Towards Understanding Generalization of Graph Neural Networks

Huayi Tang [1 2] **Yong Liu** [1 2 *]

Abstract

Graph neural networks (GNNs) are widely used in machine learning for graph-structured data. Even though GNNs have achieved remarkable success in real-world applications, understanding their working mechanism in theory is still on primary stage. In this paper, we move towards this goal from the perspective of generalization. Specifically, with consideration of stochastic optimization, we establish high probability bounds of generalization gap and gradients for transductive learning algorithms. After that, we provide high probability bounds of generalization gap for popular GNNs and analyze the factors affecting their generalization capability. These theoretical results reveal how the network architecture impacts the generalization gap. Experiments on benchmark datasets validate the theoretical findings. Our results provide new insights into understanding generalization of GNNs.

1. Introduction

Graph-structured data (Zhu et al., 2021) exists widely in real-world applications. As one of the most effective models to process graph-structured data, graph neural networks (GNNs) (Gori et al., 2005; Scarselli et al., 2009) have been widely adopted in computer vision (Qi et al., 2017; Johnson et al., 2018; Landrieu & Simonovsky, 2018; Satorras & Estrach, 2018), natural language processing (Bastings et al., 2017; Beck et al., 2018; Song et al., 2018), recommendation systems (Ying et al., 2018; Fan et al., 2019; He et al., 2020; Deng et al., 2022), and AI for science (Sanchez-Gonzalez et al., 2020; Pfaff et al., 2021; Shen et al., 2021; Han et al., 2022), among other areas. There are two main ways to view modern GNNs: spatial domain perspective (Kipf & Welling, 2017; Hamilton et al., 2017; Veličković et al., 2018; Xu et al., 2019) and spectral domain perspective (Defferrard et al., 2016; Gasteiger et al., 2019; Liao et al., 2019; Chien et al., 2021; He et al., 2021). The former regards GNNs as processes of combining and updating features based on adjacent relationships, while the latter treats GNNs as filtering functions applied to graph spectrum. Recent developments of GNNs are summarized in (Zhou et al., 2020; Wu et al., 2021; Zhang et al., 2022).

Despite the empirical success of GNNs, establishing theories to explain their behaviors is still in its infancy. Recent studies towards this direction include understanding oversmoothing (Li et al., 2018; Zhao & Akoglu, 2020; Oono & Suzuki, 2020a; Rong et al., 2020), interpretability (Ying et al., 2019; Luo et al., 2020; Vu & Thai, 2020; Yuan et al., 2020; 2021), expressiveness (Xu et al., 2019; Chen et al., 2019; Maron et al., 2019; Dehmamy et al., 2019; Feng et al., 2022), and generalization (Scarselli et al., 2018; Du et al., 2019; Verma & Zhang, 2019; Garg et al., 2020; Zhang et al., 2020; Oono & Suzuki, 2020b; Lv, 2021; Liao et al., 2021; Esser et al., 2021; Cong et al., 2021). In this work we focus on the last branch. Some prior works have used classical techniques such as Vapnik-Chervonenkis dimension (Scarselli et al., 2018), Rademacher complexity (Garg et al., 2020; Lv, 2021) and algorithm stability (Verma & Zhang, 2019) to provide generalization bounds for GCN (Kipf & Welling, 2017) and MPNN (Gilmer et al., 2017). However, these studies have to split the original graph into subgraphs composed of central nodes and their neighbors, which are then treated as independent instances. This setting significantly differs from real-world implementations (Yang et al., 2016; Kipf & Welling, 2017), where training nodes are sampled without replacement from the full set of nodes and test nodes are visible during training, leading to a gap between theory and practice.

To this end, recent studies (Oono & Suzuki, 2020b; Esser et al., 2021; Cong et al., 2021) have incorporated the learning schema of GNNs into the category of transductive learning and yielded more realistic results. However, these studies still suffer from some limitations. Firstly, the analysis in (Oono & Suzuki, 2020b) focuses on multi-scale GNNs, whose network architecture differs a lot from modern GNNs. Besides, their analysis is oriented to the AdaBoost-like optimization procedure, and it is unclear whether the tech-

*Corresponding author [1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China [2]Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China. Correspondence to: Huayi Tang <huayitang@ruc.edu.cn>, Yong Liu <liuyonggsai@ruc.edu.cn>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

nique can be applied to general optimization algorithms such as stochastic gradient descent (SGD). Secondly, the upper bound provided in (Esser et al., 2021) is of slow order and is inadequate for providing meaningful learning guarantees of node classification in large-scale scenarios. Finally, (Cong et al., 2021) only considers spectral-based GNNs with fixed coefficients, leaving spectral-based GNNs with learnable coefficients (Chien et al., 2021) unexplored.

Motivated by the aforementioned challenges, under transductive setting, we study the generalization gap of GNNs for node classification task with consideration of stochastic optimization. First, we establish high probability bounds of generalization gap and gradients under transductive setting, and derive high probability bounds of test error under gradient dominant condition. Next, we provide a comprehensive analysis on popular GNNs including both linear and non-linear models and derive the upper bound of the Lipschitz continuity and Hölder smoothness constants, by which we compare their generalization capability. These results show that SGC (Wu et al., 2019) and APPNP (Gasteiger et al., 2019) can achieve smaller generalization gap than GCN (Kipf & Welling, 2017). Besides, the unconstrained coefficients in GPRGNN (Chien et al., 2021) may lead to a large generalization gap. Our results reveal the reason why shallow models achieve comparable and even better performance than nonlinear models from the perspective of learning theory, and provide theoretical supports for widely used techniques such as early stopping and DropEdge (Rong et al., 2020). Experimental results on benchmark datasets show that the theoretical findings are generally consistent with practical evidences.

2. Related Work

2.1. Generalization Analysis of GNNs

Existing studies on generalization analysis of GNNs generally fall into two categories: graph classification tasks based studies and node classification tasks based studies.

Graph Classification Tasks. (Liao et al., 2021) is the first work to establish generalization bounds for GCN using the PAC-Bayesian approach. The authors in (Ju et al., 2023) further improve results in (Liao et al., 2021) and provide the lower bound. Besides, neural tangent kernels (Jacot et al., 2018) are also used to analyze the generalization of infinitely wide GNNs trained by gradient descent (Du et al., 2019). Different from these studies, we focus on more challenging node classification task.

Node Classification Tasks. In (Scarselli et al., 2018), the authors analyze the generalization capability of GNNs using Vapnik–Chervonenkis dimension. (Verma & Zhang, 2019) is the first work to provide generalization bounds for one-layer GCN through algorithm stability, which is later extended to multi-layer GCNs in (Zhou & Wang, 2021). The work in (Garg et al., 2020) converts the graph into individual local node-wise computation trees and establishes their generalization bound respectively via Rademacher Complexity. The aforementioned works rely on the process of converting a graph into subgraphs, which differs significantly from realistic implementations. Observing this, (Oono & Suzuki, 2020b) makes the first step that adopting the transductive learning framework to analyze multi-scale GNNs. This framework originates from (Vapnik, 1998; 2006), and is further developed in (El-Yaniv & Pechyony, 2006; 2007). The work most related to ours is (Cong et al., 2021) and (Esser et al., 2021), where the authors establish generalization bound for GNNs and its variants by transductive uniform stability and transductive Rademacher complexity respectively. However, the derived bound in (Esser et al., 2021) is of slow order, and it is still unclear whether their technique can be applied on SGD. Different from (Cong et al., 2021) that analyzing full-batch gradient descant, we consider a more complex setting, namely transductive learning under SGD, due to the involvement of randomness in optimization. Besides, there are some works orthogonal to ours, such as analyzing the generalization capability of GNNs training with topology-sampling (Li et al., 2022a) or on large random graphs (Keriven et al., 2020).

2.2. Out-of-Distribution (OOD) Generalization on Graphs

Recently, much effort has been devoted to the study of OOD generalization on graphs (Li et al., 2022b), due to the occurrence of distribution shifts in real-world scenarios. An adversarial learning schema (Wu et al., 2022) is proposed to minimize the mean and variance of risks from multiple environments. The authors propose a two-stage training schema to tackle distribution shift on molecular graphs in (Yang et al., 2022). Energy-based message passing schema is shown to be effective in enhancing the OOD detection performance of GNNs (Wu et al., 2023). The current work shows that the spurious performance of GNNs may come from its intrinsic generalization capability (Yang et al., 2023) rather than expressivity. Moreover, other recent work has also focused on reasoning (Xu et al., 2020), extrapolation ability (Xu et al., 2021; Bevilacqua et al., 2021), and generalization from small to large graphs (Yehudai et al., 2021).

3. Preliminaries

3.1. Notations

Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ be an given undirected graph with $n = |\mathcal{V}|$ nodes. Each node is an instance $z_i = (\mathbf{x}_i, y_i)$ containing feature $\mathbf{x}_i$ and label y_i from some space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $\mathbf{X}$ be the feature matrix where the i-th row $\mathbf{X}_{i*}$ is the node feature $\mathbf{x}_i$. Let $\mathbf{A}$ and $\mathbf{D}$ be the adjacency matrix and the diagonal degree matrix respectively, where $\mathbf{D}_{ii} = \sum_{j=1}^{n} \mathbf{A}_{ij}$. Denote by $\widetilde{\mathbf{A}} = (\mathbf{D}+\mathbf{I}_n)^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I}_n)(\mathbf{D}+\mathbf{I}_n)^{-\frac{1}{2}}$ the normalized adjacency matrix with self-loops and $\sqrt{|\mathcal{Y}|}$ the number of categories. We focus on the transductive learning setting where all features together with the randomly sampled labels are constructed as training set. Let $S = \{\mathbf{x}_i, y_i\}_{i=1}^{m+u}$ be the set of instances where $m + u = n$. Without loss of generality (W.l.o.g.), let $\{y_i\}_{i=1}^{m}$ be the selected labels, our task is to predict the labels of examples $\{\mathbf{x}_i\}_{i=m+1}^{m+u}$ by a learner (model) trained on $\{\mathbf{x}_i\}_{i=1}^{m+u} \bigcup \{y_i\}_{i=1}^{m}$. This setting is widely adopted in node classification task (Yang et al., 2016; Kipf & Welling, 2017) where the training and test nodes are determined by a random partition.

Now we consider a more concrete case, namely the model is set to a given GNN with learnable parameters $\{\mathbf{W}_h\}_{h=1}^{H}$. Since $\mathbb{R}^{p\times q}$ and $\mathbb{R}^{pq}$ are isomorphic, the analysis in this work is oriented to the vector space. To this end, we use a unified vector $\mathbf{w} = [\mathrm{vec}\,[\mathbf{W}_1]\,; \dots; \mathrm{vec}\,[\mathbf{W}_H]]$ to represent the collection of $\{\mathbf{W}_h\}_{h=1}^{H}$, where $\mathrm{vec}[\cdot]$ is the vectorization operator that transforms a given matrix into vector, namely $\mathrm{vec}\,[\mathbf{W}] = [\mathbf{W}_{*1}; \cdots; \mathbf{W}_{*q}]$ for $\mathbf{W} \in \mathbb{R}^{p\times q}$. Here $\mathbf{W}_{*i}$ is the i-th column of $\mathbf{W}$. For $\mathbf{w} \in \mathcal{W}$, the training and test error is defined as $R_m(\mathbf{w}) \triangleq \frac{1}{m}\sum_{i=1}^{m} \ell(\mathbf{w}; z_i)$ and $R_u(\mathbf{w}) \triangleq \frac{1}{u}\sum_{i=m+1}^{m+u} \ell(\mathbf{w}; z_i)$ respectively, where $\ell : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}_+$ is the loss function. In this work, we follow previous studies (El-Yaniv & Pechyony, 2007; Oono & Suzuki, 2020b; Esser et al., 2021) and define the transductive generalization gap by $|R_m(\mathbf{w}) - R_u(\mathbf{w})|$. Since the label of test examples are not available, the optimization process is to find parameters to minimize the the training error $R_m(\mathbf{w})$. Much efforts (Duchi et al., 2011; Kingma & Ba, 2015) are devoted to solving this stochastic optimization problem, and we mainly focus on SGD (Summarized in Algorithm 1) in this work.

We close this part by introducing additional notations used in our analysis. Denote by $\|\cdot\|_2$ and $\|\cdot\|$ the 2-norm of vector and spectral norm of matrix, respectively. Let $\mathbf{w}^{(1)}$ be the initialization weight of the model. We focus on the space $\mathcal{W} = B(\mathbf{w}^{(1)}; r), r \geq 1$ in this work, where $B(\mathbf{w}^{(1)}; r) \triangleq \left\{\mathbf{w} : \left\|\mathbf{w} - \mathbf{w}^{(1)}\right\|_2 < r\right\}$ is the ball with radius r. Denote by $\nabla\ell(\cdot; z)$ the gradient of ℓ with respective to (w.r.t.) the first argument. Denote by $b_g = \sup_{z\in\mathcal{Z}} \left\|\nabla\ell(\mathbf{w}^{(1)}; z)\right\|_2$ the supermum of gradient with initialed weight and $b_\ell = \sup_{z\in\mathcal{Z}} \left|\ell(\mathbf{w}^{(1)}; z)\right|_2$ the supermum of loss value with initialed weight. Let $\hat{\mathbf{w}} \in \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}} R_m(\mathbf{w})$ be the parameters of training error minimizer. We denote by $\sigma(\cdot)$ the activation function.

3.2. Assumptions

In this part, we present some assumptions used in this paper.

Assumption 3.1. Assume that there exists a constant $c_X > 0$ such that $\|\mathbf{x}\|_2 \leq c_X$ holds for all $\mathbf{x} \in \mathcal{X}$.

Algorithm 1 SGD for Transductive Learning

Input: Initial parameter $\mathbf{w}^{(1)}$, learning rates $\{\eta_t\}$, training set $\{\mathbf{x}_i\}_{i=1}^{m+u} \cup \{y_i\}_{i=1}^{m}$.
for $t = 1$ **to** T **do**
  Randomly draw j_t from the uniform distribution over the set $\{j : j \in [m]\}$.
  Update parameters by
  $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta_t \nabla\ell(\mathbf{w}^{(t)}; z_{j_t})$.
end for

Assumption 3.2. Assume that there exists a constant $c_W > 0$ such that $\|\mathbf{W}_h\| \leq c_W, h \in [H]$ for $\mathbf{w} \in B(\mathbf{w}^{(1)}; r)$.

Remark 3.3. Assumption 3.1 requires that input features are bounded (Verma & Zhang, 2019). This assumption can be met by applying normalization on features. Assumption 3.2 means that the spectral norm of each learnable parameter $\mathbf{W}_h$ (components of $\mathbf{w}$) is bounded during the training process, which is a common assumption in the generalization analysis of GNNs (Garg et al., 2020; Liao et al., 2021; Cong et al., 2021; Esser et al., 2021). These two assumptions are necessary to analyze the Lipschitz continuity and Hölder smoothness of the objective ℓ w.r.t. $\mathbf{w}$.

Assumption 3.4. Assume that the activation function $\sigma(\cdot)$ is α-Hölder smooth. Concretely, there exist constants $P > 0$ and $\alpha \in (0, 1]$, for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$:

$$\|\sigma'(\mathbf{u}) - \sigma'(\mathbf{v})\|_2 \leq P\|\mathbf{u} - \mathbf{v}\|_2^{\alpha}. \tag{1}$$

Remark 3.5. Assumption 3.4 implies Lipschitz continuity of the activation function if $\alpha = 0$. Moreover, it implies the smoothness of the activation function if $\alpha = 1$. Therefore, Assumption 3.4 is much milder than the assumption in previous work (Verma & Zhang, 2019; Cong et al., 2021), which requires the smoothness of activation function. For the convenience of analysis while not yielding a large gap between theory and practice, we construct a modified ReLU function (see Appendix A) with hyperparameter $q \in (1, 2]$ that satisfies Assumption 3.4 and has a tolerable approximate error to the vanilla ReLU function.

Assumption 3.6. Assume that there exists a constant $G > 0$ such that for all $z \in S$,

$$\sqrt{\eta_t}\, \|\nabla\ell(\mathbf{w}_t; z)\|_2 \leq G \tag{2}$$

holds $\forall\, t \in \mathbb{N}$, where $\{\eta_t\}_{t=1}^{T}$ is learning rates.

Remark 3.7. The formal definition of $\nabla\ell(\mathbf{w}; z)$ can be found in Lemma A.4 in the Appendix. Assumption 3.6 (Lei & Tang, 2021; Li & Liu, 2021) means that the product of gradient norm and square root of learning rate is bounded, which is milder than the widely used bounded gradient assumption (Hardt et al., 2016; Kuzborskij & Lampert, 2018), since the learning rate tends to zero during the iteration.

Assumption 3.8. Assume that there exists a constant $\sigma_0 > 0$ such that for $\forall\, t \in \mathbb{N}_+$, the following inequality holds

$$\mathbb{E}_{j_t}\left[\|\nabla \ell(\mathbf{w}_t; z_{j_t}) - \nabla R_m(\mathbf{w}_t)\|_2^2\right] \le \sigma_0^2. \tag{3}$$

Remark 3.9. Assumption 3.8 requires the boundness of variances of stochastic gradients, which is a standard assumption in stochastic optimization studies (Kuzborskij & Lampert, 2018; Lei & Tang, 2021; Li & Liu, 2021).

4. Theoretical Results

In this section, we first present the high probability bounds of generalization gap and gradients for general transductive learner trained by SGD. Afterwards, we turn to specific examples and provide results of some popular GNNs. Complete proofs can be found in the Appendix.

4.1. General Results of Transductive SGD

We first analyze properties of the objective function ℓ and provide the following proposition.

Proposition 4.1 (Informal)**.** *Suppose Assumptions 3.1, 3.2, and 3.4 hold. Denote by $\mathcal{F}$ a specific GNN, for any $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$ and $z \in S$, the objective $\ell(\mathbf{w}; z)$ satisfies*

$$|\ell(\mathbf{w}; z) - \ell(\mathbf{w}'; z)| \le L_{\mathcal{F}}\|\mathbf{w} - \mathbf{w}'\|_2, \tag{4}$$

and

$$\begin{aligned}&\|\nabla \ell(\mathbf{w}; z) - \nabla \ell(\mathbf{w}'; z)\|_2 \\ \le& P_{\mathcal{F}} \max\left\{\|\mathbf{w} - \mathbf{w}'\|_2^{\alpha}, \|\mathbf{w} - \mathbf{w}'\|_2\right\},\end{aligned} \tag{5}$$

with network dependent constants $L_{\mathcal{F}}$ and $P_{\mathcal{F}}$.

Remark 4.2. We provide a more detailed analysis of $L_{\mathcal{F}}$ and $P_{\mathcal{F}}$ in Subsection 4.2. Both $L_{\mathcal{F}}$ and $P_{\mathcal{F}}$ depend on the specific network architecture $\mathcal{F}$ of GNNs. Thus, the upper bound of generalization gap varies by the architecture.

Our first main result is high probability bounds of transductive generalization gap, as presented in Theorem 4.3.

Theorem 4.3. *Suppose Assumptions 3.1, 3.2, 3.4, 3.6, and 3.8 hold. Suppose that the learning rate $\{\eta_t\}$ satisfies $\eta_t = \frac{1}{t+t_0}$ such that $t_0 \ge \max\{(2P)^{1/\alpha}, 1\}$. For any $\delta \in (0,1)$, with probability $1-\delta$,*

(a). If $\alpha \in (0, \frac{1}{2})$, we have

$$\begin{aligned}&R_u(\mathbf{w}^{(T+1)}) - R_m(\mathbf{w}^{(T+1)}) \\ =&\mathcal{O}\left(L_{\mathcal{F}}\frac{(m+u)^{\frac{3}{2}}}{mu}\log^{\frac{1}{2}}(T)T^{\frac{1-2\alpha}{2}}\log\left(\frac{1}{\delta}\right)\right).\end{aligned}$$

(b). If $\alpha = \frac{1}{2}$, we have

$$\begin{aligned}&R_u(\mathbf{w}^{(T+1)}) - R_m(\mathbf{w}^{(T+1)}) \\ =&\mathcal{O}\left(L_{\mathcal{F}}\frac{(m+u)^{\frac{3}{2}}}{mu}\log(T)\log\left(\frac{1}{\delta}\right)\right).\end{aligned}$$

(c). If $\alpha \in (\frac{1}{2}, 1]$, we have

$$\begin{aligned}&R_u(\mathbf{w}^{(T+1)}) - R_m(\mathbf{w}^{(T+1)}) \\ =&\mathcal{O}\left(L_{\mathcal{F}}\frac{(m+u)^{\frac{3}{2}}}{mu}\log^{\frac{1}{2}}(T)\log\left(\frac{1}{\delta}\right)\right).\end{aligned}$$

Remark 4.4. Theorem 4.3 shows that the transductive generalization gap depends on the training data size m, test data size m/u, network architecture dependent Lipschitz continuity constant $L_{\mathcal{F}}$, and the number of iterations T. Our upper bounds are of order $\mathcal{O}\left((\frac{1}{m}+\frac{1}{u})\sqrt{m+u}\right)$, which is much sharper than the bound $\mathcal{O}\left((\frac{1}{m}+\frac{1}{u})(m+u)+\log(m+u)\right)$ in previous work (Esser et al., 2021). Note that as the size $m+u$ increase, the bound in (Esser et al., 2021) become increasing larger and fail to provide a reasonable generalization guarantee. This seriously restricts its application in large-scale node classification scenarios where the order of $m+u$ is usually in the millions. Our results can address this drawback and provide more applicable learning guarantees. Besides, the bound provided in (Esser et al., 2021) does not consider the specific optimization and has difficulty in revealing the influence of T on generalization gap. Our results show that the generalization gap becomes larger when the number of T increases, resulting in the over-fitting phenomenon. Thus, early stopping may be beneficial for yielding a smaller generalization gap, which is widely adopted in implementation of modern GNNs (Kipf & Welling, 2017; Chen et al., 2020). It can be seen that the generalization gap is positively related to the Lipschitz continuity constant $L_{\mathcal{F}}$ determined by specific network architecture $\mathcal{F}$. Thus, larger $L_{\mathcal{F}}$ leads to larger upper bounds of generalization gap, showing that the network architecture of GNN also have a significant influence on the generalization gap (see Subsection 4.2). The upper bound of generalization gap in (Cong et al., 2021) also increases with T when the objective is optimized by full-batch gradient descent. This is not surprise since it can be seen as a special case of SGD where the batch size is equal to the size of training examples.

Our second main result is high probability bounds of the gradients on training and test data.

Theorem 4.5. *Suppose Assumptions 3.1, 3.2, 3.4, 3.6, and 3.8 hold. Suppose that the learning rate $\{\eta_t\}$ satisfies $\eta_t = \frac{1}{t+t_0}$ such that $t_0 \ge \max\{(2P)^{1/\alpha}, 1\}$. For any $\delta \in (0,1)$, with probability $1-\delta$,*

(a). If $\alpha \in (0, \frac{1}{2})$, we have

$$\begin{aligned}&\left\|\nabla R_m(\mathbf{w}^{(T+1)}) - \nabla R_u(\mathbf{w}^{(T+1)})\right\|_2 \\ =&\mathcal{O}\left(\frac{(m+u)^{\frac{3}{2}}}{mu}\log^{\frac{1}{2}}(T)T^{\frac{1-2\alpha}{2}}\log\left(\frac{1}{\delta}\right)\right).\end{aligned}$$

(b). If $\alpha = \frac{1}{2}$, we have

$$\left\|\nabla R_m(\mathbf{w}^{(T+1)}) - \nabla R_u(\mathbf{w}^{(T+1)})\right\|_2 = \mathcal{O}\left(\frac{(m+u)^{\frac{3}{2}}}{mu}\log(T)\log\left(\frac{1}{\delta}\right)\right).$$

(c). If $\alpha \in (\frac{1}{2}, 1]$, we have

$$\left\|\nabla R_m(\mathbf{w}^{(T+1)}) - \nabla R_u(\mathbf{w}^{(T+1)})\right\|_2 = \mathcal{O}\left(\frac{(m+u)^{\frac{3}{2}}}{mu}\log^{\frac{1}{2}}(T)\log\left(\frac{1}{\delta}\right)\right).$$

Remark 4.6. Theorem 4.5 provides high probability bounds for the gap of gradients on training examples and test examples under transductive setting. Overall, the generalization gap we derive is still of order $\mathcal{O}\left((\frac{1}{m}+\frac{1}{u})\sqrt{m+u}\right)$. Besides, the generalization gap of gradients increases with the increase of T, showing that a smaller number of iterations helps achieving a smaller gap of gradients.

Since the generalization performance (often measured by test error) is determined by both training error and generalization gap, we further provide a upper bound of the test error under a special case that the objective satisfies the following PL condition.

Assumption 4.7. Suppose that there exists a constant μ such that for all $\mathbf{w} \in \mathcal{W}$,

$$R_m(\mathbf{w}) - R_m(\hat{\mathbf{w}}^*) \leq \frac{1}{2\mu}\left\|\nabla R_m(\mathbf{w})\right\|_2^2, \tag{6}$$

holds for the given set S from $\mathcal{Z}$.

Remark 4.8. Assumption 4.7 is also named as gradient dominance condition in learning theory studies, indicating that the difference between current training error and the optimal training error can be upper bounded by the quadratic function of the gradient on training instances. This assumption is widely adopted in nonconvex learning (Zhou et al., 2018; Xu & Zeevi, 2020; Lei & Tang, 2021; Li & Liu, 2021), and has been verified in over-parameterized systems including wide neural networks (Liu et al., 2020). This assumption only appears in Corollary 4.9.

Corollary 4.9. *Suppose Assumptions 3.1, 3.2, 3.4, 3.6, 3.8, and 4.7 hold. Suppose that the learning rate $\{\eta_t\}$ satisfies $\eta_t = \frac{2}{\mu(t+t_0)}$ such that $t_0 \geq \max\{\frac{2}{\mu}(2P)^{\frac{1}{\alpha}}, 1\}$. For any $\delta \in (0,1)$, with probability $1-\delta$,*

(a). If $\alpha \in (0, \frac{1}{2})$, we have

$$R_u(\mathbf{w}^{(T+1)}) - R_m(\mathbf{w}^*) = \mathcal{O}\left(L_{\mathcal{F}}\frac{(m+u)^{\frac{3}{2}}}{mu}\log^{\frac{1}{2}}(T)T^{\frac{1-2\alpha}{2}}\log(1/\delta) + \frac{1}{T^{\alpha}}\right),$$

(b). If $\alpha = \frac{1}{2}$, we have

$$R_u(\mathbf{w}^{(T+1)}) - R_m(\mathbf{w}^*) = \mathcal{O}\left(L_{\mathcal{F}}\frac{(m+u)^{\frac{3}{2}}}{mu}\log(T)\log(1/\delta) + \frac{1}{T^{\alpha}}\right).$$

(c). If $\alpha \in (\frac{1}{2}, 1)$, we have

$$R_u(\mathbf{w}^{(T+1)}) - R_m(\mathbf{w}^*) = \mathcal{O}\left(L_{\mathcal{F}}\frac{(m+u)^{\frac{3}{2}}}{mu}\log^{\frac{1}{2}}(T)\log(1/\delta) + \frac{1}{T^{\alpha}}\right).$$

(d). If $\alpha = 1$, we have

$$R_u(\mathbf{w}^{(T+1)}) - R_u(\mathbf{w}^*) = \mathcal{O}\left(L_{\mathcal{F}}\frac{(m+u)^{\frac{3}{2}}}{mu}\log^{\frac{1}{2}}(T)\log(1/\delta) + \frac{\log(T)\log^3(1/\delta)}{T}\right).$$

Remark 4.10. Theorem 4.9 shows that under Assumption 4.7, the test error are determined by the minimal training error, optimization error and generalization gap. The minimal training error $R_m(\mathbf{w}^*)$ reflects how well the model fits the data, which is a measure of the model's expressive ability. The first and the second term in the slack terms are generalization gap and optimization error, respectively. With the increase of T, the generalization gap increase while the optimization error decrease. Therefore, it is necessary to carefully choose a proper number of iterations in order to balance the trade-off between optimization and generalization. In the implementation of most GNNs studies (Kipf & Welling, 2017; Veličković et al., 2018; Chien et al., 2021; He et al., 2021), early stopping is adopted and T is determined by the performance of model on validation set. Thus, our results are consistent with real implementations.

The above analysis can be applied to any model trained with SGD in a transductive learning context. In the next part, we will narrow our focus to several popular GNN models.

4.2. Case Studies of Popular GNNs

We have established high probability bounds of transductive generalization gap in Theorem 4.3. In this part, we analyze the upper bounds of architecture dependent constant $L_{\mathcal{F}}$ and $P_{\mathcal{F}}$, with that the upper bound of generalization gap can be determined. We select five representative GNNs for case study: GCN, GCNII, SGC, APPNP, and GPRGNN. We assume that ℓ is cross-entropy loss and denote by $\hat{\mathbf{Y}}$ the prediction. For concise, we do not consider the bias term, since it can be verified that $\langle \mathbf{w}, \mathbf{x}\rangle + b = \langle \widetilde{\mathbf{w}}, \widetilde{\mathbf{x}}\rangle$ holds with $\widetilde{\mathbf{w}} = [\mathbf{w}; b]$ and $\widetilde{\mathbf{x}} = [\mathbf{x}; 1]$.

GCN. The authors in (Kipf & Welling, 2017) introduce the renormalization trick and propose to aggregate features from one-hop neighbor nodes. The feature propagation process of a two-layer GCN model is

$$\hat{\mathbf{Y}} = \mathrm{Softmax}\big(\widetilde{\mathbf{A}}\sigma(\widetilde{\mathbf{A}}\mathbf{X}\mathbf{W}_1)\mathbf{W}_2\big), \tag{7}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d\times h}, \mathbf{W}_2 \in \mathbb{R}^{h\times|\mathcal{Y}|}$ are parameters.

Proposition 4.11. *Suppose Assumptions 3.1, 3.2, and 3.4 hold, then the objective $\ell(\mathbf{w}; z)$ satisfies Eqs. (4, 5) with $\mathbf{w} = [\mathrm{vec}[\mathbf{W}_1]; \mathrm{vec}[\mathbf{W}_1]]$. Concretely, the Lipschitz continuity constant $L_\mathcal{F}$ is $L_{\mathrm{GCN}} = 2c_X c_W \big\|\widetilde{\mathbf{A}}\big\|_\infty^2$.*

Due to the tedious formulation, the concrete value of $P_\mathcal{F}$ is provided in the Appendix. Proposition 4.11 demonstrates that L_{GCN} depends on factors $\big\|\widetilde{\mathbf{A}}\big\|_\infty$, c_X, and c_W. Let $\mathrm{deg}_{\min}$ and $\mathrm{deg}_{\max}$ be the minimum and maximum node degree, respectively, $\big\|\widetilde{\mathbf{A}}\big\|_\infty$ can be further bounded by

$$\big\|\widetilde{\mathbf{A}}\big\|_\infty \leq \sqrt{\frac{\mathrm{deg}_{\max}+1}{\mathrm{deg}_{\min}+1}}. \tag{8}$$

It can be found that the generalization gap decreases with the decrease of the maximum node degree, which could be achieved by removing edges. This explains why the DropEdge (Rong et al., 2020) technique is beneficial for alleviating the over-fitting problem from the perspective of learning theory. Besides, for GCN trained on sampled sub-graphs $\{\mathcal{G}_i\}_{i=1}^n$, the Lipschitz continuity constant is $L_{\mathrm{GCN}} = 2c_X c_W \max_{i\in[n]} \big\|\widetilde{\mathbf{A}}^{[i]}\big\|_\infty^2$, where $\widetilde{\mathbf{A}}^{[i]}$ is the normalized adjacency matrix with self-loop of $\mathcal{G}_i$. Since only a portion of neighboring nodes are preserved during sub-graphs sampling (Hamilton et al., 2017; Zeng et al., 2020; 2021), the maximum node degree of each sub-graph is smaller than that of initial graph, implying $\max_{i\in[n]} \big\|\widetilde{\mathbf{A}}^{[i]}\big\|_\infty \leq \big\|\widetilde{\mathbf{A}}\big\|_\infty$ holds. Thus, Proposition 4.11 shows that training on sampled sub-graphs are beneficial to achieve smaller generalization gap. Lastly, the spectral norm of learning parameters also has an effect on the generalization gap. Thus, the commonly used L_2 regularization technique is beneficial to reduce the generalization gap.

GCNII. The authors in (Chen et al., 2020) propose to relieve over-smoothing by initial residual and identity mapping. Denote by $\mathbf{H}^{(0)} = \sigma(\mathbf{X}\mathbf{W}_0)$ the initial representation. The forward propagation of a two-layer GCNII model is

$$\begin{aligned}\mathbf{H}^{(1)} &= \sigma\Big(\Phi(\alpha_1, \mathbf{H}^{(0)})\Psi(\beta_1, \mathbf{W}_1)\Big),\\ \mathbf{H}^{(2)} &= \sigma\Big(\Phi(\alpha_2, \mathbf{H}^{(1)})\Psi(\beta_2, \mathbf{W}_2)\Big),\\ \hat{\mathbf{Y}} &= \mathrm{Softmax}\big(\mathbf{H}^{(2)}\mathbf{W}_3\big),\end{aligned} \tag{9}$$

where $\Phi(\alpha, \mathbf{H}) = (1-\alpha)\widetilde{\mathbf{A}}\mathbf{H} + \alpha\mathbf{H}^{(0)}$ and $\Psi(\beta, \mathbf{W}) = (1-\beta)\mathbf{I} + \beta\mathbf{W}$. Here $\mathbf{W}_0 \in \mathbb{R}^{d\times h}$, $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{h\times h}$ and $\mathbf{W}_3 \in \mathbb{R}^{h\times|\mathcal{Y}|}$ are parameters.

Proposition 4.12. *Suppose Assumptions 3.1, 3.2, and 3.4 hold, then the objective $\ell(\mathbf{w}; z)$ satisfies Eqs. (4, 5) with $\mathbf{w} = [\mathrm{vec}[\mathbf{W}_0]; \mathrm{vec}[\mathbf{W}_1]; \mathrm{vec}[\mathbf{W}_2]; \mathrm{vec}[\mathbf{W}_3]]$. Concretely, denote by $B_0 = c_X c_W$ and*

$$\begin{aligned}&C_\ell = 1 - \beta_\ell + \beta_\ell c_W,\\ &B_\ell = \big((1-\alpha_\ell)B_{\ell-1}\big\|\widetilde{\mathbf{A}}\big\|_\infty + \alpha_\ell c_X c_W\big)C_\ell,\end{aligned} \tag{10}$$

then the Lipschitz continuity constant $L_\mathcal{F}$ is $L_{\mathrm{GCNII}} = \sqrt{L_1 + L_2}$, where

$$\begin{aligned}&L_1 = 2\bigg(2 + \frac{c_W^2\beta_2^2}{C_2^2}\bigg)B_2^2,\\ &L_2 = 2(1-\alpha_2)^2\beta_1^2 c_W^2 \big\|\widetilde{\mathbf{A}}\big\|_\infty^2 \bigg(\frac{B_1^2 C_2^2}{C_1^2}\bigg).\end{aligned} \tag{11}$$

Proposition 4.12 shows that L_{GCNII} is a function of $\{\alpha_i\}_{i=1}^2$ and $\{\beta_i\}_{i=1}^2$. Finding the optimal value of L_{GCNII} is a quadratic programming problem with constrains $\alpha_1, \alpha_2 \in [0,1]$ and $\beta_1, \beta_2 \in [0,1]$. Now we discuss a special case that $\alpha_1 = \alpha_2 = 0$ and $\beta_1 = \beta_2 = 0$. In this case, we have $L_1 = 4c_X^2 c_W^2 \big\|\widetilde{\mathbf{A}}\big\|_\infty^4$ and $L_2 = 0$, which implies that $L_{\mathrm{GCNII}} = L_{\mathrm{GCN}}$. Note that the optimal value of L_{GCNII} is no larger than any value of objective function over the feasible region. Therefore, we conclude that $L_{\mathrm{GCNII}} \leq L_{\mathrm{GCN}}$ holds when $\{\alpha_i\}_{i=1}^2$ and $\{\beta_i\}_{i=1}^2$ take their optimal values. This result is not surprise, since GCN is a special GCNII under this setting. For proper value of $\{\alpha_i\}_{i=1}^2$ and $\{\beta_i\}_{i=1}^2$, GCNII could achieve smaller generalization gap than GCN. As GCNII can achieve lower training error by relieving the over-smoothing problem, Proposition 4.12 indicates that GCNII can achieve superior performance when hyperparameters are set properly. Due to the involve of $\{\alpha_i\}_{i=1}^2$ and $\{\beta_i\}_{i=1}^2$, the growth rate of L_{GCNII} is much smaller than L_{GCN} when propagation depth increases, which makes GCNII maintain generalization capability and achieve stable performance.

SGC. The work (Wu et al., 2019) proposes to remove all nonlinear activations in GCN. To facilitate comparison with GCN, we consider a two layers SGC model, whose propagation is given by

$$\hat{\mathbf{Y}} = \mathrm{Softmax}\big(\widetilde{\mathbf{A}}^2\mathbf{X}\mathbf{W}_1\mathbf{W}_2\big), \tag{12}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d\times h}$ and $\mathbf{W}_2 \in \mathbb{R}^{h\times|\mathcal{Y}|}$ are parameters.

Proposition 4.13. *Suppose Assumption 3.1, 3.2, and 3.4 hold, then the objective $\ell(\mathbf{w}; z)$ satisfies Eqs. (4, 5) with $\mathbf{w} = [\mathrm{vec}[\mathbf{W}_1]; \mathrm{vec}[\mathbf{W}_2]]$. Concretely, the Lipschitz continuity constant $L_\mathcal{F}$ is $L_{\mathrm{SGC}} = 2c_X c_W \big\|\widetilde{\mathbf{A}}^2\big\|_\infty$.*

Since $\big\|\widetilde{\mathbf{A}}^2\big\|_\infty \leq \big\|\widetilde{\mathbf{A}}\big\|_\infty^2$, we have $L_{\mathrm{SGC}} \leq L_{\mathrm{GCN}}$. Surprisingly, this simple linear model can achieve better smaller generalization gap than nonlinear models (Kipf & Welling, 2017; Veličković et al., 2018), even though its representation ability is inferior than them. Note that the performance on test examples is determined by both training error and generalization gap. If linear GNNs can achieve a small training error, it is natural that they can achieve comparable and even better performance than nonlinear GNNs on test examples. Therefore, Proposition 4.13 reveals why linear GNNs achieve better performance than nonlinear GNNs from learning theory, as observed in recent works (Zhu & Koniusz, 2021; Wang et al., 2021). Considering the efficiency and scalability of linear GNNs on large-scale datasets, we believe that they have much potential to be exploited.

APPNP. Multi-scale features are aggregated via personalized PageRank schema in (Gasteiger et al., 2019). Formally, the feature propagation process is formulated as

$$\hat{\mathbf{Y}} = \mathrm{Softmax}\big(g(\widetilde{\mathbf{A}})\sigma(\sigma(\mathbf{X}\mathbf{W}_1)\mathbf{W}2)\big), \tag{13}$$

where $g(\widetilde{\mathbf{A}}) = \sum_{k=0}^{K-1} \gamma(1-\gamma)^k \widetilde{\mathbf{A}}^k + (1-\gamma)^K \widetilde{\mathbf{A}}^K$. $\mathbf{W}_1 \in \mathbb{R}^{d\times h}$ and $\mathbf{W}_2 \in \mathbb{R}^{h\times|\mathcal{Y}|}$ are parameters.

Proposition 4.14. *Suppose Assumption 3.1, 3.2, and 3.4 hold, then the objective $\ell(\mathbf{w}; z)$ satisfies Eqs. (4, 5) with* $\mathbf{w} = [\mathrm{vec}[\mathbf{W}_1]; \mathrm{vec}[\mathbf{W}_2]]$. *Concretely, the Lipschitz continuity constant $L_{\mathcal{F}}$ is $L_{\mathrm{APPNP}} = 2c_X c_W \|g(\widetilde{\mathbf{A}})\|_\infty$.*

L_{APPNP} in Proposition 4.14 is positively related to the infinity matrix norm of the polynomial spectral filter. According to (Gasteiger et al., 2019), the value of γ is commonly set to be a small number, resulting in $\|g(\widetilde{\mathbf{A}})\|_\infty < \|\widetilde{\mathbf{A}}\|_\infty$ holds. Therefore, the Lipschitz continuity constant of APPNP can be smaller than that of GCN for proper value of γ, indicating that APPNP may achieve smaller generalization gap than GCN. Besides, K also affects the value of $\|g(\widetilde{\mathbf{A}})\|_\infty$, and a larger K may yield a larger generalization gap. Therefore, K is usually set as a proper value to achieve a trade-off between expressiveness and generalization performance.

GPRGNN. Compared with APPNP, the fixed coefficients are replaced by learnable weights in (Chien et al., 2021), in order to adaptively simulate both high-pass and low-pass graph filters. The feature propagation process is

$$\hat{\mathbf{Y}} = \big(g(\widetilde{\mathbf{A}}, \boldsymbol{\gamma})\sigma(\sigma(\mathbf{X}\mathbf{W}_1)\mathbf{W}2)\big), \tag{14}$$

where $g(\widetilde{\mathbf{A}}, \boldsymbol{\gamma}) = \sum_{k=0}^{K} \gamma_k \widetilde{\mathbf{A}}^k$. $\mathbf{W}_1 \in \mathbb{R}^{d\times h}$, $\mathbf{W}_2 \in \mathbb{R}^{h\times|\mathcal{Y}|}$ and $\boldsymbol{\gamma} \in \mathbb{R}^{K+1}$ are parameters.

Proposition 4.15. *Suppose Assumption 3.1, 3.2, and 3.4 hold, then the objective $\ell(\mathbf{w}; z)$ satisfies Eqs. (4, 5) with* $\mathbf{w} = [\mathrm{vec}[\mathbf{W}_1]; \mathrm{vec}[\mathbf{W}_2]; \boldsymbol{\gamma}]$. *Concretely, the Lipschitz continuity constant $L_{\mathcal{F}}$ is $L_{\mathrm{GPR}} = \sqrt{L_1^2 + L_2^2}$, where*

$$\begin{aligned} L_1 &= \sqrt{2} c_X c_W^2 \left(\sum_{k=0}^{K} \|\widetilde{\mathbf{A}}^k\|_\infty \right)^{\frac{1}{2}}, \\ L_2 &= 2c_X c_W \|g(\widetilde{\mathbf{A}}, \boldsymbol{\gamma})\|_\infty. \end{aligned} \tag{15}$$

Table 1. Loss gap comparison of different baseline models on Cora, Citeseer and Pubmed.

	Cora	Citeseer	Pubmed
GCN	0.30±0.03	0.77±0.04	0.03±0.01
GCN*	0.91±0.18	2.12±0.16	0.05±0.01
GAT	0.29±0.03	0.65±0.02	0.03±0.01
GCNII	0.19±0.03	0.43±0.02	0.02±0.01
GCNII*	0.16±0.03	0.43±0.03	0.02±0.01
SGC	0.12±0.03	0.28±0.02	0.01±0.00
APPNP	0.16±0.03	0.25±0.02	0.01±0.00
GPRGNN	0.24±0.03	0.55±0.02	0.02±0.00

Table 2. Accuracy gap comparison of different baseline models on Cora, Citeseer and Pubmed.

	Cora	Citeseer	Pubmed
GCN	9.76±1.15	22.11±1.26	1.08±0.52
GCN*	13.45±1.28	26.48±1.21	1.49±0.63
GAT	11.00±0.75	22.69±0.84	1.52±0.43
GCNII	7.69±1.48	14.85±0.80	0.88±0.52
GCNII*	6.24±1.59	13.49±1.39	0.80±0.50
SGC	5.33±1.58	11.50±1.09	0.73±0.54
APPNP	7.72±1.54	9.99±1.17	0.85±0.46
GPRGNN	8.90±1.22	19.08±0.95	0.96±0.49

Note that L_2 has similar form with L_{APPNP} and the only difference is that $g(\widetilde{\mathbf{A}})$ is replaced by $g(\widetilde{\mathbf{A}}, \boldsymbol{\gamma})$. Assume that $g(\widetilde{\mathbf{A}}, \boldsymbol{\gamma}) = g(\widetilde{\mathbf{A}})$ and note that $L_{\mathrm{GPR}} = \sqrt{L_1^2 + L_2^2} \geq L_2$, we have $L_{\mathrm{GPR}} \geq L_{\mathrm{APPNP}}$. Besides, since there is no constraint on γ, the value of $\|g(\widetilde{\mathbf{A}}, \boldsymbol{\gamma})\|_\infty$ may be larger when the norm of $\boldsymbol{\gamma}$ is large, resulting in larger generalization gap than APPNP. Therefore, adopting regularization technique on the learnable coefficients to restrict the value of $\|g(\widetilde{\mathbf{A}}, \boldsymbol{\gamma})\|_\infty$ is necessary.

To summarize, $L_{\mathcal{F}}$ and $P_{\mathcal{F}}$ are determined by the feature propagation process and graph-structured data. Estimating these constants precisely is challenging (Virmaux & Scaman, 2018; Fazlyab et al., 2019), and the upper bounds we provided are sufficient to reflect the realistic generalization gap of these models (see Section 5 for more detail). Besides, we have to emphasize that results for GCN and GCNII with more than two layers can be derived by similar techniques, yet it requires more tedious computation. Exploring new techniques to estimate these constants conveniently and precisely are left for future work.

5. Experiments

Experimental Setup. We conduct experiments on widely adopted benchmark datasets, including Cora, Citeseer, and Pubmed (Sen et al., 2008; Yang et al., 2016). The accuracy and loss gap (*i.e.*, the absolute value of difference between

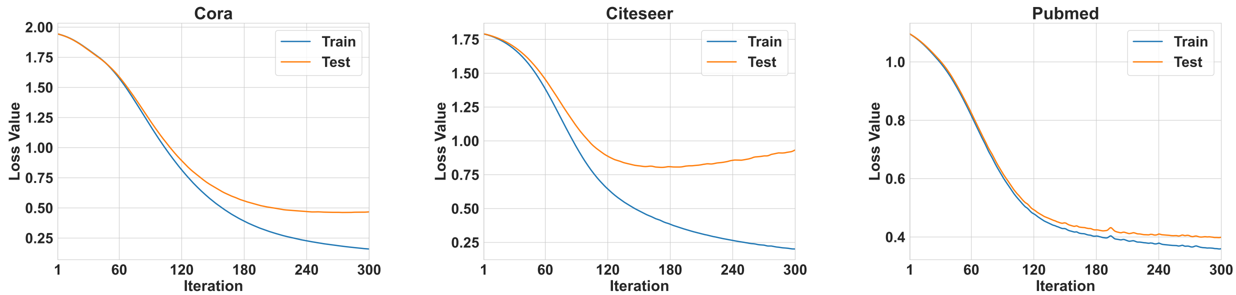

Figure 1. The loss value of GCN on training and test examples with the increase of iterations.

Table 3. Test accuracy comparison of different baseline models on Cora, Citeseer and Pubmed.

	Cora	Citeseer	Pubmed
GCN	85.91±0.53	71.78±0.72	85.29±0.19
GCN*	82.49±0.59	66.74±1.07	84.21±0.26
GAT	86.10±0.51	72.90±0.65	85.45±0.26
GCNII	82.85±2.17	73.61±0.64	84.70±0.24
GCNII*	82.85±2.44	72.89±0.96	83.67±0.46
SGC	82.39±2.48	74.37±0.56	82.00±0.27
APPNP	79.14±3.17	74.12±0.62	82.86±0.29
GPRGNN	87.24±0.71	73.79±0.67	85.07±0.34

Table 4. Loss gap, accuracy gap and test accuracy comparison of baseline models on ogbn-arxiv.

	Loss Gap	Accuracy Gap	Test Accuracy
GCN	0.04±0.01	2.39±0.48	66.92±0.57
GCNII	0.03±0.01	1.47±0.57	67.35±0.58
SGC	0.01±0.00	1.89±0.15	66.89±0.16
APPNP	0.03±0.00	1.26±0.25	68.20±0.22
GPRGNN	0.03±0.00	1.34±0.30	69.30±0.30

the loss (accuracy) on training and test instances) are used to estimate the generalization gap. Following the standard transductive learning setting, in each run, 30% sampled nodes determined by a random seed are used as training set and the rest nodes are treated as test set. The number of iterations is fixed to $T = 300$. Moreover, we also conduct experiments on large-scale dataset ogbn-arxiv (Hu et al., 2020). We set $T = 700$ and adopt the standard split. We independently repeat the experiments 10 times and report the mean value and standard deviations of all runs. Please refer to Appendix C for more detailed settings.

Experimental Results. The loss and accuracy comparisons are presented in Table 1 and Table 2, respectively. The test accuracy of baseline models are shown in Table 3. We have the following observations: (1) SGC and APPNP have smaller loss and accuracy gap than other models including GCN, which is consistent with the analysis in Subsection 4.2. Besides, the test accuracy of SGC surpass GCN on Citeseer. Therefore, the reason why linear models sometimes perform better than non-linear models is that their smaller Lipschitz continuity constants. (2) Compared with GCN, GCNII achieves smaller loss and accuracy gap with the same number of layers. We further estimate the generalization performance of GCN and GCNII with six layers (denoted as GCN* and GCNII*). Interestingly, with the increase of the number of hidden layers, the generalization gap of GCN decreases sharply. On the contrary, the loss and accuracy gap of GCNII remain unchanged. The test accuracy of GCNII also remain unchanged or only drops slightly. Therefore, the superior performance of GCNII comes from two perspectives: the first is learning non-degenerated representations by relieving over-smoothing and the second is robust generalization gap against the increase of the number of layers. (3) Although GPRGNN achieve a competitive test accuracy, it has higher accuracy and loss gap than APPNP. Therefore, the unconstrained learning coefficients improve the fitting ability but also weaken generalization capability. Designing coefficients learning schema to balance the expressive and generalization could be a direction for spectral-based GNNs. (4) The accuracy gap of GAT is slightly inferior to that of GCN. Since GAT is designed for inductive learning while our experimental setting is transducive, the superiority of GAT is not so obvious.

The results on ogbn-arxiv are presented in Table 4 and are generally consistent with those of Cora, Citeseer, and Pubmed. Additionally, loss values of GCN on training and test examples over iterations are presented in Figure 1. The loss gap increases with the increase of iterations, as demonstrated by Theorem 4.3. Overall, the theoretical results align with the experimental results. Note that our analysis is solely focused on the generalization gap. A smaller generalization gap does not necessarily mean better generalization ability, since the performance on test examples are determined by both training error and generalization gap. Estimating the training error need to consider the optimization process, which is a promising direction for future study.

6. Discussion and Conclusion

In this paper, we establish high probability learning guarantees for transductive SGD, and then derive the upper bound of generalization gap for some popular GNNs. We conduct experiments on benchmark datasets, and the results support the theoretical results. This work sheds light on understanding generalization of GNNs and provides insights into designing new GNN architecture with both expressive and generalization capabilities.

Although we have made efforts in generalization theory of GNNs, there are still some limitations in our analysis: (1) The complexity-based technique makes the dimension of parameters appear in the bounds. Further research should focus on establishing dimension-independent bounds under milder assumption and deriving the lower bound that matches the upper bound. (2) We only analyze vanilla SGD in terms of optimization algorithms. Extending our results to SGD with momentum and adaptive learning rates is worth exploring. (3) Our analysis does not explicitly consider the heterophily of graphs. Establishing heterophily-dependent generalization bounds is an interesting direction.

Acknowledgements

We sincerely appreciate the anonymous reviewers for their helpful suggestions and constructive comments. This work is supported by the National Natural Science Foundation of China (NO.62076234); the Beijing Natural Science Foundation (NO.4222029); the Intelligent Social Governance Interdisciplinary Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double First Class" Initiative, Renmin University of China; the Beijing Outstanding Young Scientist Program (NO.BJJWZYJH012019100020098); the Public Computing Cloud, Renmin University of China; the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (NO.2021030199); the Huawei-Renmin University joint program on Information Retrieval; the Unicom Innovation Ecological Cooperation Plan; the CCF Huawei Populus Grove Fund; and the National Key Research and Development Project (NO.2022YFB2703102).

References

Bastings, J., Titov, I., Aziz, W., Marcheggiani, D., and Sima'an, K. Graph convolutional encoders for syntax-aware neural machine translation. In *Conference on Empirical Methods in Natural Language Processing*, pp. 1957–1967, 2017.

Beck, D., Haffari, G., and Cohn, T. Graph-to-sequence learning using gated graph neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 273–283, 2018.

Bevilacqua, B., Zhou, Y., and Ribeiro, B. Size-invariant graph representations for graph classification extrapolations. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 837–851, 2021.

Chen, M., Wei, Z., Huang, Z., Ding, B., and Li, Y. Simple and deep graph convolutional networks. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 1725–1735, 2020.

Chen, Z., Villar, S., Chen, L., and Bruna, J. On the equivalence between graph isomorphism testing and function approximation with gnns. In *Advances in Neural Information Processing Systems*, pp. 15868–15876, 2019.

Chien, E., Peng, J., Li, P., and Milenkovic, O. Adaptive universal generalized pagerank graph neural network. In *International Conference on Learning Representations*, 2021.

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

Defferrard, M., Bresson, X., and Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, 2016.

Dehmamy, N., Barabasi, A.-L., and Yu, R. Understanding the representation power of graph neural networks in learning graph topology. In *Advances in Neural Information Processing Systems*, 2019.

Deng, L., Lian, D., Wu, C., and Chen, E. Graph convolution network based recommender systems: Learning guarantee and item mixture powered strategy. In *Advances in Neural Information Processing Systems*, 2022.

Devroye, L., Györfi, L., and Lugosi, G. *A Probablistic Theory of Pattern Recognition*. Springer, 1996.

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

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61): 2121–2159, 2011.

El-Yaniv, R. and Pechyony, D. Stable transductive learning. In *Conference on Learning Theory*, pp. 35–49, 2006.

El-Yaniv, R. and Pechyony, D. Transductive rademacher complexity and its applications. In *Conference on Learning Theory*, pp. 157–171, 2007.

Esser, P. M., Vankadara, L. C., and Ghoshdastidar, D. Learning theory can (sometimes) explain generalisation in graph neural networks. In *Advances in Neural Information Processing Systems*, pp. 27043–27056, 2021.

Fan, W., Ma, Y., Li, Q., He, Y., Zhao, Y. E., Tang, J., and Yin, D. Graph neural networks for social recommendation. In *The World Wide Web Conference*, pp. 417–426, 2019.

Fazlyab, M., Robey, A., Hassani, H., Morari, M., and Pappas, G. Efficient and accurate estimation of lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Federer, H. *Geometric Measure Theory*. Springer, 1969.

Feng, J., Chen, Y., Li, F., Sarkar, A., and Zhang, M. How powerful are k-hop message passing graph neural networks. In *Advances in Neural Information Processing Systems*, 2022.

Fey, M. and Lenssen, J. E. Fast graph representation learning with pytorch geometric. In *International Conference on Learning Representations*, 2019.

Garg, V., Jegelka, S., and Jaakkola, T. Generalization and representational limits of graph neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3419–3430, 2020.

Gasteiger, J., Bojchevski, A., and Günnemann, S. Predict then propagate: Graph neural networks meet personalized pagerank. In *International Conference on Learning Representations*, 2019.

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, 2017.

Giné, E. and Peña, V. H. *Decoupling: From Dependence to Independence*. Springer, 1999.

Gori, M., Monfardini, G., and Scarselli, F. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, pp. 729–734, 2005.

Hamilton, W., Ying, Z., and Leskovec, J. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, 2017.

Han, J., Huang, W., Ma, H., Li, J., Tenenbaum, J. B., and Gan, C. Learning physical dynamics with subequivariant graph neural networks. In *Advances in Neural Information Processing Systems*, 2022.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33nd International Conference on Machine Learning*, pp. 1225–1234, 2016.

He, M., Wei, Z., Huang, Z., and Xu, H. Bernnet: Learning arbitrary graph spectral filters via bernstein approximation. In *Advances in Neural Information Processing Systems*, 2021.

He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., and Wang, M. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 639–648, 2020.

Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 22118–22133, 2020.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Johnson, J., Gupta, A., and Fei-Fei, L. Image generation from scene graphs. In *Conference on Computer Vision and Pattern Recognition*, pp. 1219–1228, 2018.

Ju, H., Li, D., Sharma, A., and Zhang, H. R. Generalization in graph neural networks: Improved pac-bayesian bounds on graph diffusion. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pp. 6314–6341, 2023.

Keriven, N., Bietti, A., and Vaiter, S. Convergence and stability of graph convolutional networks on large random graphs. In *Advances in Neural Information Processing Systems*, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

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

Kuzborskij, I. and Lampert, C. Data-dependent stability of stochastic gradient descent. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2815–2824, 2018.

Landrieu, L. and Simonovsky, M. Large-scale point cloud semantic segmentation with superpoint graphs. In *Conference on Computer Vision and Pattern Recognition*, pp. 4558–4567, 2018.

Latała, R. and Oleszkiewicz, K. On the best constant in the khinchin-kahane inequality. *Studia Mathematica*, 109(1): 101–104, 1994.

Lei, Y. and Tang, K. Learning rates for stochastic gradient descent with nonconvex objectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12): 4505–4511, 2021.

Li, H., Wang, M., Liu, S., Chen, P., and Xiong, J. Generalization guarantee of training graph convolutional networks with graph topology sampling. In *Proceedings of The 39th International Conference on Machine Learning*, pp. 13014–13051, 2022a.

Li, H., Wang, X., Zhang, Z., and Zhu, W. Out-of-distribution generalization on graphs: A survey. *arXiv preprint arXiv:2202.07987*, 2022b.

Li, Q., Han, Z., and Wu, X. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 3538–3545, 2018.

Li, S. and Liu, Y. Improved learning rates for stochastic optimization: Two theoretical viewpoints. *arXiv preprint arXiv:2107.08686*, 2021.

Liao, R., Zhao, Z., Urtasun, R., and Zemel, R. Lanczosnet: Multi-scale deep graph convolutional networks. In *International Conference on Learning Representations*, 2019.

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

Liu, C., Zhu, L., and Belkin, M. Toward a theory of optimization for over-parameterized systems of non-linear equations: the lessons of deep learning. *arXiv preprint arXiv:2003.00307*, 2020.

Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H., and Zhang, X. Parameterized explainer for graph neural network. In *Advances in Neural Information Processing Systems*, 2020.

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

Maron, H., Ben-Hamu, H., Shamir, N., and Lipman, Y. Invariant and equivariant graph networks. In *International Conference on Learning Representations*, 2019.

Oono, K. and Suzuki, T. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020a.

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

Pfaff, T., Fortunato, M., Sanchez-Gonzalez, A., and Battaglia, P. W. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021.

Pisier, G. *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge Tracts in Mathematics. Cambridge University Press, 1989.

Qi, X., Liao, R., Jia, J., Fidler, S., and Urtasun, R. 3d graph neural networks for RGBD semantic segmentation. In *International Conference on Computer Vision*, pp. 5209–5218, 2017.

Rong, Y., Huang, W., Xu, T., and Huang, J. Dropedge: Towards deep graph convolutional networks on node classification. In *International Conference on Learning Representations*, 2020.

Sanchez-Gonzalez, A., Godwin, J., Pfaff, T., Ying, R., Leskovec, J., and Battaglia, P. W. Learning to simulate complex physics with graph networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 8459–8468, 2020.

Satorras, V. G. and Estrach, J. B. Few-shot learning with graph neural networks. In *International Conference on Learning Representations*, 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, 2009.

Scarselli, F., Tsoi, A. C., and Hagenbuchner, M. The vapnik–chervonenkis dimension of graph and recursive neural networks. *Neural Networks*, 108:248–259, 2018.

Sen, P., Namata, G., Bilgic, M., Getoor, L., Gallagher, B., and Eliassi-Rad, T. Collective classification in network data. *AI magazine*, 29(3):93–106, 2008.

Shen, Z.-A., Luo, T., Zhou, Y.-K., Yu, H., and Du, P.-F. Npi-gnn: Predicting ncrna-protein interactions with deep graph neural networks. *Briefings in bioinformatics*, 2021.

Song, L., Zhang, Y., Wang, Z., and Gildea, D. A graph-to-sequence model for amr-to-text generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 1616–1626, 2018.

Vapnik, V. *Statistical learning theory*. Wiley, 1998.

Vapnik, V. *Estimation of Dependences Based on Empirical Data, Second Editiontion*. Springer, 2006.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. In *International Conference on Learning Representations*, 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.

Virmaux, A. and Scaman, K. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems*, 2018.

Vu, M. and Thai, M. T. Pgm-explainer: Probabilistic graphical model explanations for graph neural networks. In *Advances in Neural Information Processing Systems*, pp. 12225–12235, 2020.

Wang, Y., Wang, Y., Yang, J., and Lin, Z. Dissecting the diffusion process in linear graph convolutional networks. In *Advances in Neural Information Processing Systems*, 2021.

Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., and Weinberger, K. Simplifying graph convolutional networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 6861–6871, 2019.

Wu, Q., Zhang, H., Yan, J., and Wipf, D. Handling distribution shifts on graphs: An invariance perspective. In *International Conference on Learning Representations*, 2022.

Wu, Q., Chen, Y., Yang, C., and Yan, J. Energy-based out-of-distribution detection for graph neural networks. In *The Eleventh International Conference on Learning Representations*, 2023.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2021.

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

Xu, K., Li, J., Zhang, M., Du, S. S., ichi Kawarabayashi, K., and Jegelka, S. What can neural networks reason about? In *International Conference on Learning Representations*, 2020.

Xu, K., Zhang, M., Li, J., Du, S. S., Kawarabayashi, K.-I., and Jegelka, S. How neural networks extrapolate: From feedforward to graph neural networks. In *International Conference on Learning Representations*, 2021.

Xu, Y. and Zeevi, A. Towards optimal problem dependent generalization error bounds in statistical learning theory. *arXiv preprint arXiv:2011.06186*, 2020.

Yang, C., Wu, Q., Wang, J., and Yan, J. Graph neural networks are inherently good generalizers: Insights by bridging GNNs and MLPs. In *The Eleventh International Conference on Learning Representations*, 2023.

Yang, N., Zeng, K., Wu, Q., Jia, X., and Yan, J. Learning substructure invariance for out-of-distribution molecular representations. In *Advances in Neural Information Processing Systems*, 2022.

Yang, Z., Cohen, W. W., and Salakhutdinov, R. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33nd International Conference on Machine Learning*, pp. 40–48, 2016.

Yehudai, G., Fetaya, E., Meirom, E. A., Chechik, G., and Maron, H. From local structures to size generalization in graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 11975–11986, 2021.

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

Ying, Y. and Campbell, C. Rademacher chaos complexities for learning the kernel problem. *Neural Computation*, 22 (11):2858–2886, 2010.

Ying, Z., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. Gnnexplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems*, pp. 9240–9251, 2019.

Yuan, H., Tang, J., Hu, X., and Ji, S. Xgnn: Towards model-level explanations of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 430–438, 2020.

Yuan, H., Yu, H., Wang, J., Li, K., and Ji, S. On explainability of graph neural networks via subgraph explorations. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 12241–12252, 2021.

Zeng, H., Zhou, H., Srivastava, A., Kannan, R., and Prasanna, V. Graphsaint: Graph sampling based inductive learning method. In *International Conference on Learning Representations*, 2020.

Zeng, H., Zhang, M., Xia, Y., Srivastava, A., Malevich, A., Kannan, R., Prasanna, V. K., Jin, L., and Chen, R. Decoupling the depth and scope of graph neural networks. In *Advances in Neural Information Processing Systems*, pp. 19665–19679, 2021.

Zhang, S., Wang, M., Liu, S., Chen, P.-Y., and Xiong, J. Fast learning of graph neural networks with guaranteed generalizability: One-hidden-layer case. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 11268–11277, 2020.

Zhang, Z., Cui, P., and Zhu, W. Deep learning on graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):249–270, 2022.

Zhao, L. and Akoglu, L. Pairnorm: Tackling oversmoothing in gnns. In *International Conference on Learning Representations*, 2020.

Zhou, D., Tang, Y., Yang, Z., Cao, Y., and Gu, Q. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.

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

Zhu, H. and Koniusz, P. Simple spectral graph convolution. In *International Conference on Learning Representations*, 2021.

Zhu, Y., Xu, W., Zhang, J., Du, Y., Zhang, J., Liu, Q., Yang, C., and Wu, S. A survey on graph structure learning: Progress and opportunities. *arXiv preprint arXiv:2103.03036*, 2021.

A. Notations and Lemmas

In this section, we will present some notations, definitions and lemmas that will be used in subsequent analysis. Let $f : \mathbb{R}^{m\times n} \mapsto \mathbb{R}$ be a real-value function with variable $\mathbf{W} \in \mathbb{R}^{m\times n}$. We stipulate that

$$\nabla_{\text{vec}[\mathbf{W}]} f = \left[\frac{\partial f}{\partial W_{11}}, \ldots, \frac{\partial f}{\partial W_{m1}}, \ldots, \frac{\partial f}{\partial W_{1n}}, \ldots, \frac{\partial f}{\partial W_{mn}}\right]^\top \in \mathbb{R}^{mn\times 1}.$$

Denote by $\frac{\partial f}{\partial \text{vec}[\mathbf{W}]}$ the Jacobian matrix, we have $\frac{\partial f}{\partial \text{vec}[\mathbf{W}]} = \nabla^\top_{\text{vec}[\mathbf{W}]} f \in \mathbb{R}^{1\times mn}$. Denote by $\mathbf{W}_{i*}$ the i-th row of matrix $\mathbf{W}$. We use $\odot$ and $\otimes$ to denote Hadamard product and Kronecker product, respectively. The activation function $\sigma(\cdot)$ in this work is defined as

$$\sigma(x) = \begin{cases} 0, x \leq 0, \\ x^q, 0 < x \leq \left(\frac{1}{q}\right)^{\frac{1}{q-1}}, \\ x - \left(\frac{1}{q}\right)^{\frac{1}{q-1}} + \left(\frac{1}{q}\right)^{\frac{q}{q-1}}, x > \left(\frac{1}{q}\right)^{\frac{1}{q-1}}, \end{cases}$$

where $q \in (1, 2]$. It can be verify that this activation is differential on $\mathbb{R}$, and its derivation is

$$\sigma'(x) = \begin{cases} 0, x \leq 0, \\ qx^{q-1}, 0 < x \leq \left(\frac{1}{q}\right)^{\frac{1}{q-1}}, \\ 1, x > \left(\frac{1}{q}\right)^{\frac{1}{q-1}}. \end{cases}$$

When setting $p \approx 1$ (*e.g.*, $q = 1.1$), this activation function has tolerate approximation error to vanilla ReLU function. Now we show some property of $\sigma(\cdot)$ that used in the sequential proofs.

- $\|\sigma(\mathbf{u})\|_2 < \|\mathbf{u}\|_2$ *for any* $\mathbf{u} \in \mathbb{R}^d$. We only need to show that $|\sigma(u_i)| \leq |u_i|$ holds for $i \in [d]$. The case that $u_i \in (-\infty, 0]$ is trivial. If $u_i \in (0, (1/q)^{1/(q-1)}]$, since $q > 1$ and $(1/q)^{1/(q-1)} < 1$, we have $|\sigma(u_i)| = u_i^q \leq u_i = |u_i|$. If $u_i \in ((1/q)^{1/(q-1)}, \infty)$, note that $(1/q)^{q/(q-1)} < (1/q)^{1/(q-1)}$, we have $|\sigma(u_i)| \leq u_i = |u_i|$.
- $\|\sigma'(\mathbf{u}) \odot \mathbf{v}\|_2 \leq \|\mathbf{v}\|_2$ *for any* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$. By the formulation of $\sigma'(x)$, we have $|\sigma'(x)| \leq 1$. Then

$$\|\sigma'(\mathbf{u}) \odot \mathbf{v}\|_2 = \sqrt{\sum_{i=1}^d |\sigma'(u_i)|^2 |v_i|^2} \leq \sqrt{\sum_{i=1}^d |v_i|^2} = \|\mathbf{v}\|_2.$$

- $\|\sigma'(\mathbf{u}) - \sigma'(\mathbf{v})\|_2 \leq qd^{\frac{2-q}{2}} \|\mathbf{u} - \mathbf{v}\|_2^{q-1}$ *for any* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$. We first show that for any $x, y \in \mathbb{R}$, $|\sigma'(x) - \sigma'(y)| \leq q|x - y|^{q-1}$ holds. The case that $x, y \in (-\infty, 0]$ and $x, y \in [(1/q)^{1/(q-1)}, \infty)$ are trivial. If $x, y \in (0, (1/q)^{1/(q-1)}]$, we have $|qx^{q-1} - qy^{q-1}| \leq q|x - y|^{q-1}$. If $x \in (-\infty, 0]$ and $y \in (0, (1/q)^{1/(q-1)}]$, we have $|qy^{q-1}| = qy^{q-1} \leq q|x - y|^{q-1}$. If $x \in (0, (1/q)^{1/(q-1)}]$ and $y \in ((1/q)^{1/(q-1)}, \infty)$, we have $|qx^{q-1} - 1| \leq |qx^{q-1} - qy^{q-1}| \leq q|x - y|^{q-1}$. If $x \in (-\infty, 0]$ and $y \in ((1/q)^{1/(q-1)}, \infty)$, we have $|\sigma'(x) - \sigma'(y)| \leq qy^{q-1} \leq q|x - y|^{q-1}$. Thus,

$$\begin{aligned} \|\sigma'(\mathbf{u}) - \sigma'(\mathbf{v})\|_2 =& \sqrt{\sum_{i=1}^d |\sigma'(u_i) - \sigma'(v_i)|^2} \leq \sqrt{\sum_{i=1}^d q^2 |u_i - v_i|^{2(q-1)}} \\ =& \sqrt{\left(\sum_{i=1}^d |u_i - v_i|^2\right)^{q-1} \left(\sum_{i=1}^d q^{\frac{2}{2-q}}\right)^{2-q}} \leq qd^{\frac{2-q}{2}} \|\mathbf{u} - \mathbf{v}\|_2^{q-1}. \end{aligned}$$

With the above notations, we give the following lemmas.

Lemma A.1. *Denote by* $\widetilde{\mathbf{A}}$ *the normalized adjacency matrix with self-loop, we have* $\|\widetilde{\mathbf{A}}\|_\infty \leq \sqrt{\frac{\deg_{\max}+1}{\deg_{\min}+1}}$.

Proof. By definition, $\widetilde{\mathbf{A}}_{ij} \geq 0$ holds for any $i, j \in [n]$. For any fixed $i \in [n]$, let $\mathcal{N}_i$ be the index set of the i-th nodes' one-hop neighbors, we have

$$\begin{aligned}\sum_{j=1}^{n} \widetilde{\mathbf{A}}_{ij} =&\sum_{j=1}^{n} \frac{\mathbf{A}_{ij}}{\sqrt{\deg_i + 1}\sqrt{\deg_j + 1}} \\ =&\frac{1}{\sqrt{\deg_i + 1}} \left(\frac{1}{\sqrt{\deg_i + 1}} + \sum_{j \in \mathcal{N}_i} \frac{1}{\sqrt{\deg_j + 1}} \right) \\ \leq&\frac{1}{\sqrt{\deg_i + 1}} \left(\frac{1}{\sqrt{\deg_{\min} + 1}} + \sum_{j \in \mathcal{N}_i} \frac{1}{\sqrt{\deg_{\min} + 1}} \right) \\ =&\frac{1}{\sqrt{\deg_i + 1}} \frac{\deg_i + 1}{\sqrt{\deg_{\min} + 1}} = \frac{\sqrt{\deg_i + 1}}{\sqrt{\deg_{\min} + 1}} \leq \sqrt{\frac{\deg_{\max} + 1}{\deg_{\min} + 1}}.\end{aligned}$$

Lemma A.2. *Denote by $\mathbf{u} \in \mathbb{R}^m, \mathbf{v} \in \mathbb{R}^n$, we have $\|\mathbf{u} \otimes \mathbf{v}\|_2 = \|\mathbf{u}\|_2 \|\mathbf{v}\|_2$.*

Proof. One can find that

$$\|\mathbf{u} \otimes \mathbf{v}\|_2 = \sqrt{\sum_{j=1}^{m} \|u_j \mathbf{v}\|_2^2} = \sqrt{\sum_{j=1}^{m} u_j^2 \|\mathbf{v}\|_2^2} = \sqrt{\|\mathbf{u}\|_2^2 \|\mathbf{v}\|_2^2} = \|\mathbf{u}\|_2 \|\mathbf{v}\|_2.$$

Lemma A.3. *Denote by $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{m \times n}$, we have*

$$\|\mathbf{W}_1 - \mathbf{W}_2\| \leq \|\mathrm{vec}\left[\mathbf{W}_1\right] - \mathrm{vec}\left[\mathbf{W}_2\right]\|_2.$$

Proof. We have

$$\begin{aligned}&\|\mathbf{W}_1 - \mathbf{W}_2\| = \sup_{\|\mathbf{u}\|_2 = 1} \|(\mathbf{W}_1 - \mathbf{W}_2)\mathbf{u}\|_2 \triangleq \|(\mathbf{W}_1 - \mathbf{W}_2)\mathbf{u}^*\|_2 \\ =&\sqrt{\sum_{j=1}^{m} (\langle [\mathbf{W}_1]_{j*}, \mathbf{u}^* \rangle - \langle [\mathbf{W}_2]_{j*}, \mathbf{u}^* \rangle)^2} \leq \sqrt{\sum_{j=1}^{m} \|[\mathbf{W}_1]_{j*} - [\mathbf{W}_2]_{j*}\|_2^2} = \|\mathrm{vec}(\mathbf{W}_1) - \mathrm{vec}(\mathbf{W}_2)\|_2,\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality and $\|\mathbf{u}^*\|_2 = 1$. This finishes the proof.

Lemma A.4. *Denote by $\{\mathbf{W}_h\}_{h=1}^{H}$ the learnable parameters (w.l.o.g. we assume that each parameter is matrix since vector is a special case of matrix). If for $h \in [H]$,*

$$|\ell(\mathbf{W}_1, \ldots, \mathbf{W}_h, \ldots, \mathbf{W}_H) - \ell(\mathbf{W}_1, \ldots, \mathbf{W}'_h, \ldots, \mathbf{W}_H)| \leq L_h \left\|\mathrm{vec}\left[\mathbf{W}_h\right] - \mathrm{vec}\left[\mathbf{W}'_h\right]\right\|_2,$$

and

$$\left\| \frac{\partial \ell(\mathbf{W}_1, \ldots, \mathbf{W}_H)}{\partial \mathrm{vec}\left[\mathbf{W}_h\right]} - \frac{\partial \ell(\mathbf{W}'_1, \ldots, \mathbf{W}'_H)}{\partial \mathrm{vec}\left[\mathbf{W}_h\right]} \right\|_2 \leq \sum_{i=1}^{H} \left[P_{hi} \left\|\mathrm{vec}\big[\mathbf{W}_i\big] - \mathrm{vec}\big[\mathbf{W}'_i\big]\right\|_2 + \widetilde{P}_{hi} \left\|\mathrm{vec}\big[\mathbf{W}_i\big] - \mathrm{vec}\big[\mathbf{W}'_i\big]\right\|_2^{\alpha} \right],$$

then we have$|\ell(\mathbf{w}) - \ell(\mathbf{w}')| \leq L \left\|\mathbf{w} - \mathbf{w}'\right\|_2$ and $\|\nabla \ell(\mathbf{w}) - \ell(\mathbf{w})\|_2 \leq P_{\mathcal{F}} \max\{\|\mathbf{w} - \mathbf{w}\|_2, \|\mathbf{w} - \mathbf{w}\|_2^{\alpha}\}$, where $\mathbf{w} =$ $[\mathrm{vec}\left[\mathbf{W}_1\right]; \ldots; \mathrm{vec}\left[\mathbf{W}_H\right]]$ and

$$L = \left(\sum_{h=1}^{H} L_h^2 \right)^{\frac{1}{2}},\ P = \max \left\{ \left(\sum_{i=1}^{H} P_i^2 \right)^{\frac{1}{2}} + \left(\sum_{i=1}^{H} \widetilde{P}_i^{\frac{2}{2-\alpha}} \right)^{1-\frac{\alpha}{2}} \right\}.$$

Proof. The gradient of ℓ w.r.t $\mathbf{w}$ is $\nabla \ell(\mathbf{w}) = \left[\frac{\partial \ell}{\partial \text{vec}[\mathbf{W}_1]}, \ldots, \frac{\partial \ell}{\partial \text{vec}[\mathbf{W}_H]}\right]^\top$. Then we have

$$\begin{aligned}
|\ell(\mathbf{w}) - \ell(\mathbf{w}')| \leq & \sum_{h=1}^{H} |\ell(\mathbf{W}_1, \ldots, \mathbf{W}_h, \ldots, \mathbf{W}_H) - \ell(\mathbf{W}_1, \ldots, \mathbf{W}_h', \ldots, \mathbf{W}_H)| \\
\leq & \sum_{h=1}^{H} L_h \left\|\text{vec}[\mathbf{W}_h] - \text{vec}[\mathbf{W}_h']\right\|_2 \\
\leq & \left(\sum_{h=1}^{H} L_h^2\right)^{\frac{1}{2}} \left(\sum_{h=1}^{H} \left\|\text{vec}[\mathbf{W}_h] - \text{vec}[\mathbf{W}_h']\right\|_2^2\right)^{\frac{1}{2}} \\
= & L \left\|\mathbf{w} - \mathbf{w}'\right\|_2,
\end{aligned}$$

where we obtain the last inequality by Cauchy-Schwarz inequality. Similarly, we have

$$\begin{aligned}
& \left\|\nabla \ell(\mathbf{w}) - \nabla \ell(\mathbf{w}')\right\|_2 \\
\leq & \left\| \frac{\partial \ell(\mathbf{W}_1, \ldots, \mathbf{W}_H)}{\partial \text{vec}\left[\mathbf{W}_h\right]} - \frac{\partial \ell(\mathbf{W}_1', \ldots, \mathbf{W}_H')}{\partial \text{vec}\left[\mathbf{W}_h\right]} \right\|_2 \\
\leq & \sum_{h=1}^{H} \left[\sum_{i=1}^{H} P_{hi} \left\|\text{vec}[\mathbf{W}_i] - \text{vec}[\mathbf{W}_i']\right\|_2\right] + \sum_{h=1}^{H} \left[\sum_{i=1}^{H} \widetilde{P}_{hi} \left\|\text{vec}[\mathbf{W}_i] - \text{vec}[\mathbf{W}_i']\right\|_2^{\alpha}\right] \\
= & \sum_{i=1}^{H} \left[\sum_{h=1}^{H} P_{hi} \left\|\text{vec}[\mathbf{W}_i] - \text{vec}[\mathbf{W}_i']\right\|_2\right] + \sum_{i=1}^{H} \left[\sum_{h=1}^{H} \widetilde{P}_{hi} \left\|\text{vec}[\mathbf{W}_i] - \text{vec}[\mathbf{W}_i']\right\|_2^{\alpha}\right] \\
= & \sum_{i=1}^{H} P_i \left\|\text{vec}[\mathbf{W}_i] - \text{vec}[\mathbf{W}_i']\right\|_2 + \sum_{i=1}^{H} \widetilde{P}_i \left\|\text{vec}[\mathbf{W}_i] - \text{vec}[\mathbf{W}_i']\right\|_2^{\alpha} \\
\leq & \left(\sum_{i=1}^{H} P_i^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{H} \left\|\text{vec}[\mathbf{W}_i] - \text{vec}[\mathbf{W}_i']\right\|_2^2\right)^{\frac{1}{2}} + \left(\sum_{i=1}^{H} \widetilde{P}_i^{\frac{2}{2-\alpha}}\right)^{1-\frac{\alpha}{2}} \left(\sum_{i=1}^{H} \left\|\text{vec}[\mathbf{W}_i] - \text{vec}[\mathbf{W}_i']\right\|_2^2\right)^{\frac{\alpha}{2}} \\
= & \left(\sum_{i=1}^{H} P_i^2\right)^{\frac{1}{2}} \left\|\mathbf{w} - \mathbf{w}'\right\|_2 + \left(\sum_{i=1}^{H} \widetilde{P}_i^{\frac{2}{2-\alpha}}\right)^{1-\frac{\alpha}{2}} \left\|\mathbf{w} - \mathbf{w}'\right\|_2^{\alpha} \\
\leq & \max\left\{ \left(\sum_{i=1}^{H} P_i^2\right)^{\frac{1}{2}} + \left(\sum_{i=1}^{H} \widetilde{P}_i^{\frac{2}{2-\alpha}}\right)^{1-\frac{\alpha}{2}} \right\} \max\left\{\left\|\mathbf{w} - \mathbf{w}'\right\|_2, \left\|\mathbf{w} - \mathbf{w}'\right\|_2^{\alpha}\right\} \\
= & P \max\left\{\left\|\mathbf{w} - \mathbf{w}'\right\|_2, \left\|\mathbf{w} - \mathbf{w}'\right\|_2^{\alpha}\right\}.
\end{aligned} \tag{16}$$

where we have defined $P_i = \sum_{h=1}^{H} P_{hi}$ and $\widetilde{P}_i = \sum_{h=1}^{H} \widetilde{P}_{hi}$. The third inequality is due to the Hölder inequality.

Lemma A.5. *Denote by* $\mathbf{v} \in \mathbb{R}^d$. *Let* $f : \mathbb{R}^d \mapsto \mathbb{R}^d$ *be* $f(\mathbf{v})_j = \frac{e^{v_j}}{\sum_{i=1}^{d} e^{v_i}}$. *For any* $\mathbf{v}, \mathbf{v}' \in \mathbb{R}^d$, *we have* $\|f(\mathbf{v}) - f(\mathbf{v}')\|_2 \leq 2\|\mathbf{v} - \mathbf{v}'\|_2$.

Proof. By (Federer, 1969), we have $\|f(\mathbf{v}) - f(\mathbf{v}')\|_2 \leq \sup_{\mathbf{v} \in \mathbb{R}^d} \|J(\mathbf{v})\| \|\mathbf{v} - \mathbf{v}'\|_2$, where J is the Jacobian. For the aforementioned f, we have $J(\mathbf{v}) = \text{diag}\left(f(\mathbf{v})\right) - f(\mathbf{v})f(\mathbf{v})^\top$. Then

$$\|J(\mathbf{v})\| = \|\text{diag}\left(f(\mathbf{v})\right) - f(\mathbf{v})f(\mathbf{v})^\top\| \leq \|\text{diag}\left(f(\mathbf{v})\right)\| + \|f(\mathbf{v})f(\mathbf{v})^\top\|. \tag{17}$$

First,

$$\|\text{diag}\left(f(\mathbf{v})\right)\| = \sup_{\|\mathbf{w}\|_2 = 1} \|\text{diag}\left(f(\mathbf{v})\right)\mathbf{w}\|_2 = \sup_{\|\mathbf{w}\|_2 = 1} \sqrt{\sum_{i=1}^{d} f^2(\mathbf{v})_i w_i^2} = \max_{i \in [d]} f(\mathbf{v})_i \leq 1. \tag{18}$$

Besides,

$$\|f(\mathbf{v})f(\mathbf{v})^\top\| = \sup_{\|\mathbf{w}\|_2 = 1} \|f(\mathbf{v})f(\mathbf{v})^\top \mathbf{w}\|_2 = \|f(\mathbf{v})\|_2 \sup_{\|\mathbf{w}\|_2 = 1} |f(\mathbf{v})^\top \mathbf{w}| \leq \|f(\mathbf{v})\|_2^2 \leq 1, \tag{19}$$

where the last inequality is due to $\sum_{i=1}^{d} f(\mathbf{v})_i = 1$. Plugging Eq. (18) and Eq. (19) into Eq. (17), the proof is completed.

Lemma A.6. *Denote by* $\mathbf{v} \in \mathbb{R}^d$. *Let* $f : \mathbb{R}^d \mapsto \mathbb{R}$ *be* $f(\mathbf{v})_j = -\sum_{k=1}^{K} y_k \log \hat{y}_k$, *where* $\hat{y}_k = \frac{e^{v_k}}{\sum_{i=1}^{K} e^{v_i}}$. *For any* $\mathbf{v}, \mathbf{v}' \in \mathbb{R}^d$, *we have* $|f(\mathbf{v}) - f(\mathbf{v}')| \le \sqrt{2}\|\mathbf{v} - \mathbf{v}'\|_2$.

Proof. By the chain rule, the Jacobian is $J(\mathbf{v}) = \hat{\mathbf{y}}(\mathbf{v}) - \mathbf{y}$. Note that $|f(\mathbf{v}) - f(\mathbf{v}')| \le \sup_{\mathbf{v} \in \mathbb{R}^d} \|J(\mathbf{v})\|_2 \|\mathbf{v} - \mathbf{v}'\|_2$. W.o.l.g, we assume that $y_1 = 1$, then

$$\|J(\mathbf{v})\|_2 = \sqrt{\sum_{k=1}^{K} (\hat{y}_k - y_k)^2} = \sqrt{(1 - \hat{y}_1)^2 + \sum_{k=2}^{K} \hat{y}_k^2} \le \sqrt{1 + \sum_{k=1}^{K} \hat{y}_k^2} \le \sqrt{2},$$

where the last inequality is due to $\sum_{k=1}^{K} \hat{y}_k = 1$.

Lemma A.7. *Let* $\mathcal{F} : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}$ *be a parametric function class. For* $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$, *the empirical metric defined on* $\mathcal{F}$ *is defined as*

$$d_S(\mathbf{w}, \mathbf{w}') = \left(\frac{1}{n^2} \sum_{1 \le i < j \le n} |f(\mathbf{w}; z_i, z_j) - f(\mathbf{w}'; z_i, z_j)|^2 \right)^{\frac{1}{2}}.$$

For specific $\mathbf{w}^{(1)} \in \mathcal{W}$, *assume that* $\sup_{\mathbf{w} \in \mathcal{W}} d_S(\mathbf{w}, \mathbf{w}^{(1)}) \le D$ *and* $|f(\mathbf{w}^{(1)}; z_i, z_j)| \le M_0, i, j \in [n]$ *hold. Then we have*

$$\mathcal{U}(\mathcal{F}) \triangleq \frac{1}{n} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\mathbf{w} \in \mathcal{W}} \sum_{1 \le i < j \le n} \sigma_i \sigma_j f(\mathbf{w}; z_i, z_j) \right] \le M_0 + 24e \int_0^D (1 + \log \mathcal{N}(r, \mathcal{F}_{\mathcal{W}}, d_S)) \, \mathrm{d}r,$$

where $\boldsymbol{\sigma}$ *is the transductive Rademacher variable.*

Proof. The proof extend Theorem 2 in (Ying & Campbell, 2010) to the transductive Rademancher chaos complexity. The first step is to show that the following inequality holds for $1 < p \le q < \infty$, $d \ge 1$ and $\gamma = \left(\frac{p-1}{q-1}\right)^{\frac{1}{2}}$:

$$\begin{aligned} &\left[\mathbb{E} \left\| x + \gamma \sum_{i=1}^{n} x_i \sigma_i + \gamma^2 \sum_{i_1 < i_2 \le n} x_{i_1 i_2} \sigma_{i_1} \sigma_{i_2} + \cdots + \gamma^d \sum_{i_1 < \cdots < i_d \le n} x_{i \ldots i_d} \sigma_{i_1} \cdots \sigma_{i_d} \right\|_2^q \right]^{\frac{1}{q}} \\ \le &\left[\mathbb{E} \left\| x + \sum_{i=1}^{n} x_i \epsilon_i + \sum_{i_1 < i_2 \le n} x_{i_1 i_2} \epsilon_{i_1} \epsilon_{i_2} + \cdots + \sum_{i_1 < \cdots < i_d \le n} x_{i \cdots i_d} \epsilon_{i_1} \cdots \epsilon_{i_d} \right\|_2^p \right]^{\frac{1}{p}}, \end{aligned} \tag{20}$$

where σ and ϵ are transductive and standard Rademacher variable, respectively. The process generally follows that of Theorem 3.2.2 in (Giné & Peña, 1999). First, consider the case that $n = 1$, we have to show that $(\mathbb{E}|x + \gamma \sigma y|^q)^{\frac{1}{q}} \le (\mathbb{E}|x + \epsilon y|^p)^{\frac{1}{p}}$ holds. This inequality naturally holds when $y = 0$. When $x = 0$ and $y \ne 0$, we have

$$(\mathbb{E}|x + \gamma \sigma y|^q)^{\frac{1}{q}} \Big|_{x=0} = (2p_0 |\gamma y|^q)^{\frac{1}{q}} \le (|\gamma y|^q)^{\frac{1}{q}} = |\gamma y| = (\mathbb{E}|x + \gamma \epsilon y|^p)^{\frac{1}{p}} \Big|_{x=0},$$

where the inequality is due to $p_0 \le \frac{1}{2}$. When $x \ne 0$ and $y \ne 0$, let $u = \frac{y}{x}$, then

$$(\mathbb{E}|x + \gamma \sigma y|^q)^{\frac{1}{q}} \le (\mathbb{E}|x + \gamma \epsilon y|^p)^{\frac{1}{p}} \Leftrightarrow (\mathbb{E}|1 + \gamma \sigma u|^q)^{\frac{1}{q}} \le (\mathbb{E}|1 + \gamma \epsilon u|^p)^{\frac{1}{p}}.$$

By symmetric, we only have to discuss the case that $u \geq 0$. For $0 \leq u \leq 1$:

$$\begin{aligned}
(\mathbb{E}|1+\gamma\sigma u|^q)^{\frac{1}{q}} &= (p_0|1+\gamma u|^q + p_0|1-\gamma u|^q + (1-2p_0))^{\frac{1}{q}} \\
&= \left[p_0 + \sum_{k=1}^{\infty} p_0 \binom{q}{k}\gamma^k u^k + p_0 + \sum_{k=1}^{\infty} p_0 \binom{q}{k}(-1)^k\gamma^k u^k + (1-2p_0)\right]^{\frac{1}{q}} \\
&= \left[2p_0 + 2p_0 \sum_{k=1}^{\infty}\binom{q}{2k}\gamma^{2k}u^{2k} + (1-2p_0)\right]^{\frac{1}{q}} \\
&= \left[1 + 2p_0 \sum_{k=1}^{\infty}\binom{q}{2k}\gamma^{2k}u^{2k}\right]^{\frac{1}{q}} \leq \left[1 + \sum_{k=1}^{\infty}\binom{q}{2k}\gamma^{2k}u^{2k}\right]^{\frac{1}{q}} \\
&= \left[\frac{1}{2}|1+\gamma u|^q + \frac{1}{2}|1-\gamma u|^q\right]^{\frac{1}{q}} \leq (\mathbb{E}|1+\epsilon u|^p)^{\frac{1}{p}},
\end{aligned}$$

where the first inequality is due to $p_0 \leq \frac{1}{2}$, and the last inequality is from Eq. (3.2.4') in (Giné & Peña, 1999). For $u \geq 1$, we have $|1 \pm \gamma u| \leq |u \pm \gamma|$ since $u^2(1-\gamma^2) \geq 1-\gamma^2$. Then we have

$$\begin{aligned}
&(p_0|1+\gamma u|^q + p_0|1-\gamma u|^q + (1-2p_0))^{\frac{1}{q}} \\
\leq &(p_0|u|^q|1+\gamma/u| + p_0|u|^q|1-\gamma/u| + |u|^q(1-2p_0))^{\frac{1}{q}} \\
= &|u|\left(p_0|1+\gamma/u| + p_0|1-\gamma/u| + (1-2p_0)\right)^{\frac{1}{q}} \\
\leq &|u|\left[\frac{1}{2}|1+\gamma/u|^q + \frac{1}{2}|1-\gamma/u|^q\right]^{\frac{1}{q}} \leq |u|\left[\frac{1}{2}|1+1/u|^p + \frac{1}{2}|1-1/u|^p\right]^{\frac{1}{p}} = (\mathbb{E}|1+\epsilon u|^p)^{\frac{1}{p}}.
\end{aligned}$$

where the last inequality is obtained by applying Eq. (3.2.4') in (Giné & Peña, 1999) and replacing u with $1/u$. Second, consider the case where x, y are vectors. Let $z_1 = x+y$, $z_2 = x-y$, $u = \frac{\|z_1\|+\|z_2\|}{2}$ and $v = \frac{\|z_1\|-\|z_2\|}{2}$. In the second step, let $\kappa = \frac{v}{u}$, we have

$$\begin{aligned}
(\mathbb{E}\|x+\gamma\sigma y\|^q)^{\frac{1}{q}} &= [p_0\|x+\gamma y\|^q + p_0\|x-\gamma y\|^q + (1-2p_0)\|x\|^q]^{\frac{1}{q}} \\
&\leq \left[p_0\left(\frac{1+\gamma}{2}\|z_1\| + \frac{1-\gamma}{2}\|z_2\|\right)^q + p_0\left(\frac{1-\gamma}{2}\|z_1\| + \frac{1+\gamma}{2}\|z_2\|\right)^q + (1-2p_0)\|x\|^q\right]^{\frac{1}{q}} \\
&= [p_0\,|u+\gamma v|^q + p_0\,|u-\gamma v|^q + (1-2p_0)\|x\|^q]^{\frac{1}{q}} \\
&\leq [p_0\,|u+\gamma v|^q + p_0\,|u-\gamma v|^q + (1-2p_0)u^q]^{\frac{1}{q}} \\
&= |u|\,[p_0\,|1+\gamma\kappa|^q + p_0\,|1-\gamma\kappa|^q + (1-2p_0)]^{\frac{1}{q}} \\
&\leq |u|\left[\frac{|1+\kappa|^p + |1-\kappa|^p}{2}\right]^{\frac{1}{p}} = \left[\frac{|u+v|^p + |u-v|^p}{2}\right]^{\frac{1}{p}},
\end{aligned}$$

where we use the result for the the case that $n = 1$ to obtain the last inequality. The second inequality is due to

$$\|x\| = \left\|\frac{1}{2}(x+y) + \frac{1}{2}(x-y)\right\| \leq \frac{1}{2}\|z_1\| + \frac{1}{2}\|z_2\| = u.$$

Third, we use induction to obtain the final result. Following (Giné & Peña, 1999), we only show $n = 1$ implies $n = 2$.

Denote by μ the measure on the probability space, by Fubini theorem:

$$\begin{aligned}
&\left[\mathbb{E}_{\sigma_1,\sigma_2}\left\|x+\gamma x_1\sigma_1+\gamma x_2\sigma_2+\gamma^2 x_{12}\sigma_1\sigma_2\right\|_2^q\right]^{\frac{1}{q}} \\
=&\left[\int\left(\int\left\|x+\gamma x_1\sigma_1+\gamma x_2\sigma_2+\gamma^2 x_{12}\sigma_1\sigma_2\right\|_2^q \mathrm{d}\mu(\sigma_2)\right)\mathrm{d}\mu(\sigma_1)\right]^{\frac{1}{q}} \\
\leq&\left[\int\left(\int\left\|x+\gamma x_1\sigma_1+x_2\epsilon_2+\gamma x_{12}\sigma_1\epsilon_2\right\|_2^p \mathrm{d}\mu(\epsilon_2)\right)^{\frac{q}{p}}\mathrm{d}\mu(\sigma_1)\right]^{\frac{1}{q}} \\
\leq&\left[\int\left(\int\left\|x+\gamma x_1\sigma_1+x_2\epsilon_2+\gamma x_{12}\sigma_1\epsilon_2\right\|_2^q \mathrm{d}\mu(\sigma_1)\right)^{\frac{p}{q}}\mathrm{d}\mu(\epsilon_2)\right]^{\frac{1}{p}} \\
\leq&\left[\int\int\left\|x+\gamma x_1\sigma_1+x_2\epsilon_2+x_{12}\sigma_1\epsilon_2\right\|_2^p \mathrm{d}\mu(\epsilon_1)\,\mathrm{d}\mu(\epsilon_2)\right]^{\frac{1}{p}} \\
=&\left[\mathbb{E}_{\epsilon_1,\epsilon_2}\left\|x+x_1\sigma_1+x_2\sigma_2+x_{12}\sigma_1\sigma_2\right\|_2^p\right]^{\frac{1}{p}},
\end{aligned}$$

where the first and the third inequality is due to the inducition hypothesis, and the second inequality is due to the Minkowski inequality. The remaining step is similar to the proof process in (Ying & Campbell, 2010) that bound the transductive Rademacher complexity by chaining technique. For $j \in \mathbb{N}$, let $\alpha_k = 2^{-k}D$. Denote by T_k the minimal α_k-cover of $\mathcal{F}_{\mathcal{W}}$ and $f(\mathbf{w}^k; z, z')[\mathbf{w}]$ the element in T_k that covers $f(\mathbf{w}; z, z')$. Specifically, since $\{f(\mathbf{w}^{(1)}; z, z')\}$ is a D-cover of $\mathcal{W}$, we set $f(\mathbf{w}^0; z, z')[\mathbf{w}] = f(\mathbf{w}^{(1)}; z, z')$. For arbitrary $N \in \mathbb{N}$:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}\in\mathcal{W}}\frac{1}{n}\sum_{1\leq i<j\leq n}\sigma_i\sigma_j f(\mathbf{w};z_i,z_j)\right] \\
=&\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}\in\mathcal{W}}\frac{1}{n}\Bigg(\sum_{1\leq i<j\leq n}\Big(\sigma_i\sigma_j(f(\mathbf{w};z_i,z_j)-f(\mathbf{w}^N;z_i,z_j))[\mathbf{w}]+\sum_{k=1}^{N}\sigma_i\sigma_j(f(\mathbf{w}^k;z_i,z_j)[\mathbf{w}]-f(\mathbf{w}^{k-1};z_i,z_j)[\mathbf{w}])\right. \\
&\left.+\sigma_i\sigma_j f(\mathbf{w}^{(1)};z_i,z_j)\Big)\Bigg)\right] \\
\leq&\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}\in\mathcal{W}}\left(\frac{1}{n}\sum_{1\leq i<j\leq n}\sigma_i\sigma_j(f(\mathbf{w};z_i,z_j)-f(\mathbf{w}^N;z_i,z_j)[\mathbf{w}])\right)\right]+\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}\sum_{1\leq i<j\leq n}\sigma_i\sigma_j f(\mathbf{w}^{(1)};z_i,z_j)\right] \\
&+\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\mathbf{w}\in\mathcal{W}}\left(\frac{1}{n}\sum_{1\leq i<j\leq n}\sigma_i\sigma_j(f(\mathbf{w}^k;z_i,z_j)[\mathbf{w}]-f(\mathbf{w}^{k-1};z_i,z_j)[\mathbf{w}])\right)\right].
\end{aligned} \tag{21}$$

For the first term, we apply Cauchy-Schwarz inequality and obtain

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}\in\mathcal{W}}\left(\frac{1}{n}\sum_{1\leq i<j\leq n}\sigma_i\sigma_j(f(\mathbf{w};z_i,z_j)-f(\mathbf{w}^N;z_i,z_j)[\mathbf{w}])\right)\right] \\
\leq&\left(\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{1\leq i<j\leq n}\sigma_i^2\sigma_j^2\right]\right)^{\frac{1}{2}}\left(\sup_{\mathbf{w}\in\mathcal{W}}\frac{1}{n^2}\sum_{1\leq i<j\leq n}(f(\mathbf{w};z_i,z_j)-f(\mathbf{w}^N;z_i,z_j)[\mathbf{w}])^2\right)^{\frac{1}{2}}\leq n\alpha_N.
\end{aligned} \tag{22}$$

For the second term, by and Jensen's inequality and Eq. (20),

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}\sum_{1\leq i<j\leq n}\sigma_i\sigma_j f(\mathbf{w}^{(1)};z_i,z_j)\right] \\
\leq&\left(\mathbb{E}_{\boldsymbol{\epsilon}}\left[\left|\frac{1}{n}\sum_{1\leq i<j\leq n}\epsilon_i\epsilon_j f(\mathbf{w}^{(1)};z_i,z_j)\right|^2\right]\right)^{\frac{1}{2}}=\left(\frac{1}{n^2}\sum_{1\leq i<j\leq n}f^2(\mathbf{w}^{(1)};z_i,z_j)\right)^{\frac{1}{2}}\leq M_0.
\end{aligned}\tag{23}$$

Now we handle the last term. By Jensen's inequality,

$$\begin{aligned}
&\exp\left\{\lambda\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{1}{n}\sum_{1\leq i<j\leq n}\sigma_i\sigma_j(f(\mathbf{w}^k;z_i,z_j)[\mathbf{w}]-f(\mathbf{w}^{k-1};z_i,z_j)[\mathbf{w}])\right|\right]\right\}-1\\
\leq&\mathbb{E}_{\boldsymbol{\sigma}}\left[\exp\left\{\lambda\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{1}{n}\sum_{1\leq i<j\leq n}\sigma_i\sigma_j(f(\mathbf{w}^k;z_i,z_j)[\mathbf{w}]-f(\mathbf{w}^{k-1};z_i,z_j)[\mathbf{w}])\right|\right\}-1\right]\\
=&\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}\in\mathcal{W}}\left(\exp\left\{\lambda\left|\frac{1}{n}\sum_{1\leq i<j\leq n}\sigma_i\sigma_j(f(\mathbf{w}^k;z_i,z_j)[\mathbf{w}]-f(\mathbf{w}^{k-1};z_i,z_j)[\mathbf{w}])\right|\right\}-1\right)\right]\\
\leq&|T_k||T_{k-1}|\mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{f(\mathbf{w}^{k-1})\in T_{k-1},f(\mathbf{w}^k)\in T_k}\left(\exp\left\{\lambda\left|\frac{1}{n}\sum_{1\leq i<j\leq n}\sigma_i\sigma_j(f(\mathbf{w}^k;z_i,z_j)-f(\mathbf{w}^{k-1};z_i,z_j))\right|\right\}-1\right)\right].
\end{aligned}\tag{24}$$

For any $f(\mathbf{w}^{k-1};z,z')\in T_{k-1}$ and $f(\mathbf{w}^k;z,z')\in T_k$, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma}}\left[\left(\exp\left\{\lambda\left|\frac{1}{n}\sum_{1\leq i<j\leq n}\sigma_i\sigma_j(f(\mathbf{w}^k;z_i,z_j)-f(\mathbf{w}^{k-1};z_i,z_j))\right|\right\}-1\right)\right]\\
=&\sum_{s\geq 1}\frac{1}{s!}\lambda^s\mathbb{E}_{\boldsymbol{\sigma}}\left[\left|\frac{1}{n}\sum_{1\leq i<j\leq n}\sigma_i\sigma_j(f(\mathbf{w}^k;z_i,z_j)-f(\mathbf{w}^{k-1};z_i,z_j))\right|^s\right]\\
\leq&\sum_{s\geq 1}\frac{1}{s!}\lambda^s s^s\left[\mathbb{E}_{\boldsymbol{\epsilon}}\left|\frac{1}{n}\sum_{1\leq i<j\leq n}\epsilon_i\epsilon_j(f(\mathbf{w}^k;z_i,z_j)-f(\mathbf{w}^{k-1};z_i,z_j))\right|^2\right]^{\frac{s}{2}}\\
\leq&\sum_{s\geq 1}\left(e\lambda\left[\mathbb{E}_{\boldsymbol{\epsilon}}\left|\frac{1}{n}\sum_{1\leq i<j\leq n}\epsilon_i\epsilon_j(f(\mathbf{w}^k;z_i,z_j)-f(\mathbf{w}^{k-1};z_i,z_j))\right|^2\right]^{\frac{1}{2}}\right)^s,
\end{aligned}\tag{25}$$

where the first inequality is by Eq. (20), and the second inequality is due to $e^{-s}s^s\leq s!$. Let

$$\lambda=\left(2e\max_{f(\mathbf{w}^{k-1};z,z')\in T_{k-1},f(\mathbf{w}^k;z,z')\in T_k}\left[\mathbb{E}_{\boldsymbol{\epsilon}}\left|\frac{1}{n}\sum_{1\leq i<j\leq n}\epsilon_i\epsilon_j(f(\mathbf{w}^k;z_i,z_j)-f(\mathbf{w}^{k-1};z_i,z_j))\right|^2\right]^{\frac{1}{2}}\right)^{-1},$$

we obtain $\mathbb{E}_{\boldsymbol{\sigma}}\left[\left(\exp\left\{\lambda\left|\frac{1}{n}\sum_{1\leq i<j\leq n}\sigma_i\sigma_j(f(\mathbf{w}^k;z_i,z_j)-f(\mathbf{w}^{k-1};z_i,z_j))\right|\right\}-1\right)\right]\leq 1$. Plugging this into Eq. (24)

yields

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{1}{n}\sum_{1\le i<j\le n}\sigma_i\sigma_j\big(f(\mathbf{w}^k;z_i,z_j)[\mathbf{w}]-f(\mathbf{w}^{k-1};z_i,z_j)[\mathbf{w}]\big)\right|\right]\\
\le&2e\log(1+|T_k||T_{k-1}|)\max_{f(\mathbf{w}^{k-1};z,z')\in T_{k-1},f(\mathbf{w}^k;z,z')\in T_k}\left[\mathbb{E}_{\boldsymbol{\epsilon}}\left|\frac{1}{n}\sum_{1\le i<j\le n}\epsilon_i\epsilon_j\big(f(\mathbf{w}^k;z_i,z_j)-f(\mathbf{w}^{k-1};z_i,z_j)\big)\right|^2\right]^{\frac{1}{2}}\\
=&2e\log(1+|T_k||T_{k-1}|)\max_{f(\mathbf{w}^{k-1};z,z')\in T_{k-1},f(\mathbf{w}^k;z,z')\in T_k}\left[\frac{1}{n}\sum_{1\le i<j\le n}[f(\mathbf{w}^k;z_i,z_j)-f(\mathbf{w}^{k-1};z_i,z_j)]^2\right]^{\frac{1}{2}}\\
=&2e\log(1+|T_k||T_{k-1}|)\left[\frac{1}{n}\sum_{1\le i<j\le n}[f(\mathbf{w}^k;z_i,z_j)-f(\mathbf{w};z_i,z_j)+f(\mathbf{w};z_i,z_j)-f(\mathbf{w}^{k-1};z_i,z_j)]^2\right]^{\frac{1}{2}}\\
\le&2e\log(1+|T_k||T_{k-1}|)\left[\frac{1}{n}\sum_{1\le i<j\le n}[f(\mathbf{w}^k;z_i,z_j)-f(\mathbf{w};z_i,z_j)]^2\right]^{\frac{1}{2}}\\
&+2e\log(1+|T_k||T_{k-1}|)\left[\frac{1}{n}\sum_{1\le i<j\le n}[f(\mathbf{w};z_i,z_j)-f(\mathbf{w}^{k-1};z_i,z_j)]^2\right]^{\frac{1}{2}}\\
\le&2e\log(1+|T_k||T_{k-1}|)(\alpha_{k-1}+\alpha_k)=6e\log(1+|T_k||T_{k-1}|)\alpha_k.
\end{aligned}
\tag{26}
$$

Plugging Eqs. (22, 23, 26) into Eq. (21), using the facts that $\alpha_k=2(\alpha_k-\alpha_{k+1})$ and $|T_k|\ge|T_{k-1}|$, we have

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}\in\mathcal{W}}\frac{1}{n}\sum_{1\le i<j\le n}\sigma_i\sigma_j f(\mathbf{w};z_i,z_j)\right]\\
\le&M_0+n\alpha_N+6e\sum_{k=1}^{N}\log(1+|T_k||T_{k-1}|)\alpha_k=M_0+n\alpha_N+12e\sum_{k=1}^{N}\log(1+|T_k||T_{k-1}|)(\alpha_k-\alpha_{k+1})\\
\le&M_0+n\alpha_N+24e\sum_{k=1}^{N}\log(1+|T_k|)(\alpha_k-\alpha_{k+1})\\
\le&M_0+n\alpha_N+24e\int_{\alpha_{N+1}}^{\alpha_0}\log(1+\mathcal{N}(\alpha,\mathcal{F}_{\mathcal{W}},d_{\mathcal{F}}))\,\mathrm{d}r=M_0+n\alpha_N+24e\int_{\alpha_{N+1}}^{D}\log(1+\mathcal{N}(\alpha,\mathcal{F}_{\mathcal{W}},d_{\mathcal{F}}))\,\mathrm{d}r.
\end{aligned}
\tag{27}
$$

Taking the limit as $N\to\infty$ we can obtain the result.

B. Proof of Main Results

In this part, we present detailed proof of the results in the main body.

B.1. Proof of Section 4.1

B.1.1. Proof of Theorem 4.3

Proof. Following (El-Yaniv & Pechyony, 2007), let $p_0=\frac{mu}{(m+u)^2}$, the Transductive Rademacher complexity is defined as

$$
\mathcal{R}_{m+u}(\mathbf{w})=\left(\frac{1}{m}+\frac{1}{u}\right)\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}\in B_R}\sum_{i=1}^{m+u}\sigma_i\ell(\mathbf{w};z_i)\right],
$$

where σ_i is a random variable taking value in $\{\pm 1\}$ with probability p_0 and 0 with probability $1-2p_0$. By Theorem 1 in (El-Yaniv & Pechyony, 2007), with probability at least $1-\delta/2$,

$$R_u(\mathbf{w}^{(T+1)}) \leq R_m(\mathbf{w}^{(T+1)}) + \mathcal{R}_{m+u}(\mathbf{w}) + c_0 Q\sqrt{\min(m,u)} + \sqrt{\frac{SQ}{2}\log\frac{2}{\delta}}, \tag{28}$$

where $Q \triangleq (\frac{1}{m}+\frac{1}{u})$ and $S \triangleq \frac{m+u}{(m+u-\frac{1}{2})(1-1/2(\max(m,u)))}$. $c_0 \triangleq \sqrt{\frac{32\log(4e)}{3}}$ is a constant. Applying Lemma 1 in (El-Yaniv & Pechyony, 2007) with $p_2 = \frac{1}{2}$, we obtain

$$\mathcal{R}_{m+u}(\mathbf{w}) \leq \left(\frac{1}{m}+\frac{1}{u}\right)\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\mathbf{w}\in B_R}\sum_{i=1}^{m+u}\epsilon_i \ell(\mathbf{w};z_i)\right], \tag{29}$$

where ϵ_i is the standard Rademacher random variable. Now we give an upper bound of the Transductive Rademacher Complexity by Dudley's integral technique. Denote by $d_{\mathcal{H}_S}(\mathbf{w},\widetilde{\mathbf{w}}) = \left(\frac{1}{m+u}\sum_{i=1}^{m+u}\left[\ell(\mathbf{w};z_i)-\ell(\widetilde{\mathbf{w}};z_i)\right]^2\right)^{\frac{1}{2}}$. For $j\in\mathbb{N}$, let $\alpha_j = 2^{-j}M$ with $M = \sup_{\mathbf{w}\in B_R} d_{\mathcal{H}_S}(\mathbf{w},\mathbf{w}^{(1)})$. Denote by T_j the minimal α_j-cover of B_R and $\ell(\mathbf{w}^j;z)[\mathbf{w}]$ the element in T_j that covers $\ell(\mathbf{w};z)$. Specifically, since $\{\ell(\mathbf{w}^{(1)};z)\}$ is a M-cover of B_R, we set $\ell(\mathbf{w}^0;z)[\mathbf{w}] = \ell(\mathbf{w}^{(1)};z)$ (recall that $\mathbf{w}^{(1)}$ is the initialization parameter and $\mathbf{w}^j$ is the associated parameter of ℓ in T_j). For arbitrary $N\in\mathbb{N}$:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\mathbf{w}\in B_R}\sum_{i=1}^{m+u}\epsilon_i\ell(\mathbf{w};z_i)\right]\\
=&\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\mathbf{w}\in B_R}\left(\sum_{i=1}^{m+u}\left(\epsilon_i(\ell(\mathbf{w};z_i)-\ell(\mathbf{w}^N;z_i))[\mathbf{w}]+\sum_{j=1}^{N}\epsilon_i(\ell(\mathbf{w}^j;z_i)[\mathbf{w}]-\ell(\mathbf{w}^{j-1};z_i)[\mathbf{w}])+\epsilon_i\ell(\mathbf{w}^{(1)};z_i)\right)\right)\right]\\
\leq&\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\mathbf{w}\in B_R}\left(\sum_{i=1}^{m+u}\epsilon_i(\ell(\mathbf{w};z_i)-\ell(\mathbf{w}^N;z_i)[\mathbf{w}])\right)\right]+\sum_{j=1}^{N}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\mathbf{w}\in B_R}\left(\sum_{i=1}^{m+u}\epsilon_i(\ell(\mathbf{w}^j;z_i)[\mathbf{w}]-\ell(\mathbf{w}^{j-1};z_i)[\mathbf{w}])\right)\right]\\
&+\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sum_{i=1}^{m+u}\epsilon_i\ell(\mathbf{w}^{(1)};z_i)\right].
\end{aligned} \tag{30}$$

For the first term, we apply Cauchy-Schwarz inequality and obtain

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\mathbf{w}\in B_R}\left(\sum_{i=1}^{m+u}\epsilon_i(\ell(\mathbf{w};z_i)-\ell(\mathbf{w}^N;z_i)[\mathbf{w}])\right)\right]\\
\leq&\left(\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sum_{i=1}^{m+u}\epsilon_i^2\right]\right)^{\frac{1}{2}}\left(\sup_{\mathbf{w}\in B_R}\sum_{i=1}^{m+u}(\ell(\mathbf{w};z_i)-\ell(\mathbf{w}^N;z_i)[\mathbf{w}])^2\right)^{\frac{1}{2}}\leq(m+u)\alpha_N.
\end{aligned} \tag{31}$$

By Massart's Lemma, we have

$$\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\mathbf{w}\in B_R}\left(\sum_{i=1}^{m+u}\epsilon_i(\ell(\mathbf{w}^j;z_i)[\mathbf{w}]-\ell(\mathbf{w}^{j-1};z_i)[\mathbf{w}])\right)\right]\leq\sqrt{m+u}\sup_{\mathbf{w}\in B_R}d_{\mathcal{H}_S}(\mathbf{w}^j,\mathbf{w}^{j-1})\sqrt{2\log|T_j||T_{j-1}|}. \tag{32}$$

By the Minkowski inequality,

$$\begin{aligned}
&\sup_{\mathbf{w}\in B_R}d_{\mathcal{H}_S}(\mathbf{w}^j,\mathbf{w}^{j-1})\\
=&\sup_{\mathbf{w}\in B_R}\left(\frac{1}{m+u}\sum_{i=1}^{m+u}\left[\ell(\mathbf{w}^j;z_i)[\mathbf{w}]-\ell(\mathbf{w};z)+\ell(\mathbf{w};z)-\ell(\mathbf{w}^{j-1};z_i)[\mathbf{w}]\right]^2\right)^{\frac{1}{2}}\\
\leq&\sup_{\mathbf{w}\in B_R}\left(\frac{1}{m+u}\sum_{i=1}^{m+u}\left[\ell(\mathbf{w}^j;z_i)[\mathbf{w}]-\ell(\mathbf{w};z)\right]^2\right)^{\frac{1}{2}}+\sup_{\mathbf{w}\in B_R}\left(\frac{1}{m+u}\sum_{i=1}^{m+u}\left[\ell(\mathbf{w};z)-\ell(\mathbf{w}^{j-1};z_i)[\mathbf{w}]\right]^2\right)^{\frac{1}{2}}\\
=&\sup_{\mathbf{w}\in B_R}d_{\mathcal{H}_S}(\mathbf{w}^j,\mathbf{w})+\sup_{\mathbf{w}\in B_R}d_{\mathcal{H}_S}(\mathbf{w},\mathbf{w}^{j-1})\leq\alpha_j+\alpha_{j-1}=3\alpha_j.
\end{aligned} \tag{33}$$

Plugging Eq. (33) into Eq. (32), using facts that $\alpha_j = 2(\alpha_j - \alpha_{j+1})$ and $|T_j| \geq |T_{j-1}|$, taking summation over j,

$$\begin{aligned}
&\sum_{j=1}^{N} \mathbb{E}_{\boldsymbol{\epsilon}} \left[\sup_{\mathbf{w} \in B_R} \left(\sum_{i=1}^{m+u} \epsilon_i (\ell(\mathbf{w}^j; z_i)[\mathbf{w}] - \ell(\mathbf{w}^{j-1}; z_i)[\mathbf{w}]) \right) \right] \\
\leq & 6\sqrt{m+u} \sum_{j=1}^{N} \alpha_j \sqrt{\log |T_j|} = 12\sqrt{m+u} \sum_{j=1}^{N} (\alpha_j - \alpha_{j+1}) \sqrt{\log |T_j|} \\
= & 12\sqrt{m+u} \sum_{j=1}^{N} (\alpha_j - \alpha_{j+1}) \sqrt{\log \mathcal{N}(\alpha_j, \mathcal{H}_R, d_{\mathcal{H}_S})} \\
\leq & 12\sqrt{m+u} \int_{\alpha_{N+1}}^{\alpha_0} \sqrt{\log \mathcal{N}(\alpha, \mathcal{H}_R, d_{\mathcal{H}_S})}\, \mathrm{d}\alpha \leq 12\sqrt{m+u} \int_{\alpha_{N+1}}^{\infty} \sqrt{\log \mathcal{N}(\alpha, \mathcal{H}_R, d_{\mathcal{H}_S})}\, \mathrm{d}\alpha.
\end{aligned} \tag{34}$$

For the last term, by Khintchine-Kahane inequality (Latała & Oleszkiewicz, 1994),

$$\mathbb{E}_{\boldsymbol{\epsilon}} \left[\sum_{i=1}^{m+u} \epsilon_i \ell(\mathbf{w}^{(1)}; z_i) \right] \leq \left(\sum_{i=1}^{m+u} \ell^2(\mathbf{w}^{(1)}; z_i) \right)^{\frac{1}{2}} \leq b_\ell \sqrt{m+u}. \tag{35}$$

Taking the limit as $N \to \infty$, plugging Eq. (31), Eq. (34) and Eq. (35) into Eq. (30) and combining with Eq. (29) yield

$$\mathcal{R}_{m+u}(\mathbf{w}) \leq b_\ell \frac{(m+u)^{\frac{3}{2}}}{mu} + 12 \frac{(m+u)^{\frac{3}{2}}}{mu} \int_0^\infty \sqrt{\log \mathcal{N}(r, \mathcal{H}_R, d_{\mathcal{H}_S})}\, \mathrm{d}r, \tag{36}$$

where ϵ_i is the standard Rademacher random variable. Let $\mathcal{H}_R = \left\{ z \mapsto \ell(\mathbf{w}; z) \middle| \mathbf{w} \in B_R \right\}$ be the parametric function space. One can verify that $d_{H_R}(\ell(\mathbf{w}; \cdot), \ell(\widetilde{\mathbf{w}}; \cdot)) = \max_{z \in \mathcal{Z}} |\ell(\mathbf{w}; z) - \ell(\widetilde{\mathbf{w}}; z)|$ is a metric in $\mathcal{H}_R$. we have

$$d_{\mathcal{H}_S} \leq \left(\frac{1}{m+u} \sum_{i=1}^{m+u} \left[\max_{\mathbf{w}, \widetilde{\mathbf{w}} \in B_R, z \in \mathcal{Z}} \ell(\mathbf{w}; z_i) - \ell(\widetilde{\mathbf{w}}; z_i) \right]^2 \right)^{\frac{1}{2}} \leq d_{\mathcal{H}_R}.$$

By the definition of covering number, we have $\mathcal{N}(r, \mathcal{H}_R, d_{\mathcal{H}_S}) \leq \mathcal{N}(r, \mathcal{H}_R, d_{\mathcal{H}_R})$. Besides, applying Proposition 4.1 yields

$$d_{\mathcal{H}_R} = \max_{z \in \mathcal{Z}} |\ell(\mathbf{w}; z) - \ell(\widetilde{\mathbf{w}}; z)| \leq L_{\mathcal{F}} \|\mathbf{w} - \widetilde{\mathbf{w}}\|_2.$$

By the definition of covering number, we have $\mathcal{N}(r, \mathcal{H}_R, d_{\mathcal{H}_R}) \leq \mathcal{N}\left(\frac{r}{L_{\mathcal{F}}}, B_R, d_2\right)$. According to (Pisier, 1989), $\log \mathcal{N}\left(r, B_R, d_2\right) \leq d \log(3R/r)$ holds. Therefore, we obtain

$$\log \mathcal{N}(r, \mathcal{H}_R, d_{\mathcal{H}_S}) \leq d \log \left(\frac{3L_{\mathcal{F}} R}{r} \right). \tag{37}$$

Furthermore,

$$d^2_{\mathcal{H}_S}(\mathbf{w}, \mathbf{w}^{(1)}) = \frac{1}{m+u} \sum_{i=1}^{m+u} \left[\ell(\mathbf{w}; z_i) - \ell(\mathbf{w}^{(1)}; z_i) \right]^2 \leq L^2_{\mathcal{F}} R^2,$$

where the last inequality is due to Proposition 4.1. This implies that

$$\int_0^\infty \sqrt{\log \mathcal{N}(r, \mathcal{H}_R, d_{\mathcal{H}_S})} \mathrm{d}r = \int_0^{L_{\mathcal{F}} R} \sqrt{\log \mathcal{N}(r, \mathcal{H}_R, d_{\mathcal{H}_S})}\, \mathrm{d}r. \tag{38}$$

Combining Eq. (36), Eq. (37), and Eq. (38) yields

$$\begin{aligned}
\mathcal{R}_{m+u}(\mathbf{w}) \leq & 12 \frac{(m+u)^{\frac{3}{2}}}{mu} \sqrt{d} \int_0^{L_{\mathcal{F}} R} \sqrt{\log \left(3L_{\mathcal{F}} R/r\right)}\, \mathrm{d}r \\
\leq & 12 \frac{(m+u)^{\frac{3}{2}}}{mu} \sqrt{d} \left(\sqrt{\log 3} + \frac{3}{2}\sqrt{\pi} \right) L_{\mathcal{F}} R.
\end{aligned} \tag{39}$$

Applying Lemma 43 in (Li & Liu, 2021) to bound R in Eq. (39) and plugging in Eq. (28) with probability $1-\delta/2$, we conclude that with probability at least $1-\delta$,

$$R_u(\mathbf{w}^{(T+1)}) = \begin{cases} \mathcal{O}\left(L_{\mathcal{F}} \frac{(m+u)^{\frac{3}{2}}}{mu} \log^{\frac{1}{2}}(T) T^{\frac{1}{2}-\alpha} \log\left(\frac{1}{\delta}\right)\right) & \text{If } \alpha \in \left(0, \frac{1}{2}\right) \\ \mathcal{O}\left(L_{\mathcal{F}} \frac{(m+u)^{\frac{3}{2}}}{mu} \log(T) \log(\frac{1}{\delta})\right) & \text{If } \alpha = \frac{1}{2} \\ \mathcal{O}\left(L_{\mathcal{F}} \frac{(m+u)^{\frac{3}{2}}}{mu} \log^{\frac{1}{2}}(T) \log(\frac{1}{\delta})\right) & \text{If } \alpha \in \left(\frac{1}{2}, 1\right]. \end{cases}$$

B.1.2. Proof of Theorem 4.5

This proof extends the proof of Theorem 1 in literature (El-Yaniv & Pechyony, 2007) from scalar to vector. Let $p = \frac{mu}{(m+u)^2}$, we define the vector-valued Transductive Rademacher complexities:

$$\mathcal{R}_{m+u}(\mathbf{w}; p) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}\in\mathcal{W}} \left\| \left(\frac{1}{m} + \frac{1}{u}\right) \sum_{i=1}^{m+u} \sigma_i \nabla \ell(\mathbf{w}; z_i) \right\|_2 \right],$$

where σ_i is a random variable taking value in $\{\pm 1\}$ with probability p_0 and 0 with probability $1-2p_0$. Following (El-Yaniv & Pechyony, 2007), we introduce the pairwise Rademacher variables $\tilde{\boldsymbol{\sigma}} = \{(\tilde{\sigma}_{i,1}, \tilde{\sigma}_{i,2})\}_{i=1}^{m+u}$ that satisfies: $\mathbb{P}\{(\tilde{\sigma}_{i,1}, \tilde{\sigma}_{i,2}) = (\frac{1}{m}, \frac{1}{u})\} = \frac{mu}{(m+u)^2}$, $\mathbb{P}\{(\tilde{\sigma}_{i,1}, \tilde{\sigma}_{i,2}) = (-\frac{1}{u}, -\frac{1}{m})\} = \frac{mu}{(m+u)^2}$, $\mathbb{P}\{(\tilde{\sigma}_{i,1}, \tilde{\sigma}_{i,2}) = (\frac{1}{m}, -\frac{1}{m})\} = \frac{m^2}{(m+u)^2}$, $\mathbb{P}\{(\tilde{\sigma}_{i,1}, \tilde{\sigma}_{i,2}) = (-\frac{1}{u}, \frac{1}{u})\} = \frac{u^2}{(m+u)^2}$. It can be verify that

$$\mathcal{R}_{m+u}(\mathbf{w}; p) = \mathbb{E}_{\tilde{\boldsymbol{\sigma}}}\left[\sup_{\mathbf{w}\in\mathcal{W}} \left\| \sum_{i=1}^{m+u} (\tilde{\sigma}_{i,1} + \tilde{\sigma}_{i,2}) \nabla \ell(\mathbf{w}; z_i) \right\|_2 \right].$$

Let $\Gamma : [m+u] \mapsto [m+u]$ be a symmetric group and $\pi \in \Gamma$ a specific permutation on examples. Denote by $F_m(\mathbf{w}; \pi) \triangleq \frac{1}{m}\sum_{i=1}^m \nabla \ell(\mathbf{w}; z_{\pi(i)})$ and $F_u(\mathbf{w}; \pi) \triangleq \frac{1}{u}\sum_{i=m+1}^{m+u} \nabla \ell(\mathbf{w}; z_{\pi(i)})$ the population gradient calculated on training examples and test examples determined by π. We have

$$\begin{aligned}
&\|F_m(\mathbf{w}; \pi) - F_u(\mathbf{w}; \pi)\|_2 \\
\triangleq &\sup_{\mathbf{w}\in\mathcal{W}} \left\| \frac{1}{m}\sum_{i=1}^m \nabla \ell(\mathbf{w}; z_{\pi(i)}) - \frac{1}{u}\sum_{i=m+1}^{m+u} \nabla \ell(\mathbf{w}; z_{\pi(i)}) \right\|_2 \\
= &\sup_{\mathbf{w}\in\mathcal{W}} \left\| \frac{1}{m}\sum_{i=1}^m \nabla \ell(\mathbf{w}; z_{\pi(i)}) - \frac{1}{m}\sum_{\pi'\in\Gamma}\sum_{i=1}^m \frac{\nabla \ell(\mathbf{w}; z_{\pi'(i)})}{(m+u)!} + \frac{1}{u}\sum_{\pi'\in\Gamma}\sum_{i=m+1}^{m+u} \frac{\nabla \ell(\mathbf{w}; z_{\pi'(i)})}{(m+u)!} - \frac{1}{u}\sum_{i=m+1}^{m+u} \nabla \ell(\mathbf{w}; z_{\pi(i)}) \right\|_2 \\
= &\sup_{\mathbf{w}\in\mathcal{W}} \left\| \sum_{\pi'\in\Gamma} \frac{1}{(m+u)!} \Big[\frac{1}{m}\sum_{i=1}^m \big(\nabla \ell(\mathbf{w}; z_{\pi(i)}) - \nabla \ell(\mathbf{w}; z_{\pi'(i)})\big) + \frac{1}{u}\sum_{i=m+1}^{m+u} \big(\nabla \ell(\mathbf{w}; z_{\pi'(i)}) - \nabla \ell(\mathbf{w}; z_{\pi(i)})\big) \Big] \right\|_2 \\
\leq &\sum_{\pi'\in\Gamma} \frac{1}{(m+u)!} \sup_{\mathbf{w}} \left\| \frac{1}{m}\sum_{i=1}^m \big(\nabla \ell(\mathbf{w}; z_{\pi(i)}) - \nabla \ell(\mathbf{w}; z_{\pi'(i)})\big) + \frac{1}{u}\sum_{i=m+1}^{m+u} \big(\nabla \ell(\mathbf{w}; z_{\pi'(i)}) - \nabla \ell(\mathbf{w}; z_{\pi(i)})\big) \right\|_2 \\
= &\mathbb{E}_{\pi'} \sup_{\mathbf{w}} \left\| \frac{1}{m}\sum_{i=1}^m \big(\nabla \ell(\mathbf{w}; z_{\pi(i)}) - \nabla \ell(\mathbf{w}; z_{\pi'(i)})\big) + \frac{1}{u}\sum_{i=m+1}^{m+u} \big(\nabla \ell(\mathbf{w}; z_{\pi'(i)}) - \nabla \ell(\mathbf{w}; z_{\pi(i)})\big) \right\|_2 \\
\triangleq &\Phi(\pi).
\end{aligned}$$

By Proposition 4.1,

$$\|\nabla \ell(\mathbf{w}; z)\|_2 \leq \left\|\nabla \ell(\mathbf{w}; z) - \nabla \ell(\mathbf{w}^{(1)}; z)\right\|_2 + \left\|\ell(\mathbf{w}^{(1)}; z)\right\|_2 \leq P_{\mathcal{F}} R + b_g. \tag{40}$$

where the second inequality is due to $\|\mathbf{w} - \mathbf{w}^{(1)}\|_2 \leq R$ and $R > 1$. By Lemma 2 in (El-Yaniv & Pechyony, 2007), with probability at least $1-\delta/2$ over the random permutation π, for $\mathbf{w} \in \mathcal{W}$,

$$\|F_m(\mathbf{w}) - F_u(\mathbf{w})\|_2 \leq \mathbb{E}_\pi[\Phi(\pi)] + (P_{\mathcal{F}} R + b_g)\sqrt{\frac{SQ}{2}\log\frac{2}{\delta}}. \tag{41}$$

Next we discuss how to give a upper bound of $\mathbb{E}_\pi[\Phi(\pi)]$. To achieve this, we have to build connection between $\mathbb{E}_\pi[\Phi(\pi)]$ and $\mathcal{R}_{m+u}(\mathbf{w};p)$. For a given permutation π, denote by $\boldsymbol{a} \in \mathbb{R}^{m+u}$ a random vectors where $a_i = \frac{1}{m}$ if $i \in [m]$ else $-\frac{1}{u}$ and $\boldsymbol{b} \in \mathbb{R}^{m+u}$ a random vectors where $b_i = -\frac{1}{m}$ if $i \in [m]$ else $\frac{1}{u}$. To build the connection between $\boldsymbol{a}, \boldsymbol{b}$ and $\tilde{\boldsymbol{\sigma}}$, a extra distribution that conditioned on $\tilde{\boldsymbol{\sigma}}$ is introduced. Denote by $n_1(\tilde{\boldsymbol{\sigma}}) \triangleq \sum_{i=1}^{m+u} \mathbb{I}\{(\tilde{\sigma}_{i,1}, \tilde{\sigma}_{i,2}) = (\frac{1}{m}, \frac{1}{u})\}$, $n_2(\tilde{\boldsymbol{\sigma}}) \triangleq \sum_{i=1}^{m+u} \mathbb{I}\{(\tilde{\sigma}_{i,1}, \tilde{\sigma}_{i,2}) = (\frac{1}{m}, -\frac{1}{m})\}$, and $n_3(\tilde{\boldsymbol{\sigma}}) = \sum_{i=1}^{m+u} \mathbb{I}\{(\tilde{\sigma}_{i,1}, \tilde{\sigma}_{i,2}) = (-\frac{1}{u}, -\frac{1}{m})\}$ the random variables conditioned on $\tilde{\boldsymbol{\sigma}}$, which indicate the number of pairs appearing in elements of $\tilde{\sigma}$. Let $N_1(\tilde{\boldsymbol{\sigma}}) = n_1(\tilde{\boldsymbol{\sigma}}) + n_2(\tilde{\boldsymbol{\sigma}})$ and $N_2(\tilde{\boldsymbol{\sigma}}) = n_2(\tilde{\boldsymbol{\sigma}}) + n_3(\tilde{\boldsymbol{\sigma}})$, we denote by $\mathfrak{R}(N_1, N_2)$ the distribution of $\tilde{\boldsymbol{\sigma}}$ conditioned on n_1, n_2, and n_3, which has fixed number of pairs and the randomness comes from permutations. Thus, $\boldsymbol{a} + \boldsymbol{b}$ and $\tilde{\boldsymbol{\sigma}} \sim \mathfrak{R}(N_1(\tilde{\boldsymbol{\sigma}}) = m, N_2(\tilde{\boldsymbol{\sigma}}) = m)$ have the same distribution. Then we have

$$
\begin{aligned}
&\mathbb{E}_\pi[\Phi(\pi)] \\
=&\mathbb{E}_{\pi,\pi'} \sup_{\mathbf{w}\in\mathcal{W}} \left\| \frac{1}{m}\sum_{i=1}^{m} \left(\nabla\ell(\mathbf{w}; z_{\pi(i)}) - \nabla\ell(\mathbf{w}; z_{\pi'(i)})\right) + \frac{1}{u}\sum_{i=m+1}^{m+u} \left(\nabla\ell(\mathbf{w}; z_{\pi'(i)}) - \nabla\ell(\mathbf{w}; z_{\pi(i)})\right)\right\|_2 \\
=&\mathbb{E}_{\pi,\pi'} \sup_{\mathbf{w}\in\mathcal{W}} \left\| \sum_{i=1}^{m+u} (a_{\pi(i)} + b_{\pi'(i)}) \nabla\ell(\mathbf{w}; z_i)\right\|_2 \\
=&\mathbb{E}_{\tilde{\boldsymbol{\sigma}}\sim\mathfrak{R}(m,m)} \sup_{\mathbf{w}\in\mathcal{W}} \left\| \sum_{i=1}^{m+u} (\tilde{\sigma}_{i,1} + \tilde{\sigma}_{i,2}) \nabla\ell(\mathbf{w}; z_i)\right\|_2 .
\end{aligned}
$$

Denote by

$$
\psi(N, N') = \mathbb{E}_{\tilde{\boldsymbol{\sigma}}\sim\mathfrak{R}(N_1,N_2)} \left[\sup_{\mathbf{w}\in\mathcal{W}} \left\| \sum_{i=1}^{m+u} (\tilde{\sigma}_{i,1} + \tilde{\sigma}_{i,2}) \nabla\ell(\mathbf{w}; z_i)\right\|_2\right],
$$

the Transductive Rademacher complexity where $\tilde{\boldsymbol{\sigma}}$ follows $\mathfrak{R}(N, N')$ for given N_1 and N_2. One can find that

$$
\mathcal{R}_{m+u}(\mathbf{w};p) = \mathbb{E}_{N_1(\tilde{\boldsymbol{\sigma}}),N_2(\tilde{\boldsymbol{\sigma}})}[\psi(N_1(\tilde{\boldsymbol{\sigma}}), N_2(\tilde{\boldsymbol{\sigma}}))]. \tag{42}
$$

Besides, one can find that $\mathbb{E}_{\tilde{\boldsymbol{\sigma}}}[N_1(\tilde{\boldsymbol{\sigma}})] = m$ and $\mathbb{E}_{\tilde{\boldsymbol{\sigma}}}[N_2(\tilde{\boldsymbol{\sigma}})] = m$ hold. Therefore, we have

$$
\mathbb{E}_\pi[\Phi(\pi)] = \psi(\mathbb{E}_{\tilde{\boldsymbol{\sigma}}}[N_1(\tilde{\boldsymbol{\sigma}})], \mathbb{E}_{\tilde{\boldsymbol{\sigma}}}[N_2(\tilde{\boldsymbol{\sigma}})]). \tag{43}
$$

The last step is to give a upper bound of $\psi(\mathbb{E}_{\tilde{\boldsymbol{\sigma}}}[N_1(\tilde{\boldsymbol{\sigma}})], \mathbb{E}_{\tilde{\boldsymbol{\sigma}}}[N_2(\tilde{\boldsymbol{\sigma}})]) - \mathbb{E}_{N_1(\tilde{\boldsymbol{\sigma}}),N_2(\tilde{\boldsymbol{\sigma}})}[\psi(N_1(\tilde{\boldsymbol{\sigma}}), N_2(\tilde{\boldsymbol{\sigma}}))]$. Recall the definitions of $\psi(N_1, N_2)$ and $\psi(N_1', N_2)$ are:

$$
\begin{aligned}
&\psi(N_1, N_2) \\
=&\mathbb{E}_{\pi,\pi'} \sup_{\mathbf{w}\in\mathcal{W}} \left\| \frac{1}{m}\sum_{i=1}^{N_1} \nabla\ell(\mathbf{w}; z_{\pi(i)}) - \frac{1}{m}\sum_{i=1}^{N_2} \nabla\ell(\mathbf{w}; z_{\pi'(i)}) + \frac{1}{u}\sum_{i=N_2+1}^{m+u} \nabla\ell(\mathbf{w}; z_{\pi'(i)}) - \frac{1}{u}\sum_{i=N_1+1}^{m+u} \nabla\ell(\mathbf{w}; z_{\pi(i)})\right\|_2 , \\
&\psi(N_1', N_2) \\
=&\mathbb{E}_{\pi,\pi'} \sup_{\mathbf{w}\in\mathcal{W}} \left\| \frac{1}{m}\sum_{i=1}^{N_1'} \nabla\ell(\mathbf{w}; z_{\pi(i)}) - \frac{1}{m}\sum_{i=1}^{N_2} \nabla\ell(\mathbf{w}; z_{\pi'(i)}) + \frac{1}{u}\sum_{i=N_2+1}^{m+u} \nabla\ell(\mathbf{w}; z_{\pi'(i)}) - \frac{1}{u}\sum_{i=N_1'+1}^{m+u} \nabla\ell(\mathbf{w}; z_{\pi(i)})\right\|_2 .
\end{aligned}
$$

W.o.l.g., assume that $N_1' \leq N_1$. Then we have

$$
\begin{aligned}
&|\psi(N_1, N_2) - \psi(N_1', N_2)| \\
\leq&\mathbb{E}_{\pi} \sup_{\mathbf{w}\in\mathcal{W}} \left\| \left(\frac{1}{u} + \frac{1}{m}\right) \sum_{i=N_1'+1}^{N_1} \nabla\ell(\mathbf{w}; z_{\pi(i)})\right\|_2 \leq |N_1 - N_1'|(PR + b_g)\left(\frac{1}{m} + \frac{1}{u}\right).
\end{aligned} \tag{44}
$$

Similarly, we have

$$
\begin{aligned}
&|\psi(N_1, N_2) - \psi(N_1, N_2')| \\
\leq&\mathbb{E}_{\pi} \sup_{\mathbf{w}\in\mathcal{W}} \left\| \left(\frac{1}{u} + \frac{1}{m}\right) \sum_{i=N_2'+1}^{N_2} \nabla\ell(\mathbf{w}; z_{\pi(i)})\right\|_2 \leq |N_2 - N_2'|(PR + b_g)\left(\frac{1}{m} + \frac{1}{u}\right).
\end{aligned} \tag{45}
$$

Combining Eq. (44), Eq. (45) and the inequality from (Devroye et al., 1996), we have

$$\begin{aligned}
&\mathbb{P}_{N_1(\tilde{\boldsymbol{\sigma}}),N_2(\tilde{\boldsymbol{\sigma}})}\{|\psi(N_1(\tilde{\boldsymbol{\sigma}}),N_2(\tilde{\boldsymbol{\sigma}}))-\psi(\mathbb{E}_{\tilde{\boldsymbol{\sigma}}}[N_1(\tilde{\boldsymbol{\sigma}})],\mathbb{E}_{\tilde{\boldsymbol{\sigma}}}[N_2(\tilde{\boldsymbol{\sigma}})])|\geq\epsilon\}\\
\leq&\mathbb{P}_{N_1(\tilde{\boldsymbol{\sigma}}),N_2(\tilde{\boldsymbol{\sigma}})}\{|\psi(N_1(\tilde{\boldsymbol{\sigma}}),N_2(\tilde{\boldsymbol{\sigma}}))-\psi(N_1,\mathbb{E}_{\tilde{\boldsymbol{\sigma}}}[N_2(\tilde{\boldsymbol{\sigma}})])|\geq\epsilon/2\}\\
&+\mathbb{P}_{\tilde{\boldsymbol{\sigma}}}\{|\psi(N_1,\mathbb{E}_{\tilde{\boldsymbol{\sigma}}}[N_2(\tilde{\boldsymbol{\sigma}})])-\psi(\mathbb{E}_{\tilde{\boldsymbol{\sigma}}}[N_1(\tilde{\boldsymbol{\sigma}})],\mathbb{E}_{\tilde{\boldsymbol{\sigma}}}[N_2(\tilde{\boldsymbol{\sigma}})])|\geq\epsilon/2\}\\
\leq&\mathbb{P}_{N_1(\tilde{\boldsymbol{\sigma}}),N_2(\tilde{\boldsymbol{\sigma}})}\{|N_2(\tilde{\boldsymbol{\sigma}})-\mathbb{E}_{\tilde{\boldsymbol{\sigma}}}[N_2(\tilde{\boldsymbol{\sigma}})]|(PR+b_g)Q\geq\epsilon/2\}+\mathbb{P}_{\tilde{\boldsymbol{\sigma}}}\{|N_1(\tilde{\boldsymbol{\sigma}})-\mathbb{E}_{\tilde{\boldsymbol{\sigma}}}|[N_1(\tilde{\boldsymbol{\sigma}})](PR+b_g)Q\geq\epsilon/2\}\\
\leq&4\exp\{-3\epsilon^2/(32m(PR^\alpha+b_g)^2Q)\}.
\end{aligned}$$

Applying the fact from Problem 12.1 in (Devroye et al., 1996), the following inequality holds

$$\mathbb{E}_{N_1(\tilde{\boldsymbol{\sigma}}),N_2(\tilde{\boldsymbol{\sigma}})}|\psi(N_1(\tilde{\boldsymbol{\sigma}}),N_2(\tilde{\boldsymbol{\sigma}}))-\psi(\mathbb{E}_{\tilde{\boldsymbol{\sigma}}}[N_1(\tilde{\boldsymbol{\sigma}})],\mathbb{E}_{\tilde{\boldsymbol{\sigma}}}[N_2(\tilde{\boldsymbol{\sigma}})])|\leq c_0(PR+b_g)Q\sqrt{\min(m,u)},\tag{46}$$

where $c_0=\sqrt{\frac{32\log(4e)}{3}}$. Plugging Eqs. (42), (43) and (46) into Eq. (41), with probability at least $1-\delta/2$,

$$\begin{aligned}
&\|F_m(\mathbf{w})-F_u(\mathbf{w})\|_2\\
\leq&\mathbb{E}_\pi[\Phi(\pi)]+(P_{\mathcal{F}}R+b_g)\sqrt{\frac{SQ}{2}\log\frac{2}{\delta}}\\
=&\psi(\mathbb{E}_{\tilde{\boldsymbol{\sigma}}}[N_1(\tilde{\boldsymbol{\sigma}})],\mathbb{E}_{\tilde{\boldsymbol{\sigma}}}[N_2(\tilde{\boldsymbol{\sigma}})])+(P_{\mathcal{F}}R+b_g)\sqrt{\frac{SQ}{2}\log\frac{2}{\delta}}\\
\leq&\mathbb{E}_{N_1(\tilde{\boldsymbol{\sigma}}),N_2(\tilde{\boldsymbol{\sigma}})}[\psi(N_1(\tilde{\boldsymbol{\sigma}}),N_2(\tilde{\boldsymbol{\sigma}}))]+c_0(P_{\mathcal{F}}R+b_g)Q\sqrt{\min(m,u)}+(P_{\mathcal{F}}R+b_g)\sqrt{\frac{SQ}{2}\log\frac{2}{\delta}}\\
=&\mathcal{R}_{m+u}(\mathbf{w};p)+c_0(P_{\mathcal{F}}R+b_g)Q\sqrt{\min(m,u)}+(P_{\mathcal{F}}R+b_g)\sqrt{\frac{SQ}{2}\log\frac{2}{\delta}}.
\end{aligned}$$

Till now, we have obtained the following inequality holds with probability at least $1-\delta/2$:

$$\begin{aligned}
&\sup_{\mathbf{w}\in B_R}\left\|\frac{1}{m}\sum_{i=1}^m\nabla\ell(\mathbf{w};z_i)-\frac{1}{u}\sum_{i=m+1}^{m+u}\nabla\ell(\mathbf{w};z_i)\right\|_2\\
\leq&\Big(\frac{1}{m}+\frac{1}{u}\Big)\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}\in B_R}\Big\|\sum_{i=1}^{m+u}\sigma_i\mathrm{vec}(\nabla\ell(\mathbf{w};z_i))\Big\|_2\right]+c_0(PR+b_g)Q\sqrt{\min(m,u)}+(PR+b_g)\sqrt{\frac{S}{2}(\frac{1}{m}+\frac{1}{u})\log\frac{1}{\delta}}.
\end{aligned}\tag{47}$$

Let $\mathcal{H}_R=\big\{(z,z')\mapsto\langle\nabla\ell(\mathbf{w};z),\nabla\ell(\mathbf{w};z')\rangle\big|\mathbf{w}\in B_R\big\}$ be the parametric function space, one can verify that

$$d_{\mathcal{H}_R}(\mathbf{w},\widetilde{\mathbf{w}})=\max_{z,z'\in\mathcal{Z}}|\langle\nabla\ell(\mathbf{w};z),\nabla\ell(\mathbf{w};z')\rangle-\langle\nabla\ell(\widetilde{\mathbf{w}};z),\nabla\ell(\widetilde{\mathbf{w}};z')\rangle|$$

is a metric in $\mathcal{H}_R$. Define

$$d_{\mathcal{H}_S}(\mathbf{w},\widetilde{\mathbf{w}})=\left(\frac{1}{(m+u)^2}\sum_{1\leq i<j\leq m+u}|\langle\nabla\ell(\mathbf{w};z_i),\nabla\ell(\mathbf{w};z_j)\rangle-\langle\nabla\ell(\widetilde{\mathbf{w}};z_i),\nabla\ell(\widetilde{\mathbf{w}};z_j)\rangle|^2\right)^{\frac{1}{2}},$$

we have $d_{\mathcal{H}_S}(\mathbf{w},\widetilde{\mathbf{w}})\leq d_{\mathcal{H}_R}$. By the definition of covering number, we have $\mathcal{N}(r,\mathcal{H}_R,d_{\mathcal{H}_S})\leq\mathcal{N}(r,\mathcal{H}_R,d_{\mathcal{H}_R})$. Besides, applying Proposition 4.1 yields

$$\begin{aligned}
&(m+u)^2d_S^2(\mathbf{w},\widetilde{\mathbf{w}})\\
=&\sum_{1\leq i<j\leq m+u}\big|\langle\nabla\ell(\mathbf{w};z_i),\nabla\ell(\mathbf{w};z_j)\rangle-\langle\nabla\ell(\widetilde{\mathbf{w}};z_i),\nabla\ell(\widetilde{\mathbf{w}};z_j)\rangle\big|^2\\
\leq&\sum_{1\leq i<j\leq m+u}2\big|\langle\nabla\ell(\mathbf{w};z_i)-\nabla\ell(\widetilde{\mathbf{w}};z_i),\nabla\ell(\mathbf{w};z_j)\rangle\big|^2+2\big|\langle\nabla\ell(\widetilde{\mathbf{w}};z_i),\nabla\ell(\mathbf{w};z_j)-\nabla\ell(\widetilde{\mathbf{w}};z_j)\rangle\big|^2\\
\leq&\sum_{1\leq i<j\leq m+u}2\left\|\nabla\ell(\mathbf{w};z_i)-\nabla\ell(\widetilde{\mathbf{w}};z_i)\right\|_2^2\left\|\nabla\ell(\mathbf{w};z_j)\right\|_2^2+2\left\|\nabla\ell(\widetilde{\mathbf{w}};z_i)\right\|_2^2\left\|\nabla\ell(\mathbf{w};z_j)-\nabla\ell(\widetilde{\mathbf{w}};z_j))\right\|_2^2\\
\leq&2(m+u)(m+u-1)P_{\mathcal{F}}^2(P_{\mathcal{F}}R+b_g)^2\max\left\{\|\mathbf{w}-\widetilde{\mathbf{w}}\|_2^{2\alpha},\|\mathbf{w}-\widetilde{\mathbf{w}}\|_2^2\right\}.
\end{aligned}$$

By the definition of covering number, we have

$$\mathcal{N}(r, \mathcal{H}_R, d_{\mathcal{H}_R}) \leq \mathcal{N}\left(\min\left\{\left(\frac{r}{\sqrt{2}P_{\mathcal{F}}(P_{\mathcal{F}}R+b_g)}\right)^{\frac{1}{\alpha}}, \frac{r}{\sqrt{2}P_{\mathcal{F}}(P_{\mathcal{F}}R+b_g)}\right\}, B_R, d_2\right).$$

According to (Pisier, 1989), $\log \mathcal{N}\left(r, B_R, d_2\right) \leq d\log(3R/r)$ holds. Therefore, we obtain

$$\log \mathcal{N}(r, \mathcal{H}_R, d_{\mathcal{H}_S}) \leq \max\left\{d\log\left(\frac{3R(\sqrt{2}P_{\mathcal{F}})^{\frac{1}{\alpha}}(P_{\mathcal{F}}R+b_g)^{\frac{1}{\alpha}}}{r^{\frac{1}{\alpha}}}\right), d\log\left(\frac{3\sqrt{2}P_{\mathcal{F}}R(P_{\mathcal{F}}R+b_g)}{r}\right)\right\}. \tag{48}$$

Denote by $\frac{1}{m+u}\mathbb{E}_{\boldsymbol{\epsilon}}\sum_{1\leq i<j\leq m+u}\sigma_i\sigma_j h(z_i,z_j)$ the transductive Rademacher chaos complexity, we have

$$\begin{aligned}
&\left(\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}\in B_R}\left\|\sum_{i=1}^{m+u}\sigma_i\nabla\ell(\mathbf{w};z_i)\right\|_2\right]\right)^2\\
\leq&\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}\in B_R}\left\|\sum_{i=1}^{m+u}\sigma_i\nabla\ell(\mathbf{w};z_i)\right\|_2^2\right]\\
=&\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}\in B_R}\sum_{i,j=1}^{m+u}\sigma_i\sigma_j\left\langle\nabla\ell(\mathbf{w};z_i),\nabla\ell(\mathbf{w};z_j)\right\rangle\right]\\
=&\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}\in B_R}\sum_{i=1}^{m+u}\sigma_i^2\left\|\nabla\ell(\mathbf{w};z_i)\right\|_2^2\right]+\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}\in B_R}\sum_{i,j=1,i\neq j}^{m+u}\sigma_i\sigma_j\langle\nabla\ell(\mathbf{w};z_i),\nabla\ell(\mathbf{w};z_j)\rangle\right]\\
\leq&(m+u)(P_{\mathcal{F}}R+b_g)^2+2(m+u)\mathcal{U}(\mathcal{H}_R),
\end{aligned} \tag{49}$$

Note that

$$\begin{aligned}
&(m+u)^2 d_S^2(\mathbf{w},\mathbf{w}^{(1)})\\
=&\sum_{1\leq i<j\leq m+u}\left|\langle\nabla\ell(\mathbf{w};z_i),\nabla\ell(\mathbf{w};z_j)\rangle-\langle\nabla\ell(\mathbf{w}^{(1)};z_i),\nabla\ell(\mathbf{w}^{(1)};z_j)\rangle\right|\\
\leq&\sum_{1\leq i<j\leq m+u}2\left\|\nabla\ell(\mathbf{w};z_i)-\nabla\ell(\mathbf{w}^{(1)};z_i)\right\|_2^2\left\|\nabla\ell(\mathbf{w};z_j)\right\|_2^2+2\left\|\nabla\ell(\mathbf{w}^{(1)};z_i)\right\|_2^2\left\|\nabla\ell(\mathbf{w};z_j)-\nabla\ell(\mathbf{w}^{(1)};z_j)\rangle\right\|_2^2\\
\leq&2(m+u)(m+u-1)P_{\mathcal{F}}^2R^2(P_{\mathcal{F}}R+b_g)^2.
\end{aligned}$$

By Lemma A.7, we have $\mathcal{U}(\mathcal{H}_R)\leq b_g^2+24e\int_0^{\sqrt{2}P_{\mathcal{F}}R(P_{\mathcal{F}}R+b_g)}\log(1+\mathcal{N}(r,\mathcal{H}_R,d_{\mathcal{H}_S}))\,\mathrm{d}r$. Plugging in Eq. (48) yields

$$\begin{aligned}
\mathcal{U}(\mathcal{H}_R)\leq&b_g^2+24e\int_0^{2P_{\mathcal{F}}R(P_{\mathcal{F}}R+b_g)}\log\left(1+\mathcal{N}(r,\mathcal{H}_R,d_{\mathcal{H}_S})\right)\,\mathrm{d}r\\
\leq&b_g^2+24e\int_0^{\sqrt{2}P_{\mathcal{F}}R(P_{\mathcal{F}}R+b_g)}\left(\log 2+\log\mathcal{N}(r,\mathcal{H}_R,d_{\mathcal{H}_S})\right)\,\mathrm{d}r\\
=&b_g^2+24\sqrt{2}eP_{\mathcal{F}}R(P_{\mathcal{F}}R+b_g)\log 2+24ed\int_0^{\sqrt{2}P_{\mathcal{F}}(P_{\mathcal{F}}R+b_g)}\log\left(\frac{3R(\sqrt{2}P_{\mathcal{F}})^{\frac{1}{\alpha}}(P_{\mathcal{F}}R+b_g)^{\frac{1}{\alpha}}}{r^{\frac{1}{\alpha}}}\right)\,\mathrm{d}r\\
&+24ed\int_{\sqrt{2}P_{\mathcal{F}}(P_{\mathcal{F}}R+b_g)}^{\sqrt{2}P_{\mathcal{F}}R(P_{\mathcal{F}}R+b_g)}\log\left(\frac{3\sqrt{2}P_{\mathcal{F}}R(P_{\mathcal{F}}R+b_g)}{r}\right)\,\mathrm{d}r\\
\leq&b_g^2+24\sqrt{2}eP_{\mathcal{F}}(P_{\mathcal{F}}R+b_g)\left[d\log(3e^{\frac{1}{\alpha}}R)+dR\log(3e)+R\log 2\right].
\end{aligned}$$

Applying Lemma 43 in (Li & Liu, 2021) to bound R in Eq. (39) with probability $1-\delta/2$ and combining it with Eqs. (49), (47), with probability at least $1-\delta$,

$$\begin{aligned}&\sup_{\mathbf{w}\in B_R}\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla\ell(\mathbf{w};z_i)-\frac{1}{u}\sum_{i=m+1}^{m+u}\nabla\ell(\mathbf{w};z_i)\right\|_2\\ =&\begin{cases}\mathcal{O}\Big(\frac{(m+u)^{\frac{3}{2}}}{mu}\log^{\frac{1}{2}}(T)T^{\frac{1-2\alpha}{2}}\log(\frac{1}{\delta})\Big) & \text{if } \alpha\in(0,\frac{1}{2})\\ \mathcal{O}\Big(\frac{(m+u)^{\frac{3}{2}}}{mu}\log(T)\log(\frac{1}{\delta})\Big) & \text{if } \alpha=\frac{1}{2}\\ \mathcal{O}\Big(\frac{(m+u)^{\frac{3}{2}}}{mu}\log^{\frac{1}{2}}(T)\log(\frac{1}{\delta})\Big) & \text{if } \alpha\in(\frac{1}{2},1].\end{cases}\end{aligned}$$

B.1.3. Proof of Theorem 4.9

By Lemma 43 in (Li & Liu, 2021), we have

$$R_m(\mathbf{w}^{(T+1)})-R_m(\hat{\mathbf{w}}^*)=\begin{cases}\mathcal{O}\Big(\frac{1}{T^\alpha}\Big) & \text{if } \alpha\in(0,1)\\ \mathcal{O}\Big(\frac{\log(T)\log^3(1/\delta)}{T}\Big) & \text{if } \alpha=1.\end{cases}\tag{50}$$

By Theorem 4.3,

$$R_u(\mathbf{w}^{(T+1)})-R_m(\mathbf{w}^{(T+1)})=\begin{cases}\mathcal{O}\left(L_{\mathcal{F}}\frac{(m+u)^{\frac{3}{2}}}{mu}\log^{\frac{1}{2}}(T)T^{\frac{1}{2}-\alpha}\log\left(\frac{1}{\delta}\right)\right) & \text{if } \alpha\in\left(0,\frac{1}{2}\right)\\ \mathcal{O}\Big(L_{\mathcal{F}}\frac{(m+u)^{\frac{3}{2}}}{mu}\log(T)\log(\frac{1}{\delta})\Big) & \text{if } \alpha=\frac{1}{2}\\ \mathcal{O}\Big(L_{\mathcal{F}}\frac{(m+u)^{\frac{3}{2}}}{mu}\log^{\frac{1}{2}}(T)\log(\frac{1}{\delta})\Big) & \text{if } \alpha\in\left(\frac{1}{2},1\right].\end{cases}\tag{51}$$

Combing Eq. (50) and Eq. (51) yields the result.

B.2. Proof of Section 4.2

B.2.1. Proof of Proposition 4.11

We first analyze the Lipschitz continuity. Denote by $\mathbf{Z}^{(1)}=g(\widetilde{\mathbf{A}})\mathbf{X}$, $\mathbf{H}^{(1)}=\sigma(\mathbf{Z}^{(1)}\mathbf{W}_1)$ and $\mathbf{Z}^{(2)}=g(\widetilde{\mathbf{A}})\mathbf{H}^{(1)}$, the forward process of GCN is given by $\hat{\mathbf{Y}}=\operatorname{Softmax}\left(\mathbf{Z}^{(2)}\mathbf{W}_2\right)$. First, we have

$$\begin{aligned}\max_{i\in[n]}\|\mathbf{H}_{i*}^{(1)}\|_2=&\left\|\sigma\left(\sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij}\mathbf{X}_{j*}\mathbf{W}_1\right)\right\|_2\le\left\|\sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij}\mathbf{X}_{j*}\mathbf{W}_1\right\|_2\\ \le&\sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij}\|\mathbf{X}_{j*}\mathbf{W}_1\|_2\le c_Xc_W\|g(\widetilde{\mathbf{A}})\|_\infty,\end{aligned}\tag{52}$$

where the first inequality is due to the definition of $\sigma(\cdot)$. Similarly,

$$\max_{i\in[n]}\|\mathbf{Z}_{i*}^{(1)}\|_2=\left\|\sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij}\mathbf{X}_{j*}\right\|_2\le\sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij}\|\mathbf{X}_{j*}\|_2\le c_X\|g(\widetilde{\mathbf{A}})\|_\infty\tag{53}$$

holds. Besides,

$$\max_{i\in[n]}\|\mathbf{Z}_{i*}^{(2)}\|_2=\left\|\sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij}\mathbf{H}_{j*}^{(1)}\right\|_2\le\sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left\|\mathbf{H}_{j*}^{(1)}\right\|_2\le c_Xc_W\|g(\widetilde{\mathbf{A}})\|_\infty^2.$$

Then we analyze how $\ell(\mathbf{W}_1, \mathbf{W}_2; z_i)$ change w.r.t. $\mathbf{W}_2$ for fixed $\mathbf{W}_1$ and $i \in [n]$:

$$\begin{aligned}
&|\ell(\mathbf{W}_1, \mathbf{W}_2, z_i) - \ell(\mathbf{W}_1, \mathbf{W}_2', z_i)| \\
\leq& \sqrt{2} \left\| \mathbf{Z}_{i*}^{(2)}(\mathbf{W}_2 - \mathbf{W}_2') \right\|_2 = \sqrt{2} \left\| \sum_{j=1}^n \left[g(\widetilde{\mathbf{A}}) \right]_{ij} \mathbf{H}_{j*}^{(1)}(\mathbf{W}_2 - \mathbf{W}_2') \right\|_2 \\
\leq& \sqrt{2} \sum_{j=1}^n \left[g(\widetilde{\mathbf{A}}) \right]_{ij} \left\| \mathbf{H}_{j*}^{(1)} \right\|_2 \|\mathbf{W}_2 - \mathbf{W}_2\| \leq \sqrt{2} \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty \max_{i \in [n]} \|\mathbf{H}_{i*}^{(1)}\|_2 \|\mathbf{W}_2 - \mathbf{W}_2\| \\
\leq& c_X c_W \sqrt{2} \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty^2 \|\mathbf{W}_2 - \mathbf{W}_2\| \leq c_X c_W \sqrt{2} \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty^2 \|\mathrm{vec}\,[\mathbf{W}_2] - \mathrm{vec}\,[\mathbf{W}_2']\|_2 ,
\end{aligned}$$

where the first inequality is obtained by Lemma A.6, and the last inequality is obtained by Eq. (52). Then we analyze the change of $\ell(\mathbf{W}_1, \mathbf{W}_2; z_i)$ change w.r.t. $\mathbf{W}_1$ for fixed $\mathbf{W}_2$ and $i \in [n]$. Note that $\mathbf{Z}^{(1)}$ and $\mathbf{H}^{(1)}$ are function of $\mathbf{W}_1$ in this case, which we denote by $\mathbf{Z}^{(1)}(\mathbf{W}_1)$ and $\mathbf{H}^{(1)}(\mathbf{W}_1)$, respectively. Then,

$$\begin{aligned}
&|\ell(\mathbf{W}_1, \mathbf{W}_2, z_i) - \ell(\mathbf{W}_1', \mathbf{W}_2, z_i)| \\
\leq& \sqrt{2} \left\| (\mathbf{Z}_{i*}^{(2)}(\mathbf{W}_1) - \mathbf{Z}_{i*}^{(2)}(\mathbf{W}_1'))\mathbf{W}_2 \right\|_2 \\
\leq& \sqrt{2} c_W \left\| \sum_{j=1}^n \left[g(\widetilde{\mathbf{A}}) \right]_{ij} (\mathbf{H}_{j*}^{(1)}(\mathbf{W}_1) - \mathbf{H}_{j*}^{(1)}(\mathbf{W}_1')) \right\|_2 \\
\leq& \sqrt{2} c_W \sum_{j=1}^n \left[g(\widetilde{\mathbf{A}}) \right]_{ij} \left\| \mathbf{Z}_{j*}^{(1)}(\mathbf{W}_1 - \mathbf{W}_1') \right\|_2 \\
\leq& \sqrt{2} c_W \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty \max_{i \in [n]} \left\| \mathbf{Z}_{i*}^{(1)} \right\|_2 \|\mathbf{W}_1 - \mathbf{W}_1\| \leq c_X c_W \sqrt{2} \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty^2 \|\mathrm{vec}\,[\mathbf{W}_1] - \mathrm{vec}\,[\mathbf{W}_1']\|_2 .
\end{aligned}$$

Let $L_1 = L_2 = \sqrt{2} c_X c_W \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty^2$, we conclude that $|\ell(\mathbf{w}) - \ell(\mathbf{w}')| \leq L_{\mathcal{F}} \|\mathbf{w} - \mathbf{w}'\|_2$ holds with $L_{\mathcal{F}} = 2 c_X c_W \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty^2$ by Lemma A.4. By the chain rule, we have

$$\begin{aligned}
\frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2; z_i)}{\partial \mathrm{vec}\,[\mathbf{W}_2]} &= (\hat{\mathbf{y}}_i - \mathbf{y}_i) \otimes \mathbf{Z}_{i*}^{(2)}, \\
\frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2; z_i)}{\partial \mathrm{vec}\,[\mathbf{W}_1]} &= \sum_{j=1}^n \left[g(\widetilde{\mathbf{A}}) \right]_{ij} \left(\sigma' \left(\mathbf{Z}_{j*}^{(1)} \mathbf{W}_1 \right) \odot (\hat{\mathbf{y}}_i - \mathbf{y}_i) \mathbf{W}_2^\top \right) \otimes \mathbf{Z}_{j*}^{(1)}.
\end{aligned}$$

We first analyze how $\frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2; z_i)}{\partial \mathrm{vec}[\mathbf{W}_2]}$ change w.r.t. $\mathbf{W}_1$ and $\mathbf{W}_2$. Note that

$$\begin{aligned}
&\left\| \frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2; z_i)}{\partial \mathrm{vec}\,[\mathbf{W}_2]}(\mathbf{W}_1) - \frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2; z_i)}{\partial \mathrm{vec}\,[\mathbf{W}_2]}(\mathbf{W}_1') \right\|_2 \\
\leq& \|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_2 \left\| \mathbf{Z}_{i*}^{(2)}(\mathbf{W}_1) - \mathbf{Z}_{i*}^{(2)}(\mathbf{W}_1') \right\|_2 + \|\hat{\mathbf{y}}_i(\mathbf{W}_1) - \hat{\mathbf{y}}_i(\mathbf{W}_1')\|_2 \left\| \mathbf{Z}_{i*}^{(2)} \right\|_2 \\
\leq& \left(\sqrt{2} + 2 c_X c_W^2 \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty^2 \right) \left\| \mathbf{Z}_{i*}^{(2)}(\mathbf{W}_1) - \mathbf{Z}_{i*}^{(2)}(\mathbf{W}_1') \right\|_2 \\
=& \left(\sqrt{2} + 2 c_X c_W^2 \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty^2 \right) \left\| \sum_{j=1}^n \left[g(\widetilde{\mathbf{A}}) \right]_{ij} \left(\mathbf{H}_{j*}^{(1)}(\mathbf{W}_1) - \mathbf{H}_{j*}^{(1)}(\mathbf{W}_1') \right) \right\|_2 \\
\leq& \left(\sqrt{2} \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty + 2 c_X c_W^2 \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty^3 \right) \|\mathbf{W}_1 - \mathbf{W}_1'\| \max_{i \in [n]} \left\| \mathbf{Z}_{i*}^{(1)} \right\|_2 \\
\leq& \left(\sqrt{2} c_X \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty^2 + 2 c_X^2 c_W^2 \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty^4 \right) \|\mathrm{vec}\,[\mathbf{W}_1] - \mathrm{vec}\,[\mathbf{W}_1']\| .
\end{aligned}$$

Besides,

$$\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_1,\mathbf{W}_2;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_2\right]}(\mathbf{W}_2)-\frac{\partial \ell(\mathbf{W}_1,\mathbf{W}_2;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_2\right]}(\mathbf{W}_2')\right\|_2\\
\leq&\|\hat{\mathbf{y}}_i(\mathbf{W}_2)-\hat{\mathbf{y}}_i(\mathbf{W}_2)\|_2\max_{i\in[n]}\left\|\mathbf{Z}_{i*}^{(2)}\right\|_2\leq 2\left\|\mathbf{Z}^{(2)}(\mathbf{W}_2-\mathbf{W}_2')\right\|_2\max_{i\in[n]}\left\|\mathbf{Z}_{i*}^{(2)}\right\|_2\\
\leq&2\left\|\mathbf{W}_2-\mathbf{W}_2'\right\|\max_{i\in[n]}\left\|\mathbf{Z}_{i*}^{(2)}\right\|_2^2\leq 2c_X^2c_W^2\|g(\mathbf{A})\|_\infty^4\left\|\mathrm{vec}\left[\mathbf{W}_2\right]-\mathrm{vec}\left[\mathbf{W}_2'\right]\right\|.
\end{aligned}$$

Denote by $P_{21}=\sqrt{2}c_X\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2+2c_X^2c_W^2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^4$, $P_{22}=2c_X^2c_W^2\|g(\mathbf{A})\|_\infty^4$, $\widetilde{P}_{21}=\widetilde{P}_{22}=0$, we obtain that $\left\|\frac{\partial \ell(\mathbf{w};z_i)}{\partial \mathrm{vec}[\mathbf{W}_2]}-\frac{\partial \ell(\mathbf{w}';z_i)}{\partial \mathrm{vec}[\mathbf{W}_2]}\right\|_2\leq\sum_{i=1}^2P_{2i}\left\|\mathrm{vec}\left[\mathbf{W}_i\right]-\mathrm{vec}\left[\mathbf{W}_i'\right]\right\|_2+\widetilde{P}_{2i}\left\|\mathrm{vec}\left[\mathbf{W}_i\right]-\mathrm{vec}\left[\mathbf{W}_i'\right]\right\|_2^{\alpha_{2i}}$. Then we analyze how $\frac{\partial \ell(\mathbf{W}_1,\mathbf{W}_2;z_i)}{\partial \mathrm{vec}[\mathbf{W}_1]}$ change w.r.t. $\mathbf{W}_1$ and $\mathbf{W}_2$. Note that

$$\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_1,\mathbf{W}_2;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}(\mathbf{W}_1)-\frac{\partial \ell(\mathbf{W}_1,\mathbf{W}_2;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}(\mathbf{W}_1')\right\|_2\\
=&\left\|\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left(\left(\sigma'\big(\mathbf{Z}_{j*}^{(1)}\mathbf{W}_1\big)-\sigma'\big(\mathbf{Z}_{j*}^{(1)}\mathbf{W}_1'\big)\right)\odot(\hat{\mathbf{y}}_i-\mathbf{y}_i)\mathbf{W}_2^\top\right)\otimes\mathbf{Z}_{j*}^{(1)}\right\|_2\\
&+\left\|\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left(\sigma'\big(\mathbf{Z}_{j*}^{(1)}\mathbf{W}_1\big)\odot((\hat{\mathbf{y}}_i(\mathbf{W}_1)-\hat{\mathbf{y}}_i(\mathbf{W}_1')))\mathbf{W}_2^\top\right)\otimes\mathbf{Z}_{j*}^{(1)}\right\|_2\\
\leq&\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\|(\hat{\mathbf{y}}_i-\mathbf{y}_i)\mathbf{W}_2^\top\|_2\max_{j\in[n]}\left\|\mathbf{Z}_{j*}^{(1)}\right\|_2\left\|\sigma'\big(\mathbf{Z}_{j*}^{(1)}\mathbf{W}_1\big)-\sigma'\big(\mathbf{Z}_{j*}^{(1)}\mathbf{W}_1'\big)\right\|_2\\
&+c_W\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\max_{j\in[n]}\left\|\mathbf{Z}_{j*}^{(1)}\right\|_2\|\hat{\mathbf{y}}_i(\mathbf{W}_1)-\hat{\mathbf{y}}_i(\mathbf{W}_1')\|_2\\
\leq&\sqrt{2}c_Xc_WP\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2\max_{j\in[n]}\left\|\mathbf{Z}_{j*}^{(1)}\left(\mathbf{W}_1-\mathbf{W}_1'\right)\right\|_2^\alpha+2c_Xc_W^2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2\max_{j\in[n]}\left\|\mathbf{Z}_{i*}^{(2)}(\mathbf{W}_1)-\mathbf{Z}_{i*}^{(2)}(\mathbf{W}_1')\right\|_2\\
\leq&\sqrt{2}c_Xc_WP\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2\|\mathbf{W}_1-\mathbf{W}_1'\|_2^\alpha\max_{j\in[n]}\left\|\mathbf{Z}_{j*}^{(1)}\right\|_2^\alpha+2c_Xc_W^2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2\max_{j\in[n]}\left\|\mathbf{Z}_{i*}^{(2)}(\mathbf{W}_1)-\mathbf{Z}_{i*}^{(2)}(\mathbf{W}_1')\right\|_2\\
\leq&\sqrt{2}c_X^{1+\alpha}c_WP\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^{2+\alpha}\left\|\mathrm{vec}\left[\mathbf{W}_1\right]-\mathrm{vec}\left[\mathbf{W}_1'\right]\right\|_2^\alpha+2c_X^2c_W^2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^4\left\|\mathrm{vec}\left[\mathbf{W}_1\right]-\mathrm{vec}\left[\mathbf{W}_1'\right]\right\|_2,
\end{aligned}$$

where we use the fact that the absolute value of each element of $\sigma'\big(\mathbf{Z}_{j*}^{(1)}\mathbf{W}_1\big)$ is less than 1. Similarly,

$$\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_1,\mathbf{W}_2;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}(\mathbf{W}_2)-\frac{\partial \ell(\mathbf{W}_1,\mathbf{W}_2;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_2\right]}(\mathbf{W}_2')\right\|_2\\
=&\left\|\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left(\sigma'\left(\mathbf{Z}_{j*}^{(1)}\mathbf{W}_1\right)\odot(\hat{\mathbf{y}}_i-\mathbf{y}_i)\left(\mathbf{W}_2-\mathbf{W}_2'\right)^\top\right)\otimes\mathbf{Z}_{j*}^{(1)}\right\|_2\\
&+\left\|\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left(\sigma'\big(\mathbf{Z}_{j*}^{(1)}\mathbf{W}_1\big)\odot((\hat{\mathbf{y}}_i(\mathbf{W}_2)-\hat{\mathbf{y}}_i(\mathbf{W}_2')))\mathbf{W}_2^\top\right)\otimes\mathbf{Z}_{j*}^{(1)}\right\|_2\\
\leq&\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\left\|(\hat{\mathbf{y}}_i-\mathbf{y}_i)\left(\mathbf{W}_2-\mathbf{W}_2'\right)^\top\right\|_2\max_{j\in[n]}\left\|\mathbf{Z}_{j*}^{(1)}\right\|_2+c_W\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\max_{j\in[n]}\left\|\mathbf{Z}_{j*}^{(1)}\right\|_2\|\hat{\mathbf{y}}_i(\mathbf{W}_2)-\hat{\mathbf{y}}_i(\mathbf{W}_2')\|_2\\
\leq&\left(\sqrt{2}c_X\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2+2c_X^2c_W^2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^4\right)\left\|\mathrm{vec}\left[\mathbf{W}_2\right]-\mathrm{vec}\left[\mathbf{W}_2'\right]\right\|_2.
\end{aligned}$$

Denote by

$$\begin{aligned}
&P_{11}=\sqrt{2}c_X^{1+\alpha}c_WP\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^{2+\alpha},\widetilde{P}_{11}=2c_X^2c_W^2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^4,\\
&P_{12}=\sqrt{2}c_X\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2+2c_X^2c_W^2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^4,\widetilde{P}_{12}=0,
\end{aligned}$$

we obtain $\left\|\frac{\partial \ell(\mathbf{w};z_i)}{\partial \text{vec}[\mathbf{W}_1]} - \frac{\partial \ell(\mathbf{w}';z_i)}{\partial \text{vec}[\mathbf{W}_1]}\right\|_2 \leq \sum_{i=1}^{2} P_{1i} \left\|\text{vec}\left[\mathbf{W}_i\right] - \text{vec}\left[\mathbf{W}_i'\right]\right\|_2 + \widetilde{P}_{1i} \left\|\text{vec}\left[\mathbf{W}_i\right] - \text{vec}\left[\mathbf{W}_i'\right]\right\|_2^{\alpha}$. By Lemma A.4, we conclude that $\|\nabla \ell(\mathbf{w}) - \nabla \ell(\mathbf{w}')\|_2 \leq P_{\mathcal{F}} \max\{\|\mathbf{w}-\mathbf{w}'\|_2, \|\mathbf{w}-\mathbf{w}'\|_2^{\alpha}\}$ where $\mathbf{w} = [\text{vec}\left[\mathbf{W}_1\right]; \text{vec}\left[\mathbf{W}_2\right]]$.

B.2.2. Proof of Proposition 4.12

We first analyze the Lipschitz continuity. Denote by

$$
\begin{aligned}
\mathbf{H}^{(0)} &= \sigma\left(\mathbf{X}\mathbf{W}_0\right), \\
\mathbf{H}^{(1)} &= \sigma\left(((1-\alpha_1)g(\widetilde{\mathbf{A}})\mathbf{H}^{(0)} + \alpha_1 \mathbf{H}^{(0)})((1-\beta_1)\mathbf{I} + \beta_1 \mathbf{W}_1)\right), \\
\mathbf{H}^{(2)} &= \sigma\left(((1-\alpha_2)g(\widetilde{\mathbf{A}})\mathbf{H}^{(1)} + \alpha_2 \mathbf{H}^{(0)})((1-\beta_2)\mathbf{I} + \beta_2 \mathbf{W}_2)\right),
\end{aligned}
$$

the forward process of GCNII is given by $\hat{\mathbf{Y}} = \text{Softmax}\left(\mathbf{H}^{(2)}\mathbf{W}_3\right)$. First, we have

$$
\max_{i\in[n]} \left\|\mathbf{H}_{i*}^{(0)}\right\|_2 = \max_{i\in[n]} \left\|\sigma\left(\mathbf{X}_{i*}\mathbf{W}_0\right)\right\|_2 \leq c_X c_W.
$$

Similarly, for $\ell = 1$ and $\ell = 2$, denote by $C_\ell = (1-\beta_\ell) + \beta_\ell c_W$, we have

$$
\begin{aligned}
&\max_{i\in[n]} \left\|\mathbf{H}_{i*}^{(\ell)}\right\|_2 \\
=&\max_{i\in[n]} \left\|\sigma\left(\sum_{j=1}^{n}((1-\alpha_\ell)\left[g(\widetilde{\mathbf{A}})\right]_{ij} \mathbf{H}_{j*}^{(\ell-1)} + \alpha_\ell \mathbf{H}_{i*}^{(0)})((1-\beta_\ell)\mathbf{I} + \beta_\ell \mathbf{W}_\ell)\right)\right\|_2 \\
\leq&\max_{i\in[n]} \left\{(1-\alpha_\ell)\left\|\sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij} \mathbf{H}_{j*}^{(\ell-1)}((1-\beta_\ell)\mathbf{I} + \beta_\ell \mathbf{W}_\ell)\right\|_2 + \alpha_\ell \left\|\mathbf{H}_{i*}^{(0)}((1-\beta_\ell)\mathbf{I} + \beta_\ell \mathbf{W}_\ell)\right\|_2\right\} \\
\leq&\max_{i\in[n]} \left\{(1-\alpha_\ell)\sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij} \left\|\mathbf{H}_{j*}^{(\ell-1)}((1-\beta_\ell)\mathbf{I} + \beta_\ell \mathbf{W}_\ell)\right\|_2 + \alpha_\ell \left\|\mathbf{H}_{i*}^{(0)}((1-\beta_\ell)\mathbf{I} + \beta_\ell \mathbf{W}_\ell)\right\|_2\right\} \\
\leq&(1-\alpha_\ell)\left\|g(\widetilde{\mathbf{A}})\right\|_\infty \|(1-\beta_\ell)\mathbf{I} + \beta_\ell \mathbf{W}_\ell\|_2 \max_{i\in[n]}\left\|\mathbf{H}_{i*}^{(\ell-1)}\right\|_2 + \alpha_\ell \|(1-\beta_\ell)\mathbf{I} + \beta_\ell \mathbf{W}_\ell\|_2 \max_{i\in[n]}\left\|\mathbf{H}_{i*}^{(0)}\right\|_2 \\
\leq&(1-\alpha_\ell)C_\ell \left\|g(\widetilde{\mathbf{A}})\right\|_\infty \max_{i\in[n]}\left\|\mathbf{H}_{i*}^{(\ell-1)}\right\| + \alpha_\ell c_X c_W C_\ell.
\end{aligned}
$$

Let $B_1 = \max_{i\in[n]} \left\|\mathbf{H}_{i*}^{(1)}\right\|_2$ and $B_2 = \max_{i\in[n]} \left\|\mathbf{H}_{i*}^{(2)}\right\|_2$, we have obtain $B_1 = c_X c_W C_1((1-\alpha_1)\left\|g(\widetilde{\mathbf{A}})\right\|_\infty + \alpha_1)$ and $B_2 = (1-\alpha_2)C_2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty B_1 + \alpha_2 c_X c_W C_2$. Next, we analyze the change of $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$ w.r.t. $\mathbf{W}_0$, $\mathbf{W}_1$ and $\mathbf{W}_2$.

Part A. Note that

$$
\begin{aligned}
\Delta_{11} \triangleq& \left\|\mathbf{H}_{i*}^{(1)}(\mathbf{W}_1) - \mathbf{H}_{i*}^{(1)}(\mathbf{W}_1')\right\|_2 \\
\leq&\beta_1 \left\|\left((1-\alpha_1)\sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij}\mathbf{H}_{j*}^{(0)} + \alpha_1 \mathbf{H}_{i*}^{(0)}\right)(\mathbf{W}_1 - \mathbf{W}_1')\right\|_2 \\
\leq&\beta_1 \left((1-\alpha_1)\sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left\|\mathbf{H}_{j*}^{(0)}\right\|_2 + \alpha_1 \left\|\mathbf{H}_{i*}^{(0)}\right\|_2\right)\|\mathbf{W}_1 - \mathbf{W}_1'\| \\
\leq&\beta_1 \left((1-\alpha_1)\left\|g(\widetilde{\mathbf{A}})\right\|_\infty + \alpha_1\right) c_X c_W \|\mathbf{W}_1 - \mathbf{W}_1'\| \\
\leq&\beta_1 \left((1-\alpha_1)\left\|g(\widetilde{\mathbf{A}})\right\|_\infty + \alpha_1\right) c_X c_W \left\|\text{vec}\left[\mathbf{W}_1\right] - \text{vec}\left[\mathbf{W}_1'\right]\right\|_2 = \frac{\beta_1 B_1}{C_1}\left\|\text{vec}\left[\mathbf{W}_1\right] - \text{vec}\left[\mathbf{W}_1'\right]\right\|_2.
\end{aligned}
$$

Similarly, we have

$$\begin{aligned}
\Delta_{10} \triangleq & \left\| \mathbf{H}_{i*}^{(1)}(\mathbf{W}_0) - \mathbf{H}_{i*}^{(1)}(\mathbf{W}_0') \right\|_2 \\
\leq & \left\| (1-\alpha_1) \sum_{j=1}^{n} \left[g(\widetilde{\mathbf{A}}) \right]_{ij} (\mathbf{H}_{j*}^{(0)}(\mathbf{W}_0) - \mathbf{H}_{j*}^{(0)}(\mathbf{W}_0')) + \alpha_1 (\mathbf{H}_{i*}^{(0)}(\mathbf{W}_0) - \mathbf{H}_{i*}^{(0)}(\mathbf{W}_0')) \right\|_2 \|(1-\beta_1)\mathbf{I} + \beta_1 \mathbf{W}_1\| \\
\leq & C_1 \left((1-\alpha_1) \sum_{j=1}^{n} \left[g(\widetilde{\mathbf{A}}) \right]_{ij} \left\| \mathbf{H}_{j*}^{(0)}(\mathbf{W}_0) - \mathbf{H}_{j*}^{(0)}(\mathbf{W}_0') \right\|_2 + \alpha_1 \left\| \mathbf{H}_{i*}^{(0)}(\mathbf{W}_0) - \mathbf{H}_{i*}^{(0)}(\mathbf{W}_0') \right\|_2 \right) \\
\leq & C_1 \left((1-\alpha_1) \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty + \alpha_1 \right) c_X \|\mathbf{W}_0 - \mathbf{W}_0'\| = \frac{B_1}{c_W} \|\mathbf{W}_0 - \mathbf{W}_0'\|.
\end{aligned}$$

Part B. Note that

$$\begin{aligned}
\Delta_{22} \triangleq & \left\| \mathbf{H}_{i*}^{(2)}(\mathbf{W}_2) - \mathbf{H}_{i*}^{(2)}(\mathbf{W}_2') \right\|_2 \\
\leq & \beta_2 \left\| \left((1-\alpha_2) \sum_{j=1}^{n} \left[g(\widetilde{\mathbf{A}}) \right]_{ij} \mathbf{H}_{j*}^{(1)} + \alpha_2 \mathbf{H}_{i*}^{(0)} \right) (\mathbf{W}_2 - \mathbf{W}_2') \right\|_2 \\
\leq & \beta_2 \left((1-\alpha_2) \sum_{j=1}^{n} \left[g(\widetilde{\mathbf{A}}) \right]_{ij} \left\| \mathbf{H}_{j*}^{(1)} \right\|_2 + \alpha_2 \left\| \mathbf{H}_{i*}^{(0)} \right\|_2 \right) \|\mathbf{W}_2 - \mathbf{W}_2'\| \\
\leq & \beta_2 \left((1-\alpha_2) \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty B_1 + \alpha_2 c_X c_W \right) \|\mathbf{W}_2 - \mathbf{W}_2'\| \\
\leq & \beta_2 \left((1-\alpha_2) \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty B_1 + \alpha_2 c_X c_W \right) \|\mathrm{vec}\,[\mathbf{W}_2] - \mathrm{vec}\,[\mathbf{W}_2']\|_2 = \frac{\beta_2 B_2}{C_2} \|\mathrm{vec}\,[\mathbf{W}_1] - \mathrm{vec}\,[\mathbf{W}_1']\|_2 .
\end{aligned}$$

Similarly,

$$\begin{aligned}
\Delta_{21} \triangleq & \left\| \mathbf{H}_{i*}^{(2)}(\mathbf{W}_1) - \mathbf{H}_{i*}^{(2)}(\mathbf{W}_1') \right\|_2 \\
\leq & (1-\alpha_2) \sum_{j=1}^{n} \left[g(\widetilde{\mathbf{A}}) \right]_{ij} \left\| \mathbf{H}_{j*}^{(1)}(\mathbf{W}_1) - \mathbf{H}_{j*}^{(1)}(\mathbf{W}_1') \right\|_2 \|(1-\beta_2)\mathbf{I} + \beta_2 \mathbf{W}_2\|_2 \\
\leq & (1-\alpha_2) C_2 \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty \max_{i\in[n]} \left\| \mathbf{H}_{i*}^{(1)}(\mathbf{W}_1) - \mathbf{H}_{i*}^{(1)}(\mathbf{W}_1') \right\|_2 \leq (1-\alpha_2)\beta_1 \frac{B_1 C_2}{C_1} \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty \|\mathrm{vec}\,[\mathbf{W}_1] - \mathrm{vec}\,[\mathbf{W}_1']\|_2 .
\end{aligned}$$

Besides,

$$\begin{aligned}
\Delta_{20} \triangleq & \left\| \mathbf{H}_{i*}^{(2)}(\mathbf{W}_0) - \mathbf{H}_{i*}^{(2)}(\mathbf{W}_0') \right\|_2 \\
\leq & (1-\alpha_2) \sum_{j=1}^{n} \left[g(\widetilde{\mathbf{A}}) \right]_{ij} \left\| \mathbf{H}_{j*}^{(1)}(\mathbf{W}_0) - \mathbf{H}_{j*}^{(1)}(\mathbf{W}_0') \right\|_2 \|(1-\beta_2)\mathbf{I} + \beta_2 \mathbf{W}_2\|_2 \\
& + \alpha_2 \left\| \mathbf{H}_{j*}^{(0)}(\mathbf{W}_0) - \mathbf{H}_{j*}^{(0)}(\mathbf{W}_0') \right\|_2 \|(1-\beta_2)\mathbf{I} + \beta_2 \mathbf{W}_2\|_2 \\
\leq & \left((1-\alpha_2) \frac{B_1 C_2}{c_W} \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty + \alpha_2 c_X C_2 \right) \|\mathrm{vec}\,[\mathbf{W}_0] - \mathrm{vec}\,[\mathbf{W}_0']\|_2 = \frac{B_2}{c_W} \|\mathrm{vec}\,[\mathbf{W}_0] - \mathrm{vec}\,[\mathbf{W}_0']\|_2 .
\end{aligned}$$

Now we are ready to analyze the Lipschitz continuity and Hölder smoothness. Note that

$$\begin{aligned}
& |\ell(\mathbf{W}_0, \mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3; z_i) - \ell(\mathbf{W}_0, \mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3'; z_i)| \\
\leq & \sqrt{2} \left\| \mathbf{H}_{i*}^{(2)}(\mathbf{W}_3 - \mathbf{W}_3') \right\|_2 \leq \sqrt{2} \max_{i\in[n]} \left\| \mathbf{H}_{i*}^{(2)} \right\|_2 \|\mathbf{W}_3 - \mathbf{W}_3'\|_2 \\
\leq & \sqrt{2} B_2 \|\mathbf{W}_3 - \mathbf{W}_3'\|_2 \leq \sqrt{2} B_2 \|\mathrm{vec}\,[\mathbf{W}_3] - \mathrm{vec}\,[\mathbf{W}_3']\|_2 .
\end{aligned}$$

Since $\mathbf{H}^{(2)}$ is a variable related to $\mathbf{W}_2$, we have

$$\begin{aligned}
&|\ell(\mathbf{W}_0, \mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3; z_i) - \ell(\mathbf{W}_0, \mathbf{W}_1, \mathbf{W}_2', \mathbf{W}_3; z_i)| \\
\leq& \sqrt{2} \left\| (\mathbf{H}_{i*}^{(2)}(\mathbf{W}_2) - \mathbf{H}_{i*}^{(2)}(\mathbf{W}_2'))\mathbf{W}_3 \right\|_2 \\
\leq& c_W \sqrt{2} \Delta_{22} = \frac{\beta_2 B_2}{C_2} c_W \sqrt{2} \left\| \mathrm{vec}\left[\mathbf{W}_2\right] - \mathrm{vec}\left[\mathbf{W}_2'\right] \right\|_2 .
\end{aligned}$$

Similarly, since $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$ are variables related to $\mathbf{W}_1$,

$$\begin{aligned}
&|\ell(\mathbf{W}_0, \mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3; z_i) - \ell(\mathbf{W}_0, \mathbf{W}_1', \mathbf{W}_2, \mathbf{W}_3; z_i)| \\
\leq& \sqrt{2} \left\| (\mathbf{H}_{i*}^{(2)}(\mathbf{W}_1) - \mathbf{H}_{i*}^{(2)}(\mathbf{W}_1'))\mathbf{W}_3 \right\|_2 \\
\leq& c_W \sqrt{2} \Delta_{21} = (1-\alpha_2)\beta_1 c_W \frac{B_1 C_2}{C_1} \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty \sqrt{2} \left\| \mathrm{vec}\left[\mathbf{W}_1\right] - \mathrm{vec}\left[\mathbf{W}_1'\right] \right\|_2 .
\end{aligned}$$

Lastly, since $\mathbf{H}^{(0)}$, $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$ are variables related to $\mathbf{W}_0$,

$$\begin{aligned}
&|\ell(\mathbf{W}_0, \mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3; z_i) - \ell(\mathbf{W}_0', \mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3; z_i)| \\
\leq& \sqrt{2} \left\| (\mathbf{H}_{i*}^{(2)}(\mathbf{W}_0) - \mathbf{H}_{i*}^{(2)}(\mathbf{W}_0'))\mathbf{W}_3 \right\|_2 \\
\leq& c_W \sqrt{2} \Delta_{20} = \sqrt{2} B_2 \left\| \mathrm{vec}\left[\mathbf{W}_0\right] - \mathrm{vec}\left[\mathbf{W}_0'\right] \right\|_2 .
\end{aligned}$$

Denote by

$$L_{\mathcal{F}} = \sqrt{4B_2^2 + 2c_W^2 \frac{\beta_2^2 B_2^2}{C_2^2} + 2(1-\alpha_2)^2 \beta_1^2 c_W^2 \frac{B_1^2 C_2^2}{C_1^2} \left\| g(\widetilde{\mathbf{A}}) \right\|_\infty^2},$$

by Lemma A.4, we conclude that $|\ell(\mathbf{w}) - \ell(\mathbf{w}')| \leq L_{\mathcal{F}} \|\mathbf{w} - \mathbf{w}'\|_2$ holds. Then we discuss the smoothness. By the chain rule, we have

$$\begin{aligned}
\frac{\partial \ell(\mathbf{W}_0, \mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3; z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_3\right]} &= (\hat{\mathbf{y}}_i - \mathbf{y}_i) \otimes \mathbf{H}_{i*}^{(2)}, \\
\frac{\partial \ell(\mathbf{W}_0, \mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3; z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_2\right]} &= \alpha_2 \beta_2 \delta_i \otimes \mathbf{H}_{i*}^{(0)} + (1-\alpha_2)\beta_2 \sum_{j=1}^n \left[g(\widetilde{\mathbf{A}}) \right]_{ij} \delta_i \otimes \mathbf{H}_{j*}^{(1)}, \\
\frac{\partial \ell(\mathbf{W}_0, \mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3; z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]} &= \alpha_1 \beta_1 \sum_{j=1}^n \left[g(\widetilde{\mathbf{A}}) \right]_{ij} \delta_{ij} \otimes \mathbf{H}_{j*}^{(0)} + (1-\alpha_1)\beta_1 \sum_{j=1}^n \sum_{k=1}^n \left[g(\widetilde{A}) \right]_{ij} \left[g(\widetilde{A}) \right]_{jk} \delta_{ij} \otimes \mathbf{H}_{k*}^{(0)}, \\
&\frac{\partial \ell(\mathbf{W}_0, \mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3; z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_0\right]} \\
=& \alpha_2 ((\delta_i ((1-\beta_2)\mathbf{I} + \beta_2 \mathbf{W}_2^\top)) \odot \mathbf{H}_{i*}^{(0)}) \otimes \mathbf{X}_{i*} \\
&+ \alpha_1 \sum_{j=1}^n \left[g(\widetilde{\mathbf{A}}) \right]_{ij} ((\delta_{ij}((1-\beta_1)\mathbf{I} + \beta_1 \mathbf{W}_1^\top)) \odot \mathbf{H}_{j*}^{(0)}) \otimes \mathbf{X}_{j*} \\
&+ (1-\alpha_1) \sum_{j=1}^n \sum_{k=1}^n \left[g(\widetilde{\mathbf{A}}) \right]_{ij} \left[g(\widetilde{\mathbf{A}}) \right]_{jk} ((\delta_{ij}((1-\beta_1)\mathbf{I} + \beta_1 \mathbf{W}_1^\top)) \odot \mathbf{H}_{k*}^{(0)}) \otimes \mathbf{X}_{k*},
\end{aligned}$$

where

$$\begin{aligned}
\delta_i =& ((\hat{\mathbf{y}}_i - \mathbf{y}_i)\mathbf{W}_3^\top) \odot \sigma'(\mathbf{H}_{i*}^{(2)}), \\
\delta_{ij} =& (1-\alpha_2)\sigma'(\mathbf{H}_{j*}^{(1)}) \odot (\delta_i((1-\beta_2)\mathbf{W}_2^\top + \beta_2 \mathbf{I})).
\end{aligned}$$

We first analyze how δ_i and δ_{ij} change w.r.t. $\mathbf{W}_0$, $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{W}_3$.

Part C. For $i \in [n]$, we have

$$\begin{aligned}
&\left\|\delta_i(\mathbf{W}_3)-\delta_i(\mathbf{W}_3')\right\|_2 \\
\leq &\left\|(\hat{\mathbf{y}}_i-\mathbf{y}_i)(\mathbf{W}_3-\mathbf{W}_3')^\top \odot \sigma'(\mathbf{H}_{i*}^{(2)})\right\|_2 + \left\|(\hat{\mathbf{y}}_i(\mathbf{W}_3)-\hat{\mathbf{y}}_i(\mathbf{W}_3'))\mathbf{W}_3^{'\top} \odot \sigma'(\mathbf{H}_{i*}^{(2)})\right\|_2 \\
\leq &\left\|(\hat{\mathbf{y}}_i-\mathbf{y}_i)(\mathbf{W}_3-\mathbf{W}_3')\right\|_2 + c_W\left\|(\hat{\mathbf{y}}_i(\mathbf{W}_3)-\hat{\mathbf{y}}_i(\mathbf{W}_3'))\right\|_2 \\
\leq &\sqrt{2}\left\|\mathbf{W}_3-\mathbf{W}_3'\right\|_2 + 2c_W\left\|\mathbf{W}_3-\mathbf{W}_3'\right\|_2 \max_{i\in[n]}\left\|\mathbf{H}_{i*}^{(2)}\right\|_2 = (\sqrt{2}+2c_W B_2)\left\|\mathrm{vec}\left[\mathbf{W}_3\right]-\mathrm{vec}\left[\mathbf{W}_3'\right]\right\|_2,
\end{aligned} \tag{54}$$

where we use the fact that the absolute value of each component in $\sigma'(\mathbf{H}_{i*}^{(2)})$ is less than 1. Similarly, we have

$$\begin{aligned}
&\left\|\delta_i(\mathbf{W}_2)-\delta_i(\mathbf{W}_2')\right\|_2 \\
\leq &\left\|(\hat{\mathbf{y}}_i-\mathbf{y}_i)\mathbf{W}_3^\top \odot (\sigma'(\mathbf{H}_{i*}^{(2)})(\mathbf{W}_2)-\sigma'(\mathbf{H}_{i*}^{(2)})(\mathbf{W}_2'))\right\|_2 + \left\|((\hat{\mathbf{y}}_i(\mathbf{W}_2)-\hat{\mathbf{y}}_i(\mathbf{W}_2'))\mathbf{W}_3^{'\top}) \odot \sigma'(\mathbf{H}_{i*}^{(2)})\right\|_2 \\
\leq &P\left\|(\hat{\mathbf{y}}_i-\mathbf{y}_i)\mathbf{W}_3^\top\right\|_2 \left\|\mathbf{H}_{i*}^{(2)}(\mathbf{W}_2)-\mathbf{H}_{i*}^{(2)}(\mathbf{W}_2')\right\|_2^\alpha + 2c_W^2\left\|\mathbf{H}_{i*}^{(2)}(\mathbf{W}_2)-\mathbf{H}_{i*}^{(2)}(\mathbf{W}_2')\right\|_2 \\
\leq &P\sqrt{2}c_W\Delta_{22}^\alpha + 2c_W^2\Delta_{22} \\
= &\sqrt{2}c_W P\left(\frac{\beta_2 B_2}{C_2}\right)^\alpha \left\|\mathrm{vec}\left[\mathbf{W}_2\right]-\mathrm{vec}\left[\mathbf{W}_2'\right]\right\|_2^\alpha + 2c_W^2\frac{\beta_2 B_2}{C_2}\left\|\mathrm{vec}\left[\mathbf{W}_2\right]-\mathrm{vec}\left[\mathbf{W}_2'\right]\right\|_2.
\end{aligned} \tag{55}$$

Besides,

$$\begin{aligned}
&\left\|\delta_i(\mathbf{W}_1)-\delta_i(\mathbf{W}_1')\right\|_2 \leq \sqrt{2}c_W P\Delta_{21}^\alpha + 2c_W^2\Delta_{21} \\
= &\sqrt{2}c_W P\left((1-\alpha_2)\beta_1\frac{B_1C_2}{C_1}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\right)^\alpha \left\|\mathrm{vec}\left[\mathbf{W}_1\right]-\mathrm{vec}\left[\mathbf{W}_1'\right]\right\|_2^\alpha \\
&+ 2(1-\alpha_2)\beta_1 c_W^2\frac{B_1C_2}{C_1}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty \left\|\mathrm{vec}\left[\mathbf{W}_1\right]-\mathrm{vec}\left[\mathbf{W}_1'\right]\right\|_2, \\
&\left\|\delta_i(\mathbf{W}_0)-\delta_i(\mathbf{W}_0')\right\|_2 \leq \sqrt{2}c_W P\Delta_{20}^\alpha + 2c_W^2\Delta_{20} \\
= &\sqrt{2}c_W P\left(\frac{B_2}{c_W}\right)^\alpha \left\|\mathrm{vec}\left[\mathbf{W}_0\right]-\mathrm{vec}\left[\mathbf{W}_0'\right]\right\|_2^\alpha + 2c_W B_2\left\|\mathrm{vec}\left[\mathbf{W}_0\right]-\mathrm{vec}\left[\mathbf{W}_0'\right]\right\|_2.
\end{aligned} \tag{56}$$

Part D. For $i \in [n]$, we have

$$\begin{aligned}
&\left\|\delta_{ij}(\mathbf{W}_3)-\delta_{ij}(\mathbf{W}_3')\right\|_2 \\
= &(1-\alpha_2)\left\|\sigma'(\mathbf{H}_{j*}^{(1)}) \odot ((\delta_i(\mathbf{W}_3)-\delta_i(\mathbf{W}_3'))((1-\beta_2)\mathbf{W}_2^\top + \beta_2\mathbf{I}))\right\|_2 \\
\leq &(1-\alpha_2)\left\|(\delta_i(\mathbf{W}_3)-\delta_i(\mathbf{W}_3'))((1-\beta_2)\mathbf{W}_2^\top + \beta_2\mathbf{I})\right\|_2 \\
\leq &(1-\alpha_2)C_2\left\|\delta_i(\mathbf{W}_3)-\delta_i(\mathbf{W}_3')\right\|_2 \leq (1-\alpha_2)C_2(\sqrt{2}+2c_W B_2)\left\|\mathrm{vec}\left[\mathbf{W}_3\right]-\mathrm{vec}\left[\mathbf{W}_3'\right]\right\|_2.
\end{aligned} \tag{57}$$

Similarly, we have

$$\begin{aligned}
&\left\|\delta_{ij}(\mathbf{W}_2)-\delta_{ij}(\mathbf{W}_2')\right\|_2 \\
\leq &(1-\alpha_2)\left\|\sigma'(\mathbf{H}_{j*}^{(1)}) \odot ((\delta_i(\mathbf{W}_2)-\delta_i(\mathbf{W}_2'))((1-\beta_2)\mathbf{W}_2^\top + \beta_2\mathbf{I}))\right\|_2 \\
&+ (1-\alpha_2)(1-\beta_2)\left\|\sigma'(\mathbf{H}_{j*}^{(1)}) \odot ((\delta_i(\mathbf{W}_2-\mathbf{W}_2')^\top))\right\|_2 \\
\leq &(1-\alpha_2)\left\|(\delta_i(\mathbf{W}_2)-\delta_i(\mathbf{W}_2'))((1-\beta_2)\mathbf{W}_2^\top + \beta_2\mathbf{I})\right\|_2 + (1-\alpha_2)(1-\beta_2)\left\|\delta_i(\mathbf{W}_2-\mathbf{W}_2')^\top\right\|_2 \\
\leq &(1-\alpha_2)C_2\left\|(\delta_i(\mathbf{W}_2)-\delta_i(\mathbf{W}_2'))\right\|_2 + (1-\alpha_2)(1-\beta_2)\left\|\delta_i\right\|_2\left\|\mathbf{W}_2-\mathbf{W}_2'\right\| \\
\leq &\sqrt{2}(1-\alpha_2)c_W PC_2\left(\frac{\beta_2 B_2}{C_2}\right)^\alpha \left\|\mathrm{vec}\left[\mathbf{W}_2\right]-\mathrm{vec}\left[\mathbf{W}_2'\right]\right\|_2^\alpha \\
&+ (1-\alpha_2)\left(2c_W^2\beta_2 B_2 + \sqrt{2}(1-\beta_2)c_W\right)\left\|\mathrm{vec}\left[\mathbf{W}_2\right]-\mathrm{vec}\left[\mathbf{W}_2'\right]\right\|_2.
\end{aligned} \tag{58}$$

Besides,

$$\begin{aligned}
&\|\delta_{ij}(\mathbf{W}_1)-\delta_{ij}(\mathbf{W}_1')\|_2\\
\leq&(1-\alpha_2)\left\|\sigma'(\mathbf{H}^{(1)}_{j*})\odot((\delta_i(\mathbf{W}_1)-\delta_i(\mathbf{W}_1'))((1-\beta_2)\mathbf{W}_2^\top+\beta_2\mathbf{I}))\right\|_2\\
&+(1-\alpha_2)\left\|(\sigma'(\mathbf{H}^{(1)}_{j*})(\mathbf{W}_1)-\sigma'(\mathbf{H}^{(1)}_{j*})(\mathbf{W}_1'))\odot(\delta_i((1-\beta_2)\mathbf{W}_2^\top+\beta_2\mathbf{I}))\right\|_2\\
\leq&(1-\alpha_2)C_2\|\delta_i(\mathbf{W}_1)-\delta_i(\mathbf{W}_1')\|_2+(1-\alpha_2)C_2\|\delta_i\|\left\|\sigma'(\mathbf{H}^{(1)}_{j*})(\mathbf{W}_1)-\sigma'(\mathbf{H}^{(1)}_{j*})(\mathbf{W}_1')\right\|_2\\
\leq&(1-\alpha_2)C_2\|\delta_i(\mathbf{W}_1)-\delta_i(\mathbf{W}_1')\|_2+\sqrt{2}(1-\alpha_2)c_WC_2P\left\|\mathbf{H}^{(1)}_{j*}(\mathbf{W}_1)-\mathbf{H}^{(1)}_{j*}(\mathbf{W}_1')\right\|_2^\alpha\\
\leq&\sqrt{2}c_WC_2P\left[(1-\alpha_2)^{1+\alpha}C_2^\alpha\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^\alpha+(1-\alpha_2)\right]\left(\frac{\beta_1B_1}{C_1}\right)^\alpha\|\mathrm{vec}\,[\mathbf{W}_1]-\mathrm{vec}\,[\mathbf{W}_1']\|_2^\alpha\\
&+2(1-\alpha_2)^2\beta_1c_W^2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\frac{B_1C_2^2}{C_1}\|\mathrm{vec}\,[\mathbf{W}_1]-\mathrm{vec}\,[\mathbf{W}_1']\|_2\,.
\end{aligned}\tag{59}$$

Finally,

$$\begin{aligned}
&\|\delta_{ij}(\mathbf{W}_0)-\delta_{ij}(\mathbf{W}_0')\|_2\\
\leq&(1-\alpha_2)\left\|\sigma'(\mathbf{H}^{(1)}_{j*})\odot((\delta_i(\mathbf{W}_0)-\delta_i(\mathbf{W}_0'))((1-\beta_2)\mathbf{W}_2^\top+\beta_2\mathbf{I}))\right\|_2\\
&+(1-\alpha_2)\left\|(\sigma'(\mathbf{H}^{(1)}_{j*})(\mathbf{W}_0)-\sigma'(\mathbf{H}^{(1)}_{j*})(\mathbf{W}_0'))\odot(\delta_i((1-\beta_2)\mathbf{W}_2^\top+\beta_2\mathbf{I}))\right\|_2\\
\leq&(1-\alpha_2)C_2\|\delta_i(\mathbf{W}_0)-\delta_i(\mathbf{W}_0')\|_2+\sqrt{2}(1-\alpha_2)c_WC_2P\left\|\mathbf{H}^{(1)}_{j*}(\mathbf{W}_0)-\mathbf{H}^{(1)}_{j*}(\mathbf{W}_0')\right\|_2^\alpha\\
\leq&\sqrt{2}(1-\alpha_2)C_2Pc_W\left[\left(\frac{B_2}{c_W}\right)^\alpha+\left(\frac{B_1}{c_W}\right)^\alpha\right]\|\mathrm{vec}\,[\mathbf{W}_0]-\mathrm{vec}\,[\mathbf{W}_0']\|_2^\alpha\\
&+2(1-\alpha_2)c_WB_2C_2\|\mathrm{vec}\,[\mathbf{W}_0]-\mathrm{vec}\,[\mathbf{W}_0']\|_2\,.
\end{aligned}\tag{60}$$

Now we are ready to discuss each gradient term in the following four parts.

Part F. First

$$\begin{aligned}
&\left\|\frac{\partial\ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial\mathrm{vec}\,[\mathbf{W}_3]}(\mathbf{W}_3)-\frac{\partial\ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial\mathrm{vec}\,[\mathbf{W}_3]}(\mathbf{W}_3')\right\|_2\\
\leq&\max_{i\in[n]}\left\|\mathbf{H}^{(2)}_{i*}\right\|_2\|\hat{\mathbf{y}}_i(\mathbf{W}_3)-\hat{\mathbf{y}}_i(\mathbf{W}_3')\|_2\leq 2B_2^2\|\mathrm{vec}\,[\mathbf{W}_2]-\mathrm{vec}\,[\mathbf{W}_2']\|_2\,.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\left\|\frac{\partial\ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial\mathrm{vec}\,[\mathbf{W}_3]}(\mathbf{W}_2)-\frac{\partial\ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial\mathrm{vec}\,[\mathbf{W}_3]}(\mathbf{W}_2')\right\|_2\\
\leq&\left\|\mathbf{H}^{(2)}_{i*}(\mathbf{W}_2)-\mathbf{H}^{(2)}_{i*}(\mathbf{W}_2')\right\|_2\|\hat{\mathbf{y}}_i-\mathbf{y}_i\|_2+\max_{i\in[n]}\left\|\mathbf{H}^{(2)}_{i*}\right\|_2\|\hat{\mathbf{y}}_i(\mathbf{W}_2)-\hat{\mathbf{y}}_i(\mathbf{W}_2')\|_2\\
=&(\sqrt{2}+2c_WB_2)\Delta_{22}=(\sqrt{2}+2c_WB_2)\frac{\beta_2B_2}{C_2}\|\mathrm{vec}\,[\mathbf{W}_2]-\mathrm{vec}\,[\mathbf{W}_2']\|_2\,.
\end{aligned}$$

Similarly, we can obtain

$$\begin{aligned}
&\left\|\frac{\partial\ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial\mathrm{vec}\,[\mathbf{W}_3]}(\mathbf{W}_1)-\frac{\partial\ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial\mathrm{vec}\,[\mathbf{W}_3]}(\mathbf{W}_1')\right\|_2\\
\leq&\left\|\mathbf{H}^{(2)}_{i*}(\mathbf{W}_1)-\mathbf{H}^{(2)}_{i*}(\mathbf{W}_1')\right\|_2\|\hat{\mathbf{y}}_i-\mathbf{y}_i\|_2+\max_{i\in[n]}\left\|\mathbf{H}^{(2)}_{i*}\right\|_2\|\hat{\mathbf{y}}_i(\mathbf{W}_1)-\hat{\mathbf{y}}_i(\mathbf{W}_1')\|_2\\
=&(\sqrt{2}+2c_WB_2)\Delta_{21}=(1-\alpha_2)\beta_1(\sqrt{2}+2c_WB_2)\frac{B_1C_2}{C_1}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\|\mathrm{vec}\,[\mathbf{W}_1]-\mathrm{vec}\,[\mathbf{W}_1']\|_2\,,
\end{aligned}$$

and

$$\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_3\right]}(\mathbf{W}_0)-\frac{\partial \ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_3\right]}(\mathbf{W}_0')\right\|_2\\
\le&\left\|\mathbf{H}_{i*}^{(2)}(\mathbf{W}_0)-\mathbf{H}_{i*}^{(2)}(\mathbf{W}_0')\right\|_2\|\hat{\mathbf{y}}_i-\mathbf{y}_i\|_2+\max_{i\in[n]}\left\|\mathbf{H}_{i*}^{(2)}\right\|_2\|\hat{\mathbf{y}}_i(\mathbf{W}_0)-\hat{\mathbf{y}}_i(\mathbf{W}_0')\|_2\\
=&(\sqrt{2}+2c_WB_2)\Delta_{20}=(\sqrt{2}+2c_WB_2)\frac{B_2}{c_W}\left\|\mathrm{vec}\left[\mathbf{W}_0\right]-\mathrm{vec}\left[\mathbf{W}_0'\right]\right\|_2.
\end{aligned}$$

Denote by

$$\begin{aligned}
&P_{33}=2B_2^2,\ P_{32}=(\sqrt{2}+2c_WB_2)\frac{\beta_2B_2}{C_2},\ P_{31}=(1-\alpha_2)\beta_1(\sqrt{2}+2c_WB_2)\frac{B_1C_2}{C_1}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty,\\
&P_{30}=(\sqrt{2}+2c_WB_2)\frac{B_2}{c_W},\ \widetilde{P}_{33}=\widetilde{P}_{32}=\widetilde{P}_{31}=\widetilde{P}_{30}=0,
\end{aligned}$$

we obtain that $\left\|\frac{\partial \ell(\mathbf{w};z_i)}{\partial \mathrm{vec}[\mathbf{W}_3]}-\frac{\partial \ell(\mathbf{w}';z_i)}{\partial \mathrm{vec}[\mathbf{W}_3]}\right\|_2\le\sum_{i=1}^4 P_{3i}\left\|\mathrm{vec}\left[\mathbf{W}_i\right]-\mathrm{vec}\left[\mathbf{W}_i'\right]\right\|_2+\sum_{i=1}^4\widetilde{P}_{3i}\left\|\mathrm{vec}\left[\mathbf{W}_i\right]-\mathrm{vec}\left[\mathbf{W}_i'\right]\right\|_2^\alpha$.

Part G. First,

$$\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_2\right]}(\mathbf{W}_3)-\frac{\partial \ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_2\right]}(\mathbf{W}_3')\right\|_2\\
=&\left\|\alpha_2\beta_2\mathbf{H}_{i*}^{(0)}+(1-\alpha_2)\beta_2\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\mathbf{H}_{j*}^{(1)}\right\|_2\|\delta_i(\mathbf{W}_3)-\delta_i(\mathbf{W}_3')\|_2\\
\le&(\sqrt{2}+2c_WB_2)\left(\alpha_2\beta_2\max_{i\in[n]}\left\|\mathbf{H}_{i*}^{(0)}\right\|_2+(1-\alpha_2)\beta_2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\max_{i\in[n]}\left\|\mathbf{H}_{i*}^{(1)}\right\|_2\right)\left\|\mathrm{vec}\left[\mathbf{W}_3\right]-\mathrm{vec}\left[\mathbf{W}_3'\right]\right\|_2\\
\le&(\sqrt{2}+2c_WB_2)\left(\alpha_2\beta_2c_Xc_W+(1-\alpha_2)\beta_2B_1\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\right)\left\|\mathrm{vec}\left[\mathbf{W}_3\right]-\mathrm{vec}\left[\mathbf{W}_3'\right]\right\|_2\\
=&(\sqrt{2}+2c_WB_2)\frac{\beta_2B_2}{C_2}\left\|\mathrm{vec}\left[\mathbf{W}_3\right]-\mathrm{vec}\left[\mathbf{W}_3'\right]\right\|_2.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_2\right]}(\mathbf{W}_2)-\frac{\partial \ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_2\right]}(\mathbf{W}_2')\right\|_2\\
=&\left\|\alpha_2\beta_2\mathbf{H}_{i*}^{(0)}+(1-\alpha_2)\beta_2\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\mathbf{H}_{j*}^{(1)}\right\|_2\|\delta_i(\mathbf{W}_2)-\delta_i(\mathbf{W}_2')\|_2\\
\le&\sqrt{2}Pc_W\left(\frac{\beta_2B_2}{C_2}\right)^{1+\alpha}\left\|\mathrm{vec}\left[\mathbf{W}_2\right]-\mathrm{vec}\left[\mathbf{W}_2'\right]\right\|_2^\alpha+2c_W^2\left(\frac{\beta_2B_2}{C_2}\right)^2\left\|\mathrm{vec}\left[\mathbf{W}_2\right]-\mathrm{vec}\left[\mathbf{W}_2'\right]\right\|_2.
\end{aligned}$$

Besides,

$$\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_2\right]}(\mathbf{W}_1)-\frac{\partial \ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_2\right]}(\mathbf{W}_1')\right\|_2\\
\le&\left\|\alpha_2\beta_2\mathbf{H}_{i*}^{(0)}+(1-\alpha_2)\beta_2\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\mathbf{H}_{j*}^{(1)}\right\|_2\|\delta_i(\mathbf{W}_1)-\delta_i(\mathbf{W}_1')\|_2\\
&+\left\|(1-\alpha_2)\beta_2\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}(\mathbf{H}_{j*}^{(1)}(\mathbf{W}_1)-\mathbf{H}_{j*}^{(1)}(\mathbf{W}_1))\right\|_2\|\delta_i\|_2\\
\le&\sqrt{2}Pc_W(1-\alpha_2)^\alpha\beta_1^\alpha\left(\frac{B_1C_2}{C_1}\right)^\alpha\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^\alpha\frac{\beta_2B_2}{C_2}\left\|\mathrm{vec}\left[\mathbf{W}_1\right]-\mathrm{vec}\left[\mathbf{W}_1'\right]\right\|_2^\alpha\\
&+(1-\alpha_2)c_W\frac{B_1\beta_1\beta_2}{C_1}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\left[2c_WB_2+\sqrt{2}\right]\left\|\mathrm{vec}\left[\mathbf{W}_1\right]-\mathrm{vec}\left[\mathbf{W}_1'\right]\right\|_2.
\end{aligned}$$

Finally,

$$\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_0, \mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3; z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_2\right]}(\mathbf{W}_0) - \frac{\partial \ell(\mathbf{W}_0, \mathbf{W}_0, \mathbf{W}_2, \mathbf{W}_3; z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_2\right]}(\mathbf{W}_0')\right\|_2 \\
\leq &\left\|\alpha_2\beta_2 \mathbf{H}_{i*}^{(0)} + (1-\alpha_2)\beta_2 \sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij} \mathbf{H}_{j*}^{(1)}\right\|_2 \left\|\delta_i(\mathbf{W}_0) - \delta_i(\mathbf{W}_0')\right\|_2 \\
&+ \left\|\alpha_2\beta_2(\mathbf{H}_{i*}^{(0)}(\mathbf{W}_0) - \mathbf{H}_{i*}^{(0)}(\mathbf{W}_0))\right\|_2 \max_{i\in[n]}\|\delta_i\|_2 + (1-\alpha_2)\beta_2 \left\|g(\widetilde{\mathbf{A}})\right\|_\infty \max_{i\in[n]}\|\delta_i\|_2 \left\|\mathbf{H}_{i*}^{(1)}(\mathbf{W}_0) - \mathbf{H}_{i*}^{(1)}(\mathbf{W}_0)\right\|_2 \\
\leq &\sqrt{2}Pc_W\left(\frac{B_2}{c_W}\right)^{\alpha} \frac{\beta_2 B_2}{C_2}\left\|\mathrm{vec}\left[\mathbf{W}_0\right] - \mathrm{vec}\left[\mathbf{W}_0'\right]\right\|_2^{\alpha} + \frac{\beta_2 B_2(\sqrt{2} + 2B_2 c_W)}{C_2}\left\|\mathrm{vec}\left[\mathbf{W}_0\right] - \mathrm{vec}\left[\mathbf{W}_0'\right]\right\|_2.
\end{aligned}$$

Denote by

$$\begin{aligned}
&P_{23} = (\sqrt{2} + 2c_W B_2)\frac{\beta_2 B_2}{C_2},\ P_{22} = 2c_W^2\left(\frac{\beta_2 B_2}{C_2}\right)^2, \\
&P_{21} = (1-\alpha_2)c_W\frac{B_1\beta_1\beta_2}{C_1}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\left[2c_W B_2 + \sqrt{2}\right],\ P_{20} = \frac{\beta_2 B_2(\sqrt{2} + 2B_2 c_W)}{C_2} \\
&\widetilde{P}_{23} = 0,\ \widetilde{P}_{22} = \sqrt{2}Pc_W\left(\frac{\beta_2 B_2}{C_2}\right)^{1+\alpha},\ \widetilde{P}_{21} = \sqrt{2}Pc_W(1-\alpha_2)^{\alpha}\beta_1^{\alpha}\left(\frac{B_1 C_2}{C_1}\right)^{\alpha}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^{\alpha}\frac{\beta_2 B_2}{C_2}, \\
&\widetilde{P}_{20} = \sqrt{2}Pc_W\left(\frac{B_2}{c_W}\right)^{\alpha}\frac{\beta_2 B_2}{C_2},
\end{aligned}$$

we obtain that $\left\|\frac{\partial \ell(\mathbf{w};z_i)}{\partial \mathrm{vec}[\mathbf{W}_2]} - \frac{\partial \ell(\mathbf{w}';z_i)}{\partial \mathrm{vec}[\mathbf{W}_2]}\right\|_2 \leq \sum_{i=1}^{4} P_{2i}\left\|\mathrm{vec}\left[\mathbf{W}_i\right] - \mathrm{vec}\left[\mathbf{W}_i'\right]\right\|_2 + \sum_{i=1}^{4}\widetilde{P}_{2i}\left\|\mathrm{vec}\left[\mathbf{W}_i\right] - \mathrm{vec}\left[\mathbf{W}_i'\right]\right\|_2^{\alpha}$.

Part H. Since (1) $\mathbf{H}^{(0)}$ is variable related to $\mathbf{W}_0$; (2) δ_{ij} is variable related to $\mathbf{W}_0$, $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{W}_3$, we have

$$\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_0, \mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3; z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}(\mathbf{W}_3) - \frac{\partial \ell(\mathbf{W}_0, \mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3; z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}(\mathbf{W}_3')\right\|_2 \\
=&\alpha_1\beta_1\sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left\|\mathbf{H}_{j*}^{(0)}\right\|_2\left\|\delta_{ij}(\mathbf{W}_3) - \delta_{ij}(\mathbf{W}_3')\right\|_2 \\
&+ (1-\alpha_1)\beta_1\sum_{j=1}^{n}\sum_{k=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left[g(\widetilde{\mathbf{A}})\right]_{jk}\left\|\mathbf{H}_{k*}^{(0)}\right\|_2\left\|\delta_{ij}(\mathbf{W}_3) - \delta_{ij}(\mathbf{W}_3')\right\|_2 \\
\leq &c_X c_W(1-\alpha_2)C_2\left(\alpha_1\beta_1\left\|g(\widetilde{\mathbf{A}})\right\|_\infty + (1-\alpha_1)\beta_1\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2\right)(\sqrt{2} + 2c_W B_2)\left\|\mathrm{vec}\left[\mathbf{W}_3\right] - \mathrm{vec}\left[\mathbf{W}_3'\right]\right\|_2 \\
=&(1-\alpha_2)\frac{\beta_1 B_1 C_2}{C_1}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty(\sqrt{2} + 2c_W B_2)\left\|\mathrm{vec}\left[\mathbf{W}_3\right] - \mathrm{vec}\left[\mathbf{W}_3'\right]\right\|_2,
\end{aligned}$$

where we have used

$$\alpha_1\beta_1 c_X c_W\left\|g(\widetilde{\mathbf{A}})\right\|_\infty + (1-\alpha_1)\beta_1 c_X c_W\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2 = \beta_1\frac{B_1}{C_1}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty.$$

Similarly, we have

$$
\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}(\mathbf{W}_2)-\frac{\partial \ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}(\mathbf{W}_2')\right\|_2\\
=&\alpha_1\beta_1\sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left\|\mathbf{H}_{j*}^{(0)}\right\|_2\|\delta_{ij}(\mathbf{W}_2)-\delta_{ij}(\mathbf{W}_2')\|_2\\
&+(1-\alpha_1)\beta_1\sum_{j=1}^{n}\sum_{k=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left[g(\widetilde{\mathbf{A}})\right]_{jk}\left\|\mathbf{H}_{k*}^{(0)}\right\|_2\|\delta_{ij}(\mathbf{W}_2)-\delta_{ij}(\mathbf{W}_2')\|_2\\
\leq&\sqrt{2}(1-\alpha_2)c_W P\frac{\beta_1 B_1 C_2}{C_1}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\left(\frac{\beta_2 B_2}{C_2}\right)^\alpha\|\mathrm{vec}\left[\mathbf{W}_2\right]-\mathrm{vec}\left[\mathbf{W}_2'\right]\|_2^\alpha\\
&+(1-\alpha_2)\frac{\beta_1 B_1}{C_1}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\left(2c_W^2\beta_2 B_2+\sqrt{2}(1-\beta_2)c_W\right)\|\mathrm{vec}\left[\mathbf{W}_2\right]-\mathrm{vec}\left[\mathbf{W}_2'\right]\|_2.
\end{aligned}
$$

Besides,

$$
\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}(\mathbf{W}_1)-\frac{\partial \ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}(\mathbf{W}_1')\right\|_2\\
=&\alpha_1\beta_1\sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left\|\mathbf{H}_{j*}^{(0)}\right\|_2\|\delta_{ij}(\mathbf{W}_1)-\delta_{ij}(\mathbf{W}_1')\|_2\\
&+(1-\alpha_1)\beta_1\sum_{j=1}^{n}\sum_{k=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left[g(\widetilde{\mathbf{A}})\right]_{jk}\left\|\mathbf{H}_{k*}^{(0)}\right\|_2\|\delta_{ij}(\mathbf{W}_1)-\delta_{ij}(\mathbf{W}_1')\|_2\\
\leq&\sqrt{2}c_W C_2 P\left[(1-\alpha_2)^{1+\alpha}C_2^\alpha\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^\alpha+(1-\alpha_2)\right]\left(\frac{\beta_1 B_1}{C_1}\right)^{\alpha+1}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\|\mathrm{vec}\left[\mathbf{W}_1\right]-\mathrm{vec}\left[\mathbf{W}_1'\right]\|_2^\alpha\\
&+2(1-\alpha_2)^2\beta_1^2 c_W^2\frac{B_1^2C_2^2}{C_1^2}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\|\mathrm{vec}\left[\mathbf{W}_1\right]-\mathrm{vec}\left[\mathbf{W}_1'\right]\|_2.
\end{aligned}
$$

Finally,

$$
\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}(\mathbf{W}_0)-\frac{\partial \ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}(\mathbf{W}_0')\right\|_2\\
=&\alpha_1\beta_1\sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left\|\mathbf{H}_{j*}^{(0)}\right\|_2\|\delta_{ij}(\mathbf{W}_0)-\delta_{ij}(\mathbf{W}_0')\|_2\\
&+(1-\alpha_1)\beta_1\sum_{j=1}^{n}\sum_{k=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left[g(\widetilde{\mathbf{A}})\right]_{jk}\left\|\mathbf{H}_{k*}^{(0)}\right\|_2\|\delta_{ij}(\mathbf{W}_0)-\delta_{ij}(\mathbf{W}_0')\|_2\\
&+\alpha_1\beta_1\sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij}\|\delta_{ij}\|_2\left\|\mathbf{H}_{j*}^{(0)}(\mathbf{W}_0)-\mathbf{H}_{j*}^{(0)}(\mathbf{W}_0')\right\|_2\\
&+(1-\alpha_1)\beta_1\sum_{j=1}^{n}\sum_{k=1}^{n}\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left[g(\widetilde{\mathbf{A}})\right]_{jk}\|\delta_{ij}\|_2\left\|\mathbf{H}_{k*}^{(0)}(\mathbf{W}_0)-\mathbf{H}_{k*}^{(0)}(\mathbf{W}_0')\right\|_2\\
\leq&\sqrt{2}(1-\alpha_2)c_W P\left[\left(\frac{B_2}{c_W}\right)^\alpha+\left(\frac{B_1}{c_W}\right)^\alpha\right]\frac{\beta_1 B_1 C_2}{C_1}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\|\mathrm{vec}\left[\mathbf{W}_0\right]-\mathrm{vec}\left[\mathbf{W}_0'\right]\|_2^\alpha\\
&+(1-\alpha_2)(\sqrt{2}+2B_2c_W)\frac{\beta_1 B_1 C_2}{C_1}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\|\mathrm{vec}\left[\mathbf{W}_0\right]-\mathrm{vec}\left[\mathbf{W}_0'\right]\|_2.
\end{aligned}
$$

Denote by

$$P_{13} = (1-\alpha_2)\frac{\beta_1 B_1 C_2}{C_1}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty(\sqrt{2}+2c_W B_2),\ P_{12} = (1-\alpha_2)\frac{\beta_1 B_1}{C_1}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\left(2c_W^2\beta_2 B_2+\sqrt{2}(1-\beta_2)c_W\right)$$
$$P_{11} = 2(1-\alpha_2)^2\beta_1^2 c_W^2\frac{B_1^2 C_2^2}{C_1^2}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty,\ P_{10} = (1-\alpha_2)(\sqrt{2}+2B_2 c_W)\frac{\beta_1 B_1 C_2}{C_1}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty$$
$$\widetilde{P}_{13} = 0,\ \widetilde{P}_{12} = \sqrt{2}(1-\alpha_2)c_W P\frac{\beta_1 B_1 C_2}{C_1}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\left(\frac{\beta_2 B_2}{C_2}\right)^\alpha$$
$$\widetilde{P}_{11} = \sqrt{2}c_W C_2 P\left[(1-\alpha_2)^{1+\alpha}C_2^\alpha\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^\alpha+(1-\alpha_2)\right]\left(\frac{\beta_1 B_1}{C_1}\right)^{\alpha+1}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty,$$
$$P_{10} = \sqrt{2}(1-\alpha_2)c_W P\left[\left(\frac{B_2}{c_W}\right)^\alpha+\left(\frac{B_1}{c_W}\right)^\alpha\right]\frac{\beta_1 B_1 C_2}{C_1}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty,$$

we obtain that $\left\|\frac{\partial\ell(\mathbf{w};z_i)}{\partial\mathrm{vec}[\mathbf{W}_1]}-\frac{\partial\ell(\mathbf{w}';z_i)}{\partial\mathrm{vec}[\mathbf{W}_1]}\right\|_2 \le \sum_{i=1}^4 P_{1i}\left\|\mathrm{vec}\left[\mathbf{W}_i\right]-\mathrm{vec}\left[\mathbf{W}_i'\right]\right\|_2+\sum_{i=1}^4\widetilde{P}_{1i}\left\|\mathrm{vec}\left[\mathbf{W}_i\right]-\mathrm{vec}\left[\mathbf{W}_i'\right]\right\|_2^\alpha$.

Part I. First we have

$$\begin{aligned}
&\left\|\frac{\partial\ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial\mathrm{vec}\left[\mathbf{W}_0\right]}(\mathbf{W}_3)-\frac{\partial\ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial\mathrm{vec}\left[\mathbf{W}_0\right]}(\mathbf{W}_3')\right\|_2\\
\le&\alpha_2\left\|\mathbf{X}_{i*}\right\|_2\left\|\delta_i(\mathbf{W}_3)-\delta_i(\mathbf{W}_3')\right\|_2\left\|(1-\beta_2)\mathbf{I}+\beta_2\mathbf{W}_2^\top\right\|_2\left\|\mathbf{H}_{i*}^{(0)}\right\|_2\\
&+\alpha_1\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left\|\mathbf{X}_{j*}\right\|_2\left\|\delta_{ij}(\mathbf{W}_3)-\delta_{ij}(\mathbf{W}_3')\right\|_2\left\|(1-\beta_1)\mathbf{I}+\beta_1\mathbf{W}_1^\top)\right\|_2\left\|\mathbf{H}_{j*}^{(0)}\right\|_2\\
&+(1-\alpha_1)\sum_{j=1}^n\sum_{k=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left[g(\widetilde{\mathbf{A}})\right]_{jk}\left\|\mathbf{X}_{k*}\right\|_2\left\|\delta_{ij}(\mathbf{W}_3)-\delta_{ij}(\mathbf{W}_3')\right\|_2\left\|(1-\beta_1)\mathbf{I}+\beta_1\mathbf{W}_1^\top)\right\|_2\left\|\mathbf{H}_{k*}^{(0)}\right\|_2\\
\le&\left(\alpha_2 C_2+\alpha_1(1-\alpha_2)C_1C_2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty+(1-\alpha_1)(1-\alpha_2)C_1C_2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2\right)c_X^2c_W\\
&\times(\sqrt{2}+2c_W B_2)\left\|\mathrm{vec}\left[\mathbf{W}_3\right]-\mathrm{vec}\left[\mathbf{W}_3'\right]\right\|_2\\
=&c_X B_2(\sqrt{2}+2c_W B_2)\left\|\mathrm{vec}\left[\mathbf{W}_3\right]-\mathrm{vec}\left[\mathbf{W}_3'\right]\right\|_2.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\left\|\frac{\partial\ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial\mathrm{vec}\left[\mathbf{W}_0\right]}(\mathbf{W}_2)-\frac{\partial\ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial\mathrm{vec}\left[\mathbf{W}_0\right]}(\mathbf{W}_2')\right\|_2\\
\le&\alpha_2\left\|\mathbf{X}_{i*}\right\|_2\left\|\delta_i(\mathbf{W}_2)-\delta_i(\mathbf{W}_2')\right\|_2\left\|(1-\beta_2)\mathbf{I}+\beta_2\mathbf{W}_2^\top\right\|_2\left\|\mathbf{H}_{i*}^{(0)}\right\|_2\\
&+\alpha_2\beta_2\left\|\mathbf{X}_{i*}\right\|_2\left\|\delta_i\right\|_2\left\|\mathbf{H}_{i*}^{(0)}\right\|_2\left\|\mathbf{W}_2-\mathbf{W}_2'\right\|\\
&+\alpha_1\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left\|\mathbf{X}_{j*}\right\|_2\left\|\delta_{ij}(\mathbf{W}_2)-\delta_{ij}(\mathbf{W}_2')\right\|_2\left\|(1-\beta_1)\mathbf{I}+\beta_1\mathbf{W}_1^\top)\right\|_2\left\|\mathbf{H}_{j*}^{(0)}\right\|_2\\
&+(1-\alpha_1)\sum_{j=1}^n\sum_{k=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left[g(\widetilde{\mathbf{A}})\right]_{jk}\left\|\mathbf{X}_{k*}\right\|_2\left\|\delta_{ij}(\mathbf{W}_2)-\delta_{ij}(\mathbf{W}_2')\right\|_2\left\|(1-\beta_1)\mathbf{I}+\beta_1\mathbf{W}_1^\top)\right\|_2\left\|\mathbf{H}_{k*}^{(0)}\right\|_2\\
\le&B_2c_Xc_W\sqrt{2}P\left(\frac{\beta_2B_2}{C_2}\right)^\alpha\left\|\mathrm{vec}\left[\mathbf{W}_2\right]-\mathrm{vec}\left[\mathbf{W}_2'\right]\right\|_2^\alpha\\
&+\left[2c_Xc_W^2\beta_2\frac{B_2^2}{C_2}+\sqrt{2}\alpha_2\beta_2c_X^2c_W^2+\sqrt{2}(1-\alpha_2)(1-\beta_2)c_Xc_WB_1\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\right]\left\|\mathrm{vec}\left[\mathbf{W}_2\right]-\mathrm{vec}\left[\mathbf{W}_2'\right]\right\|_2.
\end{aligned}$$

Besides,

$$
\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial \text{vec}\left[\mathbf{W}_0\right]}(\mathbf{W}_1)-\frac{\partial \ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial \text{vec}\left[\mathbf{W}_0\right]}(\mathbf{W}_1')\right\|_2\\
\leq&\alpha_2\left\|\mathbf{X}_{i*}\right\|_2\left\|\delta_i(\mathbf{W}_1)-\delta_i(\mathbf{W}_1')\right\|_2\left\|(1-\beta_2)\mathbf{I}+\beta_2\mathbf{W}_2^\top\right\|_2\left\|\mathbf{H}_{i*}^{(0)}\right\|_2\\
&+\alpha_1\beta_1\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left\|\mathbf{X}_{j*}\right\|_2\left\|\delta_{ij}\right\|_2\left\|\mathbf{H}_{j*}^{(0)}\right\|_2\left\|\mathbf{W}_1-\mathbf{W}_1'\right\|\\
&+\alpha_1\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left\|\mathbf{X}_{j*}\right\|_2\left\|\delta_{ij}(\mathbf{W}_1)-\delta_{ij}(\mathbf{W}_1')\right\|_2\left\|(1-\beta_1)\mathbf{I}+\beta_1\mathbf{W}_1^\top)\right\|_2\left\|\mathbf{H}_{j*}^{(0)}\right\|_2\\
&+(1-\alpha_1)\sum_{j=1}^n\sum_{k=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left[g(\widetilde{\mathbf{A}})\right]_{jk}\left\|\mathbf{X}_{k*}\right\|_2\left\|\delta_{ij}(\mathbf{W}_1)-\delta_{ij}(\mathbf{W}_1')\right\|_2\left\|(1-\beta_1)\mathbf{I}+\beta_1\mathbf{W}_1^\top)\right\|_2\left\|\mathbf{H}_{k*}^{(0)}\right\|_2\\
\leq&\left(B_2(1-\alpha_2)^\alpha C_2^\alpha\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^\alpha+(1-\alpha_2)B_1C_2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\right)\sqrt{2}Pc_Xc_W\left(\frac{\beta_1B_1}{C_1}\right)^\alpha\left\|\text{vec}\left[\mathbf{W}_1\right]-\text{vec}\left[\mathbf{W}_1'\right]\right\|_2^\alpha\\
&+\left(2(1-\alpha_2)\beta_1c_Xc_W^2\frac{B_1B_2C_2}{C_1}+\sqrt{2}\alpha_1(1-\alpha_2)\beta_1c_X^2c_W^2C_2\right)\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\left\|\text{vec}\left[\mathbf{W}_1\right]-\text{vec}\left[\mathbf{W}_1'\right]\right\|_2.
\end{aligned}
$$

Finally,

$$
\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial \text{vec}\left[\mathbf{W}_0\right]}(\mathbf{W}_0)-\frac{\partial \ell(\mathbf{W}_0,\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3;z_i)}{\partial \text{vec}\left[\mathbf{W}_0\right]}(\mathbf{W}_0')\right\|_2\\
\leq&\alpha_2\left\|\mathbf{X}_{i*}\right\|_2\left\|\delta_i(\mathbf{W}_0)-\delta_i(\mathbf{W}_0')\right\|_2\left\|(1-\beta_2)\mathbf{I}+\beta_2\mathbf{W}_2^\top\right\|_2\left\|\mathbf{H}_{i*}^{(0)}\right\|_2\\
&+\alpha_2\left\|\mathbf{X}_{i*}\right\|_2\left\|\delta_i((1-\beta_2)\mathbf{I}+\beta_2\mathbf{W}_2^\top)\right\|_2\left\|\mathbf{H}_{i*}^{(0)}(\mathbf{W}_0)-\mathbf{H}_{i*}^{(0)}(\mathbf{W}_0')\right\|_2\\
&+\alpha_1\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left\|\mathbf{X}_{j*}\right\|_2\left\|\delta_{ij}(\mathbf{W}_0)-\delta_{ij}(\mathbf{W}_0')\right\|_2\left\|(1-\beta_1)\mathbf{I}+\beta_1\mathbf{W}_1^\top)\right\|_2\left\|\mathbf{H}_{j*}^{(0)}\right\|_2\\
&+\alpha_1\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left\|\mathbf{X}_{j*}\right\|_2\left\|\delta_{ij}((1-\beta_1)\mathbf{I}+\beta_1\mathbf{W}_1^\top)\right\|_2\left\|\mathbf{H}_{j*}^{(0)}(\mathbf{W}_0)-\mathbf{H}_{j*}^{(0)}(\mathbf{W}_0')\right\|_2\\
&+(1-\alpha_1)\sum_{j=1}^n\sum_{k=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left[g(\widetilde{\mathbf{A}})\right]_{jk}\left\|\mathbf{X}_{k*}\right\|_2\left\|\delta_{ij}(\mathbf{W}_0)-\delta_{ij}(\mathbf{W}_0')\right\|_2\left\|(1-\beta_1)\mathbf{I}+\beta_1\mathbf{W}_1^\top)\right\|_2\left\|\mathbf{H}_{k*}^{(0)}\right\|_2\\
&+(1-\alpha_1)\sum_{j=1}^n\sum_{k=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left[g(\widetilde{\mathbf{A}})\right]_{jk}\left\|\mathbf{X}_{k*}\right\|_2\left\|\delta_{ij}((1-\beta_1)\mathbf{I}+\beta_1\mathbf{W}_1^\top)\right\|_2\left\|\mathbf{H}_{k*}^{(0)}(\mathbf{W}_0)-\mathbf{H}_{k*}^{(0)}(\mathbf{W}_0')\right\|_2\\
\leq&c_XB_2(\sqrt{2}+2c_WB_2)\left\|\text{vec}\left[\mathbf{W}_0\right]-\text{vec}\left[\mathbf{W}_0'\right]\right\|\\
&+\sqrt{2}c_Xc_WP\left[B_2\left(\frac{B_2}{c_W}\right)^\alpha+(1-\alpha_2)C_2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty B_1\left(\frac{B_1}{c_W}\right)^\alpha\right]\left\|\text{vec}\left[\mathbf{W}_0\right]-\text{vec}\left[\mathbf{W}_0'\right]\right\|_2^\alpha.
\end{aligned}
$$

Denote by

$$
\begin{aligned}
&P_{03}=c_XB_2(\sqrt{2}+2c_WB_2),\ P_{02}=2c_Xc_W^2\beta_2\frac{B_2^2}{C_2}+\sqrt{2}\alpha_2\beta_2c_X^2c_W^2+\sqrt{2}(1-\alpha_2)(1-\beta_2)c_Xc_WB_1\left\|g(\widetilde{\mathbf{A}})\right\|_\infty,\\
&P_{01}=\left(2(1-\alpha_2)\beta_1c_Xc_W^2\frac{B_1B_2C_2}{C_1}+\sqrt{2}\alpha_1(1-\alpha_2)\beta_1c_X^2c_W^2C_2\right)\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\ P_{00}=c_XB_2(\sqrt{2}+2c_WB_2),\\
&\widetilde{P}_{03}=0,\ \widetilde{P}_{02}=B_2c_Xc_W\sqrt{2}P\left(\frac{\beta_2B_2}{C_2}\right)^\alpha,\\
&\widetilde{P}_{01}=\left(B_2(1-\alpha_2)^\alpha C_2^\alpha\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^\alpha+(1-\alpha_2)B_1C_2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\right)\sqrt{2}Pc_Xc_W\left(\frac{\beta_1B_1}{C_1}\right)^\alpha,\\
&\widetilde{P}_{00}=\sqrt{2}c_Xc_WP\left[B_2\left(\frac{B_2}{c_W}\right)^\alpha+(1-\alpha_2)C_2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty B_1\left(\frac{B_1}{c_W}\right)^\alpha\right],
\end{aligned}
$$

we obtain that $\left\|\frac{\partial \ell(\mathbf{w};z_i)}{\partial \text{vec}[\mathbf{W}_0]} - \frac{\partial \ell(\mathbf{w}';z_i)}{\partial \text{vec}[\mathbf{W}_0]}\right\|_2 \le \sum_{i=1}^4 P_{0i} \left\|\text{vec}\left[\mathbf{W}_i\right] - \text{vec}\left[\mathbf{W}_i'\right]\right\|_2 + \sum_{i=1}^4 \widetilde{P}_{0i} \left\|\text{vec}\left[\mathbf{W}_i\right] - \text{vec}\left[\mathbf{W}_i'\right]\right\|_2^\alpha$. Combing the results in Part F, Part G, Part H, Part I, we conclude that $\|\nabla \ell(\mathbf{w}) - \nabla \ell(\mathbf{w}')\|_2 \le P_{\mathcal{F}} \max\{\|\mathbf{w}-\mathbf{w}'\|_2, \|\mathbf{w}-\mathbf{w}'\|_2^\alpha\}$ holds where $\mathbf{w} = [\text{vec}\left[\mathbf{W}_0\right]; \text{vec}\left[\mathbf{W}_1\right]; \text{vec}\left[\mathbf{W}_2\right]; \text{vec}\left[\mathbf{W}_3\right]]$ by Lemma A.4.

B.2.3. Proof of Proposition 4.13

For two layer SGC, we have $g(\widetilde{\mathbf{A}}) = \widetilde{\mathbf{A}}^2$. Note that

$$
\begin{aligned}
|\ell(\mathbf{W}_1\mathbf{W}_2; z_i) - \ell(\mathbf{W}_1\mathbf{W}_2'; z_i)| \le & \sqrt{2}\left\|\sum_{j=1}^n \left[g(\widetilde{\mathbf{A}})\right]_{ij} \mathbf{X}_{j*}\mathbf{W}_1(\mathbf{W}_2 - \mathbf{W}_2')\right\|_2 \\
\le & \sqrt{2}\sum_{j=1}^n \left[g(\widetilde{\mathbf{A}})\right]_{ij} \|\mathbf{X}_{j*}\mathbf{W}_1\|_2 \|\mathbf{W}_2 - \mathbf{W}_2'\| \\
\le & c_X c_W \sqrt{2}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty \|\mathbf{W}_2 - \mathbf{W}_2'\|.
\end{aligned}
$$

Similarly,

$$
|\ell(\mathbf{W}_1, \mathbf{W}_2; z_i) - \ell(\mathbf{W}_1', \mathbf{W}_2; z_i)| \le c_X c_W \sqrt{2}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty \|\mathbf{W}_1 - \mathbf{W}_1'\|.
$$

Denote by $L_1 = L_2 = c_X c_W \sqrt{2}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty$, we conclude that$|\ell(\mathbf{w}) - \ell(\mathbf{w}')| \le L_{\mathcal{F}}\|\mathbf{w}-\mathbf{w}'\|_2$ holds with $L_{\mathcal{F}} = 2c_X c_W \left\|g(\widetilde{\mathbf{A}})\right\|_\infty$ by Lemma A.4. Then we discuss the Hölder smoothness. By the chain rule, the gradients are

$$
\begin{aligned}
\frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2; z_i)}{\partial \text{vec}\left[\mathbf{W}_2\right]} &= \sum_{j=1}^n \left[g(\widetilde{\mathbf{A}})\right]_{ij} (\hat{\mathbf{y}}_i - \mathbf{y}_i) \otimes (\mathbf{X}_{j*}\mathbf{W}_1), \\
\frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2; z_i)}{\partial \text{vec}\left[\mathbf{W}_1\right]} &= \sum_{j=1}^n \left[g(\widetilde{\mathbf{A}})\right]_{ij} ((\hat{\mathbf{y}}_i - \mathbf{y}_i)\mathbf{W}_2^\top) \otimes \mathbf{X}_{j*}.
\end{aligned}
$$

First,

$$
\begin{aligned}
& \left\|\frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2; z_i)}{\partial \text{vec}\left[\mathbf{W}_2\right]}(\mathbf{W}_1) - \frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2; z_i)}{\partial \text{vec}\left[\mathbf{W}_2\right]}(\mathbf{W}_1')\right\|_2 \\
\le & \sum_{j=1}^n \left[g(\widetilde{\mathbf{A}})\right]_{ij} \|(\hat{\mathbf{y}}_i - \mathbf{y}_i) \otimes (\mathbf{X}_{j*}(\mathbf{W}_1 - \mathbf{W}_1'))\| + \sum_{j=1}^n \left[g(\widetilde{\mathbf{A}})\right]_{ij} \|(\hat{\mathbf{y}}_i(\mathbf{W}_1) - \mathbf{y}_i(\mathbf{W}_1')) \otimes (\mathbf{X}_{j*}\mathbf{W}_1)\| \\
\le & \left(\sqrt{2}c_X \left\|g(\widetilde{\mathbf{A}})\right\|_\infty + 2c_X^2 c_W^2 \left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2\right) \|\text{vec}\left[\mathbf{W}_1\right] - \text{vec}\left[\mathbf{W}_1'\right]\|.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
& \left\|\frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2; z_i)}{\partial \text{vec}\left[\mathbf{W}_2\right]}(\mathbf{W}_2) - \frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2; z_i)}{\partial \text{vec}\left[\mathbf{W}_2\right]}(\mathbf{W}_2')\right\|_2 \\
\le & \sum_{j=1}^n \left[g(\widetilde{\mathbf{A}})\right]_{ij} \|(\hat{\mathbf{y}}_i(\mathbf{W}_2) - \mathbf{y}_i(\mathbf{W}_2')) \otimes (\mathbf{X}_{j*}\mathbf{W}_1)\| \le 2c_X^2 c_W^2 \left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2 \|\text{vec}\left[\mathbf{W}_2\right] - \text{vec}\left[\mathbf{W}_2'\right]\|.
\end{aligned}
$$

Denote by $P_{21} = \sqrt{2}c_X \left\|g(\widetilde{\mathbf{A}})\right\|_\infty + 2c_X^2 c_W^2 \left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2, P_{22} = 2c_X^2 c_W^2 \left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2$ and $\widetilde{P}_{21} = \widetilde{P}_{22} = 0$, we obtain that $\left\|\frac{\partial \ell(\mathbf{w};z_i)}{\partial \text{vec}[\mathbf{W}_2]} - \frac{\partial \ell(\mathbf{w}';z_i)}{\partial \text{vec}[\mathbf{W}_2]}\right\|_2 \le \sum_{i=1}^2 P_{2i} \left\|\text{vec}\left[\mathbf{W}_i\right] - \text{vec}\left[\mathbf{W}_i'\right]\right\|_2 + \widetilde{P}_{2i} \left\|\text{vec}\left[\mathbf{W}_i\right] - \text{vec}\left[\mathbf{W}_i'\right]\right\|_2^\alpha$. By the same way, denote by $P_{11} = P_{22}, \widetilde{P}_{11} = \widetilde{P}_{22}$ and $P_{12} = P_{21}, \widetilde{P}_{12} = \widetilde{P}_{21}$ as well as $\alpha_{11} = \alpha_{12} = 1$, we obtain that $\left\|\frac{\partial \ell(\mathbf{w};z_i)}{\partial \text{vec}[\mathbf{W}_1]} - \frac{\partial \ell(\mathbf{w}';z_i)}{\partial \text{vec}[\mathbf{W}_1]}\right\|_2 \le \sum_{i=1}^2 P_{1i} \left\|\text{vec}\left[\mathbf{W}_i\right] - \text{vec}\left[\mathbf{W}_i'\right]\right\|_2 + \widetilde{P}_{1i} \left\|\text{vec}\left[\mathbf{W}_i\right] - \text{vec}\left[\mathbf{W}_i'\right]\right\|_2^\alpha$. By Lemma A.4, we conclude that $\|\nabla \ell(\mathbf{w}) - \nabla \ell(\mathbf{w}')\|_2 \le P_{\mathcal{F}} \max\{\|\mathbf{w}-\mathbf{w}'\|_2, \|\mathbf{w}-\mathbf{w}'\|_2^\alpha\}$ holds where $\mathbf{w} = [\text{vec}\left[\mathbf{W}_1\right]; \text{vec}\left[\mathbf{W}_2\right]]$.

B.2.4. PROOF OF PROPOSITION 4.14

We first show that the objective $\ell(\mathbf{W}_1, \mathbf{W}_2)$ is Lipschitz continuous w.r.t. $\mathbf{W}_1$ and $\mathbf{W}_2$. Note that

$$\begin{aligned}
&|\ell(\mathbf{W}_1, \mathbf{W}_2, z_i) - \ell(\mathbf{W}_1', \mathbf{W}_2, z_i)|_2 \\
\leq& \sqrt{2}\left\|\sum_{j=1}^n \left[g(\widetilde{\mathbf{A}})\right]_{ij} \sigma(\sigma(\mathbf{X}_{j*}\mathbf{W}_1)\mathbf{W}_2) - \sum_{j=1}^n \left[g(\widetilde{\mathbf{A}})\right]_{ij} \sigma(\sigma(\mathbf{X}_{j*}\mathbf{W}_1')\mathbf{W}_2)\right\|_2 \\
\leq& \sqrt{2}\sum_{j=1}^n \left|\left[g(\widetilde{\mathbf{A}})\right]_{ij}\right| \|(\sigma(\mathbf{X}_{j*}\mathbf{W}_1) - \sigma(\mathbf{X}_{j*}\mathbf{W}_1'))\mathbf{W}_2\|_2 \\
\leq& \sqrt{2}\sum_{j=1}^n \left|\left[g(\widetilde{\mathbf{A}})\right]_{ij}\right| \|\sigma(\mathbf{X}_{j*}\mathbf{W}_1) - \sigma(\mathbf{X}_{j*}\mathbf{W}_1')\|_2 \|\mathbf{W}_2\| \\
\leq& \sqrt{2}\sum_{j=1}^n \left|\left[g(\widetilde{\mathbf{A}})\right]_{ij}\right| \|\mathbf{X}_{j*}(\mathbf{W}_1 - \mathbf{W}_1')\|_2 \|\mathbf{W}_2\| \\
\leq& c_X c_W \sqrt{2}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty \|\mathrm{vec}(\mathbf{W}_1) - \mathrm{vec}(\mathbf{W}_1')\|.
\end{aligned}$$

Besides,

$$\begin{aligned}
&|\ell(\mathbf{W}_1, \mathbf{W}_2, z_i) - \ell(\mathbf{W}_1, \mathbf{W}_2', z_i)|_2 \\
\leq& \sqrt{2}\left\|\sum_{j=1}^n \left[g(\widetilde{\mathbf{A}})\right]_{ij} (\sigma(\sigma(\mathbf{X}_{j,:}\mathbf{W}_1)\mathbf{W}_2) - \sigma(\sigma(\mathbf{X}_{j,:}\mathbf{W}_1)\mathbf{W}_2'))\right\|_2 \\
\leq& \sqrt{2}\sum_{j=1}^n \left|\left[g(\widetilde{\mathbf{A}})\right]_{ij}\right| \|\sigma(\mathbf{X}_{j,:}\mathbf{W}_1)(\mathbf{W}_2 - \mathbf{W}_2')\|_2 \leq c_X c_W \sqrt{2}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty \|\mathbf{W}_2 - \mathbf{W}_2'\|.
\end{aligned}$$

By Lemma A.4, we conclude that $|\ell(\mathbf{w}) - \ell(\mathbf{w}')| \leq L_{\mathcal{F}}\|\mathbf{w} - \mathbf{w}'\|_2$ holds with $L_{\mathcal{F}} = 2c_X c_W \left\|g(\widetilde{\mathbf{A}})\right\|_\infty$. The gradients of ℓ w.r.t. $\mathbf{W}_1$ and $\mathbf{W}_2$ are

$$\begin{aligned}
\frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2; z_i)}{\partial \mathrm{vec}\,[\mathbf{W}_2]} &= \sum_{j=1}^n \left[g(\widetilde{\mathbf{A}})\right]_{ij} (\sigma'(\mathbf{H}^{(1)}\mathbf{W}_2)_{j*} \odot (\hat{\mathbf{y}}_i - \mathbf{y}_i)) \otimes (\mathbf{X}\mathbf{W}_1)_{j*}, \\
\frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2; z_i)}{\partial \mathrm{vec}\,[\mathbf{W}_1]} &= \sum_{j=1}^n \left[g(\widetilde{\mathbf{A}})\right]_{ij} (\sigma'(\mathbf{X}\mathbf{W}_1)_{j*} \odot (((\hat{\mathbf{y}}_i - \mathbf{y}_i) \odot \sigma'(\mathbf{H}^{(1)}\mathbf{W}_2)_{j*})\mathbf{W}_2^\top)) \otimes \mathbf{X}_{j*},
\end{aligned}$$

where $\mathbf{H}^{(1)} = \sigma(\mathbf{X}\mathbf{W}_1)$. Note that

$$\|\hat{\mathbf{y}}_i(\mathbf{W}_2) - \hat{\mathbf{y}}_i(\mathbf{W}_2')\|_2 \leq 2\sum_{j=1}^n \left|\left[g(\widetilde{\mathbf{A}})\right]_{ij}\right| \|\sigma(\mathbf{H}^{(1)}(\mathbf{W}_2 - \mathbf{W}_2'))\|_2 \leq 2c_X c_W \left\|g(\widetilde{\mathbf{A}})\right\|_\infty \|\mathrm{vec}\,[\mathbf{W}_2] - \mathrm{vec}\,[\mathbf{W}_2']\|_2 .$$

Similarly,

$$\|\hat{\mathbf{y}}_i(\mathbf{W}_1) - \hat{\mathbf{y}}_i(\mathbf{W}_1')\|_2 \leq 2c_X c_W \left\|g(\widetilde{\mathbf{A}})\right\|_\infty \|\mathrm{vec}\,[\mathbf{W}_1] - \mathrm{vec}\,[\mathbf{W}_1']\|_2 .$$

Part A. First we have

$$\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2; z_i)}{\partial \mathrm{vec}\,[\mathbf{W}_2]}(\mathbf{W}_2) - \frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2; z_i)}{\partial \mathrm{vec}\,[\mathbf{W}_2]}(\mathbf{W}_2')\right\|_2 \\
\leq& \sum_{j=1}^n \left|\left[g(\widetilde{\mathbf{A}})\right]_{ij}\right| \sqrt{2}\|\mathbf{X}_{j*}\mathbf{W}_1\|_2 \|\sigma'(\mathbf{H}^{(1)}\mathbf{W}_2)_{j*} - \sigma'(\mathbf{H}^{(1)}\mathbf{W}_2')_{j*}\|_2 \\
&+ \sum_{j=1}^n \left|\left[g(\widetilde{\mathbf{A}})\right]_{ij}\right| \|\mathbf{X}_{j*}\mathbf{W}_1\|_2 \|\hat{\mathbf{y}}_i(\mathbf{W}_2) - \hat{\mathbf{y}}_i(\mathbf{W}_2')\|_2 \\
\leq& \sqrt{2}P\left\|g(\widetilde{\mathbf{A}})\right\|_\infty c_X^{1+\alpha} c_W^{1+\alpha} \|\mathrm{vec}\,[\mathbf{W}_2] - \mathrm{vec}\,[\mathbf{W}_2']\|_2^\alpha + 2c_X^2 c_W^2 \left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2 \|\mathrm{vec}\,[\mathbf{W}_2] - \mathrm{vec}\,[\mathbf{W}_2']\|_2 .
\end{aligned}$$

Also,

$$\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_1,\mathbf{W}_2;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_2\right]}(\mathbf{W}_1)-\frac{\partial \ell(\mathbf{W}_1,\mathbf{W}_2;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_2\right]}(\mathbf{W}_1')\right\|_2\\
\leq&\sqrt{2}\sum_{j=1}^n\left|\left[g(\widetilde{\mathbf{A}})\right]_{ij}\right|\|\mathbf{X}_{j*}\mathbf{W}_1\|_2\|\sigma'(\mathbf{H}^{(1)}(\mathbf{W}_1)\mathbf{W}_2)_{j*}-\sigma'(\mathbf{H}^{(1)}(\mathbf{W}_1')\mathbf{W}_2)_{j*}\|_2\\
&+\sqrt{2}\sum_{j=1}^n\left|\left[g(\widetilde{\mathbf{A}})\right]_{ij}\right|\|\mathbf{X}_{j*}(\mathbf{W}_1-\mathbf{W}_1')\|_2+\sum_{j=1}^n\left|\left[g(\widetilde{\mathbf{A}})\right]_{ij}\right|\|\hat{\mathbf{y}}_i(\mathbf{W}_1)-\hat{\mathbf{y}}_i(\mathbf{W}_1')\|_2\|\mathbf{X}_{j*}\mathbf{W}_1\|_2\\
\leq&\sqrt{2}P\left\|g(\widetilde{\mathbf{A}})\right\|_\infty c_X^{1+\alpha}c_W^{1+\alpha}\|\mathrm{vec}\left[\mathbf{W}_1\right]-\mathrm{vec}\left[\mathbf{W}_1'\right]\|_2^\alpha\\
&+\left(\sqrt{2}c_X\left\|g(\widetilde{\mathbf{A}})\right\|_\infty+2c_X^2c_W^2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2\right)\|\mathrm{vec}\left[\mathbf{W}_1\right]-\mathrm{vec}\left[\mathbf{W}_1'\right]\|_2.
\end{aligned}$$

Denote by

$$P_{21}=\left(\sqrt{2}c_X\left\|g(\widetilde{\mathbf{A}})\right\|_\infty+2c_X^2c_W^2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2\right),P_{22}=2c_X^2c_W^2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2,\widetilde{P}_{21}=\widetilde{P}_{22}=\sqrt{2}P\left\|g(\widetilde{\mathbf{A}})\right\|_\infty c_X^{1+\alpha}c_W^{1+\alpha},$$

we obtain that

$$\left\|\frac{\partial \ell(\mathbf{w};z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_2\right]}-\frac{\partial \ell(\mathbf{w}';z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_2\right]}\right\|_2\leq\sum_{i=1}^2 P_{2i}\left\|\mathrm{vec}\left[\mathbf{W}_i\right]-\mathrm{vec}\left[\mathbf{W}_i'\right]\right\|_2+\widetilde{P}_{2i}\left\|\mathrm{vec}\left[\mathbf{W}_i\right]-\mathrm{vec}\left[\mathbf{W}_i'\right]\right\|_2^\alpha.$$

Part B. We have

$$\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_1,\mathbf{W}_2;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}(\mathbf{W}_2)-\frac{\partial \ell(\mathbf{W}_1,\mathbf{W}_2;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}(\mathbf{W}_2')\right\|_2\\
\leq&\sqrt{2}c_Xc_W\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left\|\sigma'(\mathbf{H}^{(1)}\mathbf{W}_2)_{j*}-\sigma'(\mathbf{H}^{(1)}\mathbf{W}_2')_{j*}\right\|_2+\sqrt{2}c_X\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\|\mathbf{W}_2-\mathbf{W}_2'\|_2\\
&+c_W\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\|\mathbf{X}_{j*}\|_2\|\hat{\mathbf{y}}_i(\mathbf{W}_2)-\hat{\mathbf{y}}_i(\mathbf{W}_2')\|_2\\
\leq&\sqrt{2}Pc_X^{1+\alpha}c_W^{1+\alpha}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\|\mathrm{vec}\left[\mathbf{W}_2\right]-\mathrm{vec}\left[\mathbf{W}_2'\right]\|_2^\alpha\\
&+\left(c_X\sqrt{2}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty+2c_X^2c_W^2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2\right)\|\mathrm{vec}\left[\mathbf{W}_2\right]-\mathrm{vec}\left[\mathbf{W}_2'\right]\|_2.
\end{aligned}$$

Also, one can find that

$$\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_1,\mathbf{W}_2;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}(\mathbf{W}_1)-\frac{\partial \ell(\mathbf{W}_1,\mathbf{W}_2;z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}(\mathbf{W}_1')\right\|_2\\
\leq&c_Xc_W\sqrt{2}\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\|\sigma'(\mathbf{X}\mathbf{W}_1)_{j*}-\sigma'(\mathbf{X}\mathbf{W}_1')_{j*}\|_2+c_W\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\|\mathbf{X}_{j*}\|_2\|\hat{\mathbf{y}}_i(\mathbf{W}_1)-\hat{\mathbf{y}}_i(\mathbf{W}_1')\|_2\\
&+c_Xc_W\sqrt{2}\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left\|\sigma'(\mathbf{H}^{(1)}(\mathbf{W}_1)\mathbf{W}_2)-\sigma'(\mathbf{H}^{(1)}(\mathbf{W}_1')\mathbf{W}_2)\right\|_2\\
\leq&c_Xc_WP\sqrt{2}\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\|\mathbf{X}_{j*}(\mathbf{W}_1-\mathbf{W}_1')\|_2^\alpha+c_Xc_WP\sqrt{2}\sum_{j=1}^n\left[g(\widetilde{\mathbf{A}})\right]_{ij}\left\|(\mathbf{H}^{(1)}(\mathbf{W}_1)-\mathbf{H}^{(1)}(\mathbf{W}_1))\mathbf{W}_2\right\|_2^\alpha\\
\leq&\left[c_X^{1+\alpha}c_W+c_Xc_W^{1+\alpha}\right]P\sqrt{2}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty\|\mathrm{vec}\left[\mathbf{W}_1\right]-\mathrm{vec}\left[\mathbf{W}_1'\right]\|_2^\alpha+2c_X^2c_W^2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2\|\mathrm{vec}\left[\mathbf{W}_1\right]-\mathrm{vec}\left[\mathbf{W}_1'\right]\|_2.
\end{aligned}$$

Denote by

$$\begin{aligned}
&P_{11}=2c_X^2c_W^2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2,\ P_{12}=\left(c_X\sqrt{2}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty+2c_X^2c_W^2\left\|g(\widetilde{\mathbf{A}})\right\|_\infty^2\right),\\
&\widetilde{P}_{11}=\left[c_X^{1+\alpha}c_W+c_Xc_W^{1+\alpha}\right]P\sqrt{2}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty,\ \widetilde{P}_{12}=c_X^{1+\alpha}c_W^{1+\alpha}P\sqrt{2}\left\|g(\widetilde{\mathbf{A}})\right\|_\infty,
\end{aligned}$$

we obtain that

$$\left\|\frac{\partial \ell(\mathbf{w}; z_i)}{\partial \text{vec}\,[\mathbf{W}_1]} - \frac{\partial \ell(\mathbf{w}'; z_i)}{\partial \text{vec}\,[\mathbf{W}_1]}\right\|_2 \le \sum_{i=1}^{2} P_{1i} \left\|\text{vec}\,[\mathbf{W}_i] - \text{vec}\,[\mathbf{W}_i']\right\|_2 + \sum_{i=1}^{2} \widetilde{P}_{1i} \left\|\text{vec}\,[\mathbf{W}_i] - \text{vec}\,[\mathbf{W}_i']\right\|_2^{\alpha}.$$

By Lemma A.4, we conclude that $\|\nabla \ell(\mathbf{w}) - \nabla \ell(\mathbf{w}')\|_2 \le P_{\mathcal{F}} \max\{\|\mathbf{w} - \mathbf{w}'\|_2, \|\mathbf{w} - \mathbf{w}'\|_2^{\alpha}\}$ holds where $\mathbf{w} = [\text{vec}\,[\mathbf{W}_1]\,; \text{vec}\,[\mathbf{W}_2]]$.

B.2.5. Proof of Proposition 4.15

We first show that the objective $\ell(\mathbf{W}_1, \mathbf{W}_2, \gamma)$ is Lipschitz continuous w.r.t. $\mathbf{W}_1$, $\mathbf{W}_2$ and γ. Note that

$$\begin{aligned}
&|\ell(\mathbf{W}_1, \mathbf{W}_2, \boldsymbol{\gamma}) - \ell(\mathbf{W}_1, \mathbf{W}_2, \boldsymbol{\gamma}')| \\
\le&\sqrt{2}\left\|\sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}}, \boldsymbol{\gamma})\right]_{ij} \sigma(\sigma(\mathbf{X}_{j*}\mathbf{W}_1)\mathbf{W}_2) - \sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}}, \boldsymbol{\gamma}')\right]_{ij} \sigma(\sigma(\mathbf{X}_{j*}\mathbf{W}_1)\mathbf{W}_2)\right\|_2 \\
=&\sqrt{2}\left\|\sum_{k=0}^{K}(\gamma_k - \gamma_k')\left(\sum_{j=1}^{n}\left[\widetilde{\mathbf{A}}^k\right]_{ij} \sigma(\sigma(\mathbf{X}_{j*}\mathbf{W}_1)\mathbf{W}_2)\right)\right\|_2 \\
\le&\sqrt{2}\|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\|_2 \sqrt{\sum_{k=0}^{K}\left\|\sum_{j=1}^{n}\left[\widetilde{\mathbf{A}}^k\right]_{ij} \sigma(\sigma(\mathbf{X}_{j*}\mathbf{W}_1)\mathbf{W}_2)\right\|_2^2} \\
\le&\sqrt{2} c_X c_W^2 \left(\sum_{k=0}^{K}\left\|\widetilde{\mathbf{A}}^k\right\|_\infty\right)^{\frac{1}{2}} \|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\|_2.
\end{aligned}$$

From the proof of Proposition 4.14, we have

$$\begin{aligned}
|\ell(\mathbf{W}_1, \mathbf{W}_2, \boldsymbol{\gamma}) - \ell(\mathbf{W}_1', \mathbf{W}_2, \boldsymbol{\gamma})| \le \sqrt{2} c_X c_W \left\|g(\widetilde{\mathbf{A}}, \boldsymbol{\gamma})\right\|_\infty, \\
|\ell(\mathbf{W}_1, \mathbf{W}_2, \boldsymbol{\gamma}) - \ell(\mathbf{W}_1, \mathbf{W}_2', \boldsymbol{\gamma})| \le \sqrt{2} c_X c_W \left\|g(\widetilde{\mathbf{A}}, \boldsymbol{\gamma})\right\|_\infty.
\end{aligned}$$

Denote by

$$L_{\mathcal{F}} = \sqrt{2 c_X^2 c_W^4 \left(\sum_{k=0}^{K}\left\|\widetilde{\mathbf{A}}^k\right\|_\infty\right) + 4 c_X^2 c_W^2 \left\|g(\widetilde{\mathbf{A}}, \boldsymbol{\gamma})\right\|_\infty^2},$$

we conclude that $|\ell(\mathbf{w}) - \ell(\mathbf{w}')| \le L_{\mathcal{F}}\|\mathbf{w} - \mathbf{w}'\|_2$ holds. Then we discuss the smoothness of this model. The gradients of $\ell(\mathbf{W}_1, \mathbf{W}_2, \gamma)$ w.r.t. $\mathbf{W}_1$, $\mathbf{W}_2$, and $\boldsymbol{\gamma}$ are

$$\begin{aligned}
\frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2, \boldsymbol{\gamma}; z_i)}{\partial \text{vec}\,[\mathbf{W}_2]} &= \sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}}, \boldsymbol{\gamma})\right]_{ij} (\sigma'(\sigma(\mathbf{X}\mathbf{W}_1)\mathbf{W}_2)_{j*} \odot (\hat{\mathbf{y}}_i - \mathbf{y}_i)) \otimes (\mathbf{X}\mathbf{W}_1)_{j*}, \\
\frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2, \boldsymbol{\gamma}; z_i)}{\partial \text{vec}\,[\mathbf{W}_1]} &= \sum_{j=1}^{n}\left[g(\widetilde{\mathbf{A}}, \boldsymbol{\gamma})\right]_{ij} (\sigma'(\mathbf{X}\mathbf{W}_1)_{j,:} \odot (((\hat{\mathbf{y}}_i - \mathbf{y}_i) \odot \sigma'(\sigma(\mathbf{X}\mathbf{W}_1)\mathbf{W}_2)_{j*})\mathbf{W}_2^\top)) \otimes \mathbf{X}_{j*}, \\
\frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2, \boldsymbol{\gamma}; z_i)}{\partial \boldsymbol{\gamma}} &= (\hat{\mathbf{y}}_i - \mathbf{y}_i)\left[\sum_{j=1}^{n}\left[\widetilde{\mathbf{A}}^0\right]_{ij} \mathbf{H}_{j*}^\top, \ldots, \sum_{j=1}^{n}\left[\widetilde{\mathbf{A}}^K\right]_{ij} \mathbf{H}_{j*}^\top\right],
\end{aligned}$$

where $\mathbf{H} = \sigma(\sigma(\mathbf{X}\mathbf{W}_1)\mathbf{W}_2)$. Note that

$$\|\hat{\mathbf{y}}_i(\boldsymbol{\gamma}) - \hat{\mathbf{y}}_i(\boldsymbol{\gamma}')\|_2 \le 2 c_X c_W^2 \left(\sum_{k=0}^{K}\left\|\widetilde{\mathbf{A}}^k\right\|_\infty\right)^{\frac{1}{2}} \|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\|_2.$$

Then one can find that

$$\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_1,\mathbf{W}_2,\boldsymbol{\gamma};z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_2\right]}(\boldsymbol{\gamma}) - \frac{\partial \ell(\mathbf{W}_1,\mathbf{W}_2,\boldsymbol{\gamma};z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_2\right]}(\boldsymbol{\gamma}')\right\|_2 \\
\leq &\left\|\sum_{k=0}^{K}(\gamma_k-\gamma_k')\left(\sum_{j=1}^{n}\left[\widetilde{\mathbf{A}}^k\right]_{ij}\left(\sigma'(\sigma(\mathbf{X}\mathbf{W}_1)\mathbf{W}_2)_{j*}\odot(\hat{\mathbf{y}}_i-\mathbf{y}_i)\right)\otimes(\mathbf{X}\mathbf{W}_1)_{j*}\right)\right\| \\
&+\sum_{j=1}^{n}\left|\left[g(\widetilde{\mathbf{A}},\boldsymbol{\gamma})\right]_{ij}\right|\|\hat{\mathbf{y}}_i(\boldsymbol{\gamma})-\hat{\mathbf{y}}_i(\boldsymbol{\gamma}')\|_2\|(\mathbf{X}\mathbf{W}_1)_{j*}\|_2 \\
\leq &\left(\sqrt{2}+2c_Xc_W\left\|g(\widetilde{\mathbf{A}},\boldsymbol{\gamma})\right\|_\infty\right)c_Xc_W^2\left(\sum_{k=0}^{K}\left\|\widetilde{\mathbf{A}}^k\right\|_\infty\right)^{\frac{1}{2}}\|\boldsymbol{\gamma}-\boldsymbol{\gamma}'\|_2.
\end{aligned}$$

Denote by

$$\begin{aligned}
&P_{21}=\left(\sqrt{2}c_X\left\|g(\widetilde{\mathbf{A}},\boldsymbol{\gamma})\right\|_\infty+2c_X^2c_W^2\left\|g(\widetilde{\mathbf{A}},\boldsymbol{\gamma})\right\|_\infty^2\right),\ P_{22}=c_X^2c_W^2\left\|g(\widetilde{\mathbf{A}},\boldsymbol{\gamma})\right\|_\infty^2, \\
&P_{23}=\left(\sqrt{2}+2c_Xc_W\left\|g(\widetilde{\mathbf{A}},\boldsymbol{\gamma})\right\|_\infty\right)c_Xc_W^2\left(\sum_{k=0}^{K}\left\|\widetilde{\mathbf{A}}^k\right\|_\infty\right)^{\frac{1}{2}},\widetilde{P}_{21}=\widetilde{P}_{22}=\sqrt{2}P\left\|g(\widetilde{\mathbf{A}},\boldsymbol{\gamma})\right\|_\infty c_X^{1+\alpha}c_W^{1+\alpha},\ \widetilde{P}_{23}=0,
\end{aligned}$$

we obtain that

$$\begin{aligned}
\left\|\frac{\partial \ell(\mathbf{w};z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}-\frac{\partial \ell(\mathbf{w}';z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}\right\|_2 \leq &\sum_{i=1}^{2}P_{2i}\left\|\mathrm{vec}\left[\mathbf{W}_i\right]-\mathrm{vec}\left[\mathbf{W}_i'\right]\right\|_2+\sum_{i=1}^{2}\widetilde{P}_{2i}\left\|\mathrm{vec}\left[\mathbf{W}_i\right]-\mathrm{vec}\left[\mathbf{W}_i'\right]\right\|_2^\alpha \\
&+P_{23}\|\boldsymbol{\gamma}-\boldsymbol{\gamma}'\|_2+\widetilde{P}_{23}\|\boldsymbol{\gamma}-\boldsymbol{\gamma}'\|_2^\alpha.
\end{aligned}$$

Besides,

$$\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_1,\mathbf{W}_2,\boldsymbol{\gamma};z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}(\boldsymbol{\gamma}) - \frac{\partial \ell(\mathbf{W}_1,\mathbf{W}_2,\boldsymbol{\gamma};z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}(\boldsymbol{\gamma}')\right\| \\
= &\left\|\sum_{k=0}^{K}(\gamma_k-\gamma_k')\left(\sum_{j=1}^{n}\left[\widetilde{\mathbf{A}}^k\right]_{ij}\left(\sigma'(\mathbf{X}\mathbf{W}_1)_{j,:}\odot\left(((\hat{\mathbf{y}}_i-\mathbf{y}_i)\odot\sigma'(\sigma(\mathbf{X}\mathbf{W}_1)\mathbf{W}_2)_{j*})\mathbf{W}_2^\top\right)\right)\otimes\mathbf{X}_{j*}\right)\right\|_2 \\
&+\sum_{j=1}^{n}c_Xc_W\left|\left[g(\widetilde{\mathbf{A}},\boldsymbol{\gamma})\right]_{ij}\right|\|\hat{\mathbf{y}}_i(\boldsymbol{\gamma})-\hat{\mathbf{y}}_i(\boldsymbol{\gamma}')\|_2 \\
\leq &\left(\sqrt{2}+2c_Xc_W\left\|g(\widetilde{\mathbf{A}},\boldsymbol{\gamma})\right\|_\infty\right)c_Xc_W^2\left(\sum_{k=0}^{K}\left\|\widetilde{\mathbf{A}}^k\right\|_\infty\right)^{\frac{1}{2}}\|\boldsymbol{\gamma}-\boldsymbol{\gamma}'\|_2.
\end{aligned}$$

Denote by

$$\begin{aligned}
&P_{11}=2c_X^2c_W^2\left\|g(\widetilde{\mathbf{A}},\boldsymbol{\gamma})\right\|_\infty^2,\ P_{12}=\sqrt{2}c_X\left\|g(\widetilde{\mathbf{A}},\boldsymbol{\gamma})\right\|_\infty+2c_X^2c_W^2\left\|g(\widetilde{\mathbf{A}},\boldsymbol{\gamma})\right\|_\infty^2, \\
&P_{13}=\left(\sqrt{2}+c_Xc_W\left\|g(\widetilde{\mathbf{A}},\boldsymbol{\gamma})\right\|_\infty\right)c_Xc_W^2\left(\sum_{k=0}^{K}\left\|\widetilde{\mathbf{A}}^k\right\|_\infty\right)^{\frac{1}{2}} \\
&\widetilde{P}_{11}=\sqrt{2}P\left[c_X^{1+\alpha}c_W+c_Xc_W^{1+\alpha}\right],\ \widetilde{P}_{12}=c_X^{1+\alpha}c_W^{1+\alpha}P\sqrt{2}\left\|g(\widetilde{\mathbf{A}},\boldsymbol{\gamma})\right\|_\infty,\widetilde{P}_{13}=0,
\end{aligned}$$

we obtain that

$$\begin{aligned}
\left\|\frac{\partial \ell(\mathbf{w};z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}-\frac{\partial \ell(\mathbf{w}';z_i)}{\partial \mathrm{vec}\left[\mathbf{W}_1\right]}\right\|_2 \leq &\sum_{i=1}^{2}P_{1i}\left\|\mathrm{vec}\left[\mathbf{W}_i\right]-\mathrm{vec}\left[\mathbf{W}_i'\right]\right\|_2+\sum_{i=1}^{2}\widetilde{P}_{1i}\left\|\mathrm{vec}\left[\mathbf{W}_i\right]-\mathrm{vec}\left[\mathbf{W}_i'\right]\right\|_2^\alpha \\
&+P_{13}\|\boldsymbol{\gamma}-\boldsymbol{\gamma}'\|_2+\widetilde{P}_{13}\|\boldsymbol{\gamma}-\boldsymbol{\gamma}'\|_2^\alpha.
\end{aligned}$$

Lastly, since

$$\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2, \boldsymbol{\gamma}; z_i)}{\partial \boldsymbol{\gamma}}(\mathbf{W}_2) - \frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2, \boldsymbol{\gamma}; z_i)}{\partial \boldsymbol{\gamma}}(\mathbf{W}_2')\right\|_2 \\
\leq & \|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_2 \sqrt{\sum_{k=0}^{K} \left\| \sum_{j=1}^{n} [\widetilde{\mathbf{A}}^k]_{ij} (\sigma(\sigma(\mathbf{X}_{j*}\mathbf{W}_1)\mathbf{W}_2) - \sigma(\sigma(\mathbf{X}_{j*}\mathbf{W}_1)\mathbf{W}_2')) \right\|_2^2} \\
&+ \|\hat{\mathbf{y}}_i(\mathbf{W}_2) - \hat{\mathbf{y}}_i(\mathbf{W}_2)\|_2 \sqrt{\sum_{k=0}^{K} \left\| \sum_{j=1}^{n} [\widetilde{\mathbf{A}}^k]_{ij} \sigma(\sigma(\mathbf{X}_{j*}\mathbf{W}_1)\mathbf{W}_2) \right\|_2^2} \\
\leq & \left(\sqrt{2} + 2c_X c_W^2 \left\|g(\widetilde{\mathbf{A}}, \boldsymbol{\gamma})\right\|_\infty\right) c_X c_W \left(\sum_{k=0}^{K} \left\|\widetilde{\mathbf{A}}^k\right\|_\infty\right)^{\frac{1}{2}} \|\mathrm{vec}\,[\mathbf{W}_2] - \mathrm{vec}\,[\mathbf{W}_2']\,\|_2.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2, \boldsymbol{\gamma}; z_i)}{\partial \boldsymbol{\gamma}}(\mathbf{W}_1) - \frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2, \boldsymbol{\gamma}; z_i)}{\partial \boldsymbol{\gamma}}(\mathbf{W}_1')\right\|_2 \\
\leq & \left(\sqrt{2} + 2c_X c_W^2 \left\|g(\widetilde{\mathbf{A}}, \boldsymbol{\gamma})\right\|_\infty\right) c_X c_W \left(\sum_{k=0}^{K} \left\|\widetilde{\mathbf{A}}^k\right\|_\infty\right)^{\frac{1}{2}} \|\mathrm{vec}\,[\mathbf{W}_1] - \mathrm{vec}\,[\mathbf{W}_1']\,\|_2,
\end{aligned}$$

and

$$\begin{aligned}
&\left\|\frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2, \boldsymbol{\gamma}; z_i)}{\partial \boldsymbol{\gamma}}(\boldsymbol{\gamma}) - \frac{\partial \ell(\mathbf{W}_1, \mathbf{W}_2, \boldsymbol{\gamma}; z_i)}{\partial \boldsymbol{\gamma}}(\boldsymbol{\gamma}')\right\|_2 \\
\leq & \|\hat{\mathbf{y}}_i(\boldsymbol{\gamma}) - \hat{\mathbf{y}}_i(\boldsymbol{\gamma}')\|_2 \sqrt{\sum_{k=0}^{K} \left\| \sum_{j=1}^{n} \left[\widetilde{\mathbf{A}}^k\right]_{ij} \sigma(\sigma(\mathbf{X}_{j*}\mathbf{W}_1)\mathbf{W}_2) \right\|_2^2} \leq 2c_X^2 c_W^4 \left(\sum_{k=0}^{K} \left\|\widetilde{\mathbf{A}}^k\right\|_\infty\right) \|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\|_2.
\end{aligned}$$

Denote by

$$\begin{aligned}
&P_{31} = P_{32} = \left(\sqrt{2} + 2c_X c_W^2 \left\|g(\widetilde{\mathbf{A}}, \boldsymbol{\gamma})\right\|_\infty\right) c_X c_W \left(\sum_{k=0}^{K} \left\|\widetilde{\mathbf{A}}^k\right\|_\infty\right)^{\frac{1}{2}},\ P_{33} = 2c_X^2 c_W^4 \left(\sum_{k=0}^{K} \left\|\widetilde{\mathbf{A}}^k\right\|_\infty\right), \\
&\widetilde{P}_{31} = \widetilde{P}_{32} = \widetilde{P}_{33} = 0,
\end{aligned}$$

we obtain that

$$\begin{aligned}
\left\|\frac{\partial \ell(\mathbf{w}; z_i)}{\partial \boldsymbol{\gamma}} - \frac{\partial \ell(\mathbf{w}'; z_i)}{\partial \boldsymbol{\gamma}}\right\|_2 \leq & \sum_{i=1}^{2} P_{3i} \|\mathrm{vec}\,[\mathbf{W}_i] - \mathrm{vec}\,[\mathbf{W}_i']\|_2 + \sum_{i=1}^{2} \widetilde{P}_{3i} \|\mathrm{vec}\,[\mathbf{W}_i] - \mathrm{vec}\,[\mathbf{W}_i']\|_2^\alpha \\
&+ P_{33}\|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\|_2 + \widetilde{P}_{33}\|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\|_2^\alpha.
\end{aligned}$$

By Lemma A.4, we conclude that $\|\nabla \ell(\mathbf{w}) - \nabla \ell(\mathbf{w}')\|_2 \leq P_{\mathcal{F}} \max\{\|\mathbf{w} - \mathbf{w}'\|_2, \|\mathbf{w} - \mathbf{w}'\|_2^\alpha\}$ holds with $\mathbf{w} = [\mathrm{vec}\,[\mathbf{W}_1]\,; \mathrm{vec}\,[\mathbf{W}_2]\,; \boldsymbol{\gamma}]$.

C. Experiments Details

For GCN, GAT, SGC, APPNP and GCNII, we adopt the official PyTorch Geometric library implementations (Fey & Lenssen, 2019). For GPRGNN, we adopt the released codes [1] with commit number 2507f10. Following (Cong et al., 2021), we remove all dropout layers and adopt the Adam optimizer with default setting. The batch size is set to 512 and the number of hidden units are set to 64 for all baseline models. K is set to 10 for APPNP and GPRGNN. For ogbn-arxiv, following the official implementation in (Hu et al., 2020), we adopt the Adam optimizer with learning rate 0.01.

[1] https://github.com/jianhao2016/GPRGNN