^T\boldsymbol{X}$ belongs to the Gaussian distribution and thus is a sub-Gaussian random vector as well. Then, from Definition 1, we have

$$\begin{aligned}&\Big(\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\big|^{2p}\Big)^{1/2p} \le (2p)^{1/2}\|\boldsymbol{X}^T\boldsymbol{a}_n\|_{\psi_2} \le (2p)^{1/2}\|\boldsymbol{a}_n\|_2,\\ \text{and}\quad&\Big(\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{a}_n^T\boldsymbol{X}(\boldsymbol{w}_j^* - \boldsymbol{w}_j)\big|^{2p}\Big)^{1/2p} \le (2p)^{1/2}\|\boldsymbol{a}_n\|_2\cdot\|\boldsymbol{w}_j^* - \boldsymbol{w}_j\|_2.\end{aligned} \tag{62}$$

Then, we have

$$\begin{aligned}&p^{-1}\Big(\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_k)\big(\phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j^*) - \phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j)\big)\big|^p\Big)^{1/p}\\ \le& p^{-1}\cdot 2p\|\boldsymbol{a}_n\|_2^2\cdot\|\boldsymbol{w}_j^* - \boldsymbol{w}_j\|_2\\ \le& 2\sigma_1^2(\boldsymbol{A})\cdot\|\boldsymbol{w}_j^* - \boldsymbol{w}_j\|_2.\end{aligned} \tag{63}$$

Therefore, from Definition 2, $\boldsymbol{z}_n(k,j)$ belongs to the sub-exponential distribution with

$$\|\boldsymbol{z}_n\|_{\phi_1} \le 2\sigma_1^2(\boldsymbol{A})\cdot\|\boldsymbol{w}_j^* - \boldsymbol{w}_j\|_2. \tag{64}$$

Recall that each node is connected with at most δ other nodes. Hence, for any fixed $\boldsymbol{z}_n$, there are at most $(1+\delta^2)$ (including $\boldsymbol{z}_n$ itself) elements in $\{\boldsymbol{z}_l | l\in\Omega_t\}$ are dependant with $\boldsymbol{z}_n$. From Lemma 7, the moment generation function of $\sum_{n\in\Omega_t}(\boldsymbol{z}_n - \mathbb{E}_{\boldsymbol{X}}\boldsymbol{z}_n)$ satisfies

$$\mathbb{E}_{\boldsymbol{X}}e^{s\sum_{n\in\Omega_t}(\boldsymbol{z}_n - \mathbb{E}_{\boldsymbol{X}}\boldsymbol{z}_n)} \le e^{C(1+\delta^2)|\Omega_t|s^2}. \tag{65}$$

By Chernoff inequality, we have

$$\text{Prob}\Big\{\Big\|\frac{1}{|\Omega_t|}\sum_{n\in\Omega_t}\big(\boldsymbol{z}_n(k,j) - \mathbb{E}_{\boldsymbol{X}}\boldsymbol{z}_n(k,j)\big)\Big\|_2 > t\Big\} \le \frac{e^{C(1+\delta^2)|\Omega_t|s^2}}{e^{|\Omega_t|ts}} \tag{66}$$

for any $s>0$.

Let $s = t/\big(C(1+\delta^2)\|\boldsymbol{z}_n\|_{\phi_1}^2\big)$ and $t = \sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}}\|\boldsymbol{z}_n\|_{\phi_1}$, we have

$$\begin{aligned}\Big\|\frac{1}{|\Omega_t|}\sum_{n\in\Omega_t}\big(\boldsymbol{z}_n(k,j)-\mathbb{E}_{\boldsymbol{X}}\boldsymbol{z}_n(k,j)\big)\Big\|_2 \leq & C\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}}\sigma_1^2(\boldsymbol{A})\cdot\|\boldsymbol{w}_j^*-\boldsymbol{w}_j\|_2\\ \leq & C\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}}\cdot\|\boldsymbol{W}^*-\boldsymbol{W}\|_2\end{aligned} \tag{67}$$

with probability at least $1-N^{-d}$.

In conclusion, by selecting $\xi=\frac{1}{2}$ in Lemmas 8 and 9, we have

$$\begin{aligned}\Big\|\frac{\partial\hat{f}_{\Omega_t}}{\partial\boldsymbol{w}_k}(\boldsymbol{W})-\frac{\partial f_{\Omega_t}}{\partial\boldsymbol{w}_k}(\boldsymbol{W})\Big\|_2 \leq & \sum_{k=1}^K\sum_{j=1}^K\frac{1}{K^2}\Big\|\frac{1}{|\Omega_t|}\sum_{n\in\Omega_t}\boldsymbol{z}_n(k,j)-\mathbb{E}_{\boldsymbol{X}}\boldsymbol{z}_n(k,j)\Big\|_2\\ \leq & C\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}}\cdot\|\boldsymbol{W}^*-\boldsymbol{W}\|_2\end{aligned} \tag{68}$$

with probability at least $1-\big(\frac{5}{N}\big)^d$. □

A.3. Proof of auxiliary lemmas for regression problems

A.3.1. Proof of Lemma 5

Proof of Lemma 5 . For any normalized $\boldsymbol{\alpha}\in\mathbb{R}^{Kd}$, the lower bound of $\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)$ is derived from

$$\begin{aligned}\boldsymbol{\alpha}^T\nabla^2 f(\boldsymbol{W}^*)\boldsymbol{\alpha} = & \frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\mathbb{E}_{\boldsymbol{X}}\Big[\Big(\sum_{j=1}^K\boldsymbol{\alpha}_j^T\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j^*)\Big)^2\Big]\\ \geq & \frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\frac{\|\boldsymbol{a}_n\|_2^2}{11\kappa^2\gamma}\|\boldsymbol{\alpha}\|_2^2=\frac{\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2},\end{aligned} \tag{69}$$

where the last inequality can be derived from Lemma D.6 in (Zhong et al., 2017c). In spite that the error bound in (Zhong et al., 2017c) is given in terms of $\boldsymbol{x}_n$ instead of $\boldsymbol{X}^T\boldsymbol{a}_n$, both $\boldsymbol{x}_n$ and $\boldsymbol{X}^T\boldsymbol{a}_n$ belong to Gaussian distribution. Hence, we can follow the similar steps in (Zhong et al., 2017c) to derive the results for Gaussian random variable $\boldsymbol{X}^T\boldsymbol{a}_n$ with 0 mean and $\|\boldsymbol{a}_n\|_2^2$ variance.

Next, the upper bound of $\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)$ is derived from

$$\begin{aligned}&\boldsymbol{\alpha}^T\nabla^2 f(\boldsymbol{W}^*)\boldsymbol{\alpha}\\ =&\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\mathbb{E}_{\boldsymbol{X}}\Big[\Big(\sum_{j=1}^K\boldsymbol{\alpha}_j^T\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j^*)\Big)^2\Big]\\ =&\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\sum_{j_1=1}^K\sum_{j_2=1}^K\mathbb{E}_{\boldsymbol{X}}\Big[\boldsymbol{\alpha}_{j_1}^T\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\boldsymbol{\alpha}_{j_2}^T\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2}^*)\Big]\\ \leq&\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\sum_{j_1=1}^K\sum_{j_2=1}^K\Big[\mathbb{E}_{\boldsymbol{X}}|\boldsymbol{\alpha}_{j_1}^T\boldsymbol{X}^T\boldsymbol{a}_n|^4\cdot\mathbb{E}_{\boldsymbol{X}}|\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)|^4\cdot\mathbb{E}_{\boldsymbol{X}}|\boldsymbol{\alpha}_{j_2}^T\boldsymbol{X}^T\boldsymbol{a}_n|^4\cdot\mathbb{E}_{\boldsymbol{X}}|\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2}^*)|^4\Big]^{\frac{1}{4}}\\ \leq&\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\sum_{j_1=1}^K\sum_{j_2=1}^K 3\sigma_1^2(\boldsymbol{A})\|\boldsymbol{\alpha}_{j_1}\|_2\|\boldsymbol{\alpha}_{j_2}\|_2\\ \leq&3\sigma_1^2(\boldsymbol{A})\frac{\|\boldsymbol{\alpha}\|_2^2}{K},\end{aligned} \tag{70}$$

which completes the proof. □

A.3.2. PROOF OF LEMMA 6

Proof of Lemma 6 . The second-order derivative of f_{Ω_t} is written as

$$\begin{aligned}
&\frac{\partial^2 f_{\Omega_t}}{\partial \boldsymbol{w}_{j_1}\partial \boldsymbol{w}_{j_2}}(\boldsymbol{W}) - \frac{\partial^2 f_{\Omega_t}}{\partial \boldsymbol{w}_{j_1}\partial \boldsymbol{w}_{j_2}}(\boldsymbol{W}^*)\\
=&\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\mathbb{E}_{\boldsymbol{X}}(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T\Big[\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1})\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2}) - \phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2}^*)\Big]\\
=&\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\mathbb{E}_{\boldsymbol{X}}(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T\big(\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}) - \phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\big)\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2})\\
&-\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\mathbb{E}_{\boldsymbol{X}}(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\big(\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2}^*) - \phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2})\big).
\end{aligned} \tag{71}$$

For any normalized $\boldsymbol{\alpha}\in\mathbb{R}^d$, we have

$$\begin{aligned}
&\Big|\boldsymbol{\alpha}^T\Big[\frac{\partial^2 f_{\Omega_t}}{\partial \boldsymbol{w}_{j_1}\partial \boldsymbol{w}_{j_2}}(\boldsymbol{W}) - \frac{\partial^2 f_{\Omega_t}}{\partial \boldsymbol{w}_{j_1}\partial \boldsymbol{w}_{j_2}}(\boldsymbol{W}^*)\Big]\boldsymbol{\alpha}\Big|\\
\le&\Big|\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\mathbb{E}_{\boldsymbol{X}}(\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2\big(\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}) - \phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\big)\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2})\Big|\\
&+\Big|\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\mathbb{E}_{\boldsymbol{X}}(\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\big(\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2}^*) - \phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2})\big)\Big|\\
\le&\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\big|^2\cdot\Big|\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}) - \phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\Big|\\
&+\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\big|^2\cdot\Big|\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2}^*) - \phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2})\Big|.
\end{aligned} \tag{72}$$

It is easy to verify there exists a basis such that $\mathcal{B}=\{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\alpha}_4^\perp,\cdots,\boldsymbol{\alpha}_d^\perp\}$ with $\{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}\}$ spanning a subspace that contains $\boldsymbol{\alpha}$, $\boldsymbol{w}_{j_1}$ and $\boldsymbol{w}_{j_1}^*$. Then, for any $\boldsymbol{X}^T\boldsymbol{a}_n\in\mathbb{R}^d$, we have a unique $\boldsymbol{z}=\begin{bmatrix}z_1 & z_2 & \cdots & z_d\end{bmatrix}^T$ such that

$$\boldsymbol{X}^T\boldsymbol{a}_n = z_1\boldsymbol{\alpha}+z_2\boldsymbol{\beta}+z_3\boldsymbol{\gamma}+\cdots+z_d\boldsymbol{\alpha}_d^\perp.$$

Also, since $\boldsymbol{X}^T\boldsymbol{a}_n\sim\mathcal{N}(\boldsymbol{0},\|\boldsymbol{a}_n\|_2^2\boldsymbol{I}_d)$, we have $\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{0},\|\boldsymbol{a}_n\|_2^2\boldsymbol{I}_d)$. Then, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\big|^2\cdot\Big|\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}) - \phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\Big|\\
=&\mathbb{E}_{z_1,z_2,z_3}|\phi'\big(\boldsymbol{w}_{j_1}^T\widetilde{\boldsymbol{x}}\big) - \phi'\big({\boldsymbol{w}_{j_1}^*}^T\widetilde{\boldsymbol{x}}\big)|\cdot|\boldsymbol{a}^T\widetilde{\boldsymbol{x}}|^2\\
=&\int|\phi'\big(\boldsymbol{w}_{j_1}^T\widetilde{\boldsymbol{x}}\big) - \phi'\big({\boldsymbol{w}_{j_1}^*}^T\widetilde{\boldsymbol{x}}\big)|\cdot|\boldsymbol{a}^T\widetilde{\boldsymbol{x}}|^2\cdot f_Z(z_1,z_2,z_3)dz_1dz_2dz_3,
\end{aligned}$$

where $\widetilde{\boldsymbol{x}}=z_1\boldsymbol{\alpha}+z_2\boldsymbol{\beta}+z_3\boldsymbol{\gamma}$ and $f_Z(z_1,z_2,z_3)$ is the probability density function of (z_1,z_2,z_3). Next, we consider spherical coordinates with $z_1=r\cos\phi_1, z_2=r\sin\phi_1\sin\phi_2, z_3=z_2=r\sin\phi_1\cos\phi_2$. Hence,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\big|^2\cdot\Big|\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}) - \phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\Big|\\
=&\int\int\int|\phi'\big(\boldsymbol{w}_{j_1}^T\widetilde{\boldsymbol{x}}\big) - \phi'\big({\boldsymbol{w}_{j_1}^*}^T\widetilde{\boldsymbol{x}}\big)|\cdot|r\cos\phi_1|^2\cdot f_Z(r,\phi_1,\phi_2)r^2\sin\phi_1 dr d\phi_1 d\phi_2.
\end{aligned} \tag{73}$$

It is easy to verify that $\phi'\big(\boldsymbol{w}_{j_1}^T\widetilde{\boldsymbol{x}}\big)$ only depends on the direction of $\widetilde{\boldsymbol{x}}$ and

$$f_Z(r,\phi_1,\phi_2)=\frac{1}{(2\pi\|\boldsymbol{a}_n\|_2^2)^{\frac{3}{2}}}e^{-\frac{x_1^2+x_2^2+x_3^2}{2\|\boldsymbol{a}_n\|_2^2}}=\frac{1}{(2\pi\|\boldsymbol{a}_n\|_2^2)^{\frac{3}{2}}}e^{-\frac{r^2}{2\|\boldsymbol{a}_n\|_2^2}}$$

only depends on r. Then, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}}\left|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\right|^2\cdot\left|\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1})-\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\right|\\
=&\int\int\int|\phi'\big(\boldsymbol{w}_{j_1}^T(\widetilde{\boldsymbol{x}}/r)\big)-\phi'\big(\boldsymbol{w}_{j_1}^{*\,T}(\widetilde{\boldsymbol{x}}/r)\big)|\cdot|r\cos\phi_1|^2\cdot f_Z(r)r^2\sin\phi_1 dr d\phi_1 d\phi_2\\
=&\int_0^\infty r^4 f_z(r)dr\int_0^\pi\int_0^{2\pi}|\cos\phi_1|^2\cdot\sin\phi_1\cdot|\phi'\big(\boldsymbol{w}_{j_2}^T(\widetilde{\boldsymbol{x}}/r)\big)-\phi'\big(\boldsymbol{w}_{j_2}^{*\,T}(\widetilde{\boldsymbol{x}}/r)\big)|d\phi_1 d\phi_2\\
\leq&3\|\boldsymbol{a}_n\|_2^2\cdot\int_0^\infty r^2 f_z(r)dr\int_0^\pi\int_0^{2\pi}\sin\phi_1\cdot|\phi'\big(\boldsymbol{w}_{j_2}^T(\widetilde{\boldsymbol{x}}/r)\big)-\phi'\big(\boldsymbol{w}_{j_2}^{*\,T}(\widetilde{\boldsymbol{x}}/r)\big)|d\phi_1 d\phi_2\\
=&3\|\boldsymbol{a}_n\|_2^2\cdot\mathbb{E}_{z_1,z_2,z_3}\left|\phi'\big(\boldsymbol{w}_{j_1}^T\widetilde{\boldsymbol{x}}\big)-\phi'\big(\boldsymbol{w}_{j_1}^{*\,T}\widetilde{\boldsymbol{x}}\big)\right|\\
=&3\|\boldsymbol{a}_n\|_2^2\cdot\mathbb{E}_{\boldsymbol{X}}\left|\phi'\big(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}\big)-\phi'\big(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*\big)\right|
\end{aligned}\tag{74}$$

Define a set $\mathcal{A}_1=\{\boldsymbol{x}|(\boldsymbol{w}_{j_1}^{*\,T}\boldsymbol{x})(\boldsymbol{w}_{j_1}{}^T\boldsymbol{x})<0\}$. If $\boldsymbol{x}\in\mathcal{A}_1$, then $\boldsymbol{w}_{j_1}^{*\,T}\boldsymbol{x}$ and $\boldsymbol{w}_{j_1}{}^T\boldsymbol{x}$ have different signs, which means the value of $\phi'(\boldsymbol{w}_{j_1}^T\boldsymbol{x})$ and $\phi'(\boldsymbol{w}_{j_1}^{*\,T}\boldsymbol{x})$ are different. This is equivalent to say that

$$|\phi'(\boldsymbol{w}_{j_1}^T\boldsymbol{x})-\phi'(\boldsymbol{w}_{j_1}^{*\,T}\boldsymbol{x})|=\begin{cases}1,\ \text{if}\ \boldsymbol{x}\in\mathcal{A}_1\\0,\ \text{if}\ \boldsymbol{x}\in\mathcal{A}_1^c\end{cases}.\tag{75}$$

Moreover, if $\boldsymbol{x}\in\mathcal{A}_1$, then we have

$$|\boldsymbol{w}_{j_1}^{*\,T}\boldsymbol{x}|\leq|\boldsymbol{w}_{j_1}^{*\,T}\boldsymbol{x}-\boldsymbol{w}_{j_1}{}^T\boldsymbol{x}|\leq\|\boldsymbol{w}_{j_1}^*-\boldsymbol{w}_{j_1}\|\cdot\|\boldsymbol{x}\|.\tag{76}$$

Define a set $\mathcal{A}_2$ such that

$$\mathcal{A}_2=\Big\{\boldsymbol{x}\Big|\frac{|\boldsymbol{w}_{j_1}^{*\,T}\boldsymbol{x}|}{\|\boldsymbol{w}_{j_1}^*\|\|\boldsymbol{x}\|}\leq\frac{\|\boldsymbol{w}_{j_1}^*-\boldsymbol{w}_{j_1}\|}{\|\boldsymbol{w}_{j_1}^*\|}\Big\}=\Big\{\theta_{\boldsymbol{x},\boldsymbol{w}_{j_1}^*}\Big||\cos\theta_{\boldsymbol{x},\boldsymbol{w}_{j_1}^*}|\leq\frac{\|\boldsymbol{w}_{j_1}^*-\boldsymbol{w}_{j_1}\|}{\|\boldsymbol{w}_{j_1}^*\|}\Big\}.\tag{77}$$

Hence, we have that

$$\mathbb{E}_{\boldsymbol{x}}|\phi'(\boldsymbol{w}_{j_1}^T\boldsymbol{x})-\phi'(\boldsymbol{w}_{j_1}^{*\,T}\boldsymbol{x}_{i_2})|=\text{Prob}(\boldsymbol{x}\in\mathcal{A}_1)\leq\text{Prob}(\boldsymbol{x}\in\mathcal{A}_2).\tag{78}$$

Since $\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$, $\theta_{\boldsymbol{x},\boldsymbol{w}_{j_1}^*}$ belongs to the uniform distribution on $[-\pi,\pi]$, we have

$$\begin{aligned}
\text{Prob}(\boldsymbol{x}\in\mathcal{A}_2)=&\frac{\pi-\arccos\frac{\|\boldsymbol{w}_{j_1}^*-\boldsymbol{w}_{j_1}\|}{\|\boldsymbol{w}_{j_1}^*\|}}{\pi}\\
\leq&\frac{1}{\pi}\tan(\pi-\arccos\frac{\|\boldsymbol{w}_{j_1}^*-\boldsymbol{w}_{j_1}\|}{\|\boldsymbol{w}_{j_1}^*\|})\\
=&\frac{1}{\pi}\cot(\arccos\frac{\|\boldsymbol{w}_{j_1}^*-\boldsymbol{w}_{j_1}\|}{\|\boldsymbol{w}_{j_1}^*\|})\\
\leq&\frac{2}{\pi}\frac{\|\boldsymbol{w}_{j_1}^*-\boldsymbol{w}_{j_1}\|}{\|\boldsymbol{w}_{j_1}^*\|}.
\end{aligned}\tag{79}$$

Hence, (81) and (79) suggest that

$$\mathbb{E}_{\boldsymbol{X}}\left|\phi'\big(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}\big)-\phi'\big(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*\big)\right|\leq\frac{6}{\pi}\frac{\|\boldsymbol{w}_{j_1}^*-\boldsymbol{w}_{j_1}\|}{\|\boldsymbol{w}_{j_1}^*\|}.\tag{80}$$

Then, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}}\left|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\right|^2\cdot\left|\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1})-\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\right|\\
=&3\|\boldsymbol{a}_n\|_2^2\cdot\mathbb{E}_{\boldsymbol{X}}\left|\phi'\big(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}\big)-\phi'\big(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*\big)\right|\\
\leq&\frac{6\|\boldsymbol{a}_n\|_2^2}{\pi}\cdot\frac{\|\boldsymbol{w}_{j_1}-\boldsymbol{w}_{j_1}^*\|_2}{\|\boldsymbol{w}_{j_1}^*\|_2},
\end{aligned}\tag{81}$$

All in all, we have

$$\begin{aligned}\left\|\nabla^2 f_{\Omega_t}(\boldsymbol{W})-\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\right\|_2 \leq & \sum_{j_1}^{K}\sum_{j_2}^{K}\left\|\frac{\partial^2 f_{\Omega_t}}{\partial \boldsymbol{w}_{j_1}\partial \boldsymbol{w}_{j_2}}(\boldsymbol{W})-\frac{\partial^2 f_{\Omega_t}}{\partial \boldsymbol{w}_{j_1}\partial \boldsymbol{w}_{j_2}}(\boldsymbol{W}^*)\right\|_2 \\ \leq & K^2\max_{j_1,j_2}\left\|\frac{\partial^2 f_{\Omega_t}}{\partial \boldsymbol{w}_{j_1}\partial \boldsymbol{w}_{j_2}}(\boldsymbol{W})-\frac{\partial^2 f_{\Omega_t}}{\partial \boldsymbol{w}_{j_1}\partial \boldsymbol{w}_{j_2}}(\boldsymbol{W}^*)\right\|_2 \\ \leq & K^2\cdot\frac{12\|\boldsymbol{a}_n\|_2^2}{\pi}\max_j\frac{\|\boldsymbol{w}_j-\boldsymbol{w}_j^*\|_2}{\|\boldsymbol{w}_j^*\|_2} \\ \leq & 4\sigma_1^2(\boldsymbol{A})\frac{\|\boldsymbol{W}^*-\boldsymbol{W}\|_2}{\sigma_K}.\end{aligned} \tag{82}$$

□

A.3.3. Proof of Lemma 7

Proof of Lemma 7. According to the Definitions in (Janson, 2004), there exists a family of $\{(\mathcal{X}_j, w_j)\}_j$, where $\mathcal{X}_j \subseteq \mathcal{X}$ and $w_j \in [0,1]$, such that $\sum_j w_j \sum_{x_{n_j}\in\mathcal{X}_j} x_{n_j} = \sum_{n=1}^N x_n$, and $\sum_j w_j \leq d_{\mathcal{X}}$ by equations (2.1) and (2.2) in (Janson, 2004). Then, let p_j be any positive numbers with $\sum_j p_j = 1$. By Jensen's inequality, for any $s \in \mathbb{R}$, we have

$$e^{s\sum_{n=1}^N x_n} = e^{\sum_j p_j \frac{sw_j}{p_j}X_j} \leq \sum_j p_j e^{\frac{sw_j}{p_j}X_j}, \tag{83}$$

where $X_j = \sum_{x_{n_j}\in\mathcal{X}_j} x_{n_j}$.

Then, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{X}} e^{s\sum_{n=1}^N x_n} \leq & \mathbb{E}_{\mathcal{X}}\sum_j p_j e^{\frac{sw_j}{p_j}X_j} = \sum_j p_j \prod_{\mathcal{X}_j}\mathbb{E}_{\mathcal{X}} e^{\frac{sw_j}{p_j}x_{n_j}} \\ \leq & \sum_j p_j \prod_{\mathcal{X}_j} e^{\frac{Cw_j^2}{p_j^2}s^2} \\ \leq & \sum_j p_j e^{\frac{C|\mathcal{X}_j|w_j^2}{p_j^2}s^2}.\end{aligned} \tag{84}$$

Let $p_j = \frac{w_j|\mathcal{X}_j|^{1/2}}{\sum_j w_j|\mathcal{X}_j|^{1/2}}$, then we have

$$\mathbb{E}_{\mathcal{X}} e^{s\sum_{n=1}^N x_n} \leq \sum_j p_j e^{C\left(\sum_j w_j|\mathcal{X}_j|^{1/2}\right)^2 s^2} = e^{C\left(\sum_j w_j|\mathcal{X}_j|^{1/2}\right)^2 s^2}. \tag{85}$$

By Cauchy-Schwarz inequality, we have

$$\left(\sum_j w_j|\mathcal{X}_j|^{1/2}\right)^2 \leq \sum_j w_j \sum_j w_j|\mathcal{X}_j| \leq d_{\mathcal{X}}N. \tag{86}$$

Hence, we have

$$\mathbb{E}_{\mathcal{X}} e^{s\sum_{n=1}^N x_n} \leq e^{Cd_{\mathcal{X}}Ns^2}. \tag{87}$$

□

B. Proof of Theorem 2

Recall that the empirical risk function in (4) is defined as

$$\min_{\boldsymbol{W}}: \quad \hat{f}_{\Omega}(\boldsymbol{W}) = \frac{1}{|\Omega|}\sum_{n\in\Omega} -y_n \log\big(g(\boldsymbol{W};\boldsymbol{a}_n^T\boldsymbol{X})\big) - (1-y_n)\log\big(1-g(\boldsymbol{W};\boldsymbol{a}_n^T\boldsymbol{X})\big). \tag{88}$$

The population risk function is defined as

$$\begin{aligned} f_{\Omega}(\boldsymbol{W}) :=& \mathbb{E}_{\boldsymbol{X},y_n}\hat{f}_{\Omega}(\boldsymbol{W}) \\ =& \mathbb{E}_{\boldsymbol{X}}\mathbb{E}_{y_n|\boldsymbol{X}}\Big[\frac{1}{|\Omega|}\sum_{n\in\Omega} -y_n \log\big(g(\boldsymbol{W};\boldsymbol{a}_n^T\boldsymbol{X})\big) - (1-y_n)\log\big(1-g(\boldsymbol{W};\boldsymbol{a}_n^T\boldsymbol{X})\big)\Big] \\ =& \mathbb{E}_{\boldsymbol{X}}\frac{1}{|\Omega|}\sum_{n\in\Omega} -g(\boldsymbol{W}^*;\boldsymbol{a}_n^T\boldsymbol{X})\log\big(g(\boldsymbol{W};\boldsymbol{a}_n^T\boldsymbol{X})\big) - (1-g(\boldsymbol{W}^*;\boldsymbol{a}_n^T\boldsymbol{X}))\log\big(1-g(\boldsymbol{W};\boldsymbol{a}_n^T\boldsymbol{X})\big). \end{aligned} \tag{89}$$

The road-map of proof for Theorem 2 follows the similar three steps as those for Theorem 1. The major differences lie in three aspects: (i) in the second step, the objective function $\hat{f}_{\Omega_t}$ is smooth since the activation function $\phi(\cdot)$ is sigmoid. Hence, we can directly apply the mean value theorem as $\nabla\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)}) = \langle\nabla^2\hat{f}_{\Omega_t}(\widehat{\boldsymbol{W}}^{(t)}), \boldsymbol{W}^{(t)}-\boldsymbol{W}^*\rangle$ to characterize the effects of the gradient descent term in each iteration, and the error bound of $\nabla^2\hat{f}_{\Omega_t}$ is provided in Lemma 10; (ii) the objective function is the sum of cross-entry loss functions, which have more complex structure of derivatives than those of square loss functions; (iii) as the convergent point may not be the critical point of empirical loss function, we need to provide the distance from the convergent point to the ground-truth parameters additionally, where Lemma 11 is used.

Lemmas 10 and 11 are summarized in the following contents. Also, the notations $\lesssim$ and $\gtrsim$ follow the same definitions as in (27). The proofs of Lemmas 10 and 11 can be found in Appendix B.1 and B.2, respectively.

Lemma 10. *For any $\boldsymbol{W}$ that satisfies*

$$\|\boldsymbol{W}-\boldsymbol{W}^*\| \le \frac{2\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2} \tag{90}$$

then the second-order derivative of the empirical risk function in (88) for binary classification problems is bounded as

$$\frac{2(1-\varepsilon_0)}{11\kappa^2\gamma K^2}\sigma_1^2(\boldsymbol{A}) \preceq \nabla^2\hat{f}_{\Omega_t}(\boldsymbol{W}) \preceq \sigma_1^2(\boldsymbol{A}). \tag{91}$$

provided the number of samples satisfies

$$|\Omega_t| \gtrsim \varepsilon_0^{-2}(1+\delta^2)\kappa^2\gamma\sigma_1^4(\boldsymbol{A})K^6 d\log N. \tag{92}$$

Lemma 11. *Let $\hat{f}_{\Omega_t}$ and f_{Ω_t} be the empirical and population risk function in (88) and (89) for binary classification problems, respectively, then the first-order derivative of $\hat{f}_{\Omega_t}$ is close to its expectation f_{Ω_t} with an upper bound as*

$$\|\nabla f_{\Omega_t}(\boldsymbol{W}) - \nabla\hat{f}_{\Omega_t}(\boldsymbol{W})\|_2 \lesssim K^2\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log d}{|\Omega_t|}} \tag{93}$$

with probability at least $1-K^2N^{-10}$.

With these preliminary lemmas, the proof of Theorem 2 is formally summarized in the following contents.

Proof of Theorem 2. The update rule of $\boldsymbol{W}^{(t)}$ is

$$\boldsymbol{W}^{(t+1)} = \boldsymbol{W}^{(t)} - \eta\nabla\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)}) + \beta(\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(t-1)}) \tag{94}$$

Since $\widehat{\boldsymbol{W}}$ is a critical point, then we have $\nabla\hat{f}_{\Omega_t}(\widehat{\boldsymbol{W}}) = 0$. By the intermediate value theorem, we have

$$\begin{aligned} \boldsymbol{W}^{(t+1)} = \boldsymbol{W}^{(t)} - \eta\nabla^2\hat{f}_{\Omega_t}(\widehat{\boldsymbol{W}}^{(t)})(\boldsymbol{W}^{(t)}-\widehat{\boldsymbol{W}}) \\ + \beta(\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(t-1)}) \end{aligned} \tag{95}$$

where $\widehat{\boldsymbol{W}}^{(t)}$ lies in the convex hull of $\boldsymbol{W}^{(t)}$ and $\widehat{\boldsymbol{W}}$.

Next, we have

$$\begin{bmatrix} \boldsymbol{W}^{(t+1)} - \boldsymbol{W}^* \\ \boldsymbol{W}^{(t)} - \boldsymbol{W}^* \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} - \eta \nabla^2 \hat{f}_{\Omega_t}(\widehat{\boldsymbol{W}}^{(t)}) + \beta \boldsymbol{I} & \beta \boldsymbol{I} \\ \boldsymbol{I} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{W}^{(t)} - \boldsymbol{W}^* \\ \boldsymbol{W}^{(t-1)} - \boldsymbol{W}^* \end{bmatrix}. \tag{96}$$

Let $\boldsymbol{P}(\beta) = \begin{bmatrix} \boldsymbol{I} - \eta \nabla^2 \hat{f}_{\Omega_t}(\widehat{\boldsymbol{W}}^{(t)}) + \beta \boldsymbol{I} & \beta \boldsymbol{I} \\ \boldsymbol{I} & 0 \end{bmatrix}$, so we have

$$\left\| \begin{bmatrix} \boldsymbol{W}^{(t+1)} - \boldsymbol{W}^* \\ \boldsymbol{W}^{(t)} - \boldsymbol{W}^* \end{bmatrix} \right\|_2 = \|\boldsymbol{P}(\beta)\|_2 \left\| \begin{bmatrix} \boldsymbol{W}^{(t)} - \boldsymbol{W}^* \\ \boldsymbol{W}^{(t-1)} - \boldsymbol{W}^* \end{bmatrix} \right\|_2.$$

Then, we have

$$\|\boldsymbol{W}^{(t+1)} - \boldsymbol{W}^*\|_2 \lesssim \|\boldsymbol{P}(\beta)\|_2 \|\boldsymbol{W}^{(t)} - \boldsymbol{W}^*\|_2 \tag{97}$$

Let λ_i be the i-th eigenvalue of $\nabla^2 \hat{f}_{\Omega_t}(\widehat{\boldsymbol{W}}^{(t)})$, and δ_i be the i-th eigenvalue of matrix $\boldsymbol{P}(\beta)$. Following the similar analysis in proof of Theorem 1, we have

$$\delta_i(0) > \delta_i(\beta), \quad \text{for} \quad \forall \beta \in \big(0, (1 - \eta\lambda_i)^2\big). \tag{98}$$

Moreover, δ_i achieves the minimum $\delta_i^* = |1 - \sqrt{\eta\lambda_i}|$ when $\beta = \big(1 - \sqrt{\eta\lambda_i}\big)^2$.

Let us first assume $\boldsymbol{W}^{(t)}$ satisfies (90) and the number of samples satisfies (92), then from Lemma 10, we know that

$$0 < \frac{2(1-\varepsilon_0)\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2} \le \lambda_i \le \sigma_1^2(\boldsymbol{A}).$$

We define $\gamma_1 = \frac{2(1-\varepsilon_0)\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2}$ and $\gamma_2 = \sigma_1^2(\boldsymbol{A})$. Also, for any $\varepsilon_0 \in (0,1)$, we have

$$\nu(\beta^*) = \|\boldsymbol{P}(\beta^*)\|_2 = 1 - \sqrt{\frac{\gamma_1}{2\gamma_2}} = 1 - \sqrt{\frac{1-\varepsilon_0}{11\kappa^2\gamma K}} \tag{99}$$

Let $\beta = 0$, we have

$$\nu(0) = \|\boldsymbol{A}(0)\|_2 = 1 - \frac{1-\varepsilon_0}{11\kappa^2\gamma K}.$$

Hence, with probability at least $1 - K^2 \cdot N^{-10}$, we have

$$\|\boldsymbol{W}^{(t+1)} - \boldsymbol{W}^*\|_2 \le \Big(1 - \sqrt{\frac{1-\varepsilon_0}{11\kappa^2\gamma K}}\Big)\|\boldsymbol{W}^{(t)} - \boldsymbol{W}^*\|_2, \tag{100}$$

provided that $\boldsymbol{W}^{(t)}$ satisfies (25), and

$$|\Omega_t| \gtrsim \varepsilon_0^{-2}\kappa^2\gamma(1+\delta^2)\sigma_1^4(\boldsymbol{A})K^6 d \log N. \tag{101}$$

According to Lemma 4, we know that (90) holds for $\boldsymbol{W}^{(0)}$ if

$$|\Omega_t| \gtrsim \varepsilon_0^{-2}\kappa^8\gamma^2(1+\delta^2)K^8 d \log N. \tag{102}$$

Combining (101) and (102), we need $|\Omega_t| \gtrsim \varepsilon_0^{-2}\kappa^8\gamma^2(1+\delta^2)\sigma_1^4(\boldsymbol{A})K^8 d \log N$.

Finally, by the mean value theorem, we have

$$\hat{f}_{\Omega_t}(\widehat{\boldsymbol{W}}) \le \hat{f}_{\Omega_t}(\boldsymbol{W}^*) + \nabla \hat{f}_{\Omega_t}(\boldsymbol{W}^*)^T(\widehat{\boldsymbol{W}} - \boldsymbol{W}^*) + \frac{1}{2}(\widehat{\boldsymbol{W}} - \boldsymbol{W}^*)^T \nabla^2 \hat{f}_{\Omega_t}(\widetilde{\boldsymbol{W}})(\widehat{\boldsymbol{W}} - \boldsymbol{W}^*) \tag{103}$$

for some $\widetilde{\boldsymbol{W}}$ between $\widehat{\boldsymbol{W}}$ and $\boldsymbol{W}^*$. Since $\widehat{\boldsymbol{W}}$ is the local minima, we have $\hat{f}_{\Omega_t}(\widehat{\boldsymbol{W}}) \leq \hat{f}_{\Omega_t}(\boldsymbol{W}^*)$. That is to say

$$\nabla \hat{f}_{\Omega_t}(\boldsymbol{W}^*)^T(\widehat{\boldsymbol{W}} - \boldsymbol{W}^*) + \frac{1}{2}(\widehat{\boldsymbol{W}} - \boldsymbol{W}^*)^T \nabla^2 \hat{f}_{\Omega_t}(\widetilde{\boldsymbol{W}})(\widehat{\boldsymbol{W}} - \boldsymbol{W}^*) \leq 0 \tag{104}$$

which implies

$$\frac{1}{2}\|\nabla^2 \hat{f}_{\Omega_t}(\widetilde{\boldsymbol{W}})\|_2 \|\widehat{\boldsymbol{W}} - \boldsymbol{W}^*\|_2^2 \leq \|\nabla \hat{f}_{\Omega_t}(\boldsymbol{W}^*)\|_2 \|\widehat{\boldsymbol{W}} - \boldsymbol{W}^*\|_2. \tag{105}$$

From Lemma 10, we know that

$$\|\nabla^2 \hat{f}_{\Omega_t}(\widetilde{\boldsymbol{W}})\|_2 \geq \frac{2(1-\varepsilon_0)}{11\kappa^2\gamma K^2}\sigma^2(\boldsymbol{A}). \tag{106}$$

From Lemma 11, we know that

$$\|\nabla \hat{f}_{\Omega_t}(\boldsymbol{W}^*)\|_2 = \|\nabla \hat{f}_{\Omega_t}(\boldsymbol{W}^*) - \nabla f_{\Omega_t}(\boldsymbol{W}^*)\|_2 \lesssim K^2\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}}. \tag{107}$$

Plugging inequalities (106) and (107) back into (105), we have

$$\|\widehat{\boldsymbol{W}} - \boldsymbol{W}^*\|_2 \lesssim (1-\varepsilon_0)^{-1}\kappa^2\gamma K^4\sqrt{\frac{(1+\delta^2)d\log d}{|\Omega_t|}}. \tag{108}$$

□

B.1. Proof of Lemma 10

The roadmap of proof for Lemma 10 follows the similar steps as those of Lemma 2 for regression problems. Lemmas 12, 13 and 14 are the preliminary lemmas, and their proofs can be found in Appendix B.2. The proof of Lemma 10 is summarized after these preliminary lemmas.

Lemma 12. *The second-order derivative of f_{Ω_t} at the ground truth $\boldsymbol{W}^*$ satisfies*

$$\frac{4\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2}\boldsymbol{I} \preceq \nabla^2 f_{\Omega_t}(\boldsymbol{W}^*) \preceq \frac{\sigma_1^2(\boldsymbol{A})}{4}\boldsymbol{I}. \tag{109}$$

Lemma 13. *Suppose f_{Ω_t} is the population loss function with respect to binary classification problems, then we have*

$$\|\nabla^2 f_{\Omega_t}(\boldsymbol{W}) - \nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\|_2 \lesssim \|\boldsymbol{W} - \boldsymbol{W}^*\|_2. \tag{110}$$

Lemma 14. *Suppose $\hat{f}_{\Omega_t}$ is the empirical loss function with respect to binary classification problems, then the second-order derivative of $\hat{f}_{\Omega_t}$ is close to its expectation with an upper bound as*

$$\|\nabla^2 f_{\Omega_t}(\boldsymbol{W}) - \nabla^2 \hat{f}_{\Omega_t}(\boldsymbol{W})\|_2 \lesssim K^2\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log d}{|\Omega_t|}} \tag{111}$$

with probability at least $1 - K^2N^{-10}$.

Proof of Lemma 10 . For any $\boldsymbol{W}$, we have

$$\Big|\|\nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2 - \|\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\|_2\Big| \leq \|\nabla^2 f_{\Omega_t}(\boldsymbol{W}) - \nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\|_2. \tag{112}$$

That is

$$\begin{aligned} &\|\nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2 \leq \|\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\|_2 + \|\nabla^2 f_{\Omega_t}(\boldsymbol{W}) - \nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\|_2 \\ \text{and}\quad &\|\nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2 \geq \|\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\|_2 - \|\nabla^2 f_{\Omega_t}(\boldsymbol{W}) - \nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\|_2 \end{aligned} \tag{113}$$

Then, for any $\boldsymbol{W}$ that satisfies $\|\boldsymbol{W}-\boldsymbol{W}^*\| \le \frac{2\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2}$, from Lemmas 12 and 13, we have

$$\frac{2}{11\kappa^2\gamma K^2}\sigma_1^2(\boldsymbol{A}) \preceq \nabla^2 f_{\Omega_t}(\boldsymbol{W}) \preceq \frac{1}{2}\sigma_1^2(\boldsymbol{A}). \tag{114}$$

Next, we have

$$\begin{aligned} &\|\nabla^2 \hat{f}_{\Omega_t}(\boldsymbol{W})\|_2 \le \|\nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2 + \|\nabla^2 \hat{f}_{\Omega_t}(\boldsymbol{W}) - \nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2 \\ \text{and}\quad &\|\nabla^2 \hat{f}_{\Omega_t}(\boldsymbol{W})\|_2 \ge \|\nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2 - \|\nabla^2 \hat{f}_{\Omega_t}(\boldsymbol{W}) - \nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2 \end{aligned} \tag{115}$$

Then, from (114) and Lemma 14, we have

$$\frac{2(1-\varepsilon_0)}{11\kappa^2\gamma K^2}\sigma_1^2(\boldsymbol{A}) \preceq \nabla^2 \hat{f}_{\Omega_t}(\boldsymbol{W}) \preceq \sigma_1^2(\boldsymbol{A}) \tag{116}$$

provided that the sample size $|\Omega_t| \gtrsim \varepsilon_0^{-2}(1+\delta^2)\kappa^2\gamma\sigma_1^4(\boldsymbol{A})K^6 d\log N$. □

B.2. Proof of auxiliary lemmas for binary classification problems

B.2.1. Proof of Lemma 12

Proof of Lemma 12 . Since $\mathbb{E}_{\boldsymbol{X}} y_n = g_n(\boldsymbol{W}^*;\boldsymbol{a}_n)$, then we have

$$\begin{aligned} \frac{\partial^2 f_{\Omega_t}(\boldsymbol{W}^*)}{\partial \boldsymbol{w}_j^* \partial \boldsymbol{w}_k^*} =& \mathbb{E}_{\boldsymbol{X}} \frac{\partial^2 \hat{f}_{\Omega_t}(\boldsymbol{W}^*)}{\partial \boldsymbol{w}_j^* \partial \boldsymbol{w}_k^*} \\ =& \mathbb{E}_{\boldsymbol{X}} \frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t} \frac{1}{g(\boldsymbol{W};\boldsymbol{a}_n)\big(1-g(\boldsymbol{W};\boldsymbol{a}_n)\big)} \phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi'(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T, \end{aligned} \tag{117}$$

for any $j,k\in[K]$.

Then, for any $\boldsymbol{\alpha} = \begin{bmatrix}\boldsymbol{\alpha}_1^T, & \boldsymbol{\alpha}_2^T, & \cdots, & \boldsymbol{\alpha}_K^T\end{bmatrix}^T \in \mathbb{R}^{dk}$ with $\boldsymbol{\alpha}_j \in \mathbb{R}^d$, the lower bound can be obtained from

$$\begin{aligned} \boldsymbol{\alpha}^T\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\boldsymbol{\alpha} =& \mathbb{E}_{\boldsymbol{X}} \frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t} \frac{\Big(\sum_{j=1}^K \boldsymbol{\alpha}_j^T\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{w}_j^{*T}\boldsymbol{X}^T\boldsymbol{a}_n)\Big)^2}{g(\boldsymbol{W}^*;\boldsymbol{a}_n)\big(1-g(\boldsymbol{W}^*;\boldsymbol{a}_n)\big)} \\ \ge& \mathbb{E}_{\boldsymbol{X}} \frac{4}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\Big(\sum_{j=1}^K \boldsymbol{\alpha}_j^T\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{w}_j^{*T}\boldsymbol{X}^T\boldsymbol{a}_n)\Big)^2 \\ \ge& \frac{4\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2}. \end{aligned} \tag{118}$$

Also, for the upper bound, we have

$$\begin{aligned} \boldsymbol{\alpha}^T\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\boldsymbol{\alpha} =& \mathbb{E}_{\boldsymbol{X}} \frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t} \frac{\Big(\sum_{j=1}^K \boldsymbol{\alpha}_j^T\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{w}_j^{*T}\boldsymbol{X}^T\boldsymbol{a}_n)\Big)^2}{g(\boldsymbol{W}^*;\boldsymbol{a}_n)\big(1-g(\boldsymbol{W}^*;\boldsymbol{a}_n)\big)} \\ =& \mathbb{E}_{\boldsymbol{X}} \frac{1}{|\Omega_t|}\sum_{n\in\Omega_t} \frac{\Big(\sum_{j=1}^K \boldsymbol{\alpha}_j^T\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{w}_j^{*T}\boldsymbol{X}^T\boldsymbol{a}_n)\Big)^2}{\sum_{j_1=1}^K \phi(\boldsymbol{w}_{j_1}^{*T}\boldsymbol{X}^T\boldsymbol{a}_n)\sum_{j_2=1}^K\big(1-\phi(\boldsymbol{w}_{j_2}^{*T}\boldsymbol{X}^T\boldsymbol{a}_n)\big)} \\ \le& \mathbb{E}_{\boldsymbol{X}} \frac{1}{|\Omega_t|}\sum_{n\in\Omega_t} \frac{\sum_{j=1}^K\big(\boldsymbol{\alpha}_j^T\boldsymbol{X}^T\boldsymbol{a}_n\big)^2\sum_{j=1}^K\big(\phi'(\boldsymbol{w}_j^{*T}\boldsymbol{X}^T\boldsymbol{a}_n)\big)^2}{\sum_{j_1=1}^K \phi(\boldsymbol{w}_{j_1}^{*T}\boldsymbol{X}^T\boldsymbol{a}_n)\sum_{j_2=1}^K\big(1-\phi(\boldsymbol{w}_{j_2}^{*T}\boldsymbol{X}^T\boldsymbol{a}_n)\big)}. \end{aligned} \tag{119}$$

For the denominator item, we have

$$\begin{aligned}\sum_{j_1=1}^{K}\phi({\boldsymbol{w}_{j_1}^*}^T\boldsymbol{X}^T\boldsymbol{a}_n)\sum_{j_2=1}^{K}\big(1-\phi({\boldsymbol{w}_{j_2}^*}^T\boldsymbol{X}^T\boldsymbol{a}_n)\big)\geq&\sum_{j=1}^{K}\phi({\boldsymbol{w}_j^*}^T\boldsymbol{X}^T\boldsymbol{a}_n)\big(1-\phi({\boldsymbol{w}_j^*}^T\boldsymbol{X}^T\boldsymbol{a}_n)\big)\\=&\sum_{j=1}^{K}\phi'({\boldsymbol{w}_j^*}^T\boldsymbol{X}^T\boldsymbol{a}_n)\\\geq&4\sum_{j=1}^{K}\phi'({\boldsymbol{w}_j^*}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2.\end{aligned}\tag{120}$$

Hence, we have

$$\boldsymbol{\alpha}^T\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\boldsymbol{\alpha}\leq\mathbb{E}_{\boldsymbol{X}}\frac{1}{4|\Omega_t|}\sum_{n\in\Omega_t}\sum_{j=1}^{K}(\boldsymbol{\alpha}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)^2\leq\frac{1}{4}\sigma_1^2(\boldsymbol{A}).\tag{121}$$

□

B.2.2. PROOF OF LEMMA 13

Proof of Lemma 13 . Recall that

$$\begin{aligned}&\frac{\partial^2 f_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}\\=&\mathbb{E}_{\boldsymbol{X}}\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\left(\frac{g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{g^2(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{(1-g(\boldsymbol{W};\boldsymbol{a}_n))^2}\right)\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi'(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T,\end{aligned}\tag{122}$$

and

$$\begin{aligned}\frac{\partial^2 f_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j^2}=&\mathbb{E}_{\boldsymbol{X}}\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\left(\frac{g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{g^2(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{(1-g(\boldsymbol{W};\boldsymbol{a}_n))^2}\right)\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)^2(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T\\&-\mathbb{E}_{\boldsymbol{X}}\frac{1}{K|\Omega_t|}\sum_{n\in\Omega_t}\left(-\frac{g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{g(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{1-g(\boldsymbol{W};\boldsymbol{a}_n)}\right)\phi''(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T.\end{aligned}\tag{123}$$

Let us denote $A_{j,k}(\boldsymbol{W};\boldsymbol{a}_n)$ as

$$A_{j,k}(\boldsymbol{W};\boldsymbol{a}_n)=\begin{cases}\frac{1}{K^2}\big(\frac{g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{g^2(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{(1-g(\boldsymbol{W};\boldsymbol{a}_n))^2}\big)\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi'(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)\\\qquad-\frac{1}{K}\big(-\frac{g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{g(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{1-g(\boldsymbol{W};\boldsymbol{a}_n)}\big)\phi''(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n),\quad\text{when}\quad j=k;\\\frac{1}{K^2}\big(\frac{g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{g^2(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{(1-g(\boldsymbol{W};\boldsymbol{a}_n))^2}\big)\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi'(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n),\quad\text{when}\quad j\neq k.\end{cases}\tag{124}$$

Further, let us define $M(\boldsymbol{W};\boldsymbol{a}_n)=\max\left\{\frac{2}{K^3}\frac{1}{g^3(\boldsymbol{W};\boldsymbol{a}_n)},\frac{2}{K^3}\frac{1}{(1-g(\boldsymbol{W};\boldsymbol{a}_n))^3},\frac{1}{K^2}\frac{1}{g^2(\boldsymbol{W};\boldsymbol{a}_n)},\frac{1}{K^2}\frac{1}{((1-g(\boldsymbol{W};\boldsymbol{a}_n))^2}\right\}$.

Then, by the mean value theorem, we have

$$A_{j,k}(\boldsymbol{W};\boldsymbol{a}_n)-A_{j,k}(\boldsymbol{W};\boldsymbol{a}_n)=\sum_{l=1}^{K}\langle\frac{\partial A_{j,k}}{\partial\boldsymbol{w}_l}(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n),\boldsymbol{w}_l-\boldsymbol{w}_l^*\rangle.\tag{125}$$

For $\frac{\partial A_{j,k}}{\partial\boldsymbol{w}_l}$, we have

$$\frac{\partial A_{j,k}}{\partial\boldsymbol{w}_l}(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n)=B_{j,k,l}(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n)\boldsymbol{X}^T\boldsymbol{a}_n\tag{126}$$

with

$$\begin{aligned}|B_{j,k,l}(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n)|\leq&\frac{2}{K^3}\frac{1}{g^3(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n)}+\frac{2}{K^3}\frac{1}{(1-g(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n))^3}+\frac{1}{K^2}\frac{1}{g(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n)}+\frac{1}{K^2}\frac{1}{(1-g(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n))^2}\\ \leq&4M(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n).\end{aligned}\tag{127}$$

for all $j\in[K],k\in[K],l\in[K]$.

Therefore, for any $\boldsymbol{\alpha}\in\mathbb{R}^{Kd}$, we have

$$\begin{aligned}&\boldsymbol{\alpha}^T\nabla^2 f_{\Omega_t}(\boldsymbol{W})\boldsymbol{\alpha}\\ \leq&\frac{1}{|\Omega_t|}\sum_{n\in|\Omega_t|}\sum_{j=1}^K\sum_{k=1}^K\mathbb{E}_{\boldsymbol{X}}\Big|\boldsymbol{\alpha}_j^T\frac{\partial f_{\Omega_t}}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}(\boldsymbol{W})\boldsymbol{\alpha}_k\Big|\\ =&\frac{1}{|\Omega_t|}\sum_{n\in|\Omega_t|}\sum_{j=1}^K\sum_{k=1}^K\mathbb{E}_{\boldsymbol{X}}\Big|\sum_{l=1}^K|B_{j,k,l}(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n)|\langle\boldsymbol{w}_l-\boldsymbol{w}_l^*,\boldsymbol{X}^T\boldsymbol{a}_n\rangle\langle\boldsymbol{\alpha}_j,\boldsymbol{X}^T\boldsymbol{a}_n\rangle\langle\boldsymbol{\alpha}_k,\boldsymbol{X}^T\boldsymbol{a}_n\rangle\Big|\\ =&\frac{1}{|\Omega_t|}\sum_{n\in|\Omega_t|}\sum_{j=1}^K\sum_{k=1}^K\Big(\sum_{l=1}^K\mathbb{E}_{\boldsymbol{X}}|B_{j,k,l}(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n)|^2\Big)^{\frac{1}{2}}\Big(\sum_{l=1}^K\mathbb{E}_{\boldsymbol{X}}\big|\langle\boldsymbol{w}_l-\boldsymbol{w}_l^*,\boldsymbol{X}^T\boldsymbol{a}_n\rangle\langle\boldsymbol{\alpha}_j,\boldsymbol{X}^T\boldsymbol{a}_n\rangle\langle\boldsymbol{\alpha}_k,\boldsymbol{X}^T\boldsymbol{a}_n\rangle\big|^2\Big)^{\frac{1}{2}}\\ \leq&\frac{1}{|\Omega_t|}\sum_{n\in|\Omega_t|}\sum_{j=1}^K\sum_{k=1}^K36K^{\frac{1}{2}}\Big(\mathbb{E}_{\boldsymbol{X}}M^2(\widehat{\boldsymbol{W}};\boldsymbol{a}_n)\Big)^{\frac{1}{2}}\cdot\Big(\sum_{l=1}^K\|\boldsymbol{w}_l-\boldsymbol{w}_l^*\|_2^2\Big)^{\frac{1}{2}}\|\boldsymbol{a}_j\|_2\|\boldsymbol{a}_k\|_2\\ \leq&\frac{1}{|\Omega_t|}\sum_{n\in|\Omega_t|}36K^3\Big(\mathbb{E}_{\boldsymbol{X}}M^2(\widehat{\boldsymbol{W}};\boldsymbol{a}_n)\Big)^{\frac{1}{2}}\|\boldsymbol{W}-\boldsymbol{W}^*\|_2\\ \overset{(a)}{\lesssim}&e^{\sigma_1^2(\boldsymbol{A})}\|\boldsymbol{W}-\boldsymbol{W}^*\|_2\\ \lesssim&\|\boldsymbol{W}-\boldsymbol{W}^*\|_2,\end{aligned}\tag{128}$$

where (a) comes from Lemma 5 in (Fu et al., 2018). □

B.2.3. Proof of Lemma 14

Proof of Lemma 14 . Recall that

$$\begin{aligned}&\frac{\partial^2\hat{f}_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}\\ =&\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\Big(\frac{y_n}{g^2(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-y_n}{(1-g(\boldsymbol{W};\boldsymbol{a}_n))^2}\Big)\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi'(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T,\end{aligned}\tag{129}$$

and

$$\begin{aligned}\frac{\partial^2\hat{f}_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j^2}=&\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\Big(\frac{y_n}{g^2(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-y_n}{(1-g(\boldsymbol{W};\boldsymbol{a}_n))^2}\Big)\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)^2(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T\\ &-\frac{1}{K|\Omega_t|}\sum_{n\in\Omega_t}\Big(-\frac{y_n}{g(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-y_n}{1-g(\boldsymbol{W};\boldsymbol{a}_n)}\Big)\phi''(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T.\end{aligned}\tag{130}$$

When $y_n=1$ and $j\neq k$, we have

$$\frac{\partial^2\hat{f}_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}=\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\frac{\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi'(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)}{g^2(\boldsymbol{W};\boldsymbol{a}_n)}(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T,\tag{131}$$

and

$$\begin{aligned}\frac{\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi'(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)}{g^2(\boldsymbol{W};\boldsymbol{a}_n)}=&\frac{\phi(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)(1-\phi(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n))\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)(1-\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n))}{\big(\frac{1}{K}\sum_{l=1}^K\phi(\boldsymbol{w}_l^T\boldsymbol{X}^T\boldsymbol{a}_n)\big)^2}\\ \le&K^2\frac{\phi(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)(1-\phi(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n))\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)(1-\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n))}{\phi(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)}\\ =&K^2(1-\phi(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n))(1-\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n))\le K^2.\end{aligned}\tag{132}$$

When $y_n=1$ and $j=k$, we have

$$\frac{\partial^2\hat{f}_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}=\frac{1}{|\Omega_t|}\sum_{n\in\Omega_t}\Big[\frac{1}{K^2}\frac{\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi'(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)}{g^2(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1}{K}\frac{\phi''(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)}{g(\boldsymbol{W};\boldsymbol{a}_n)}\Big](\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T,\tag{133}$$

and

$$\Big|\frac{\phi''(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)}{g(\boldsymbol{W};\boldsymbol{a}_n)}\Big|=\frac{\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)(1-\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n))\cdot\big|1-2\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)\big|}{\frac{1}{K}\sum_{l=1}^K\phi(\boldsymbol{w}_l^T\boldsymbol{X}^T\boldsymbol{a}_n)}\le K.\tag{134}$$

Similar to (132) and (134), we can obtain the following inequality for $y_n=0$.

$$\frac{\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi'(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)}{\big(1-g(\boldsymbol{W};\boldsymbol{a}_n)\big)^2}\le K^2,\quad\text{and}\quad\Big|\frac{\phi''(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)}{1-g(\boldsymbol{W};\boldsymbol{a}_n)}\Big|\le K.\tag{135}$$

Then, for any $\boldsymbol{\alpha}\in\mathbb{R}^d$, we have

$$\begin{aligned}\boldsymbol{\alpha}^T\frac{\partial^2\hat{f}_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}\boldsymbol{\alpha}=&\frac{1}{|\Omega_t|}\sum_{n\in\Omega_t}\bigg[\frac{1}{K^2}\bigg(\frac{y_n}{g^2(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-y_n}{(1-g(\boldsymbol{W};\boldsymbol{a}_n))^2}\bigg)\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi'(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)\\ &-\frac{\mathbb{1}_{\{j=k\}}}{K}\bigg(-\frac{y_n}{g(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-y_n}{1-g(\boldsymbol{W};\boldsymbol{a}_n)}\bigg)\phi''(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\bigg](\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2\\ :=&\frac{1}{|\Omega_t|}\sum_{n\in\Omega_t}H_{j,k}(\boldsymbol{a}_n)\cdot(\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2.\end{aligned}\tag{136}$$

Next, we show that $H_{j,k}(\boldsymbol{a}_n)\cdot(\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2$ belongs to the sub-exponential distribution. For any $p\in\mathbb{N}^+$, we have

$$\begin{aligned}\Big(\mathbb{E}_{\boldsymbol{X},\boldsymbol{y}_n}\Big[\big|H_{j,k}(\boldsymbol{a}_n)\cdot(\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2\big|^p\Big]\Big)^{1/p}\le&\Big(\mathbb{E}_{\boldsymbol{X}}\Big[\big|4(\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2\big|^p\Big]\Big)^{1/p}\\ \le&8\|\boldsymbol{a}_n\|_2^2p\le 8\sigma_1^2(\boldsymbol{A})p\end{aligned}\tag{137}$$

Hence, $H_{j,k}(\boldsymbol{a}_n)\cdot(\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2$ belongs to the sub-exponential distribution with $\|H_{j,k}(\boldsymbol{a}_n)(\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2\|_{\psi_1}=8\sigma_1^2(\boldsymbol{A})$. Then, the moment generation function of $H_{j,k}(\boldsymbol{a}_n)\cdot(\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2$ can be bounded as

$$\mathbb{E}e^{sH_{j,k}(\boldsymbol{a}_n)\cdot(\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2}\le e^{C\sigma_1^2(\boldsymbol{A})s^2}\tag{138}$$

for some positive constant C and any $s\in\mathbb{R}$. From Lemma 7 and Chernoff bound, we have

$$\boldsymbol{\alpha}^T\Big(\frac{\partial^2\hat{f}_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}-\frac{\partial^2f_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}\Big)\boldsymbol{\alpha}\le C\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}}\tag{139}$$

with probability at least $1-N^{-d}$. By selecting $\xi=\frac{1}{2}$ in Lemmas 8 and 9, we have

$$\bigg\|\frac{\partial^2\hat{f}_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}-\frac{\partial^2f_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}\bigg\|_2\le C\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}}\tag{140}$$

with probability at least $1-\left(\frac{5}{N}\right)^d$.

In conclusion, we have

$$\begin{aligned}\|\nabla^2 f_{\Omega_t}(\boldsymbol{W})-\nabla^2\hat{f}_{\Omega_t}(\boldsymbol{W})\|_2 \leq & \sum_{j=1}^{K}\sum_{k=1}^{K}\left\|\frac{\partial^2\hat{f}_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}-\frac{\partial^2 f_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}\right\|_2 \\ \leq & CK^2\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log d}{|\Omega_t|}}\end{aligned} \tag{141}$$

with probability at least $1-\left(\frac{5}{d}\right)^d$. □

B.2.4. Proof of Lemma 11

Proof of Lemma 11. Recall that the first-order derivative of $\hat{f}_{\Omega_t}(\boldsymbol{W})$ is calculated from

$$\frac{\partial\hat{f}_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j}=-\frac{1}{K|\Omega_t|}\sum_{n\in\Omega}\frac{y_n-g(\boldsymbol{W};\boldsymbol{a}_n)}{g(\boldsymbol{W};\boldsymbol{a}_n)\big(1-g(\boldsymbol{W};\boldsymbol{a}_n)\big)}\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\boldsymbol{X}^T\boldsymbol{a}_n. \tag{142}$$

Similar to (134), we have

$$\left|\frac{\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)}{g(\boldsymbol{W};\boldsymbol{a}_n)}\right|=\frac{\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)(1-\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n))}{\frac{1}{K}\sum_{l=1}^{K}\phi(\boldsymbol{w}_l^T\boldsymbol{X}^T\boldsymbol{a}_n)}\leq K. \tag{143}$$

Similar to (137), for any fixed $\boldsymbol{\alpha}\in\mathbb{R}^{dK}$, we can show that random variable $\boldsymbol{\alpha}^T\frac{\partial\hat{f}_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j}$ belongs to sub-exponential distribution with the same bounded norm up to a constant. Hence, by applying Lemma 7 and the Chernoff bound, we have

$$\left\|\nabla f_{\Omega_t}(\boldsymbol{W})-\nabla\hat{f}_{\Omega_t}(\boldsymbol{W})\right\|_2\lesssim K^2\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}} \tag{144}$$

with probability at least $1-\left(\frac{5}{N}\right)^d$. □

C. Proof of Lemma 1

Proof of Lemma 1. Let $\widetilde{\boldsymbol{A}}$ denote the adjacency matrix, then we have

$$\sigma_1(\widetilde{\boldsymbol{A}})=\max_{\boldsymbol{z}}\frac{\boldsymbol{z}^T\widetilde{\boldsymbol{A}}\boldsymbol{z}}{\boldsymbol{z}^T\boldsymbol{z}}\geq\frac{\mathbf{1}^T\widetilde{\boldsymbol{A}}\mathbf{1}}{\mathbf{1}^T\mathbf{1}}=1+\frac{\sum_{n=1}^{N}\delta_n}{N}, \tag{145}$$

where δ_n denotes the degree of node v_n. Let $\boldsymbol{z}$ be the eigenvetor of the maximum eigenvalue $\sigma_1(\boldsymbol{A})$. Since $\sigma_1(\boldsymbol{A})=\boldsymbol{D}^{-1/2}\widetilde{\boldsymbol{A}}\boldsymbol{D}^{-1/2}$ and $\boldsymbol{D}$ is diagonal matrix, then $\boldsymbol{z}$ is the eigenvector to $\sigma_1(\widetilde{\boldsymbol{A}})$ as well. Then, let $n\in[N]$ be the index of the largest value of vector $\boldsymbol{z}_n$ as $z_n=\|\boldsymbol{z}\|_\infty$, we have

$$\sigma_1(\widetilde{\boldsymbol{A}})=\frac{(\widetilde{\boldsymbol{A}}\boldsymbol{z})_n}{z_n}=\frac{\widetilde{\boldsymbol{a}}_n^T\boldsymbol{z}}{z_n}\leq\frac{\|\boldsymbol{a}_n\|_1\|\boldsymbol{z}\|_\infty}{z_n}=1+\delta. \tag{146}$$

where $\widetilde{\boldsymbol{a}}_n$ is the n-th row of $\widetilde{\boldsymbol{A}}$.

Since $\boldsymbol{D}$ is a diagonal matrix with $\|\boldsymbol{D}\|_2\leq 1+\delta$, then we can conclude the inequality in this lemma. □

D. Proof of Lemma 4

The proof of Lemma 4 is divided into three major parts to bound I_1, I_2 and I_3 in (153). Lemmas 15, 16 and 17 provide the error bounds for I_1, I_2 and I_3, respectively. The proofs of these preliminary lemmas are similar to those of Theorem 5.6 in (Zhong et al., 2017b), the difference is to apply Lemma 7 plus Chernoff inequality instead of standard Hoeffding inequality, and we skip the details of the proofs of Lemmas 15, 16 and 17 here.

Lemma 15. *Suppose $\boldsymbol{M}_2$ is defined as in (7) and $\widehat{\boldsymbol{M}}_2$ is the estimation of $\boldsymbol{M}_2$ by samples. Then, with probability $1-N^{-10}$, we have*

$$\|\widehat{\boldsymbol{M}}_2-\boldsymbol{M}_2\|\lesssim\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega|}}, \tag{147}$$

provided that $|\Omega|\gtrsim(1+\delta^2)d\log^4 N$.

Lemma 16. *Let $\widehat{\boldsymbol{V}}$ be generated by step 4 in Subroutine 1. Suppose $\boldsymbol{M}_3(\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}})$ is defined as in (9) and $\widehat{\boldsymbol{M}}_3(\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}})$ is the estimation of $\boldsymbol{M}_3(\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}})$ by samples. Further, we assume $\boldsymbol{V}\in\mathbb{R}^{d\times K}$ is an orthogonal basis of $\boldsymbol{W}^*$ and satisfies $\|\boldsymbol{V}\boldsymbol{V}^T-\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^T\|\le 1/4$. Then, provided that $N\gtrsim K^5\log^6 d$, with probability at least $1-N^{-10}$, we have*

$$\|\widehat{\boldsymbol{M}}_3(\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}})-\boldsymbol{M}_3(\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}})\|\lesssim\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)K^3\log N}{|\Omega|}}. \tag{148}$$

Lemma 17. *Suppose $\boldsymbol{M}_1$ is defined as in (6) and $\widehat{\boldsymbol{M}}_1$ is the estimation of $\boldsymbol{M}_1$ by samples. Then, with probability $1-N^{-10}$, we have*

$$\|\widehat{\boldsymbol{M}}_1-\boldsymbol{M}_1\|\lesssim\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega|}} \tag{149}$$

provided that $|\Omega|\gtrsim(1+\delta^2)d\log^4 N$.

Lemma 18 ((Zhong et al., 2017b), Lemma E.6)**.** *Let $\boldsymbol{V}\in\mathbb{R}^{d\times K}$ be an orthogonal basis of $\boldsymbol{W}^*$ and $\widehat{\boldsymbol{V}}$ be generated by step 4 in Subroutine 1. Assume $\|\widehat{\boldsymbol{M}}_2-\boldsymbol{M}_2\|_2\le\sigma_K(\boldsymbol{M}_2)/10$. Then, for some small ε_0, we have*

$$\|\boldsymbol{V}\boldsymbol{V}^T-\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^T\|_2\le\frac{\|\boldsymbol{M}_2-\widehat{\boldsymbol{M}}_2\|}{\sigma_K(\boldsymbol{M}_2)}. \tag{150}$$

Lemma 19 ((Zhong et al., 2017b), Lemma E.13)**.** *Let $\boldsymbol{V}\in\mathbb{R}^{d\times K}$ be an orthogonal basis of $\boldsymbol{W}^*$ and $\widehat{\boldsymbol{V}}$ be generated by step 4 in Subroutine 1. Assume $\boldsymbol{M}_1$ can be written in the form of (6) with some homogeneous function ϕ_1, and let $\widehat{\boldsymbol{M}}_1$ be the estimation of $\boldsymbol{M}_1$ by samples. Let $\widehat{\alpha}$ be the optimal solution of (11) with $\widehat{\overline{\boldsymbol{w}}}_j=\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j$. Then, for each $j\in[K]$, if*

$$\begin{aligned}
T_1&:=\|\boldsymbol{V}\boldsymbol{V}^T-\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^T\|_2\le\frac{1}{\kappa^2\sqrt{K}},\\
T_2&:=\|\widehat{\boldsymbol{u}}_j-\widehat{\boldsymbol{V}}^T\overline{\boldsymbol{w}}_j\|_2\le\frac{1}{\kappa^2\sqrt{K}},\\
T_3&:=\|\widehat{\boldsymbol{M}}_1-\boldsymbol{M}_1\|_2\le\frac{1}{4}\|\boldsymbol{M}_1\|_2,
\end{aligned} \tag{151}$$

then we have

$$\Big|\|\boldsymbol{w}_j\|_2-\widehat{\alpha}_j\Big|\le\Big(\kappa^4K^{\frac{3}{2}}\big(T_1+T_2\big)+\kappa^2K^{\frac{1}{2}}T_3\Big)\|\boldsymbol{W}^*\|_2. \tag{152}$$

Proof of Lemma 4. we have

$$\begin{aligned}
\|\boldsymbol{w}_j^*-\widehat{\alpha}_j\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j\|_2\le&\Big\|\boldsymbol{w}_j^*-\|\boldsymbol{w}_j\|_2\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j+\|\boldsymbol{w}_j\|_2\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j-\widehat{\alpha}_j\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j\Big\|_2\\
\le&\Big\|\boldsymbol{w}_j^*-\|\boldsymbol{w}_j\|_2\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j\|_2\Big\|_2\Big\|\|\boldsymbol{w}_j\|_2\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j-\widehat{\alpha}_j\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j\Big\|_2\\
\le&\|\boldsymbol{w}_j^*\|_2\|\overline{\boldsymbol{w}}_j^*-\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j\|_2+\Big|\|\boldsymbol{w}_j\|_2-\widehat{\alpha}_j\Big\|_2\|\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j\|_2\\
\le&\sigma_1\big(\|\overline{\boldsymbol{w}}_j^*-\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^T\overline{\boldsymbol{w}}_j^*\|_2+\|\widehat{\boldsymbol{V}}^T\overline{\boldsymbol{w}}_j^*-\widehat{\boldsymbol{u}}_j\|_2\big)+\Big|\|\boldsymbol{w}_j\|_2-\widehat{\alpha}_j\Big|\\
:=&\sigma_1\big(I_1+I_2\big)+I_3.
\end{aligned} \tag{153}$$

From Lemma 18, we have

$$I_1=\|\overline{\boldsymbol{w}}_j^*-\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^T\overline{\boldsymbol{w}}_j^*\|_2\le\|\boldsymbol{V}\boldsymbol{V}^T-\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^T\|_2\le\frac{\|\widehat{\boldsymbol{M}}_2-\boldsymbol{M}_2\|_2}{\sigma_K(\boldsymbol{M}_2)}, \tag{154}$$

where the last inequality comes from Lemma 15. Then, from (7), we know that

$$\sigma_K(\boldsymbol{M}_2) \lesssim \min_{1\le j\le K} \|\boldsymbol{w}_j^*\|_2 \lesssim \sigma_K(\boldsymbol{W}^*). \tag{155}$$

From Theorem 3 in (Kuleshov et al., 2015), we have

$$I_2 = \|\widehat{\boldsymbol{V}}^T \overline{\boldsymbol{w}}_j^* - \widehat{\boldsymbol{u}}_j\|_2 \lesssim \frac{\kappa}{\sigma_K(\boldsymbol{W}^*)} \|\widehat{\boldsymbol{M}}_3(\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}}) - \boldsymbol{M}_3(\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}})\|_2. \tag{156}$$

To guarantee the condition (151) in Lemma 19 hold, according to Lemmas 15 and 16, we need $|\Omega| \gtrsim \kappa^3(1+\delta^2)Kd\log N$. Then, from Lemma 19, we have

$$I_3 = \Big(\kappa^4 K^{3/2}(I_1+I_2) + \kappa^2 K^{1/2}\|\widehat{\boldsymbol{M}}_1 - \boldsymbol{M}_1\|\Big)\|\boldsymbol{W}^*\|_2. \tag{157}$$

Since $d \gg K$, according to Lemmas 15, 16 and 17, we have

$$\|\boldsymbol{w}_j^* - \widehat{\alpha}_j \widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j\|_2 \lesssim \kappa^6 \sigma_1^2(\boldsymbol{A}) \sqrt{\frac{K^3(1+\delta^2)d\log N}{|\Omega|}} \|\boldsymbol{W}^*\|_2 \tag{158}$$

provided $|\Omega| \gtrsim (1+\delta^2)d\log^4 N$. □