Generalization Analysis of Message Passing Neural Networks on Large Random Graphs

Sohir Maskey[*]
Ludwig-Maximilian University of Munich[†]
maskey@math.lmu.de

Ron Levie[*]
Technion - Israel Institute of Technology [‡]
levieron@technion.ac.il

Yunseok Lee
Ludwig-Maximilian University of Munich[†]
ylee@math.lmu.de

Gitta Kutyniok
Ludwig-Maximilian University of Munich[†]
University of Tromsø[§]
kutyniok@math.lmu.de

Abstract

Message passing neural networks (MPNN) have seen a steep rise in popularity since their introduction as generalizations of convolutional neural networks to graph structured data, and are now considered state-of-the-art tools for solving a large variety of graph-focused problems. We study the generalization error of MPNNs in graph classification and regression. We assume that graphs of different classes are sampled from different random graph models. We show that, when training a MPNN on a dataset sampled from such a distribution, the generalization gap increases in the complexity of the MPNN, and decreases, not only with respect to the number of training samples, but also with the average number of nodes in the graphs. This shows how a MPNN with high complexity can generalize from a small dataset of graphs, as long as the graphs are large. The generalization bound is derived from a uniform convergence result, that shows that any MPNN, applied on a graph, approximates the MPNN applied on the geometric model that the graph discretizes.

1 Introduction

A graph is an abstract structure that represents a set of objects along with the connections that exist between those objects. In many important fields, such as chemistry, biology, social networks, or computer graphics, data can be described by graphs. This has led to a tremendous interest in the development of machine learning models for graph-structured data in recent years. A ubiquitous tool for processing such data are graph convolutional neural networks (GCNNs), which extend standard Euclidean convolutional neural networks (CNNs) to graph-structured data.

Most GCNNs used in practice can be described using the general architecture of *Message Passing Neural Networks (MPNNs)*. MPNNs generalize the convolution operator to graph domains by a neighborhood aggregation or message passing scheme. By $\mathbf{f}_i^{(t-1)}$ denoting the feature of node i in layer $t-1$ and $\mathbf{e}_{j,i}$ denoting edge features from node j to i, one layer in a message passing graph

[*]The first two authors contributed equally to this work
[†]Department of Mathematics, LMU Munich, 80333 Munich, Germany
[‡]Faculty of Mathematics, Technion - Israel Institute of Technology, 32000 Haifa, Israel
[§]Department of Physics and Technology, University of Tromsø, 9019 Tromsø, Norway

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

neural network is given by

$$\mathbf{f}_i^{(t)} = \Psi^{(t)}\Big(\mathbf{f}_i^{(t-1)}, \mathbf{AGG}\big\{\Phi^{(t)}(\mathbf{f}_i^{(t-1)}, \mathbf{f}_j^{(t-1)}, \mathbf{e}_{j,i})\big\}_{j\in\mathcal{N}(i)}\Big), \tag{1}$$

where $\mathcal{N}(i)$ is the set of nodes connected to node i, $\mathbf{AGG}$ denotes a differentiable and permutation invariant function, e.g., sum, mean, or max, and $\Psi^{(t)}$ and $\Phi^{(t)}$ denote differentiable functions such as MLPs (Multi-Layer Perceptrons) [Fey and Lenssen, 2019].

MPNNs have shown state-of-the-art performance in many graph machine learning tasks such as node or graph classification. As such, MPNNs had a tremendous impact to the applied sciences, with promising achievements such as discovering a new class of antibiotics [Stokes et al., 2020], and has impacted the industry with applications in social media, recommendation systems, and 3D reconstruction, among others (see, e.g., [Ying et al., 2018, Wang et al., 2018a,b, Monti et al., 2019, Fan et al., 2019]). The practical success of MPNNs led to a significant boost in research aimed at understanding the theoretical properties of MPNNs. See, e.g., the variational inference point of view of MPNNs [Dai et al., 2016], and algorithmic alignment of MPNNs with combinatorial algorithms [Xu et al., 2019, Morris et al., 2019].

In this paper we study the generalization capabilities of MPNNs with mean aggregation in a graph classification task. Previous works developed generalization bounds that do not depend on any model of the data, namely, graphs in these works can be generated and labeled in any arbitrary way [Scarselli et al., 2018, Garg et al., 2020, Liao et al., 2021]. In this work, we consider a generative model for the graphs which is theoretically powerful and general on the one hand, and allows much tighter generalization bounds on the other hand.

Formally, we are given pairs of graphs and graph signals $\mathbf{x} = (G, \mathbf{f})$ and a target output $\mathbf{y}$, where $(\mathbf{x}, \mathbf{y})$ are jointly drawn from a distribution $\mu_{\mathcal{G}}(\mathbf{x}, \mathbf{y})$. The goal is to learn a MPNN Θ that approximates $\mathbf{y}$ by $\Theta(\mathbf{x})$. For this, one uses a loss function $\mathcal{L}$, which measures the discrepancy between the true label $\mathbf{y}$ and the output of the MPNN $\Theta(\mathbf{x})$. The aim of a machine learning algorithm is to minimize the statistical loss (also called expected loss)

$$R_{exp}(\Theta) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mu_{\mathcal{G}}}\Big[\mathcal{L}(\Theta(\mathbf{x}), \mathbf{y})\Big].$$

In (data-driven) machine learning one has only access to a training set instead of knowing the distribution $\mu_{\mathcal{G}}$. Namely, we consider a multi-graph setting, where the training set $\mathcal{T} = (\mathbf{x}^i = (G^i, \mathbf{f}^i), \mathbf{y}^i)_{i=1}^m$ is a collection of m samples drawn i.i.d. from the distribution $\mu_{\mathcal{G}}(\mathbf{x}, \mathbf{y})$. Then, instead of minimizing the statistical loss, one minimizes the empirical loss, given by

$$R_{\text{emp}}(\Theta) = \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(\Theta(\mathbf{x}^i), \mathbf{y}^i).$$

The optimized MPNN then depends on the dataset, and is hence denoted by $\Theta_{\mathcal{T}}$. The *generalization error* is defined to be

$$GE(\Theta_{\mathcal{T}}) = |R_{\text{exp}}(\Theta_{\mathcal{T}}) - R_{\text{emp}}(\Theta_{\mathcal{T}})|. \tag{2}$$

One then usually bounds (2) by the *uniform generalization error*

$$GE = \sup_{\Theta}|R_{\text{exp}}(\Theta) - R_{\text{emp}}(\Theta)|, \tag{3}$$

where the supremum is taken over some space of MPNNs. Bounds of GE typically take the form $GE^2 \leq \frac{C}{m}q(N)$, where C is a constant that describes the complexity of the model class (e.g., number of parameters), m is the size of the training set, and $q(N)$ is a constant that depends on the (average) size of the graphs. For such bounds, see, e.g., VC-dimension based bounds [Scarselli et al., 2018], Rademacher complexity based bounds [Garg et al., 2020], and PAC-Bayesian based bounds [Liao et al., 2021].

While in previous bounds from the literature $q(N)$ either increases in N or in the average degree, in this paper we develop a generalization bound that decays in the average number of nodes N. The idea is to treat the nodes of each graph as randomly sampled from some random graph model. In this point of view, not only the different graphs $\mathbf{x}^i$ are seen as random samples, but the union of all nodes of all graphs comprise together the random samples of the empirical loss. In the spirit of Monte Carlo theory, such a point of view should lead to a decay of the error between the empirical

and statistical losses as N increases. As opposed to graphs, nodes cannot be seen as independent, due to the correlations entailed by the graph structure. Hence, our analysis focuses on developing Monte Carlo error bounds in a correlated nodes regime.

Since in our approach we model graphs as randomly sampled from underlying continuous models, we define the application of message passing neural networks, not only on graphs, but also on the underlying space from which graphs are sampled. We then formulate and prove the following convergence result, that we write here informally. Let $\mathbf{x} = (G, \mathbf{f})$ be drawn from the model χ, then with high probability, we have for all MPNNs Θ

$$\|\Theta(\mathbf{x}) - \Theta(\chi)\| = O(N^{-\alpha}),$$

where N is the number of nodes in $\mathbf{x}$ and $\alpha > 0$. Based on this convergence result, we are able prove a generalization bound that decays in N.

1.1 Validity of the Proposed Model

The random graph models in our work are graphons [Lovász, 2012] with associated graphon signals (see Definition 2.3). The main assumption in our analysis is that graphs that are sampled from the same graphon belong to the same class. While this may seem like a limitation, it is actually a very mild and reasonable assumption. It is well known that equivalence classes of isomorphic graphs can be characterized by homomorphism densities [Lovász, 1967]. Namely, given two graphs G_1, G_2, if (and only if) for every simple graph F the number of homomorphisms from F to G_1 is equal to the number of homomorphisms from F to G_2, then G_1 is isomorphic to G_2. Graphon analysis relaxes this observation to a continuous similarity measure. A sequence of graphs $\{G_j\}_{j\in\mathbb{N}}$ is said to converge in the graphon sense, if for every simple graph F the homomorphism densities of F in the graphs $\{G_j\}_{j\in\mathbb{N}}$ converge to some value. Graphs from such a sequence can be thought of as being similar in some sense which relaxes the combinatorial notion of graph isomorphism. Moreover, for each such converging sequence, there is a unique (up to some symmetry) limit object, called a graphon. This graphon is also seen as a generative model for graphs in the respective sequence, where graphs are generated by randomly sampling the graphon (see Definition 2.3). Now, since it is well known that MPNNs cannot distinguish between isomorphic graphs, it is also unreasonable to expect them to separate two graphs that are sampled from the same graphon. We hence assume that two graphs that are sampled from the same graphon belong to the same class (but not necessarily vice versa). This assumption allows us to derive a generalization bound that is much tighter than previously proposed bounds (see Figure 1 for comparison).

From a practical stance, our graphon assumption is reasonable since many graph models are special cases of graphons, like Erdős–Rényi, stochastic block model, and random geometric graphs [Penrose, 2003]. Moreover, the decoder of a graph variational autoencoder [Kipf and Welling, 2016] can be seen as a graphon.

1.2 Related Work

In this subsection we briefly survey different approaches for studying the convergence and generalization capabilities of GCNNs that were introduced in previous contributions. We give a comparison with our results in Section 3.

Levie et al. [2021] introduce the notion of GCNN transferability – the ability to transfer a GCNN between different graphs, which is closely related to generalization. For example, Levie et al. [2019], Gama et al. [2020], Kenlay et al. [2021] show that the output of spectral-based GCNNs is linearly stable with respect to perturbations of the input graphs. Levie et al. [2021] prove that spectral-based methods are transferable under graphs and graph signals that are sampled from the same latent space. Keriven et al. [2020], Ruiz et al. [2021a,b], Maskey et al. [2021] show that spectral-based GCNNs are transferable under graphs that approximate the same limit object – the so called graphon. Cervino et al. [2021] show that gradients of spectral-based GCNNs are transferable under graphs that approximate the same graphon.

Scarselli et al. [2018] provide generalization bounds that are comparable to VC-dimension bounds known for CNNs. These bounds are improved by Garg et al. [2020], who provide the first data dependent generalization bounds for MPNNs with sum aggregation that are comparable to Rademacher bounds for recurrent neural networks. Liao et al. [2021] derive a generalization bound via a PAC-Bayesian approach that is governed by the maximum node degree and spectral norms of the weights.

Verma and Zhang [2019a] consider generalization abilities of single-layer spectral GCNNs for node-classification task and provide a generalization bound that is directly proportional to the largest eigenvalue of the graph Laplacian. Another work along these lines comes from Yehudai et al. [2021], who show that certain MPNNs (with sum aggregation) do not generalize from small to large graphs.

1.3 Main Contributions

We follow the route of Keriven et al. [2020] and consider graphs as discretizations of continuous spaces in our analysis, called random graph models (RGM, see Definition 2.3). We introduce a continuous version of message passing neural networks – the realization of MPNNs on random graph models, which we call cMPNNs. Such cMPNNs are seen as limit objects of graph MPNNs, when the number of graph nodes goes to infinity. We prove, up to our knowledge, the first convergence result of the graph MPNN to the corresponding cMPNN as the number of nodes increases, which is uniform in the choice of the MPNN.

For the generalization analysis, we assume that the data distribution $\mu_{\mathcal{G}}$ represents graphs which are randomly sampled from a collection of template RGMs, with a random number of nodes. Using our convergence results, we can then prove that the generalization error between the training set and the true distribution is small. Here, we give the following informal version of Theorem 3.3.

Theorem 1.1 (Informal version of Theorem 3.3)**.** *Consider a graph classification task with m training samples $\mathcal{T} = (\mathbf{x}^i = (G^i, \mathbf{f}^i), \mathbf{y}^i)_{i=1}^m$ drawn i.i.d. from the data distribution $\mu_{\mathcal{G}}(\mathbf{x}, \mathbf{y})$ on a metric-measure space χ of dimension D_χ. Suppose that the size N of each graph in $\mathcal{T}$ is drawn from a distribution ν. Then*

$$\mathbb{E}_{\mathcal{T}\sim\mu_{\mathcal{G}}^m}\left[\sup_{\Theta}\left(R_{emp}(\Theta) - R_{exp}(\Theta)\right)^2\right] \leq \frac{C}{m}\mathbb{E}_{N\sim\nu}\left[N^{-\frac{1}{D_\chi+1}}\right].$$

The constant C represents the complexity of the hypothesis space of the network, via the Lipschitz constants of the message and update functions and the depth of the MPNNs.

Theorem 3.3 shows how we can use fewer graphs m than model complexity C when training MPNNs if the graphs are sufficiently large.

2 Preliminaries

A weighted graph $G = (V, \mathbf{W}, E)$ with N nodes is a tuple, where $V = \{1, \ldots, N\}$ is the node set. The edge set is given by $E \subset V \times V$, where $(i, j) \in E$ if node i and j are connected by an edge. $\mathbf{W} = (w_{k,l})_{k,l}$ is the weight matrix, assigning the weight $w_{i,j}$ to the edge $(i, j) \in E$, and assigning zero if (i, j) is not an edge. The degree d_i of a node i is defined as $\mathrm{d}_i = \sum_{j=1}^N w_{i,j}$. If G is a simple graph, i.e., a weighted graph with $\mathbf{W} \in \{0, 1\}^{N\times N}$, the degree d_i is the number of nodes connected to node i by an edge. We define a *graph signal* $\mathbf{f} : V \to \mathbb{R}^F$ as a function that maps nodes to their features in $\mathbb{R}^F$, where $F \in \mathbb{N}$ is the feature dimension. The signal $\mathbf{f}$ can be represented by a matrix $\mathbf{f} = (\mathbf{f}_1, \ldots, \mathbf{f}_N) \in \mathbb{R}^{N\times F}$, where $\mathbf{f}_i \in \mathbb{R}^F$ is the feature at node i. We also call $\mathbf{f}$ a *(graph) feature map*.

For a random variable Y distributed according to κ, and a function F of Y, we denote by $\mathbb{E}_{Y\sim\kappa}[F(Y)]$ the expected value of $F(Y)$. Similarly, we denote by $\mathrm{Var}_{Y\sim\kappa}[F(Y)]$ the variance of $F(Y)$.

2.1 Message Passing Graph Neural Networks

Message passing graph neural networks (gMPNNs) are defined by realizing an architecture of a *message passing neural network (MPNN)* on a graph. MPNNs are defined independently of a particular graph.

Definition 2.1. *Let $T \in \mathbb{N}$ denote the number of layers. For $t = 1, \ldots, T$, let $\Phi^{(t)} : \mathbb{R}^{2F_{t-1}} \to \mathbb{R}^{H_{t-1}}$ and $\Psi^{(t)} : \mathbb{R}^{F_{t-1}+H_{t-1}} \to \mathbb{R}^{F_t}$ be functions that we call the* message *and* update *functions, where $F_t \in \mathbb{N}$ is called the feature dimension of layer t. The corresponding* message passing neural network (MPNN) Θ *is defined to be the sequence*

$$\Theta = ((\Phi^{(t)})_{t=1}^T, (\Psi^{(t)})_{t=1}^T).$$

The message and the update function in Definition 2.1 are often defined as multi-layer-perceptrons (MLPs). In a MPNNs, messages are sent between nodes and aggregated. An *aggregation scheme* is a permutation invariant function that takes the collection of features in the edges of each node and computes a new nodes feature. In this paper, we consider MPNNs with *mean aggregation*. Then, a gMPNN processes graph signals by realizing a MPNN on the graph as follows.

Definition 2.2. *Let $G = (V, \mathbf{W})$ be a weighted graph and Θ be a MPNN, as defined in Definition 2.1. For each $t \in \{1, \dots, T\}$, we define the* gMPNN $\Theta_G^{(t)}$ *as the mapping that maps input graph signals $\mathbf{f} = \mathbf{f}^{(0)} \in \mathbb{R}^{N \times F_0}$ to the features in the t-th layer by*

$$\Theta_G^{(t)} : \mathbb{R}^{N \times F_0} \to \mathbb{R}^{N \times F_t}, \quad \mathbf{f} \mapsto \mathbf{f}^{(t)} = (\mathbf{f}_i^{(t)})_{i=1}^N,$$

where $\mathbf{f}^{(t)} \in \mathbb{R}^{N \times F_t}$ are defined sequentially by

$$\mathbf{m}_i^{(t)} := \frac{1}{\mathrm{d}_i} \sum_{j=1}^N w_{i,j} \Phi^{(t)}(\mathbf{f}_i^{(t-1)}, \mathbf{f}_j^{(t-1)})$$
$$\mathbf{f}_i^{(t)} := \Psi^{(t)}(\mathbf{f}_i^{(t-1)}, \mathbf{m}_i^{(t)}),$$

for every $i \in V$. We call $\Theta_G := \Theta_G^{(T)}$ a message passing graph neural network (gMPNN).

Given a MPNN Θ as defined in Definition 2.1, the output $\Theta_G(\mathbf{f}) \in \mathbb{R}^{N \times F_T}$ is a graph signal. In graph classification or regression, the network should output a single feature for the whole graph. Hence, the output of a gMPNN after *global pooling* is a single vector $\Theta_G^P(\mathbf{f}) \in \mathbb{R}^{F_T}$, defined by

$$\Theta_G^P(\mathbf{f}) = \frac{1}{N} \sum_{i=1}^N \Theta_G(\mathbf{f})_i.$$

For brevity, in this paper we typically do not distinguish between a MPNN and its realization on a graph.

2.2 Random Graph Models

Let (χ, d, μ) be a metric-measure space, where χ is a set, d is a metric and μ is a probability Borel measure.

A *kernel* (also called a *graphon*), is a measurable mapping $W : \chi \times \chi \to \mathbb{R}$. The points $x \in \chi$ of the metric space are seen as the nodes of a continuous model, and the kernel is seen as a continuous version of a weight matrix. Kernels are treated as generative graph models using the following definition.

Definition 2.3. *A* random graph model (RGM) *on (χ, d, μ) is defined as a pair (W, f) of a kernel $W : \chi \times \chi \to \mathbb{R}$ and a measurable function $f : \chi \to \mathbb{R}$ called a* metric-space signal*. We define a* random graph *with corresponding node features $(G, \mathbf{f})$ by sampling N i.i.d. random points $X_1, \dots, X_N$ from χ, with probability density μ, as the nodes of G. The weight matrix $\mathbf{W} = (w_{i,j})_{i,j}$ of G is defined by $w_{i,j} = W(X_i, X_j)$ for $i, j = 1, \dots, N$. The graph signal $\mathbf{f}$ is defined by $\mathbf{f}_i = f(X_i)$. We say that $(G, \mathbf{f})$ is* drawn *from W, and denote $(G, \mathbf{f}) \sim (W, f)$.*

2.3 Continuous Message Passing Neural Networks

Given a MPNN, we define *continuous message passing neural networks (cMPNNs)* that act on kernels and metric-space signals $f : \chi \to \mathbb{R}^F$, by replacing the graph node features and the aggregation scheme in (2.2) by continuous counterparts. Let W be a kernel. We define the *kernel degree* of W at $x \in \chi$ by

$$\mathrm{d}_W(x) = \int_\chi W(x, y) d\mu(y). \tag{4}$$

Consider a message signal $U : \chi \times \chi \to \mathbb{R}^H$, where $U(x, y)$ is interpreted as a message sent from the point y to the point x in χ. We define the continuous mean aggregation of U by

$$M_W(U)(x) = \int_\chi \frac{W(x, y)}{\mathrm{d}_W(x)} U(x, y) d\mu(y).$$

Given the messages $U(x,y) = \Phi(f(x), f(y))$, where $\Phi : \mathbb{R}^{2F} \to \mathbb{R}^H$, we have

$$M_W(U)(x) = M_W\Big(\Phi\big(f(\cdot), f(\cdot\cdot)\big)\Big)(x) = \int_\chi \frac{W(x,y)}{\mathrm{d}_W(x)} \Phi\big(f(x), f(y)\big) d\mu(y).$$

By abuse of notation, we often denote in short $\Phi(f,f) := \Phi\big(f(\cdot), f(\cdot\cdot)\big)$.

By replacing mean aggregation by continuous mean aggregation in Definition 2.2, the same message and update functions that define a graph MPNN can also process metric-space signals.

Definition 2.4. *Let W be a kernel and Θ be a MPNN, as defined in Definition 2.1. For each $t \in \{1, \dots, T\}$, we define $\Theta_W^{(t)}$ as the mapping that maps the input signal to the signal in the t-th layer by*

$$\Theta_W^{(t)} : L^2(\chi) \to L^2(\chi), \quad f \mapsto f^{(t)}, \tag{5}$$

where $f^{(t)}$ are defined sequentially by

$$\begin{aligned} g^{(t)}(x) &= M_W\Big(\Phi^{(t)}\big(f^{(t-1)}, f^{(t-1)}\big)\Big)(x) \\ f^{(t)}(x) &= \Psi^{(t)}\Big(f^{(t-1)}(x), g^{(t)}(x)\Big) \end{aligned} \tag{6}$$

and $f^{(0)} = f : \chi \to \mathbb{R}^{F_0}$ is the input metric-space signal. We call $\Theta_W := \Theta_W^{(T)}$ a continuous message passing neural network (cMPNN).

As with graphs, the output of a cMPNN Θ_W on a metric-space signal $f : \chi \to \mathbb{R}^{F_0}$ is another metric-space signal $\Theta_W(f) : \chi \to \mathbb{R}^{F_T}$. The output of a cMPNN after *global pooling* is a single vector $\Theta_W^P(f) \in \mathbb{R}^{F_T}$, defined by $\Theta_W^P(\mathbf{f}) = \int_\chi \Theta_W(f)(x) d\mu(x)$.

2.4 Data Distribution for Graph Classification Tasks

In the following, we consider a training data $\mathcal{T} = \big(\mathbf{x}^i = (G^i, \mathbf{f}^i), \mathbf{y}^i\big)_{i=1}^m$ of graphs G^i, graph signals $\mathbf{f}^i$, and corresponding values $\mathbf{y}^i$ that can represent the classes of the graph-signal pairs. The training data is assumed to be drawn i.i.d. from a distribution $\mu_{\mathcal{G}}(\mathbf{x}, \mathbf{y})$ that we describe next.

In this paper, we focus on classification tasks. More precisely we have classes $j = 1, \dots, \Gamma$, each represented by a RGM (W^j, f^j) on a metric-measure space (χ^j, d^j, μ^j). In fact, we suppose that each class corresponds to a set of metric spaces. For example, a graph representing a chair can be sampled from a template of either an office chair, a garden chair, a bar stool, etc., and each of these is represented by a metric space. For simplicity of the exposition, we however treat every template metric space as its own class. This does not affect our analysis.

The distribution $\mu_{\mathcal{G}}(\mathbf{x}, \mathbf{y})$ is defined via the following procedure of data sampling. For sampling one graph, first, choose a class with probability γ_j, i.e., for $(\mathbf{x}, \mathbf{y}) \sim \mu_{\mathcal{G}}$ and $j = 1, \dots, \Gamma$, $\gamma_j = \mathbb{P}(\mathbf{y} = j)$. Independently of the choice of the class, choose the number of nodes $N \sim \nu$, where ν is a discrete distribution on $N \in \mathbb{N}$. After choosing a class $\mathbf{y} \in \{1, \dots, \Gamma\}$ and the graph size N, a random graph $(G, \mathbf{f}) \sim (W^{\mathbf{y}}, f^{\mathbf{y}})$ with N nodes is drawn from the space $\chi^{\mathbf{y}}$ with probability density of the nodes $(\mu^{\mathbf{y}})^N$.

The notation $\mathcal{T} \sim \mu_{\mathcal{G}}^m$ describes a dataset $\mathcal{T}$ consisting of m samples $(\mathbf{x}^1, \mathbf{y}^1), \dots, (\mathbf{x}^m, \mathbf{y}^m)$ drawn i.i.d. from $\mu_{\mathcal{G}}$. We refer to Subsection C.1 in the appendix for a detailed definition of the distribution $\mu_{\mathcal{G}}$.

3 Convergence and Generalization of MPNNs

In this section, we provide our main results on convergence (Subsection 3.1) and generalization (Subsection 3.2) of MPNNs. For $z \in \mathbb{R}^F$, we define $\|z\|_\infty = \max_{j=1,\dots,F} |z_j|$. Given a metric space $(\mathcal{Y}, d_{\mathcal{Y}})$, we define the infinity norm of a vector valued function $g : \mathcal{Y} \to \mathbb{R}^F$ by $\|g\|_\infty = \max_{j=1,\dots,F} \operatorname{ess\,sup}_{y \in \mathcal{Y}} |(g(y))_j|$. The function g is called *Lipschitz continuous* if there exists a constant $L_g \in \mathbb{R}$ such that for all $y, y' \in \mathcal{Y}$,

$$\|g(y) - g(y')\|_\infty \le L_g d_{\mathcal{Y}}(y, y').$$

If the domain $\mathcal{Y}$ is Euclidean, we always endow it with the L^∞-metric.

We measure the error between the output of a continuous MPNN and a gMPNN after pooling as follows. Given a graph signal $\mathbf{f} \in \mathbb{R}^{N\times F}$ and a metric-space signal $f : \chi \to \mathbb{R}^F$, both the graph and the continuous MPNN map to the same output space, i.e, $\Theta^P_W(f), \Theta^P_G(\mathbf{f}) \in \mathbb{R}^{F_T}$. Namely, the output dimension of Θ^P is independent of the random graph model it is realized on and also independent of the graph. Hence, we define the error to be the supremum norm $\|\Theta^P_W(f) - \Theta^P_G(\mathbf{f})\|_\infty$. We define the ε-covering numbers of the metric space χ, denoted by $\mathcal{C}(\chi, \varepsilon, d)$, as the minimal number of balls of radius ε required to cover χ.

For every $j = 1, \dots, \Gamma$, we make the following assumptions, which hold for the remainder of the paper. We assume that there exist constants $C_{\chi^j}, D_{\chi^j} > 0$ such that

$$\mathcal{C}(\chi^j, \varepsilon, d) \;\leq\; C_{\chi^j}\, \varepsilon^{-D_{\chi^j}} \tag{7}$$

for every $\varepsilon > 0$. Denote $D_\chi = \max_j D_{\chi^j}$ and $C_\chi = \max_j C_{\chi^j}$ Such constants exist for every metric space with finite Minkowski dimension (see Appendix A). We assume that $\operatorname{diam}(\chi^j) := \sup_{x,y\in\chi^j}\{d(x,y)\} \leq 1$. Further, we only consider kernels W^j such that there exists a constant $\mathrm{d}_{\min} > 0$ satisfying

$$\mathrm{d}_{W^j}(x) \geq \mathrm{d}_{\min}, \tag{8}$$

where the kernel degree d_{W^j} is defined in (4). We moreover assume that $W^j(x,\cdot)$ and $W^j(\cdot,x)$ are Lipschitz continuous (with respect to its second and first variable, respectively) with Lipschitz constant L_{W^j} for every $x \in \chi$. We also assume that the metric-space signal $f^j : \chi \to \mathbb{R}^F$ is Lipschitz continuous. Since the diameter of χ^j is finite, this means that $f^j \in L^\infty(\chi)$. We consider the following class of MPNNs

$$\begin{aligned}&\mathrm{Lip}_{L,B} := \\ &\Big\{\Theta = \big((\Phi^{(l)})_{l=1}^T, (\Psi^{(l)})_{l=1}^T\big) \;\Big|\; \forall l = 1,\dots,T,\;\; \Phi^{(l)} : \mathbb{R}^{F_l} \to \mathbb{R}^{H_l} \text{ and } \Psi^{(l)} : \mathbb{R}^{F_l+H_l} \to \mathbb{R}^{F_{l+1}} \\ &\qquad\qquad \text{satisfy } \; L_{\Phi^{(l)}}, L_{\Psi^{(l)}} \leq L \text{ and } \|\Phi^{(l)}(0,0)\|_\infty, \|\Psi^{(l)}(0,0)\|_\infty \leq B\Big\}.\end{aligned}$$

3.1 Convergence

In this subsection we show that the error between the cMPNN and the according gMPNN decays when the number of nodes increases.

Theorem 3.1. *Let $W : \chi^2 \to \mathbb{R}$ be a Lipschitz continuous kernel with Lipschitz constant L_W, where the metric space χ satisfies (7) with respect to the constants $C_\chi, D_\chi > 0$, and W satisfies (8). Consider a graph $(G, \mathbf{f}) \sim (W, f)$ with N nodes $X_1, \dots, X_N$ drawn i.i.d. from χ with probability density μ. Then, for every Lipschitz continuous $f : \chi \to \mathbb{R}^F$,*

$$\mathbb{E}_{X_1,\dots,X_N\sim\mu^N}\left[\sup_{\Theta\in\mathrm{Lip}_{L,B}} \left\|\Theta^P_G(\mathbf{f}) - \Theta^P_W(f)\right\|_\infty^2\right] \leq C'\big(1 + \|f\|_\infty^2 + L_f^2\big)\frac{\log(N)}{N^{1/(D_\chi+1)}} + \mathcal{O}(N^{-1}),$$

where C' is defined in Subsection B.2 of the appendix.

Remark 3.2. *The constant C' in Theorem 3.1 depends polynomially on the Lipschitz constants $L_{\Phi^{(l)}}$ and $L_{\Psi^{(l)}}$ of the message and update functions $\Phi^{(l)}$ and $\Psi^{(l)}$, on the so called* formal biases *$\|\Phi^{(l)}(0,0)\|_\infty$ and $\|\Psi^{(l)}(0,0)\|_\infty$, on $\|W\|_\infty$, on the Lipschitz constant L_W of W, on $\sqrt{\log(C_\chi)} + \sqrt{D_\chi}$, and on $\frac{1}{\mathrm{d}_{\min}}$, where the degree of the polynomial is T. A regularization of these constants can alleviate the exponential dependency of the bound on T.*

The proof of Theorem 3.1 is given in Subsection B.2 of the appendix.

Discussion and Comparison to other Convergence Results The work closest related to our convergence results is by Keriven et al. [2020], where the authors show convergence of a fixed spectral GCNN to its continuous counterpart with comparable regularity assumptions as in Theorem 3.1. Our result holds for MPNNs, which are more general than spectral GCNNs. Moreover, our bound is uniform in the choice of the MPNN Θ. This last property is essential for leveraging the convergence result to derive a generalization bound. Indeed, using the bound from Keriven et al.

[2020], for each MPNN Θ there is a different high probability event $\mathcal{E}_\Theta$ where the convergence error is small. However, the trained MPNN $\Theta = \Theta_\mathcal{T}$ depends on the dataset $\mathcal{T}$ and cannot be fixed in the analysis. Hence, we would need to intersect all events $\bigcap_\Theta \mathcal{E}_\Theta$ to guarantee a small convergence error of the trained network $\Theta_\mathcal{T}$, which would not result in an event of high probability.

3.2 Generalization

In this subsection, we state the main result of our paper, which provides a non-asymptotic bound on the generalization error of MPNNs, as defined in (3). We consider a graph classification task with a training set $\mathcal{T} = (\mathbf{x}^i = (G^i, \mathbf{f}^i), \mathbf{y}^i)_{i=1}^m$ and Γ classes. The graphs and graph features in $\mathcal{T}$ are drawn i.i.d. from a probability distribution $\mu_\mathcal{G}(\mathbf{x}, \mathbf{y})$ as described in Subsection 2.4. We recall that the distribution that samples the size of the graph is denote by ν.

Given a MPNN with pooling, Θ^P, and its output dimension $\mathbb{R}^{F_T}$, we consider a non-negative loss function $\mathcal{L} : \mathbb{R}^{F_T} \times \{1, \dots, \Gamma\} \to [0, \infty)$. Additionally, we assume that $\mathcal{L}$ is Lipschitz continuous with Lipschitz constant $L_\mathcal{L}$. Note that although the cross-entropy loss, a popular choice for loss function in classification tasks, is not Lipschitz-continuous, cross-entropy composed on softmax is.

Theorem 3.3. *There exists a constant $C > 0$ such that*

$$\mathbb{E}_{\mathcal{T}\sim p^m}\left[\sup_{\Theta\in\mathrm{Lip}_{L,B}}\left(R_{emp}(\Theta^P) - R_{exp}(\Theta^P)\right)^2\right] \leq \frac{2^\Gamma 8\|\mathcal{L}\|_\infty^2\pi}{m}$$
$$+ \frac{2^\Gamma L_\mathcal{L}^2 C}{m}\sum_j \gamma_j\left(1 + \|f^j\|_\infty^2 + L_{f^j}^2\right)\cdot\left(\mathbb{E}_{N\sim\nu}\left[\frac{1}{N} + \frac{1+\log(N)}{N^{1/(D_{\chi^j}+1)}} + \mathcal{O}\left(\exp(-N)N^{3\frac{T-1}{2}}\right)\right]\right),$$

where C is specified in Subsection C.2 of the appendix.

The proof of Theorem 3.3 is given in Subsection C.2 of the appendix.

Remark 3.4. *The constant C in Theorem 3.3 represents the complexity of the class $\mathrm{Lip}_{L,B}$ and can be bounded similarly to the constant C' from Theorem 3.1, as described in Remark 3.2. We summarize its dependencies on the parameters of the MPNN and the RGM by $\sqrt{C} \lesssim BL^{2T}\frac{1}{\mathrm{d}_{\min}^{T+1}}\max_{j=1,\dots,\Gamma}\left(\sqrt{\log(C_{\chi^j})} + \sqrt{D_{\chi^j}}\right)L_{W^j}\|W^j\|_\infty^T$ and refer to Subsection C.3 of the appendix for more details. Similarly to Remark 3.2 the exponential dependency of the constant C in Theorem 3.3 on the depth T and the polynomial dependency on the uniform Lipschitz bound L can be alleviated by regularizing the latter. We also note that the exponential dependency on the number of classes Γ in Theorem 3.3 can be eliminated by assuming that the data is representative, i.e., if the number of training samples that fall into class $j = 1, \dots, \Gamma$ is deterministically $\gamma_j m$.*

The term $\frac{2^\Gamma 8\|\mathcal{L}\|_\infty^2\pi}{m}$ in Theorem 3.3 does not depend on the model complexity and is typically much smaller than the second term. Hence, it does not affect bias–variance tradeoff considerations, and can be ignored in the situation where $m \gg C\mathbb{E}_{N\sim\nu}[\log(N)N^{-\frac{1}{D_\chi+1}}] \gg 1$. Theorem 3.3 allows us to think not just about graphs as samples, but also about individual nodes as samples. However, nodes are correlated with their neighbors, and the higher the dimension D_χ is, the larger the neighborhoods are. This is why the dependency on the number of nodes is $N^{-\frac{1}{2(D_\chi+1)}}$ and not $N^{-1/2}$. Still, this dependency of the bound on N explains one way in which we train on less graphs than model complexity and still generalize well. Another insight is that the generalization bound becomes smaller the smaller the Lipschitz constants of the message and update functions (see Remark 3.4). This indicates that regularization methods like weight decay promote generalization.

Comparison to other generalization bounds in graph classification We compare our generalization bound with other generalization bounds derived by bounding the VC-dimension [Scarselli et al., 2018], the Rademacher complexity [Garg et al., 2020], and using a PAC-Bayesian approach [Liao et al., 2021]. We do not compare with Verma and Zhang [2019b] since they derive generalization bounds for single-layered MPNNs in node-classification tasks. Hence, the role of depth is unexplored. Furthermore, their bound scales as $\mathcal{O}(\lambda_{\max}^{2T}/m)$, where T is the number of SGD steps and $\lambda_{\max}$ is the largest eigenvalue of the graph Laplacian. Hence, the generalization bound can increase monotonically for increasing T (see [Liao et al., 2021] for more details). We summarize the comparison in Table 1 and provide more details, specially on the comparability, in Subsection C.4 of the appendix.

Table 1: Comparison of generalization bounds for GNNs. We consider the following formula for a generic generalization bound: $GE \leq m^{-1/2}A(d,N)B(h)C(L,T) + Em^{-1/2}$, where m is the samples size, T is the depth, L is the bound of the Lipschitz constants of the message and update functions, h is the maximum hidden dimension, d is the average node degree and N is the graphs size and E is a term that does not depend on the model complexity.

	$A(d,N)$	$B(h)$	$C(L)$
VC-Dimension [Scarselli et al., 2018]	$\mathcal{O}(\log(N)N)$	$\mathcal{O}(h^4)$	-
Rademacher Complexity [Garg et al., 2020]	$\mathcal{O}(d^{T-1}\sqrt{\log(d^{2T-3})})$	$\mathcal{O}(h\sqrt{\log(h)})$	$\mathcal{O}(L^{2T})$
PAC-Bayesian [Liao et al., 2021]	$\mathcal{O}(d^{T-1})$	$\mathcal{O}(\sqrt{h\log(h)})$	$\mathcal{O}(L^{2T})$
Ours	$\mathcal{O}(\mathbb{E}_{N\sim\nu}[\log(N)N^{-\frac{1}{2(D_X+1)}}])$	$\mathcal{O}(1)$	$\mathcal{O}(L^{2T})$

Our analysis derives a generalization bound on MPNNs that has essentially the same dependency on the sample size m (up to a logarithmic factor), but does not directly depend on the number of hidden units. We emphasize that our bound depends on negative moments of the expected node size N. In contrast, the VC-dimension based bound [Scarselli et al., 2018] scales as $\mathcal{O}(\log(N)N)$, the Rademacher complexity based bound [Garg et al., 2020] scales as $\mathcal{O}(d^{T-1}\sqrt{\log(d^{2T-3})})$, and the PAC-Bayesian approach based bound [Liao et al., 2021] scales as $\mathcal{O}(d^{T-1})$, where d denotes the maximum node degree.

4 Numerical Experiments

We give empirical evaluations of our generalization bound in comparison to the PAC-Bayesian based bound [Liao et al., 2021] and the Rademacher complexity based bound [Garg et al., 2020]. We note that the VC dimension bound by Scarselli et al. [2018] is written in O notations and hence cannot be quantitatively evaluated. We experiment on a synthetic dataset of 100K random graphs of 50 nodes, sampled from three different RGMs: the Erdös-Rényi model (ERM) with edge probability 0.4, a smooth version of a stochastic block model (SBM), based on the kernel $K(x,y) = \sin(2\pi x)\sin(2\pi y)/2\pi + 0.25$ on $[0,1]^2$, and a geometric graph with kernel $K(x,y) = \exp(-|x-y|^2)$. The corresponding signals are given in Appendix D.2.1. Each RGM represents one class in three binary classification problems, comparing all pairs of RGMs. For the MPNN we consider GraphSAGE [Hamilton et al., 2017] with mean aggregation, and number of layers $T = 1, 2$ or 3, implemented using Pytorch Geometric [Fey and Lenssen, 2019]. We consider a maximal hidden dimension of 128. In Appendix D we give more details and also consider synthetic data sampled from additional RGMs.

Our generalization bound becomes smaller the smaller the Lipschitz constants of the message and update functions are. To control the Lipschitz constants, we consider two learning settings. First, we train with weight decay regularization, which decreases the Lipschitz bounds, and second, we train with no regularization. For each setting (each choice of the number of layers and regularization) we train the MPNN, and read the resulting Lipschitz constants of the network. We then plug all constants into our generalization bound formula (see Theorem C.7 in the appendix for the full formula), and into the generalization bound formulas of the PAC-Bayes and Rademacher bounds (see Appendix C.4 for the formulas). The results are reported in Figure 1. We observe that our generalization bounds are orders of magnitude smaller than the other works. In fact, theoretical generalization bounds typically teach us about the asymptotic behavior of generalization, and about the hyperparameters that affect generalization, but rarely give realistic numerical bounds (less than 1) that guarantee generalization. Nevertheless, in one of the scenarios (one layer MPNN) our theory gives the bounds 0.08837 and 0.1325 (respectively in the two datasets of Figure 1), which guarantees generalization in practice.

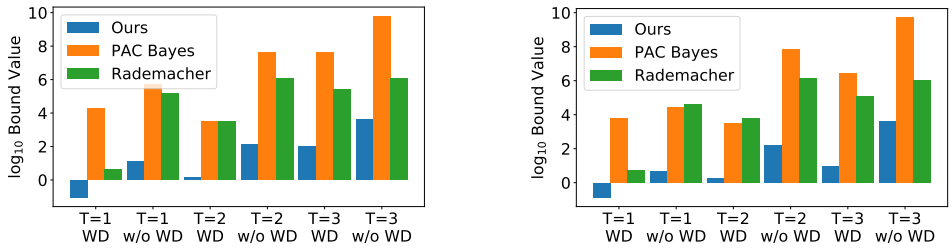


Figure 1: Generalization bounds given by our theory, PAC-Bayes [Liao et al., 2021] and Rademacher complexity [Garg et al., 2020] on a binary classification problem over Erdös-Rényi and SBM (left) and Erdös-Rényi and a geometric graph (right). Training is done with weight decay (WD) and without weight decay (w/o WD), and on three models with $T = 1, 2$ and 3 layers.

5 Conclusion

In this paper we proved that MPNNs with mean aggregation generalize from training to test data in classification tasks, if the graphs are sampled from RGMs that represent the different classes. This follows from the fact that the MPNN on sampled graphs converges to the MPNN on the RGM when the number of nodes goes to infinity. Our generalization bounds become smaller the larger the graphs, which gives one explanation to how MPNNs with high complexity can generalize well from a relatively small dataset of large graphs. We observe two main limitations of our current model. First, the dependency of the generalization bound on the size of the graph N is $\mathcal{O}(N^{-\frac{1}{2(D_X+1)}})$, which is typically slower than the observed decay in experiments (See Appendix D.1). One potential future direction is to improve this dependency using a more sophisticated models of the trained network and of the message and update functions. Secondly, our model of the data is somewhat limited. One future direction is to allow deformations of the RGMs, to consider a continuum of RGMs instead of a finite set, and to consider sparse graphs.

Acknowledgments

S.M acknowledges partial support by the NSF–Simons Research Collaboration on the Mathematical and Scientific Foundations of Deep Learning (MoDL) (NSF DMS 2031985) and DFG SPP 1798, KU 1446/27-2.

Y.L. acknowledges support by DFG-SPP-2298, KU 1446/32-1.

G.K. acknowledges support from the ONE Munich Strategy Forum (LMU Munich, TU Munich, and the Bavarian Ministery for Science and Art), the Konrad Zuse School of Excellence in Reliable AI (DAAD), the Munich Center for Machine Learning (BMBF) as well as the German Research Foundation under Grants DFG-SPP-2298, KU 1446/31-1 and KU 1446/32-1 and under Grant DFG-SFB/TR 109, Project C09 and the Federal Ministry of Education and Research under Grant MaGriDo.

References

Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, Jul 2017. ISSN 1558-0792. doi: 10.1109/msp.2017.2693418.

Juan Cervino, Luana Ruiz, and Alejandro Ribeiro. Learning by transference: Training graph neural networks on growing graphs, 2021. URL https://arxiv.org/abs/2106.03693.

Hanjun Dai, Bo Dai, and Le Song. Discriminative embeddings of latent variable models for structured data. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 2702–2711. JMLR.org, 2016.

Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. Graph neural networks for social recommendation. In *The World Wide Web Conference*, WWW '19, page 417–426,

New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313488. URL https://doi.org/10.1145/3308558.3313488.

Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Fernando Gama, Joan Bruna, and Alejandro Ribeiro. Stability properties of graph neural networks. *IEEE Transactions on Signal Processing*, 68:5680–5695, 2020. ISSN 1941-0476. doi: 10.1109/tsp.2020.3026980. URL http://dx.doi.org/10.1109/TSP.2020.3026980.

Vikas Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3419–3430. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/garg20c.html.

William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1025–1035, 2017.

Henry Kenlay, Dorina Thanou, and Xiaowen Dong. Interpretable stability bounds for spectral graph filters. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021. URL http://proceedings.mlr.press/v139/kenlay21a.html.

Nicolas Keriven, Alberto Bietti, and Samuel Vaiter. Convergence and stability of graph convolutional networks on large random graphs. *stat*, 1050:23, 2020.

Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

Ron Levie, Elvin Isufi, and Gitta Kutyniok. On the transferability of spectral graph filters. In *13th International conference on Sampling Theory and Applications (SampTA)*. IEEE, 2019. doi: 10.1109/SampTA45681.2019.9030932.

Ron Levie, Wei Huang, Lorenzo Bucci, Michael Bronstein, and Gitta Kutyniok. Transferability of spectral graph convolutional neural networks. *Journal of Machine Learning Research*, 22(272): 1–59, 2021.

Renjie Liao, Raquel Urtasun, and Richard Zemel. A {pac}-bayesian approach to generalization bounds for graph neural networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=TR-Nj6nFx42.

L. Lovász. Operations with structures. *Acta Mathematica Academiae Scientiarum Hungarica*, 18(3): 321–328, 1967. doi: 10.1007/BF02280291. URL https://doi.org/10.1007/BF02280291.

László Lovász. *Large networks and graph limits*. Colloquium Publications, Budapest, 2012.

Sohir Maskey, Ron Levie, and Gitta Kutyniok. Transferability of graph neural networks: an extended graphon approach. *arXiv preprint arXiv:2109.10096*, 2021.

Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*, 2019.

Christopher Morris, Martin Ritzert, Matthias Fey, William L. Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4602–4609, Jul. 2019. doi: 10.1609/aaai.v33i01.33014602. URL https://ojs.aaai.org/index.php/AAAI/article/view/4384.

Mathew Penrose. *Random Geometric Graphs*. Oxford Scholarship Online, 2003. ISBN 9780198506263.

Luana Ruiz, Luiz F. O. Chamon, and Alejandro Ribeiro. Transferability properties of graph neural networks, 2021a. URL https://arxiv.org/abs/2112.04629.

Luana Ruiz, Fernando Gama, and Alejandro Ribeiro. Graph neural networks: Architectures, stability, and transferability. *Proceedings of the IEEE*, 109(5):660–682, 2021b. doi: 10.1109/JPROC.2021.3055400.

Franco Scarselli, Ah Chung Tsoi, and Markus Hagenbuchner. The vapnik–chervonenkis dimension of graph and recursive neural networks. *Neural Networks*, 108:248–259, 2018.

Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackermann, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702.e13, 2020. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2020.01.021. URL https://www.sciencedirect.com/science/article/pii/S0092867420301021.

Aad W. Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer New York, NY, 1996.

Saurabh Verma and Zhi-Li Zhang. Stability and generalization of graph convolutional neural networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1539–1548, 2019a.

Saurabh Verma and Zhi-Li Zhang. Stability and generalization of graph convolutional neural networks, 2019b.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Jizhe Wang, Pipei Huang, Huan Zhao, Zhibo Zhang, Binqiang Zhao, and Dik Lun Lee. Billion-scale commodity embedding for e-commerce recommendation in alibaba. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 839–848, New York, NY, USA, 2018a. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3219869. URL https://doi.org/10.1145/3219819.3219869.

Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–67, 2018b.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Gilad Yehudai, Ethan Fetaya, Eli Meirom, Gal Chechik, and Haggai Maron. From local structures to size generalization in graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11975–11986. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/yehudai21a.html.

Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 974–983, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3219890. URL https://doi.org/10.1145/3219819.3219890.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See the conclusion in Subsection 5.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This is theoretical work with no foreseeable societal consequences
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Yes
   (b) Did you include complete proofs of all theoretical results? [Yes] The appendix includes complete proofs.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A] The experiments are very simple which is why we decided to not include the code.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A] We did not perform training
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We report the standard error with respect to the randomness of our experiments.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] We did not perform training, i.e., the experiments were toy examples and not computation heavy.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We refer to [Fey and Lenssen, 2019] for using Pytorch Geometric. We do not use any existing code, data or model.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We do not use real data. Only toy experiments on synthetic data.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We do not use real data. Only toy experiments on synthetic data.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not crowdsource or conduct research with human subjects.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

In Appendix A, we introduce notations that we use throughout the rest of the appendix. In Appendix B, we study the convergence of MPNNs and give the proof of Theorem 3.1. In Appendix C, we analyze generalization properties of MPNNs and prove our main contribution, Theorem 3.3 from Section 3. We give some details on the numerical experiments from Section 4 in Appendix D. For completeness, we recall in Appendix E well-known results that we frequently use.

A Definitions and Notation

We denote metric spaces by (χ, d), where $d : \chi \times \chi \to [0, \infty)$ denotes the metric in the space χ. The ball around $x \in \chi$ of radius $\epsilon > 0$ is defined to be $B_\epsilon(x) = \{y \in \chi \mid d(x, y) < \epsilon\}$. Since, in our analysis, the nodes of the graph are taken as the sample points $X = (X_1, \ldots, X_N)$ in χ, we identify node i of the graph G with the point X_i, for every $i = 1, \ldots, N$. Moreover, since graph signals $\mathbf{f} = (\mathbf{f}_1, \ldots, \mathbf{f}_N)$ represent mappings from nodes in V to feature values, we denote, by abuse of notation, $\mathbf{f}(X_i) := \mathbf{f}_i$ for $i = 1, \ldots, N$.

Definition A.1 ([Vershynin, 2018]). *Let (χ, d) be a compact metric space.*

1. *The ε-covering numbers of χ, denoted by $\mathcal{C}(\chi, \varepsilon, d)$, is the minimal number of balls of radius ε required to cover χ.*
2. *The* Minkowski dimension *of χ is defined to be*

$$\dim(\chi) = \inf\{D \geq 0 \mid \forall \varepsilon \in (0, 1)\, \mathcal{C}(\chi, \varepsilon, d) \leq \varepsilon^{-D}\}.$$

Next, we define various notions of degree.

Definition A.2. *Let $W : \chi \times \chi \to [0, \infty)$ be a kernel , $X = (X_1, \ldots, X_N)$ sample points, and G the corresponding sampled graph.*

1. *We define the* kernel degree *of W at $x \in \chi$ by*

$$\mathrm{d}_W(x) = \int_\chi W(x, y) d\mu(y). \tag{9}$$

2. *Given a point $x \in \chi$ that need not be in X, we define the* graph-kernel degree *of X at x by*

$$\mathrm{d}_X(x) = \frac{1}{N} \sum_{i=1}^{N} W(x, X_i). \tag{10}$$

3. *The* normalized degree *of G at the node $X_c \in X$ is defined by*

$$\mathrm{d}_G(X_c) = \frac{1}{N} \sum_{i=1}^{N} W(X_c, X_i). \tag{11}$$

When $x \notin X$, $d_X(x)$ is interpreted as the degree of the node x in the graph $(x, X_1, \ldots, X_n)$ with edge weights sampled from W.

Based on the different version of degrees in Definition A.2, we define the corresponding three versions of mean aggregation.

Definition A.3. *Given the kernel W, we define the* continuous mean aggregation *of the metric-space message signal $U : \chi \times \chi \to \mathbb{R}^F$ by*

$$M_W U = \int_\chi \frac{W(\cdot, y)}{\mathrm{d}_W(\cdot)} U(\cdot, y) d\mu(y).$$

In Definition A.3, $U(x, y)$ represents a message sent from the point y to the point x in the metric space. Given a metric-space signal $f : \chi \to \mathbb{R}^{F'}$ and a message function Φ, we have

$$M_W \Phi(f, f) = \int_\chi \frac{W(\cdot, y)}{\mathrm{d}_W(\cdot)} \Phi\big(f(\cdot), f(y)\big) d\mu(y).$$

Definition A.4. *Let W be a kernel $X = X_1, \ldots, X_N$ sample points. For a metric-space message signal $U : \chi \times \chi \to \mathbb{R}^F$, we define the* graph-kernel mean aggregation *by*

$$M_X U = \frac{1}{N} \sum_j \frac{W(\cdot, X_j)}{\mathrm{d}_X(\cdot)} U(\cdot, X_j).$$

Note that in the definition of M_X, messages are sent from graph nodes to arbitrary points in the metric space. Hence, $M_X U : \chi \to \mathbb{R}^F$ is a metric-space signal.

Definition A.5. *Let G be a graph with nodes $X = X_1, \ldots, X_N$. For a graph message signal $\mathbf{U} : X \times X \to \mathbb{R}^F$, where $\mathbf{U}(X_i, X_j)$ represents a message sent from the node X_j to the node X_i, we define the* mean aggregation *by*

$$(M_G \mathbf{U})(X_i) = \frac{1}{N} \sum_j \frac{W(X_i, X_j)}{\mathrm{d}_X(X_i)} \mathbf{U}(X_i, X_j).$$

Note that $M_G \mathbf{U} : X \to \mathbb{R}^F$ is a graph signal.

Remark A.6. *Given a graph signal $\mathbf{f} : X \to \mathbb{R}^F$, which can be written as a finite sequence $\mathbf{f} = (\mathbf{f}_i)_i$, and a message function $\Phi : \mathbb{R}^{2F} \to \mathbb{R}^H$, we define*

$$\Phi(\mathbf{f}, \mathbf{f}) := \big(\Phi(\mathbf{f}_i, \mathbf{f}_j)\big)_{i,j=1}^N.$$

Hence, given a graph signal $\mathbf{f} : X \to \mathbb{R}^F$ and the graph messages $\mathbf{U}(X_i, X_j) = \Phi(\mathbf{f}(X_i), \mathbf{f}(X_j))$, we have

$$M_G \mathbf{U} = M_G \Phi(\mathbf{f}, \mathbf{f}) = \frac{1}{N} \sum_j \frac{W(\cdot, X_j)}{\mathrm{d}_X(\cdot)} \Phi\big(\mathbf{f}(\cdot), \mathbf{f}(X_j)\big).$$

Next, we define the different norms used in our analysis.

Definition A.7.

1. *For a vector $\mathbf{z} = (z_1, \ldots, z_F) \in \mathbb{R}^F$, we define as usual*

$$\|\mathbf{z}\|_\infty = \max_{1 \le k \le F} |z_k|.$$

2. *For a function $g : \chi \to \mathbb{R}^F$, we define*

$$\|g\|_\infty = \max_{1 \le k \le F} \sup_{x \in \chi} \big|\big(g(x)\big)_k\big|,$$

3. *Given a graph with N nodes, we define the norm $\|\mathbf{f}\|_{2;\infty}$ of graph feature maps $\mathbf{f} = (\mathbf{f}_1, \ldots, \mathbf{f}_N) \in \mathbb{R}^{N \times F}$, with feature dimension F, as the root mean square over the infinity norms of the node features, i.e.,*

$$\|\mathbf{f}\|_{2;\infty} = \sqrt{\frac{1}{N} \sum_{i=1}^N \|\mathbf{f}_i\|_\infty^2}.$$

Definition A.8. *For a metric-space signal $f : \chi \to \mathbb{R}^F$ and samples $X = (X_1, \ldots, X_N)$ in χ, we define the sampling operator S^X by*

$$S^X f = \big(f(X_i)\big)_{i=1}^N \in \mathbb{R}^{N \times F}.$$

For a metric-space signal $f : \chi \to \mathbb{R}^F$ and a graph signal $\mathbf{f} \in \mathbb{R}^{N \times F}$, we define the distance dist as $\mathrm{dist}(\mathbf{f}, f) = \|\mathbf{f} - S^X f\|_{2;\infty}$., i.e,

$$\mathrm{dist}(f, \mathbf{f}) = \left(\frac{1}{N} \sum_{i=1}^N \|\mathbf{f}_i - (S^X f)_i\|_\infty^2 \right)^{1/2}. \tag{12}$$

Given a MPNN, we define the *formal bias* of the update and message functions by $\|\Psi^{(l)}(0,0)\|_\infty$ and $\|\Phi^{(l)}(0,0)\|_\infty$ respectively. Furthermore, we say that a function $\Phi : \mathbb{R}^F \to \mathbb{R}^H$ is *Lipschitz continuous* if there exists a $L_\Phi > 0$ such that for every $x, x' \in \mathbb{R}^H$, we have

$$\|\Phi(x) - \Phi(x')\|_\infty \leq L_\Phi \|x - x'\|_\infty.$$

Similarly, a function $f : \chi \to \mathbb{R}^F$ is Lipschitz continuous if there exists a $L_f > 0$ such that for every $x, x' \in \chi$, we have

$$\|\Phi(x) - \Phi(x')\|_\infty \leq L_f d(x, x').$$

Next we introduce notations for the mappings between consecutive layers of a MPNN.

Definition A.9. *Let $\Theta = ((\Phi^{(l)})_{l=1}^T, (\Psi^{(l)})_{l=1}^T)$ be a MPNN with T layers and feature dimensions $(F_l)_{l=1}^T$. For $l = 1, \ldots, T$, we define the mapping from the $(l-1)$'th layer to the l'th layer of the gMPNN as*

$$\Lambda_{\Theta_G}^{(l)} : \mathbb{R}^{N \times F_{l-1}} \to \mathbb{R}^{N \times F_l}$$
$$\mathbf{f}^{(l-1)} \mapsto \mathbf{f}^{(l)}.$$

Similarly, we define $\Lambda_{\Theta_W}^{(l)}$ as the mapping from the $(l-1)$'th layer to the l'th layer of the cMPNN $f^{(l-1)} \mapsto f^{(l)}$.

Definition A.9 leads to the following,

$$\Theta_G^{(T)} = \Lambda_{\Theta_G}^{(T)} \circ \Lambda_{\Theta_G}^{(T-1)} \circ \ldots \circ \Lambda_{\Theta_G}^{(1)}$$

and

$$\Theta_W^{(T)} = \Lambda_{\Theta_W}^{(T)} \circ \Lambda_{\Theta_W}^{(T-1)} \circ \ldots \circ \Lambda_{\Theta_W}^{(1)}$$

Lastly, we formulate the following assumption on the space χ, the kernel W, and the MPNN Θ, to which we will refer often in Appendix B.

Assumption A.10. *Let (χ, d) be a metric space and $W : \chi \times \chi \to [0, \infty)$. Let Θ be a MPNN with message and update functions $\Phi^{(l)} : \mathbb{R}^{2F_l} \to \mathbb{R}^{H_l}$ and $\Psi^{(l)} : \mathbb{R}^{F_l + H_l} \to \mathbb{R}^{F_{l+1}}$, $l = 1, \ldots, T-1$.*

1. *The space χ is compact, and there exist $D_\chi, C_\chi \geq 0$ such that $\mathcal{C}(\chi, \varepsilon, d) \leq C_\chi \varepsilon^{-D_\chi}$ for every $\varepsilon > 0$.* [5]
2. *The diameter of χ is bounded by 1. Namely, $\mathrm{diam}(\chi) := \sup_{x,y \in \chi} d(x, y) \leq 1$.*
3. *The kernel satisfies $\|W\|_\infty < \infty$.*
4. *For every $y \in \chi$, the function $W(\cdot, y)$ is Lipschitz continuous (with respect to its first variable) with Lipschitz constant L_W.*
5. *For every $x \in \chi$, the function $W(x, \cdot)$ is Lipschitz continuous (with respect to its second variable) with Lipschitz constant L_W.*
6. *There exists a constant $\mathrm{d}_{\min} > 0$ such that for every $x \in \chi$, we have $d_W(x) \geq \mathrm{d}_{\min}$.*
7. *For every $l = 1, \ldots, T$, the message function $\Phi^{(l)}$ and update function $\Psi^{(l)}$ are Lipschitz continuous with Lipschitz constants $L_{\Phi^{(l)}}$ and $L_{\Psi^{(l)}}$ respectively.*
8. *There exists a constant $\mathrm{W}_{\mathrm{diag}} > 0$ such that for every $x \in \chi$, we have $W(x, x) \geq \mathrm{W}_{\mathrm{diag}} > 0$.*

B Convergence Analysis

In this section we provide the proofs for Theorem 3.1 from Section 3.

[5] The Minkowski dimension $\dim(\chi)$ is a lower bound for all such possible D_χ.

B.1 Preparation

This section is a preparation for the upcoming proof of Theorem 3.1 from Section 3. An important goal of this section is to formulate and prove Lemma B.5, which provides a uniform concentration of measure of the uniform error between the continuous mean aggregation M_W and the graph-kernel mean aggregation M_X. We then show in Corollary B.6 that this uniform bound is preserved by application of an update function. We begin with the following concentration of error lemma which is a slight modification of [Keriven et al., 2020, Lemma 4], and can be derived directly from [Keriven et al., 2020, Lemma 4], by using the assumption $\mathcal{C}(\chi, \varepsilon, d) \leq C_\chi \varepsilon^{-D_\chi}$ instead of $\mathcal{C}(\chi, \varepsilon, d) \leq \varepsilon^{-\dim(\chi)}$.

Lemma B.1 (Lemma 4, [Keriven et al., 2020].). *Let (χ, d, μ) be a metric-measure space and W be a kernel s.t. Assumptions A.10.1-4. are satisfied. Consider a metric-space signal $f : \chi \to \mathbb{R}$ with $\|f\|_\infty < \infty$. Suppose that $X_1, \ldots, X_N$ are drawn i.i.d. from μ on χ and let $p \in (0, 1)$. Then, with probability at least $1 - p$, we have*

$$\left\| \frac{1}{N} \sum_{i=1}^{N} W(\cdot, X_i) f(X_i) - \int_\chi W(\cdot, x) f(x) d\mu(x) \right\|_\infty \leq \frac{\|f\|_\infty \Big(\zeta L_W (\sqrt{\log(C_\chi)} + \sqrt{D_\chi}) + (\sqrt{2}\|W\|_\infty + \zeta L_W) \sqrt{\log 2/p} \Big)}{\sqrt{N}},$$

where

$$\zeta := \frac{2}{\sqrt{2}} e \Big(\frac{2}{\ln(2)} + 1 \Big) \frac{1}{\sqrt{\ln(2)}} C \tag{13}$$

and C is the universal constant from Dudley's inequality (see Theorem 8.1.6 [Vershynin, 2018]).

As a consequence of Lemma B.1, we can derive a sufficient condition on the sample size N which ensures that the graph-kernel degrees are uniformly bounded from below.

Lemma B.2. *Let (χ, d, μ) be a metric-measure space and W be a kernel s.t. Assumptions A.10.1-4. and A.10.6. are satisfied. Suppose that $X_1, \ldots, X_N$ are drawn i.i.d. from μ on χ and let $p \in (0, 1)$. Let*

$$\sqrt{N} \geq 2 \Big(\zeta \frac{L_W}{\mathrm{d}_{\min}} \big(\sqrt{\log(C_\chi)} + \sqrt{D_\chi} \big) + \frac{\sqrt{2}\|W\|_\infty + \zeta L_W}{\mathrm{d}_{\min}} \sqrt{\log 2/p} \Big), \tag{14}$$

where ζ is defined in (13). Then, with probability at least $1 - p$ the following two inequalities hold: For every $x \in \chi$,

$$\mathrm{d}_X(x) \geq \frac{\mathrm{d}_{\min}}{2} \tag{15}$$

and

$$\left\| \frac{1}{N} \sum_{i=1}^{N} W(\cdot, X_i) f(X_i) - \int_\chi W(\cdot, x) f(x) d\mu(x) \right\|_\infty \leq \frac{\|f\|_\infty \Big(\zeta L_W (\sqrt{\log(C_\chi)} + \sqrt{D_\chi}) + (\sqrt{2}\|W\|_\infty + \zeta L_W) \sqrt{\log 2/p} \Big)}{\sqrt{N}}. \tag{16}$$

Proof. By Lemma B.1, with $f = 1$, with probability at least $1 - p$ we have

$$\|\mathrm{d}_X(\cdot) - \mathrm{d}_W(\cdot)\|_\infty \leq \frac{\Big(\zeta L_W \big(\sqrt{\log(C_\chi)} + \sqrt{D_\chi} \big) + \big(\sqrt{2}\|W\|_\infty + \zeta L_W \big) \sqrt{\log 2/p} \Big)}{\sqrt{N}}.$$

By using the lower bound (14) of $\sqrt{N}$, we have $\|\mathrm{d}_X(\cdot) - \mathrm{d}_W(\cdot)\|_\infty \leq \frac{\mathrm{d}_{\min}}{2}$. Let $x \in \chi$. By Assumption A.10.6, we have $|\mathrm{d}_W(x)| \geq \mathrm{d}_{\min}$, hence $|\mathrm{d}_X(x)| \geq \mathrm{d}_{\min}/2$. □

The following lemma is a uniform concentration of measure of the Monte Carlo approximation of Lipschitz functions. Related results about uniform law of large numbers for Lipschitz functions can be found in [Vershynin, 2018, Chapter 8.2]. Our result holds for general metric spaces with finite Minkowski dimension.

Lemma B.3. *Let (χ, d, μ) be a metric-measure space s.t. Assumption A.10.1. is satisfied. Suppose that $X_1, \dots, X_N$ are drawn i.i.d. from μ on χ. For every $p > 0$, there exists an event $\mathcal{E}^p_{\text{Lip}} \subset \chi^N$ regarding the choice of $(X_1, \dots, X_N) \in \chi^N$, with probability $\mu^N(\mathcal{E}^p_{\text{Lip}}) \geq 1 - p$, such that the following uniform bound is satisfied: For every Lipschitz continuous function $F : \chi \to \mathbb{R}^F$ with Lipschitz constant L_F, we have*

$$\left\| \frac{1}{N} \sum_{i=1}^{N} F(X_i) - \int_\chi F(x) d\mu(x) \right\|_\infty$$
$$\leq N^{-\frac{1}{2(D_\chi+1)}} \left(2L_F + \frac{C_\chi}{\sqrt{2}} \|F\|_\infty \sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)} \log(N) + \log(2/p)} \right).$$

For completion, we provide a proof of Lemma B.3.

Proof. Let $r > 0$. By Assumption A.10.1, there exists an open covering $(B_j)_{j \in \mathcal{J}}$ of χ by a family of balls with radius r such that $|\mathcal{J}| \leq C_\chi r^{-D_\chi}$. For $j = 2, \dots, |\mathcal{J}|$, we define $I_j := B_j \setminus \cup_{i<j} B_i$, and define $I_1 = B_1$. Hence, $(I_j)_{j \in \mathcal{J}}$ is a family of measurable sets such that $I_j \cap I_i = \emptyset$ for all $i \neq j \in \mathcal{J}$, $\bigcup_{j \in \mathcal{J}} I_j = \chi$, and $\text{diam}(I_j) \leq 2r$ for all $j \in \mathcal{J}$, where by convention $\text{diam}(\emptyset) = 0$. For each $j \in \mathcal{J}$, let z_j be the center of the ball B_j.

Next, we compute a concentration of error bound on the difference between the measure of I_j and its Monte Carlo approximation, which is uniform in $j \in \mathcal{J}$. Let $j \in \mathcal{J}$ and $q \in (0, 1)$. By Hoeffding's inequality, there is an event $\mathcal{E}^q_j$ with probability $\mu(\mathcal{E}_j) \geq 1 - q$, in which

$$\left\| \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{I_j}(X_i) - \mu(I_k) \right\|_\infty \leq \frac{1}{\sqrt{2}} \frac{\sqrt{\log(2/q)}}{\sqrt{N}}. \tag{17}$$

Consider the event

$$\mathcal{E}^{|\mathcal{J}|q}_{\text{Lip}} = \bigcap_{j=1}^{|\mathcal{J}|} \mathcal{E}^q_j,$$

with probability $\mu^N(\mathcal{E}^{|\mathcal{J}|q}_{\text{Lip}}) \geq 1 - |\mathcal{J}|q$. In this event, (17) holds for all $j \in \mathcal{J}$. We change the failure probability variable $p = |\mathcal{J}|q$, and denote $\mathcal{E}^p_{\text{Lip}} = \mathcal{E}^{|\mathcal{J}|q}_{\text{Lip}}$.

Next we bound uniformly the Monte Carlo approximation error of the integral of bounded Lipschitz continuous functions $F : \chi \to \mathbb{R}^F$. Let $F : \chi \to \mathbb{R}^F$ be a bounded Lipschitz continuous function with Lipschitz constant L_F. We define the step function

$$F^r(y) = \sum_{j \in \mathcal{J}} F(z_j) \mathbb{1}_{I_j}(y).$$

Then,

$$\begin{aligned}
\left\| \frac{1}{N} \sum_{i=1}^{N} F(X_i) - \int_\chi F(y) d\mu(y) \right\|_\infty &\leq \left\| \frac{1}{N} \sum_{i=1}^{N} F(X_i) - \frac{1}{N} \sum_{i=1}^{N} F^r(X_i) \right\|_\infty \\
&+ \left\| \frac{1}{N} \sum_{i=1}^{N} F^r(X_i) - \int_\chi F^r(y) d\mu(y) \right\|_\infty \\
&+ \left\| \int_\chi F^r(y) d\mu(y) - \int_\chi F(y) d\mu(y) \right\|_\infty \\
&=: (1) + (2) + (3).
\end{aligned} \tag{18}$$

To bound (1), we define for each X_i the unique index $j_i \in \mathcal{J}$ s.t. $X_i \in I_{j_i}$. We calculate,

$$\begin{aligned}\left\|\frac{1}{N}\sum_{i=1}^{N} F(X_i) - \frac{1}{N}\sum_{i=1}^{N} F^r(X_i)\right\|_\infty \leq & \frac{1}{N}\sum_{i=1}^{N}\left\|F(X_i) - \sum_{j\in\mathcal{J}} F(z_j)\mathbb{1}_{I_j}(X_i)\right\|_\infty \\ = & \frac{1}{N}\sum_{i=1}^{N}\|F(X_i) - F(z_{j_i})\|_\infty \\ \leq & rL_F.\end{aligned}$$

We proceed by bounding (2). In the event of $\mathcal{E}^p_{\text{Lip}}$, which holds with probability at least $1-p$, equation (17) holds for all $j \in \mathcal{J}$. In this event, we get

$$\begin{aligned}\left\|\frac{1}{N}\sum_{i=1}^{N} F^r(X_i) - \int_\chi F^r(y)d\mu(y)\right\|_\infty &= \left\|\sum_{j\in\mathcal{J}}\left(\frac{1}{N}\sum_{i=1}^{N} F(z_j)\mathbb{1}_{I_j}(X_i) - \int_{I_j} F(z_j)dy\right)\right\|_\infty \\ &\leq \sum_{j\in\mathcal{J}}\|F\|_\infty\left|\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{I_j}(X_i) - \mu(I_j)\right| \\ &\leq |\mathcal{J}|\|F\|_\infty\frac{1}{\sqrt{2}}\frac{\sqrt{\log(2|\mathcal{J}|/p)}}{\sqrt{N}}.\end{aligned}$$

Recall that $|\mathcal{J}| \leq C_\chi r^{-D_\chi}$. Then, with probability at least $1-p$

$$\begin{aligned}&\left\|\frac{1}{N}\sum_{i=1}^{N} F^r(X_i) - \int_\chi F^r(y)d\mu(y)\right\|_\infty \\ &\leq C_\chi r^{-D_\chi}\|F\|_\infty\frac{1}{\sqrt{2}}\frac{\sqrt{\log(C_\chi) - D_\chi\log(r) + \log(2/p)}}{\sqrt{N}}.\end{aligned}$$

To bound (3), we calculate

$$\begin{aligned}\left\|\int_\chi F^r(y)d\mu(y) - \int_\chi F(y)d\mu(y)\right\|_\infty &= \left\|\int_\chi\sum_{j\in\mathcal{J}} F(z_j)\mathbb{1}_{I_j}d\mu(y) - \int_\chi F(y)d\mu(y)\right\|_\infty \\ &\leq \sum_{j\in\mathcal{J}}\int_{I_j}\|F(z_j) - F(y)\|_\infty d\mu(y) \\ &\leq rL_F.\end{aligned}$$

By plugging the bounds of (1), (2) and (3) into (18), we get

$$\left\|\frac{1}{N}\sum_{i=1}^{N} F(X_i) - \int_\chi F(y)d\mu(y)\right\|_\infty \leq 2rL_F + C_\chi r^{-D_\chi}\|F\|_\infty\frac{1}{\sqrt{2}}\frac{\sqrt{\log(C_\chi) - D_\chi\log(r) + \log(2/p)}}{\sqrt{N}}.$$

Lastly, choosing $r = N^{-\frac{1}{2(D_\chi+1)}}$ gives us an overall error of

$$\begin{aligned}&\left\|\frac{1}{N}\sum_{i=1}^{N} F(X_i) - \int_\chi F(y)d\mu(y)\right\|_\infty \\ &\leq N^{-\frac{1}{2(D_\chi+1)}}\left(2L_F + C_\chi\|F\|_\infty\frac{1}{\sqrt{2}}\sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)}\log(N) + \log(2/p)}\right)\end{aligned}$$

Since the event $\mathcal{E}^p_{\text{Lip}}$ is independent of the choice of $F : \chi \to \mathbb{R}^F$, the proof is finished. □

The next lemma is based on Lemma B.3, and provides a uniform concentration of measure on the L^∞-error between a non-normalized version of the kernel mean aggregation from Definition A.3 and a non-normalized version of the graph-kernel mean aggregation from Definition A.4.

Lemma B.4. *Let (χ, d, μ) be a metric-measure space and W be a kernel s.t. Assumptions A.10.1-3 and A.10.5. are satisfied. Let $p \in (0,1)$. Suppose that $X_1, \dots, X_N$ are drawn i.i.d. from μ on χ such that $(X_1, \dots, X_N) \in \mathcal{E}^p_{\text{Lip}}$, where the event $\mathcal{E}^p_{\text{Lip}}$ is defined in Lemma B.3. Then, for every $x \in \chi$, $f : \chi \to \mathbb{R}^F$ with Lipschitz constant L_f, and $\Phi : \mathbb{R}^{2F} \to \mathbb{R}^H$ with Lipschitz constant L_Φ, we have*

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{i=1}^N W(x, X_i)\Phi\big(f(x), f(X_i)\big) - \int_\chi W(x,y)\Phi\big(f(x), f(y)\big) d\mu(y) \right\|_\infty \\
&\leq N^{-\frac{1}{2(D_\chi+1)}} \Bigg(2\Big(\|W\|_\infty L_\Phi L_f + L_W \big(L_\Phi \|f\|_\infty + \|\Phi(0,0)\|_\infty\big)\Big) \\
&+ C_\chi \Big(\|W\|_\infty \big(L_\Phi \|f\|_\infty + \|\Phi(0,0)\|_\infty\big)\Big) \frac{1}{\sqrt{2}} \sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)} \log(N) + \log(2/p)} \Bigg).
\end{aligned} \tag{19}$$

Proof. For any $x \in \chi$, $f : \chi \to \mathbb{R}^F$ and $\Phi : \mathbb{R}^{2F} \to \mathbb{R}^H$, we define the random variable

$$Y_{x;\Phi} = \frac{1}{N} \sum_{i=1}^N W(x, X_i)\Phi\big(f(x), f(X_i)\big) - \int_\chi W(x,y)\Phi\big(f(x), f(y)\big) d\mu(y)$$

on the sample space χ^N. Applying Lemma B.3 on the integrand $F_x(y) := W(x,y)\Phi\big(f(x), f(y)\big)$, uniformly on the choice of the parameter $x \in \chi$, yields in the event $\mathcal{E}^p_{\text{Lip}}$:

$$\|Y_{x;\Phi}\|_\infty \leq N^{-\frac{1}{2(D_\chi+1)}} \left(2L_{F_x} + C_\chi \|F_x\|_\infty \frac{1}{\sqrt{2}} \sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)} \log(N) + \log(2/p)} \right). \tag{20}$$

So it remains to calculate the Lipschitz constant and the infinity-norm of F_x. For this, calculate for $y, y' \in \chi$

$$\begin{aligned}
\|F_x(y) - F_x(y')\|_\infty =& \|W(x,y)\Phi\big(f(x), f(y)\big) - W(x,y')\Phi\big(f(x), f(y')\big)\|_\infty \\
\leq& \|W(x,y)\Phi\big(f(x), f(y)\big) - W(x,y)\Phi\big(f(x), f(y')\big)\|_\infty \\
+& \|W(x,y)\Phi\big(f(x), f(y')\big) - W(x,y')\Phi\big(f(x), f(y')\big)\|_\infty \\
\leq& \big(\|W\|_\infty L_\Phi L_f + L_W (L_\Phi \|f\|_\infty + \|\Phi(0,0)\|_\infty)\big) d(y, y')
\end{aligned}$$

and

$$\begin{aligned}
\|F_x(\cdot)\|_\infty &= \|W(x,\cdot)\Phi\big(f(x), f(\cdot)\big)\|_\infty \\
&\leq \|W\|_\infty (L_\Phi \|f\|_\infty + \|\Phi(0,0)\|_\infty).
\end{aligned}$$

□

The next lemma provides a uniform concentration of measure bound on the error between the graph-kernel mean aggregation M_X and the continuous mean aggregation M_W .

Lemma B.5. *Let (χ, d, μ) be a metric-measure space and W be a kernel s.t. Assumptions A.10.1-6. are satisfied. Let $N \in \mathbb{N}$ satisfy (14). Let $\mathcal{E}^p_{\text{Lip}}$ be the event defined in Lemma B.3. There exists an event $\mathcal{F}^p_{\text{Lip}} \subset \mathcal{E}^p_{\text{Lip}}$ regarding the choice of i.i.d $X_1, \dots, X_N$ from μ in χ, with probability $\mu(\mathcal{F}^p_{\text{Lip}}) \geq 1 - 2p$, such that condition (15) together with (21) below are satisfied: for every $f : \chi \to \mathbb{R}^F$ with Lipschitz constant L_f and $\Phi : \mathbb{R}^{2F} \to \mathbb{R}^H$ with Lipschitz constant L_Φ*

$$\begin{aligned}
&\|(M_X - M_W)\big(\Phi(f,f)\big)\|_\infty \leq 4 \frac{\varepsilon_1}{\sqrt{N} \mathrm{d}^2_{min}} \|W\|_\infty (L_\Phi \|f\|_\infty + \|\Phi(0,0)\|_\infty) \\
&+ N^{-\frac{1}{2(D_\chi+1)}} \Bigg(2\Big(\frac{\|W\|_\infty}{\mathrm{d}_{min}} L_\Phi L_f + \frac{L_W}{\mathrm{d}_{min}} \big(L_\Phi \|f\|_\infty + \|\Phi(0,0)\|_\infty\big)\Big) \\
&+ C_\chi \Big(\frac{\|W\|_\infty}{\mathrm{d}_{min}} \big(L_\Phi \|f\|_\infty + \|\Phi(0,0)\|_\infty\big)\Big) \frac{1}{\sqrt{2}} \sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)} \log(N) + \log(2/p)} \Bigg),
\end{aligned} \tag{21}$$

where

$$\varepsilon_1 = L_W\big(\sqrt{\log(C_\chi)} + \sqrt{D_\chi}\big) + \big(\sqrt{2}\|W\|_\infty + L_W\big)\sqrt{\log 2/p}. \tag{22}$$

Proof. By Lemma B.2, we have with probability at least $1-p$

$$\begin{aligned}\|\mathrm{d}_X - \mathrm{d}_W\|_\infty &\le \frac{\varepsilon_1}{\sqrt{N}} = \zeta\frac{L_W\big(\sqrt{\log(C_\chi)} + \sqrt{D_\chi}\big) + \big(\sqrt{2}\|W\|_\infty + L_W\big)\sqrt{\log 2/p}}{\sqrt{N}} \\ &\le \frac{\mathrm{d}_{\min}}{2},\end{aligned} \tag{23}$$

where the second inequality follows from (14). Furthermore, in the same event we have

$$|\mathrm{d}_X(x)|_\infty \ge \frac{\mathrm{d}_{\min}}{2}$$

for all $x \in \chi$. Moreover, $|\mathrm{d}_W(x)|_\infty \ge \mathrm{d}_{\min}$ by Assumption A.10.6. Hence, for all $x \in \chi$, we have

$$\begin{aligned}\left|\frac{1}{\mathrm{d}_X(x)} - \frac{1}{\mathrm{d}_W(x)}\right| &= \frac{|\mathrm{d}_W(x) - \mathrm{d}_X(x)|}{|\mathrm{d}_X(x)\mathrm{d}_W(x)|} \\ &\le 4\frac{\varepsilon_1}{\sqrt{N}\mathrm{d}_{\min}^2}.\end{aligned} \tag{24}$$

Denote that intersection of $\mathcal{E}^p_{\mathrm{Lip}}$ and the event in which (23) occur by $\mathcal{F}^p_{\mathrm{Lip}}$. Let $(X_1, \dots, X_N)$ be i.i.d samples in $\mathcal{F}^p_{\mathrm{Lip}}$. Define $\tilde{W}(x,y) = \frac{W(x,y)}{\mathrm{d}_W(x)}$. Next we apply Lemma B.4 on the kernel $\tilde{W}$. For this, note that for $x \in \chi$ the kernel $\tilde{W}(x,\cdot)$ is Lipschitz continuous (with respect to the second variable) with Lipschitz constant $L_{\tilde{W}} = \frac{L_W}{\mathrm{d}_{\min}}$, since for $y, y' \in \chi$, we have

$$\left|\frac{W(x,y)}{\mathrm{d}_W(x)} - \frac{W(x,y')}{\mathrm{d}_W(x)}\right| \le \frac{L_W}{\mathrm{d}_{min}} d(y,y').$$

Moreover, for all $y \in \chi$ we have $\|\tilde{W}(\cdot,y)\|_\infty \le \frac{\|W\|_\infty}{\mathrm{d}_{\min}}$.

Then, we use Lemma B.4 to obtain, for every $f : \chi \to \mathbb{R}^F$ and $\Phi : \mathbb{R}^{2F} \to \mathbb{R}^H$ as specified in the lemma,

$$\begin{aligned}&\left\|\frac{1}{N}\sum_{i=1}^N \tilde{W}(\cdot,X_i)\Phi\big(f(\cdot),f(X_i)\big) - \int_\chi \tilde{W}(\cdot,y)\Phi\big(f(\cdot),f(y)\big)d\mu(y)\right\|_\infty \\ &\le N^{-\frac{1}{2(D_\chi+1)}}\Bigg(2\Big(\|\tilde{W}\|_\infty L_\Phi L_f + L_{\tilde{W}}\big(L_\Phi\|f\|_\infty + \|\Phi(0,0)\|_\infty\big)\Big) \\ &+ C_\chi\Big(\|\tilde{W}\|_\infty\big(L_\Phi\|f\|_\infty + \|\Phi(0,0)\|_\infty\big)\Big)\frac{1}{\sqrt{2}}\sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)}\log(N) + \log(2/p)}\Bigg) \\ &\le N^{-\frac{1}{2(D_\chi+1)}}\Bigg(2\Big(\frac{\|W\|_\infty}{\mathrm{d}_{\min}}L_\Phi L_f + \frac{L_W}{\mathrm{d}_{\min}}\big(L_\Phi\|f\|_\infty + \|\Phi(0,0)\|_\infty\big)\Big) \\ &+ C_\chi\Big(\frac{\|W\|_\infty}{\mathrm{d}_{\min}}\big(L_\Phi\|f\|_\infty + \|\Phi(0,0)\|_\infty\big)\Big)\frac{1}{\sqrt{2}}\sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)}\log(N) + \log(2/p)}\Bigg).\end{aligned} \tag{25}$$

Then, by (24) and (25), for every $f : \chi \to \mathbb{R}^F$ and $\Phi : \mathbb{R}^{2F} \to \mathbb{R}^H$ as specified in the lemma,

$$
\begin{aligned}
&\|(M_X - M_W)\Phi(f,f)\|_\infty \\
&= \left\| \frac{1}{N}\sum_{i=1}^N \frac{W(\cdot, X_i)}{\mathrm{d}_X(\cdot)} \Phi\big(f(\cdot), f(X_i)\big) - \int_\chi \frac{W(\cdot,x)}{\mathrm{d}_W(\cdot)} \Phi\big(f(\cdot), f(x)\big) d\mu(x) \right\|_\infty \\
&\le \frac{1}{N}\sum_{i=1}^N \big\| W(x, X_i)\Phi\big(f(\cdot), f(X_i)\big)\big\|_\infty \left\| \frac{1}{\mathrm{d}_X(\cdot)} - \frac{1}{\mathrm{d}_W(\cdot)} \right\|_\infty \\
&+ \left\| \frac{1}{N}\sum_{i=1}^N \tilde{W}(\cdot, X_i)\Phi\big(f(\cdot), f(X_i)\big) - \int_\chi \tilde{W}(\cdot,x)\Phi\big(f(\cdot), f(x)\big) d\mu(x) \right\|_\infty \\
&\le 4\frac{\varepsilon_1}{\sqrt{N}\mathrm{d}_{min}^2} \|W\|_\infty (L_\Phi \|f\|_\infty + \|\Phi(0,0)\|_\infty) \\
&+ N^{-\frac{1}{2(D_\chi+1)}} \Bigg(2\Big(\frac{\|W\|_\infty}{\mathrm{d}_{min}} L_\Phi L_f + \frac{L_W}{\mathrm{d}_{min}} \big(L_\Phi \|f\|_\infty + \|\Phi(0,0)\|_\infty\big)\Big) \\
&+ C_\chi \Big(\frac{\|W\|_\infty}{\mathrm{d}_{min}} \big(L_\Phi \|f\|_\infty + \|\Phi(0,0)\|_\infty\big)\Big) \frac{1}{\sqrt{2}} \sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)}\log(N) + \log(2/p)} \Bigg).
\end{aligned}
$$

□

The next corollary shows that Lemma B.5 is preserved by the application of an update function.

Corollary B.6. *Let (χ, d, μ) be a metric-measure space and W be a kernel s.t. Assumptions A.10.1-6. are satisfied. Let $p > 0$ and $N \in \mathbb{N}$ satisfy (14). Suppose that $X_1, \dots, X_N$ are drawn i.i.d. from μ on χ. If the event $\mathcal{F}^p_{\mathrm{Lip}}$ from Lemma B.5 occurs, then condition (15) together with (26) below are satisfied: for every $f : \chi \to \mathbb{R}^F$ with Lipschitz constant L_f, $\Phi : \mathbb{R}^{2F} \to \mathbb{R}^H$ with Lipschitz constant L_Φ and $\Psi : \mathbb{R}^{F+H} \to \mathbb{R}^{F'}$ with Lipschitz constant L_Ψ*

$$
\begin{aligned}
&\Big\| \Psi\Big(f(\cdot), M_X\big(\Phi(f,f)\big)(\cdot)\Big) - \Psi\Big(f(\cdot), M_W\big(\Phi(f,f)\big)(\cdot)\Big) \Big\|_\infty \\
&\le L_\Psi \Bigg(4\frac{\varepsilon_1}{\sqrt{N}\mathrm{d}_{min}^2} \|W\|_\infty (L_\Phi \|f\|_\infty + \|\Phi(0,0)\|_\infty) \\
&+ N^{-\frac{1}{2(D_\chi+1)}} \Bigg(2\Big(\frac{\|W\|_\infty}{\mathrm{d}_{min}} L_\Phi L_f + \frac{L_W}{\mathrm{d}_{min}} \big(L_\Phi \|f\|_\infty + \|\Phi(0,0)\|_\infty\big)\Big) \\
&+ \frac{C_\chi}{\sqrt{2}} \Big(\frac{\|W\|_\infty}{\mathrm{d}_{min}} \big(L_\Phi \|f\|_\infty + \|\Phi(0,0)\|_\infty\big)\Big) \sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)}\log(N) + \log(2/p)} \Bigg)\Bigg),
\end{aligned}
\tag{26}
$$

where ε_1 is defined in (22).

Proof. We calculate,

$$
\begin{aligned}
&\Big\| \Psi\Big(f(\cdot), M_X\big(\Phi(f,f)\big)(\cdot)\Big) - \Psi\Big(f(\cdot), M_W\big(\Phi(f,f)\big)(\cdot)\Big) \Big\|_\infty \\
\le& L_\Psi \big\| M_X\big(\Phi(f,f)\big)(\cdot) - M_W\big(\Phi(f,f)\big)(\cdot) \big\|_\infty ,
\end{aligned}
$$

and apply Lemma B.5 to the right-hand-side. □

We continue by providing three lemmas which capture deterministic properties of cMPNNs and gMPNNs. We start by showing that the infinity norm of the output of the l-th layer of a cMPNN $f^{(l)}$ can be bounded in terms of the infinity norm of its input f.

Lemma B.7. *Let (χ, d, μ) be a metric-measure space, W be a kernel and $\Theta = ((\Phi^{(l)})_{l=1}^T, (\Psi^{(l)})_{l=1}^T)$ be a MPNN s.t. Assumptions A.10.1-7. are satisfied. Consider a metric-space signal $f : \chi \to \mathbb{R}^F$ with $\|f\|_\infty < \infty$. Then, for $l = 0, \dots, T-1$, the cMPNN output $f^{(l+1)}$ satisfies*

$$
\|f^{(l+1)}\|_\infty \le B_1^{(l+1)} + \|f\|_\infty B_2^{(l+1)},
$$

where

$$B_1^{(l+1)} = \sum_{k=1}^{l+1} \left(L_{\Psi^{(k)}} \|\Phi^{(k)}(0,0)\|_\infty + \|\Psi^{(k)}(0,0)\|_\infty \right) \prod_{l'=k+1}^{l+1} L_{\Psi^{(l')}} \left(1 + L_{\Phi^{(l')}}\right) \tag{27}$$

and

$$B_2^{(l+1)} = \prod_{k=1}^{l+1} L_{\Psi^{(k)}} \left(1 + L_{\Phi^{(k)}}\right). \tag{28}$$

Proof. Let $l = 0, \dots, T-1$. Then, for $k = 0, \dots, l$, we have

$$\begin{aligned}
\|f^{(k+1)}(\cdot)\|_\infty &= \left\| \Psi^{(k+1)}\Big(f^{(k)}(\cdot), M_W\big(\Phi^{(k+1)}(f^{(k)}, f^{(k)})\big)(\cdot)\Big) \right\|_\infty \\
&\leq \left\| \Psi^{(k+1)}\Big(f^{(k)}(\cdot), M_W\big(\Phi^{(k+1)}(f^{(k)}, f^{(k)})\big)(\cdot)\Big) - \Psi^{(k+1)}(0,0) \right\|_\infty + \|\Psi^{(k+1)}(0,0)\|_\infty \\
&\leq L_{\Psi^{(k+1)}} \left(\|f^{(k)}\|_\infty + \left\| M_W\big(\Phi^{(k+1)}(f^{(k)}, f^{(k)})\big)(\cdot) \right\|_\infty \right) + \|\Psi^{(k+1)}(0,0)\|_\infty.
\end{aligned}$$

For the message term, we have

$$\begin{aligned}
\left\| M_W\big(\Phi^{(k+1)}(f^{(k)}, f^{(k)})\big)(\cdot) \right\|_\infty &= \left\| \int_\chi \frac{W(\cdot, y)}{\mathrm{d}_W(\cdot)} \Phi^{(k+1)}\big(f^{(k)}(\cdot), f^{(k)}(y)\big) d\mu(y) \right\|_\infty \\
&\leq \left\| \int_\chi \left| \frac{W(\cdot, y)}{d_W(\cdot)} \right| d\mu(y) \cdot \sup_{y \in \chi} \left| \Phi^{(k+1)}(f^{(k)}(\cdot), f^{(k)}(y)) \right| \right\|_\infty \\
&\leq L_{\Phi^{(k+1)}} \|f^{(k)}\|_\infty + \|\Phi^{(k+1)}(0,0)\|_\infty.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\|f^{(k+1)}(\cdot)\|_\infty \\
&\leq L_{\Psi^{(k+1)}} \left(\|f^{(k)}\|_\infty + (L_{\Phi^{(k+1)}} \|f^{(k)}\|_\infty + \|\Phi^{(k+1)}(0,0)\|_\infty) \right) + \|\Psi^{(k+1)}(0,0)\|_\infty,
\end{aligned}$$

which we can reorder to

$$\begin{aligned}
&\|f^{(k+1)}(\cdot)\|_\infty \\
&\leq L_{\Psi^{(k+1)}} \left(1 + L_{\Phi^{(k+1)}}\right) \|f^{(k)}\|_\infty + L_{\Psi^{(k+1)}} \|\Phi^{(k+1)}(0,0)\|_\infty + \|\Psi^{(k+1)}(0,0)\|_\infty.
\end{aligned}$$

We apply Lemma B.11 to solve this recurrence relation which finishes the proof. □

In the following, we denote by $L_{f^{(l)}}$ the Lipschitz constant of $f^{(l)}$. The next lemma bounds $L_{f^{(l+1)}}$ in terms of L_f.

Lemma B.8. *Let (χ, d, μ) be a metric-measure space, W be a kernel and $\Theta = ((\Phi^{(l)})_{l=1}^T, (\Psi^{(l)})_{l=1}^T)$ be a MPNN s.t. Assumptions A.10.1-7. are satisfied. Consider a Lipschitz continuous metric-space signal $f : \chi \to \mathbb{R}^F$ with $\|f\|_\infty < \infty$ and Lipschitz constant L_f. Then, for $l = 0, \dots, T-1$, the cMPNN output $f^{(l+1)}$ is Lipschitz continuous with Lipschitz constant $L_{f^{(l+1)}}$ satisfying*

$$\begin{aligned}
L_{f^{(l+1)}} \leq \sum_{k=1}^{l+1} \Bigg(&\Big(L_{\Psi^{(k)}} \frac{L_W}{\mathrm{d}_{\min}} (\|\Phi^{(k)}(0,0)\|_\infty + L_{\Phi^{(k)}} \|f^{(k-1)}\|_\infty) + L_{\Psi^{(k)}} \|W\|_\infty (\|\Phi^{(k)}(0,0)\|_\infty \\
&+ L_{\Phi^{(k)}} \|f^{(k-1)}\|_\infty) \frac{L_W}{\mathrm{d}_{\min}^2} \Big) \prod_{l'=k+1}^{l+1} L_{\Psi^{(l')}} \Big(1 + \frac{\|W\|_\infty}{\mathrm{d}_{\min}} L_{\Phi^{(l')}}\Big) \Bigg) \\
&+ L_f \prod_{k=1}^{l+1} L_{\Psi^{(k)}} \Big(1 + \frac{\|W\|_\infty}{\mathrm{d}_{\min}} L_{\Phi^{(k)}}\Big).
\end{aligned}$$

Proof. Let $l = 0, \dots, T-1$ and consider $k = 0, \dots, l$. For $x, x' \in \chi$, we have

$$
\begin{aligned}
&\|f^{(k+1)}(x) - f^{(k+1)}(x')\|_\infty \\
&= \Big\| \Psi^{(k+1)}\Big(f^{(k)}(x), M_W\big(\Phi^{(k+1)}(f^{(k)}, f^{(k)})\big)(x)\Big) \\
&- \Psi^{(k+1)}\Big(f^{(k)}(x'), M_W\big(\Phi^{(k+1)}(f^{(k)}, f^{(k)})\big)(x')\Big)\Big\|_\infty \\
&\le L_{\Psi^{(k+1)}}\Big(\Big\|f^{(k)}(x) - f^{(k)}(x')\Big\|_\infty \\
&+ \Big\|M_W\big(\Phi^{(k+1)}(f^{(k)}, f^{(k)})\big)(x) - M_W\big(\Phi^{(k+1)}(f^{(k)}, f^{(k)})\big)(x')\Big\|_\infty\Big) \\
&\le L_{\Psi^{(k+1)}}\Big(L_{f^{(k)}} d(x, x') + \|M_W\big(\Phi^{(k+1)}(f^{(k)}, f^{(k)})\big)(x) - M_W\big(\Phi^{(k+1)}(f^{(k)}, f^{(k)})\big)(x')\|_\infty\Big).
\end{aligned} \tag{29}
$$

For the second term, we have

$$
\begin{aligned}
&\|M_W\big(\Phi^{(k+1)}(f^{(k)}, f^{(k)})\big)(x) - M_W\big(\Phi^{(k+1)}(f^{(k)}, f^{(k)})\big)(x')\|_\infty \\
&= \| \int_\chi \frac{W(x,y)}{\mathrm{d}_W(x)} \Phi^{(k+1)}\big(f^{(k)}(x), f^{(k)}(y)\big) - \frac{W(x',y)}{\mathrm{d}_W(x')} \Phi^{(k+1)}\big(f^{(k)}(x'), f^{(k)}(y)\big) d\mu(y)\|_\infty \\
&\le \int_\chi \Big\| \frac{W(x,y)}{\mathrm{d}_W(x)} \Phi^{(k+1)}\big(f^{(k)}(x), f^{(k)}(y)\big) - \frac{W(x',y)}{\mathrm{d}_W(x)} \Phi^{(k+1)}\big(f^{(k)}(x), f^{(k)}(y)\big)\Big\|_\infty d\mu(y) \\
&+ \int_\chi \Big\| \frac{W(x',y)}{\mathrm{d}_W(x)} \Phi^{(k+1)}\big(f^{(k)}(x), f^{(k)}(y)\big) - \frac{W(x',y)}{\mathrm{d}_W(x)} \Phi^{(k+1)}\big(f^{(k)}(x'), f^{(k)}(y)\big)\Big\|_\infty d\mu(y) \\
&+ \int_\chi \Big\| \frac{W(x',y)}{\mathrm{d}_W(x)} \Phi^{(k+1)}\big(f^{(k)}(x'), f^{(k)}(y)\big) - \frac{W(x',y)}{\mathrm{d}_W(x')} \Phi^{(k+1)}\big(f^{(k)}(x'), f^{(k)}(y)\big)\Big\|_\infty d\mu(y) \\
&= (A) + (B) + (C)
\end{aligned} \tag{30}
$$

For (A), we have

$$
\begin{aligned}
(A) &= \int_\chi \Big\| \frac{W(x,y)}{\mathrm{d}_W(x)} \Phi^{(k+1)}\big(f^{(k)}(x), f^{(k)}(y)\big) - \frac{W(x',y)}{\mathrm{d}_W(x)} \Phi^{(k+1)}\big(f^{(k)}(x), f^{(k)}(y)\big)\Big\|_\infty d\mu(y) \\
&= \int_\chi \frac{|W(x,y) - W(x',y)|}{\mathrm{d}_W(x)} \Big\|\Phi^{(k+1)}\big(f^{(k)}(x), f^{(k)}(y)\big)\Big\|_\infty d\mu(y) \\
&\le L_W \frac{d(x,x')}{\mathrm{d}_{\min}} \int_\chi \Big\|\Phi^{(k+1)}\big(f^{(k)}(x), f^{(k)}(y)\big)\Big\|_\infty d\mu(y) \\
&\le \frac{L_W}{\mathrm{d}_{\min}} \big(\|\Phi^{(k+1)}(0,0)\|_\infty + L_{\Phi^{(k+1)}} \|f^{(k)}\|_\infty\big) d(x,x').
\end{aligned}
$$

For (B), we have

$$
\begin{aligned}
(B) &= \int_\chi \Big\| \frac{W(x',y)}{\mathrm{d}_W(x)} \Phi^{(k+1)}\big(f^{(k)}(x), f^{(k)}(y)\big) - \frac{W(x',y)}{\mathrm{d}_W(x)} \Phi^{(k+1)}\big(f^{(k)}(x'), f^{(k)}(y)\big)\Big\|_\infty d\mu(y) \\
&= \int_\chi \frac{|W(x',y)|}{|\mathrm{d}_W(x)|} \Big\|\Phi^{(k+1)}\big(f^{(k)}(x), f^{(k)}(y)\big) - \Phi^{(k+1)}\big(f^{(k)}(x'), f^{(k)}(y)\big)\Big\|_\infty d\mu(y) \\
&\le \frac{\|W\|_\infty}{\mathrm{d}_{\min}} L_{\Phi^{(k+1)}} \int_\chi \big\|\big(f^{(k)}(x), f^{(k)}(y)\big) - \big(f^{(k)}(x'), f^{(k)}(y)\big)\big\|_\infty d\mu(y) \\
&\le \frac{\|W\|_\infty}{\mathrm{d}_{\min}} L_{\Phi^{(k+1)}} \|f^{(k)}(x) - f^{(k)}(x')\big)\|_\infty \\
&\le \frac{\|W\|_\infty}{\mathrm{d}_{\min}} L_{\Phi^{(k+1)}} L_{f^{(k)}} d(x,x').
\end{aligned}
$$

For (C), we have

$$
\begin{aligned}
(C) &= \int_{\chi} \Big\| \frac{W(x',y)}{\mathrm{d}_W(x)} \Phi^{(k+1)}\big(f^{(k)}(x'), f^{(k)}(y)\big) - \frac{W(x',y)}{\mathrm{d}_W(x')} \Phi^{(k+1)}\big(f^{(k)}(x'), f^{(k)}(y)\big) \Big\|_{\infty} d\mu(y) \\
&= \int_{\chi} |W(x',y)| \Big| \frac{1}{\mathrm{d}_W(x)} - \frac{1}{\mathrm{d}_W(x')} \Big| \Big\| \Phi^{(k+1)}\big(f^{(k)}(x'), f^{(k)}(y)\big) \Big\|_{\infty} d\mu(y) \\
&\leq \|W\|_{\infty} (\|\Phi^{(k+1)}(0,0)\|_{\infty} + L_{\Phi^{(k+1)}} \|f^{(k)}\|_{\infty}) \frac{L_W}{\mathrm{d}_{\min}^2} d(x,x'),
\end{aligned}
$$

where the last inequality holds since

$$
\begin{aligned}
\left| \frac{1}{\mathrm{d}_W(x)} - \frac{1}{\mathrm{d}_W(x')} \right| &\leq \frac{|\mathrm{d}_W(x') - \mathrm{d}_W(x)|}{|\mathrm{d}_W(x)\mathrm{d}_W(x')|} \\
&\leq \frac{1}{\mathrm{d}_{\min}^2} |\mathrm{d}_W(x') - \mathrm{d}_W(x)| \\
&\leq \frac{1}{\mathrm{d}_{\min}^2} \int_{\chi} |W(x',y) - W(x,y)| d\mu(y) \\
&\leq \frac{1}{\mathrm{d}_{\min}^2} \int_{\chi} L_W d(x,x') d\mu(y) \\
&\leq \frac{L_W}{\mathrm{d}_{\min}^2} d(x,x').
\end{aligned}
$$

Hence, by plugging (30) and our bounds for $(A), (B)$ and (C) into (29), we have

$$
\begin{aligned}
&\|f^{(k+1)}(x) - f^{(k+1)}(x')\|_{\infty} \\
&\leq L_{\Psi^{(k+1)}} \Big(L_{f^{(k)}} d(x,x') + \|M_W\big(\Phi^{(k+1)}(f^{(k)}, f^{(k)})\big)(x) - M_W\big(\Phi^{(k+1)}(f^{(k)}, f^{(k)})\big)(x')\|_{\infty} \Big) \\
&\leq L_{\Psi^{(k+1)}} \Big(L_{f^{(k)}} d(x,x') + (A) + (B) + (C) \Big) \\
&\leq L_{\Psi^{(k+1)}} \Big(L_{f^{(k)}} + \frac{L_W}{\mathrm{d}_{\min}} \big(\|\Phi^{(k+1)}(0,0)\|_{\infty} + L_{\Phi^{(k+1)}} \|f^{(k)}\|_{\infty}\big) \\
&+ \frac{\|W\|_{\infty}}{\mathrm{d}_{\min}} L_{\Phi^{(k+1)}} L_{f^{(k)}} + \|W\|_{\infty} (\|\Phi^{(k+1)}(0,0)\|_{\infty} + L_{\Phi^{(k+1)}} \|f^{(k)}\|_{\infty}) \frac{L_W}{\mathrm{d}_{\min}^2} \Big) d(x,x').
\end{aligned}
$$

Hence,

$$
\begin{aligned}
L_{f^{(k+1)}} \leq\ & L_{\Psi^{(k+1)}} \frac{L_W}{\mathrm{d}_{\min}} (\|\Phi^{(k+1)}(0,0)\|_{\infty} + L_{\Phi^{(k+1)}} \|f^{(k)}\|_{\infty}) + L_{\Psi^{(k+1)}} \left(1 + \frac{\|W\|_{\infty}}{\mathrm{d}_{\min}} L_{\Phi^{(k+1)}} \right) L_{f^{(k)}} \\
&+ L_{\Psi^{(k+1)}} \|W\|_{\infty} (\|\Phi^{(k+1)}(0,0)\|_{\infty} + L_{\Phi^{(k+1)}} \|f^{(k)}\|_{\infty}) \frac{L_W}{\mathrm{d}_{\min}^2}.
\end{aligned}
$$

We finish the proof by solving the recurrence relation with Lemma B.11. □

Corollary B.9. *Consider the same setting as in Lemma B.8. Then, for $l = 0, \ldots, T-1$,*

$$
L_{f^{(l)}} \leq Z_1^{(l)} + Z_2^{(l)} \|f\|_{\infty} + Z_3^{(l)} L_f,
$$

where $Z_1^{(l)}$, $Z_2^{(l)}$ and $Z_3^{(l)}$ are independent of f and defined as

$$
\begin{aligned}
Z_1^{(l)} &= \sum_{k=1}^{l} \Bigg(\Big(L_{\Psi^{(k)}} \frac{L_W}{\mathrm{d}_{\min}} \|\Phi^{(k)}(0,0)\|_{\infty} + L_{\Psi^{(k)}} \|W\|_{\infty} \|\Phi^{(k)}(0,0)\|_{\infty} \frac{L_W}{\mathrm{d}_{\min}^2} \Big) \\
&+ B_1^{(k-1)} \Big(L_{\Psi^{(k)}} \frac{L_W}{\mathrm{d}_{\min}} L_{\Phi^{(k)}} + L_{\Psi^{(k)}} \|W\|_{\infty} L_{\Phi^{(k)}} \frac{L_W}{\mathrm{d}_{\min}^2} \Big) \Bigg) \prod_{l'=k+1}^{l} L_{\Psi^{(l')}} \Big(1 + \frac{\|W\|_{\infty}}{\mathrm{d}_{\min}} L_{\Phi^{(l')}} \Big), \\
Z_2^{(l)} &= \sum_{k=1}^{l} B_2^{(k-1)} \Big(L_{\Psi^{(k)}} \frac{L_W}{\mathrm{d}_{\min}} L_{\Phi^{(k)}} + L_{\Psi^{(k)}} \|W\|_{\infty} L_{\Phi^{(k)}} \frac{L_W}{\mathrm{d}_{\min}^2} \Big) \prod_{l'=k+1}^{l} L_{\Psi^{(l')}} \Big(1 + \frac{\|W\|_{\infty}}{\mathrm{d}_{\min}} L_{\Phi^{(l')}} \Big), \\
Z_3^{(l)} &= \prod_{k=1}^{l} L_{\Psi^{(k)}} \Big(1 + \frac{\|W\|_{\infty}}{\mathrm{d}_{\min}} L_{\Phi^{(k)}} \Big),
\end{aligned}
\tag{31}
$$

where $B_1^{(k)}$ and $B_2^{(k)}$ are defined in (27) and (28).

Proof. By Lemma B.8, we have

$$
\begin{aligned}
L_{f^{(l)}} \leq & \sum_{k=1}^{l} \Bigg(\Big(L_{\Psi^{(k)}} \frac{L_W}{\mathrm{d}_{\min}} (\|\Phi^{(k)}(0,0)\|_\infty + L_{\Phi^{(k)}} \|f^{(k-1)}\|_\infty) + L_{\Psi^{(k)}} \|W\|_\infty (\|\Phi^{(k)}(0,0)\|_\infty \\
& + L_{\Phi^{(k)}} \|f^{(k-1)}\|_\infty) \frac{L_W}{\mathrm{d}_{\min}^2} \Big) \prod_{l'=k+1}^{l} L_{\Psi^{(l')}} \Big(1 + \frac{\|W\|_\infty}{\mathrm{d}_{\min}} L_{\Phi^{(l')}} \Big) \Bigg) \\
& + L_f \prod_{k=1}^{l} L_{\Psi^{(k)}} \Big(1 + \frac{\|W\|_\infty}{\mathrm{d}_{\min}} L_{\Phi^{(k)}} \Big) \\
= & \sum_{k=1}^{l} \Big(L_{\Psi^{(k)}} \frac{L_W}{\mathrm{d}_{\min}} \|\Phi^{(k)}(0,0)\|_\infty + L_{\Psi^{(k)}} \|W\|_\infty \|\Phi^{(k)}(0,0)\|_\infty \frac{L_W}{\mathrm{d}_{\min}^2} \Big) \\
& \prod_{l'=k+1}^{l} L_{\Psi^{(l')}} \Big(1 + \frac{\|W\|_\infty}{\mathrm{d}_{\min}} L_{\Phi^{(l')}} \Big) \\
& + \sum_{k=1}^{l} \|f^{(k-1)}\|_\infty \Big(L_{\Psi^{(k)}} \frac{L_W}{\mathrm{d}_{\min}} L_{\Phi^{(k)}} + L_{\Psi^{(k)}} \|W\|_\infty L_{\Phi^{(k)}} \frac{L_W}{\mathrm{d}_{\min}^2} \Big) \\
& \prod_{l'=k+1}^{l} L_{\Psi^{(l')}} \Big(1 + \frac{\|W\|_\infty}{\mathrm{d}_{\min}} L_{\Phi^{(l')}} \Big) \\
& + L_f \prod_{k=1}^{l} L_{\Psi^{(k)}} \Big(1 + \frac{\|W\|_\infty}{\mathrm{d}_{\min}} L_{\Phi^{(k)}} \Big) \\
\leq & \sum_{k=1}^{l} \Big(L_{\Psi^{(k)}} \frac{L_W}{\mathrm{d}_{\min}} \|\Phi^{(k)}(0,0)\|_\infty + L_{\Psi^{(k)}} \|W\|_\infty \|\Phi^{(k)}(0,0)\|_\infty \frac{L_W}{\mathrm{d}_{\min}^2} \Big) \\
& \prod_{l'=k+1}^{l} L_{\Psi^{(l')}} \Big(1 + \frac{\|W\|_\infty}{\mathrm{d}_{\min}} L_{\Phi^{(l')}} \Big) \\
& + \sum_{k=1}^{l} (B_1^{(k-1)} + B_2^{(k-1)} \|f\|_\infty) \Big(L_{\Psi^{(k)}} \frac{L_W}{\mathrm{d}_{\min}} L_{\Phi^{(k)}} + L_{\Psi^{(k)}} \|W\|_\infty L_{\Phi^{(k)}} \frac{L_W}{\mathrm{d}_{\min}^2} \Big) \\
& \prod_{l'=k+1}^{l} L_{\Psi^{(l')}} \Big(1 + \frac{\|W\|_\infty}{\mathrm{d}_{\min}} L_{\Phi^{(l')}} \Big) \\
& + L_f \prod_{k=1}^{l} L_{\Psi^{(k)}} \Big(1 + \frac{\|W\|_\infty}{\mathrm{d}_{\min}} L_{\Phi^{(k)}} \Big),
\end{aligned}
$$

where the last inequality holds by Lemma B.7. □

We continue with the following simple lemma which bounds the infinity norm of the output of a gMPNN.

Lemma B.10. *Let (χ, d, μ) be a metric-measure space, W be a kernel and $\Theta = \big((\Phi^{(l)})_{l=1}^T, (\Psi^{(l)})_{l=1}^T\big)$ be a MPNN s.t. Assumptions A.10.1-8. are satisfied. Consider a metric-space signal $f : \chi \to \mathbb{R}^F$ with $\|f\|_\infty < \infty$. Consider a graph $(G, \mathbf{f}) \sim (W, f)$ with N nodes and corresponding graph features. Then,*

$$
\|\Theta_G(\mathbf{f})\|_{2;\infty}^2 \leq N^T (A' + A'' \|f\|_\infty^2),
$$

where

$$A' = \sum_{l=1}^{T} \Big(2(L_{\Psi^{(l)}})^2 \frac{2}{\mathrm{W}_{\mathrm{diag}}} \|W\|_\infty \|\Phi^{(l)}(0,0)\|_\infty^2 + 2\|\Psi^{(l)}(0,0)\|_\infty^2 \Big)$$
$$\prod_{l'=l+1}^{T} 2(L_{\Psi^{(l')}})^2 \big(\frac{2}{\mathrm{W}_{\mathrm{diag}}} \|W\|_\infty (L_{\Phi^{(l')}})^2 + 1 \big)$$

and

$$A'' = \prod_{l=1}^{T} 2(L_{\Psi^{(l)}})^2 \big(\frac{2}{\mathrm{W}_{\mathrm{diag}}} \|W\|_\infty (L_{\Phi^{(l)}})^2 + 1 \big).$$

Proof. Let $l = 0, \ldots, T-1$. We have

$$\|\mathbf{f}^{(l+1)}\|_{2;\infty}^2 = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{f}_i^{(l+1)}\|_\infty^2,$$

where $\mathbf{f}_i^{(l+1)} = \Psi^{(l+1)}(\mathbf{f}_i^{(l)}, \mathbf{m}_i^{(l+1)})$ with $\mathbf{m}_i^{(l+1)} = M_G\big(\Phi^{(l+1)}(\mathbf{f}^{(l)}, \mathbf{f}^{(l)})\big)(X_i)$. By using the Lipschitz continuity of $\Psi^{(l+1)}$, we get

$$\begin{aligned} \|\mathbf{f}_i^{(l+1)}\|_\infty^2 &\leq 2\big(\|\Psi^{(l+1)}(\mathbf{f}_i^{(l)}, \mathbf{m}_i^{(l+1)}) - \Psi^{(l+1)}(0,0)\|_\infty^2 + \|\Psi^{(l+1)}(0,0)\|_\infty^2 \big) \\ &\leq 2\big((L_{\Psi^{(l+1)}})^2 (\|\mathbf{f}_i^{(l)}\|_\infty^2 + \|\mathbf{m}_i^{(l+1)}\|_\infty^2) + \|\Psi^{(l+1)}(0,0)\|_\infty^2 \big) \end{aligned} \tag{32}$$

For the message term we calculate

$$\begin{aligned} \|\mathbf{m}_i^{(l+1)}\|_\infty^2 &= \left\| \frac{1}{\sum_{j=1}^N W(X_i, X_j)} \sum_{j=1}^{N} W(X_i, X_j) \Phi^{(l+1)}(\mathbf{f}_i^{(l)}, \mathbf{f}_j^{(l)}) \right\|_\infty^2 \\ &= \left| \frac{1}{\sum_{j=1}^N W(X_i, X_j)} \right|^2 \left\| \sum_{j=1}^{N} \sqrt{W(X_i, X_j)} \cdot \sqrt{W(X_i, X_j)} \Phi^{(l+1)}(\mathbf{f}_i^{(l)}, \mathbf{f}_j^{(l)}) \right\|_\infty^2 \\ &\leq \left| \frac{1}{\sum_{j=1}^N W(X_i, X_j)} \right|^2 \sum_{j=1}^{N} |W(X_i, X_j)| \sum_{j=1}^{N} |W(X_i, X_j)| \|\Phi^{(l+1)}(\mathbf{f}_i^{(l)}, \mathbf{f}_j^{(l)})\|_\infty^2 \\ &\leq \left| \frac{1}{\sum_{j=1}^N W(X_i, X_j)} \right| \|W\|_\infty \sum_{j=1}^{N} \|\Phi^{(l+1)}(\mathbf{f}_i^{(l)}, \mathbf{f}_j^{(l)})\|_\infty^2 \\ &\leq \frac{\|W\|_\infty}{\mathrm{W}_{\mathrm{diag}}} \sum_{j=1}^{N} \|\Phi^{(l+1)}(\mathbf{f}_i^{(l)}, \mathbf{f}_j^{(l)})\|_\infty^2 \end{aligned}$$

where the inequality follows from Cauchy-Schwarz inequality. Per assumption, we have $|W(X_i, X_i)| \geq \mathrm{W}_{\mathrm{diag}}$ and for every $i = 1, \ldots, N$,

$$\begin{aligned} \|\Phi^{(l+1)}(\mathbf{f}_i^{(l)}, \mathbf{f}_j^{(l)})\|_\infty^2 &= \|\Phi^{(l+1)}(\mathbf{f}_i^{(l)}, \mathbf{f}_j^{(l)}) - \Phi^{(l+1)}(0,0) + \Phi^{(l+1)}(0,0)\|_\infty^2 \\ &\leq 2\Big(\|\Phi^{(l+1)}(\mathbf{f}_i^{(l)}, \mathbf{f}_j^{(l)}) - \Phi^{(l+1)}(0,0)\|_\infty^2 + \|\Phi^{(l+1)}(0,0)\|_\infty^2 \Big) \\ &\leq 2\Big((L_{\Phi^{(l+1)}})^2 (\|\mathbf{f}_i^{(l)}\|_\infty^2 + \|\mathbf{f}_j^{(l)}\|_\infty^2) + \|\Phi^{(l+1)}(0,0)\|_\infty^2 . \Big) \end{aligned}$$

Hence,

$$\begin{aligned} \|\mathbf{m}_i^{(l+1)}\|_\infty^2 &\leq \frac{2\|W\|_\infty}{\mathrm{W}_{\mathrm{diag}}} \frac{N}{N} \sum_{j=1}^{N} \Big((L_{\Phi^{(l+1)}})^2 (\|\mathbf{f}_i^{(l)}\|_\infty^2 + \|\mathbf{f}_j^{(l)}\|_\infty^2) + \|\Phi^{(l+1)}(0,0)\|_\infty^2 \Big) \\ &\leq N \frac{2\|W\|_\infty}{\mathrm{W}_{\mathrm{diag}}} \Big((L_{\Phi^{(l+1)}})^2 \|\mathbf{f}_i^{(l)}\|_\infty^2 + (L_{\Phi^{(l+1)}})^2 \|\mathbf{f}^{(l)}\|_{2;\infty}^2 + \|\Phi^{(l+1)}(0,0)\|_\infty^2 \Big). \end{aligned} \tag{33}$$

By (32) and (33), we have

$$
\begin{aligned}
\|\mathbf{f}^{(l+1)}\|_{2;\infty}^2 &\leq \frac{1}{N}\sum_{i=1}^{N} 2\Big((L_{\Psi^{(l+1)}})^2(\|\mathbf{f}_i^{(l)}\|_\infty^2 + \|\mathbf{m}_i^{(l+1)}\|_\infty^2) + \|\Psi^{(l+1)}(0,0)\|_\infty^2\Big) \\
&\leq \frac{1}{N}\sum_{i=1}^{N} 2\Bigg((L_{\Psi^{(l+1)}})^2\Big(\|\mathbf{f}_i^{(l)}\|_\infty^2 + N\frac{2\|W\|_\infty}{\mathrm{W}_{\mathrm{diag}}}\big((L_{\Phi^{(l+1)}})^2\|\mathbf{f}_i^{(l)}\|_\infty^2 \\
&+ (L_{\Phi^{(l+1)}})^2\|\mathbf{f}^{(l)}\|_{2;\infty}^2 + \|\Phi^{(l+1)}(0,0)\|_\infty^2\big)\Big) + \|\Psi^{(l+1)}(0,0)\|_\infty^2\Bigg) \\
&= 2(L_{\Psi^{(l+1)}})^2\Big(\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{f}_i^{(l)}\|_\infty^2 + N\frac{2\|W\|_\infty}{\mathrm{W}_{\mathrm{diag}}}\big((L_{\Phi^{(l+1)}})^2\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{f}_i^{(l)}\|_\infty^2 \\
&+ (L_{\Phi^{(l+1)}})^2\|\mathbf{f}^{(l)}\|_{2;\infty}^2 + \|\Phi^{(l+1)}(0,0)\|_\infty^2\big)\Big) + 2\|\Psi^{(l+1)}(0,0)\|_\infty^2 \\
&= 2(L_{\Psi^{(l+1)}})^2\Big(\|\mathbf{f}^{(l)}\|_{2;\infty}^2 + N\frac{2\|W\|_\infty}{\mathrm{W}_{\mathrm{diag}}}\big((L_{\Phi^{(l+1)}})^2\|\mathbf{f}^{(l)}\|_{2;\infty}^2 + \|\Phi^{(l+1)}(0,0)\|_\infty^2\big)\Big) \\
&+ 2\|\Psi^{(l+1)}(0,0)\|_\infty^2 \\
&= 2(L_{\Psi^{(l+1)}})^2\big(N\frac{2\|W\|_\infty}{\mathrm{W}_{\mathrm{diag}}}(L_{\Phi^{(l+1)}})^2 + 1\big)\|\mathbf{f}^{(l)}\|_{2;\infty}^2 \\
&+ 2(L_{\Psi^{(l+1)}})^2 N\frac{2\|W\|_\infty}{\mathrm{W}_{\mathrm{diag}}}\|\Phi^{(l+1)}(0,0)\|_\infty^2 + 2\|\Psi^{(l+1)}(0,0)\|_\infty^2
\end{aligned}
$$

Hence, by $\|\mathbf{f}\|_{2;\infty}^2 \leq \|f\|_\infty^2$ and Lemma B.11, we have

$$
\begin{aligned}
\|\mathbf{f}^{(T)}\|_{2;\infty}^2 &\leq \sum_{l=1}^{T}\Big(2(L_{\Psi^{(l)}})^2 N\frac{2\|W\|_\infty}{\mathrm{W}_{\mathrm{diag}}}\|\Phi^{(l)}(0,0)\|_\infty^2 + 2\|\Psi^{(l)}(0,0)\|_\infty^2\Big) \\
&\prod_{l'=l+1}^{T} 2(L_{\Psi^{(l')}})^2\big(N\frac{2\|W\|_\infty}{\mathrm{W}_{\mathrm{diag}}}(L_{\Phi^{(l')}})^2 + 1\big) \\
&+ \|f\|_\infty^2\prod_{l=1}^{T}\Big(2(L_{\Psi^{(l)}})^2\big(N\frac{2\|W\|_\infty}{\mathrm{W}_{\mathrm{diag}}}(L_{\Phi^{(l)}})^2 + 1\big)\Big) \\
&\leq N^T\sum_{l=1}^{T}\Big(2(L_{\Psi^{(l)}})^2\frac{2\|W\|_\infty}{\mathrm{W}_{\mathrm{diag}}}\|\Phi^{(l)}(0,0)\|_\infty^2 + 2\|\Psi^{(l)}(0,0)\|_\infty^2\Big) \\
&\prod_{l'=l+1}^{T} 2(L_{\Psi^{(l')}})^2\big(\frac{2}{\mathrm{W}_{\mathrm{diag}}^2}\|W\|_\infty^2(L_{\Phi^{(l')}})^2 + 1\big) \\
&+ \|f\|_\infty^2 N^T\prod_{l}^{T}\Big(2(L_{\Psi^{(l)}})^2\big(\frac{2\|W\|_\infty}{\mathrm{W}_{\mathrm{diag}}}(L_{\Phi^{(l)}})^2 + 1\big)\Big).
\end{aligned}
$$

□

We finish this subsection with the following easily verifiable lemma that provides a general solution for certain recurrence relations.

Lemma B.11. *Let $(\eta^{(l)})_{l=0}^T$ be a sequence of real numbers satisfying $\eta^{(l+1)} \leq a^{(l+1)}\eta^{(l)} + b^{(l+1)}$ for $l = 0, \ldots, T-1$, for some real numbers $a^{(l)}, b^{(l)}$, $l = 1, \ldots, T$. Then*

$$
\eta^{(T)} \leq \sum_{l=1}^{T} b^l \prod_{l'=l+1}^{T} a^{(l')} + \eta^{(0)}\prod_{l=1}^{T} a^{(l)},
$$

where we define the product $\prod_{T+1}^{T}$ as 1.

B.2 Proof of Theorem 3.1

The idea of the Proof of Theorem 3.1 is as follows. We first use Corollary B.6 to bound the error between a cMPNN and a gMPNN layer-wise, when the input of layer l of the gMPNN is exactly the sampled graph signal from the output of layer $l-1$ of the cMPNN. This is shown in Corollary B.12. Then, we use this to provide a recurrence relation for the true error between a cMPNN and the corresponding gMPNN in Lemma B.13. We solve this recurrence relation in Corollary B.14, where we have an error bound that depends only on the parameters of the MPNN, the regularity of the kernel and the regularity of the continuous output metric-space signals of the layers of the cMPNN. We remove the last dependency in Theorem B.15. We then analyze the additional error by a final pooling layer, leading to the formulation and final proof of Theorem 3.1, rewritten as Theorem B.18.

Corollary B.12. *Let (χ, d, μ) be a metric-measure space and W be a kernel s.t. Assumptions A.10.1-6 are satisfied. Let $p \in (0, \frac{1}{2})$. Consider a graph $(G, \mathbf{f}) \sim (W, f)$ with N nodes and corresponding graph features, where N satisfies (14). If the event $\mathcal{F}^p_{\text{Lip}}$ from Lemma B.5 occurs, then condition (15) together with (34) below are satisfied: For every MPNN Θ satisfying Assumption A.10.7. and $f : \chi \to \mathbb{R}^F$ with Lipschitz constant L_f, we have*

$$\text{dist}\left(\Lambda^{(l+1)}_{\Theta_G}(S^X f^{(l)}), \Lambda^{(l+1)}_{\Theta_W}(f^{(l)})\right) \leq Q^{(l+1)} \tag{34}$$

for all $l = 0, \dots, T-1$, where $f^{(l)} = \Theta^{(l)}_W f$ as defined in (5), and $\Lambda^{(l+1)}_{\Theta_G}$ and $\Lambda^{(l+1)}_{\Theta_W}$ are defined in Definition A.9. Here,

$$\begin{aligned}
Q^{(l+1)} = L_{\Psi^{(l+1)}} \Bigg(& 4\frac{\varepsilon_1}{\sqrt{N}\mathrm{d}^2_{min}} \|W\|_\infty (L_{\Phi^{(l+1)}} \|f^{(l)}\|_\infty + \|\Phi^{(l+1)}(0,0)\|_\infty) \\
& + N^{-\frac{1}{2(D_\chi+1)}} \Bigg(2\Big(\frac{\|W\|_\infty}{\mathrm{d}_{min}} L_{\Phi^{(l+1)}} L_{f^{(l)}} + \frac{L_W}{\mathrm{d}_{min}} \big(L_{\Phi^{(l+1)}} \|f^{(l)}\|_\infty + \|\Phi^{(l+1)}(0,0)\|_\infty \big) \Big) \\
& + \frac{C_\chi}{\sqrt{2}} \Big(\frac{\|W\|_\infty}{\mathrm{d}_{min}} \big(L_{\Phi^{(l+1)}} \|f^{(l)}\|_\infty + \|\Phi^{(l+1)}(0,0)\|_\infty \big) \Big) \\
& \cdot \sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)} \log(N) + \log(2/p)} \Bigg) \Bigg),
\end{aligned} \tag{35}$$

and dist *is defined in (12).*

Proof. Let $l = 0, \dots, T-1$. We have,

$$
\begin{aligned}
&\Big(\mathrm{dist}\big(\Lambda^{(l+1)}_{\Theta_G}(S^X f^{(l)}), \Lambda^{(l+1)}_{\Theta_W}(f^{(l)})\big)\Big)^2 \\
&= \|\Lambda^{(l+1)}_{\Theta_G}(S^X f^{(l)}) - S^X \Lambda^{(l+1)}_{\Theta_W}(f^{(l)})\|^2_{2;\infty} \\
&= \frac{1}{N}\sum_{i=1}^{N} \|\Lambda^{(l+1)}_{\Theta_G}(S^X f^{(l)})(X_i) - S^X \Lambda^{(l+1)}_{\Theta_W}(f^{(l)})(X_i)\|^2_{\infty} \\
&= \frac{1}{N}\sum_{i=1}^{N} \Big\| \Psi^{(l+1)}\Big(f^{(l)}(X_i), M_G\big(\Phi^{(l+1)}(S^X f^{(l)}, S^X f^{(l)})\big)(X_i)\Big) \\
&- \Psi^{(l+1)}\Big(f^{(l)}(X_i), M_W\big(\Phi^{(l+1)}(f^{(l)}, f^{(l)})\big)(X_i)\Big)\Big\|^2_{\infty} \\
&= \frac{1}{N}\sum_{i=1}^{N} \Big\| \Psi^{(l+1)}\Big(f^{(l)}(X_i), M_X\big(\Phi^{(l+1)}(f^{(l)}, f^{(l)})\big)(X_i)\Big) \\
&- \Psi^{(l+1)}\Big(f^{(l)}(X_i), M_W\big(\Phi^{(l+1)}(f^{(l)}, f^{(l)})\big)(X_i)\Big)\Big\|^2_{\infty} \\
&\leq L^2_{\Psi^{(l+1)}}\Bigg(4\frac{\varepsilon_1}{\sqrt{N}\mathrm{d}^2_{min}}\|W\|_\infty (L_{\Phi^{(l+1)}}\|f^{(l)}\|_\infty + \|\Phi^{(l+1)}(0,0)\|_\infty) \\
&+ N^{-\frac{1}{2(D_\chi+1)}}\Bigg(2\Big(\frac{\|W\|_\infty}{\mathrm{d}_{min}} L_{\Phi^{(l+1)}} L_{f^{(l)}} + \frac{L_W}{\mathrm{d}_{min}}\big(L_{\Phi^{(l+1)}}\|f^{(l)}\|_\infty + \|\Phi^{(l+1)}(0,0)\|_\infty\big)\Big) \\
&+ \frac{C_\chi}{\sqrt{2}}\Big(\frac{\|W\|_\infty}{\mathrm{d}_{min}}\big(L_{\Phi^{(l+1)}}\|f^{(l)}\|_\infty + \|\Phi^{(l+1)}(0,0)\|_\infty\big)\Big) \\
&\cdot \sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)}\log(N) + \log(2/p)}\Bigg)\Bigg)^2,
\end{aligned}
$$

where the final inequality holds, by applying Corollary B.6. □

Lemma B.13. *Let (χ, d, μ) be a metric-measure space and W be a kernel s.t. Assumptions A.10.1-6. are satisfied. Let $p \in (0, \frac{1}{2})$. Consider a graph $(G, \mathbf{f}) \sim (W, f)$ with N nodes and corresponding graph features, where N satisfies (14). Denote, for $l = 1, \dots, T$,*

$$\varepsilon^{(l)} = \mathrm{dist}(\Theta^{(l)}_G(\mathbf{f}), \Theta^{(l)}_W(f)),$$

and $\varepsilon^{(0)} = \mathrm{dist}(\mathbf{f}, f)$. If the event $\mathcal{F}^p_{\mathrm{Lip}}$ from Lemma B.5 occurs, then, for every MPNN Θ satisfying Assumption A.10.7. and $f : \chi \to \mathbb{R}^F$ with Lipschitz constant L_f, the following recurrence relation holds:

$$\varepsilon^{(l)} \leq K^{(l+1)}\varepsilon^{(l)} + Q^{(l+1)}$$

for $l = 0, \dots, T-1$. Here, $Q^{(l+1)}$ is defined in (35), and

$$K^{(l+1)} = \sqrt{(L_{\Psi^{(l+1)}})^2 + \frac{8\|W\|^2_\infty}{\mathrm{d}^2_{\min}}(L_{\Phi^{(l+1)}})^2(L_{\Psi^{(l+1)}})^2}. \tag{36}$$

Proof. In the event $\mathcal{F}^p_{\mathrm{Lip}}$, by Corollary B.12, we have for every MPNN Θ satisfying Assumption A.10.7. and $f : \chi \to \mathbb{R}^F$ with Lipschitz constant L_f,

$$\mathrm{dist}\left(\Lambda^{(l+1)}_{\Theta_G}(S^X f^{(l)}), \Lambda^{(l+1)}_{\Theta_W}(f^{(l)})\right) \leq Q^{(l+1)} \tag{37}$$

for $l = 0, \dots, T-1$, and

$$|\mathrm{d}_X(x)| \geq \frac{\mathrm{d}_{\min}}{2} \tag{38}$$

for all $x \in \chi$. Let $l = 0, \ldots, T-1$. We have

$$
\begin{aligned}
&\operatorname{dist}(\Theta_G^{(l+1)}(\mathbf{f}), \Theta_W^{(l+1)}(f)) \\
&= \|\Theta_G^{(l+1)}(\mathbf{f}) - S^X \Theta_W^{(l+1)}(f)\|_{2;\infty} \\
&\leq \|\Theta_G^{(l+1)}(\mathbf{f}) - \Lambda_{\Theta_G}^{(l+1)}(S^X f^{(l)})\|_{2;\infty} + \|\Lambda_{\Theta_G}^{(l+1)}(S^X f^{(l)}) - S^X \Theta_{\Theta_W}^{(l+1)}(f)\|_{2;\infty} \\
&= \|\Lambda_G^{(l+1)}(\mathbf{f}^{(l)}) - \Lambda_G^{(l+1)}(S^X f^{(l)})\|_{2;\infty} + \|\Lambda_{\Theta_G}^{(l+1)}(S^X f^{(l)}) - S^X \Lambda_{\Theta_W}^{(l+1)}(f^{(l)})\|_{2;\infty} \\
&\leq \|\Lambda_{\Theta_G}^{(l+1)}(\mathbf{f}^{(l)}) - \Lambda_{\Theta_G}^{(l+1)}(S^X f^{(l)})\|_{2;\infty} + Q^{(l+1)}.
\end{aligned}
\tag{39}
$$

We bound the first term on the right-hand-side of (39) as follows.

$$
\begin{aligned}
&\|\Lambda_{\Theta_G}^{(l+1)}(\mathbf{f}^{(l)}) - \Lambda_{\Theta_G}^{(l+1)}(S^X f^{(l)})\|_{2;\infty}^2 \\
&= \frac{1}{N} \sum_{i=1}^{N} \Big\| \Psi^{(l+1)}\Big(\mathbf{f}_i^{(l)}, M_G\big(\Phi^{(l+1)}(\mathbf{f}^{(l)}, \mathbf{f}^{(l)})\big)(X_i)\Big) \\
&- \Psi^{(l+1)}\Big((S^X f^{(l)})_i, M_G\big(\Phi^{(l+1)}(S^X f^{(l)}, S^X f^{(l)})\big)(X_i)\Big)\Big\|_\infty^2 \\
&\leq \frac{1}{N} (L_{\Psi^{(l+1)}})^2 \sum_{i=1}^{N} \Big\| \Big(\mathbf{f}_i^{(l)}, M_G\big(\Phi^{(l+1)}(\mathbf{f}^{(l)}, \mathbf{f}^{(l)})\big)(X_i)\Big) \\
&- \Big((S^X f^{(l)})_i, M_G\big(\Phi^{(l+1)}(S^X f^{(l)}, S^X f^{(l)})\big)(X_i)\Big)\Big\|_\infty^2 \\
&\leq \frac{1}{N} (L_{\Psi^{(l+1)}})^2 \Big(\sum_{i=1}^{N} \Big\| \mathbf{f}_i^{(l)} - (S^X f^{(l)})_i \Big\|_\infty^2 \\
&+ \sum_{i=1}^{N} \Big\| M_G\big(\Phi^{(l+1)}(\mathbf{f}^{(l)}, \mathbf{f}^{(l)})\big)(X_i) - M_G\big(\Phi^{(l+1)}(S^X f^{(l)}, S^X f^{(l)})\big)(X_i) \Big\|_\infty^2 \Big) \\
&\leq (L_{\Psi^{(l+1)}})^2 \Big(\big(\operatorname{dist}(\mathbf{f}^{(l)}, f^{(l)})\big)^2 \\
&+ \frac{1}{N} \sum_{i=1}^{N} \big\| M_G\big(\Phi^{(l+1)}(\mathbf{f}^{(l)}, \mathbf{f}^{(l)})\big)(X_i) - M_G\big(\Phi^{(l+1)}(S^X f^{(l)}, S^X f^{(l)})\big)(X_i) \big\|_\infty^2 \Big) \\
&\leq (L_{\Psi^{(l+1)}})^2 \Big((\varepsilon^{(l)})^2 \\
&+ \frac{1}{N} \sum_{i=1}^{N} \big\| M_G\big(\Phi^{(l+1)}(\mathbf{f}^{(l)}, \mathbf{f}^{(l)})\big)(X_i) - M_G\big(\Phi^{(l+1)}(S^X f^{(l)}, S^X f^{(l)})\big)(X_i) \big\|_\infty^2 \Big).
\end{aligned}
\tag{40}
$$

Now, for every $i = 1, \dots, N$, we have

$$\begin{aligned}
&\Big\| M_G\Big(\Phi^{(l+1)}\big(\mathbf{f}^{(l)}, \mathbf{f}^{(l)}\big)\Big)(X_i) - M_G\Big(\Phi^{(l+1)}\big(S^X f^{(l)}, S^X f^{(l)}\big)\Big)(X_i)\Big\|_\infty^2 \\
&= \Big\| \frac{1}{N}\sum_{j=1}^N \frac{W(X_i, X_j)}{\mathrm{d}_X(X_i)} \Phi^{(l+1)}\big(\mathbf{f}^{(l)}(X_i), \mathbf{f}^{(l)}(X_j)\big) \\
&- \frac{1}{N}\sum_{j=1}^N \frac{W(X_i, X_j)}{\mathrm{d}_X(X_i)} \Phi^{(l+1)}\big(S^X f^{(l)}(X_i), S^X f^{(l)}(X_j)\big)\Big\|_\infty^2 \\
&= \Big\| \frac{1}{N}\sum_{j=1}^N \frac{W(X_i, X_j)}{\mathrm{d}_X(X_i)} \Big(\Phi^{(l+1)}\big(\mathbf{f}^{(l)}(X_i), \mathbf{f}^{(l)}(X_j)\big) - \Phi^{(l+1)}\big(S^X f^{(l)}(X_i), S^X f^{(l)}(X_j)\big)\Big)\Big\|_\infty^2 \\
&\leq \frac{1}{N^2}\sum_{j=1}^N \Big|\frac{W(X_i, X_j)}{\mathrm{d}_X(X_i)}\Big|^2 \sum_{j=1}^N \Big\|\Big(\Phi^{(l+1)}\big(\mathbf{f}^{(l)}(X_i), \mathbf{f}^{(l)}(X_j)\big) - \Phi^{(l+1)}\big(S^X f^{(l)}(X_i), S^X f^{(l)}(X_j)\big)\Big)\Big\|_\infty^2 \\
&\leq \frac{4\|W\|_\infty^2}{\mathrm{d}_{\min}^2} \frac{1}{N}\sum_{j=1}^N \Big\|\Big(\Phi^{(l+1)}\big(\mathbf{f}^{(l)}(X_i), \mathbf{f}^{(l)}(X_j)\big) - \Phi^{(l+1)}\big(S^X f^{(l)}(X_i), S^X f^{(l)}(X_j)\big)\Big)\Big\|_\infty^2,
\end{aligned} \tag{41}$$

where the second-to-last inequality holds by the Cauchy–Schwarz inequality and the last inequality holds by (38). Now, for the term on the right-hand-side of (41), we have

$$\begin{aligned}
&\frac{1}{N}\sum_{j=1}^N \Big\|\Phi^{(l+1)}\big(\mathbf{f}^{(l)}(X_i), \mathbf{f}^{(l)}(X_j)\big) - \Phi^{(l+1)}\big(S^X f^{(l)}(X_i), S^X f^{(l)}(X_j)\big)\Big\|_\infty^2 \\
&\leq (L_{\Phi^{(l+1)}})^2 \frac{1}{N}\sum_{j=1}^N \Big(\big\|\mathbf{f}^{(l)}(X_i) - S^X f^{(l)}(X_i)\big\|_\infty^2 + \big\|\mathbf{f}^{(l)}(X_j) - S^X f^{(l)}(X_j)\big\|_\infty^2\Big) \\
&\leq (L_{\Phi^{(l+1)}})^2 \big\|\mathbf{f}^{(l)}(X_i) - S^X f^{(l)}(X_i)\big\|_\infty^2 + (L_{\Phi^{(l+1)}})^2 (\varepsilon^{(l)})^2.
\end{aligned} \tag{42}$$

Hence, by inserting (42) into (41) and (41) into (40), we have

$$\begin{aligned}
&\|\Lambda_{\Theta_G}^{(l+1)}(\mathbf{f}^{(l)}) - \Lambda_{\Theta_G}^{(l+1)}(S^X f^{(l)})\|_{2;\infty}^2 \\
&\leq (L_{\Psi^{(l+1)}})^2 \Big((\varepsilon^{(l)})^2 + \frac{1}{N}\sum_{i=1}^N \big\|M_G\big(\Phi^{(l)}(\mathbf{f}^{(l)}, \mathbf{f}^{(l)})\big)(X_i) - M_G\big(\Phi^{(l)}(S^X f^{(l)}, S^X f^{(l)})\big)(X_i)\big\|_\infty^2\Big) \\
&\leq (L_{\Psi^{(l+1)}})^2 \Big((\varepsilon^{(l)})^2 + \frac{4\|W\|_\infty^2}{\mathrm{d}_{\min}^2}(L_{\Phi^{(l+1)}})^2 \big(\frac{1}{N}\sum_{i=1}^N \big\|\mathbf{f}^{(l)}(X_i) - S^X f^{(l)}(X_i)\big\|_\infty^2 + (\varepsilon^{(l)})^2\big)\Big) \\
&\leq (L_{\Psi^{(l+1)}})^2 \Big((\varepsilon^{(l)})^2 + \frac{4\|W\|_\infty^2}{\mathrm{d}_{\min}^2}(L_{\Phi^{(l+1)}})^2 \big((\varepsilon^{(l)})^2 + (\varepsilon^{(l)})^2\big)\Big) \\
&\leq (L_{\Psi^{(l+1)}})^2 \Big((\varepsilon^{(l)})^2 + \frac{8\|W\|_\infty^2}{\mathrm{d}_{\min}^2}(L_{\Phi^{(l+1)}})^2 (\varepsilon^{(l)})^2\Big).
\end{aligned}$$

By inserting this into (39), we conclude

$$\mathrm{dist}(\Theta_G^{(l+1)}(\mathbf{f}), \Theta_W^{(l+1)}(f)) \leq (L_{\Psi^{(l+1)}})^2 \big(1 + \frac{8\|W\|_\infty^2}{\mathrm{d}_{\min}^2}(L_{\Phi^{(l+1)}})^2\big)(\varepsilon^{(l)})^2 + Q^{(l+1)}.$$

□

Corollary B.14. *Let (χ, d, μ) be a metric-measure space and W be a kernel s.t. Assumptions A.10.1-6. are satisfied. Let $p \in (0, \frac{1}{2})$. Consider a graph $(G, \mathbf{f}) \sim (W, f)$ with N nodes and corresponding graph features, where N satisfies (14). If the event $\mathcal{F}_{\mathrm{Lip}}^p$ from Lemma B.5 occurs, then, for every MPNN Θ satisfying Assumption A.10.7. and every Lipschitz continuous $f : \chi \to \mathbb{R}^F$ with Lipschitz constant L_f,*

$$\mathrm{dist}\big(\Theta_G(f(X)), \Theta_W(f)\big) \leq \sum_{l=1}^T Q^{(l)} \prod_{l'=l+1}^T K^{(l')},$$

where $Q^{(l)}$ and $K^{(l')}$ are defined in (35) and (36), respectively.

Proof. By Lemma B.13, for every MPNN Θ satisfying Assumption A.10.7. and every Lipschitz continuous $f : \chi \to \mathbb{R}^F$ with Lipschitz constant L_f, the recurrence relation

$$\varepsilon^{(l+1)} \leq K^{(l+1)}\varepsilon^{(l)} + Q^{(l+1)}$$

holds for $l = 0, \ldots, T-1$. We use that $\varepsilon^{(0)} = 0$ and $\varepsilon^{(T)} = \mathrm{dist}\big(\Theta_G(f(X)), \Theta_W(f)\big)$, and solve this recurrence relation by Lemma B.11 to finish the proof. □

Theorem B.15. *Let (χ, d, μ) be a metric-measure space and W be a kernel s.t. Assumptions A.10.1-6. are satisfied. Let $p \in (0, \frac{1}{2})$. Consider a graph $(G, \mathbf{f}) \sim (W, f)$ with N nodes and corresponding graph features, where N satisfies (14). If the event $\mathcal{F}^p_{\mathrm{Lip}}$ from Lemma B.5 occurs, then for every MPNN Θ satisfying Assumption A.10.7 and $f : \chi \to \mathbb{R}^F$ with Lipschitz constant L_f,*

$$\begin{aligned}
&\mathrm{dist}\big(\Theta_G(f(X)), \Theta_W(f)\big) \\
&\leq N^{-\frac{1}{2}}\left(\Omega_1 + \Omega_2 \log(2/p) + \Omega_3\|f\|_\infty + \Omega_4\|f\|_\infty \log(2/p)\right) \\
&+ N^{-\frac{1}{2(D_\chi+1)}}\big(\Omega_5 + \Omega_6\|f\|_\infty + \Omega_7 L_f\big) \\
&+ N^{-\frac{1}{2(D_\chi+1)}}\sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)}\log(N) + \log(2/p)} \cdot (\Omega_8 + \Omega_9\|f\|_\infty),
\end{aligned}$$

where Ω_i, for $i = 1, \ldots, 9$, are constants of the MPNN Θ, defined in (48), which depend only on the Lipschitz constants of the message and update functions $\{L_{\Phi^{(l)}}, L_{\Psi^{(l)}}\}_{l=1}^T$, and the formal biases $\{\|\Phi^{(l)}(0,0)\|_\infty\}_{l=1}^T$.

Proof. In the event $\mathcal{F}^p_{\mathrm{Lip}}$, by Corollary B.14, for every MPNN Θ satisfying Assumption A.10.7. and $f : \chi \to \mathbb{R}^F$ with Lipschitz constant L_f,

$$\mathrm{dist}\big(\Theta_G(f(X)), \Theta_W(f)\big) \leq \sum_{l=1}^{T} Q^{(l)} \prod_{l'=l+1}^{T} K^{(l')}, \tag{43}$$

where

$$\begin{aligned}
Q^{(l)} &= L_{\Psi^{(l)}}\Bigg(4\frac{\varepsilon_1}{\sqrt{N}\mathrm{d}^2_{min}}\|W\|_\infty(L_{\Phi^{(l)}}\|f^{(l-1)}\|_\infty + \|\Phi^{(l)}(0,0)\|_\infty) \\
&+ N^{-\frac{1}{2(D_\chi+1)}}\Bigg(2\Big(\frac{\|W\|_\infty}{\mathrm{d}_{min}}L_{\Phi^{(l)}}L_{f^{(l-1)}} + \frac{L_W}{\mathrm{d}_{min}}\big(L_{\Phi^{(l)}}\|f^{(l-1)}\|_\infty + \|\Phi^{(l)}(0,0)\|_\infty\big)\Big) \\
&+ \frac{C_\chi}{\sqrt{2}}\Big(\frac{\|W\|_\infty}{\mathrm{d}_{min}}\big(L_{\Phi^{(l)}}\|f^{(l-1)}\|_\infty + \|\Phi^{(l)}(0,0)\|_\infty\big)\Big) \\
&\cdot \sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)}\log(N) + \log(2/p)}\Bigg)\Bigg),
\end{aligned}$$

and

$$(K^{(l')})^2 = (L_{\Psi^{(l')}})^2 + \frac{8\|W\|^2_\infty}{\mathrm{d}^2_{\min}}(L_{\Phi^{(l')}})^2(L_{\Psi^{(l')}})^2.$$

We plug the definition of $Q^{(l)}$ into the right-hand-side of (43), to get

$$\begin{aligned}
&\mathrm{dist}\big(\Theta_G(f(X)), \Theta_W(f)\big) \\
&\leq \sum_{l=1}^{T} L_{\Psi^{(l)}} \Bigg(4\frac{\varepsilon_1}{\sqrt{N}\mathrm{d}_{min}^2} \|W\|_\infty (L_{\Phi^{(l)}} \|f^{l-1)}\|_\infty + \|\Phi^{(l)}(0,0)\|_\infty) \\
&+ N^{-\frac{1}{2(D_\chi+1)}} \Big(\frac{2\|W\|_\infty}{\mathrm{d}_{min}} L_{\Phi^{(l)}} L_{f^{(l-1)}} + \frac{2L_W}{\mathrm{d}_{min}} (L_{\Phi^{(l)}} \|f^{(l-1)}\|_\infty + \|\Phi^{(l)}(0,0)\|_\infty) \\
&+ \frac{C_\chi}{\sqrt{2}} \big(\frac{\|W\|_\infty}{\mathrm{d}_{min}} (L_{\Phi^{(l)}} \|f^{(l-1)}\|_\infty + \|\Phi^{(l)}(0,0)\|_\infty) \big) \\
&\sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)} \log(N) + \log(2/p)} \Big) \Bigg) \prod_{l'=l+1}^{T} K^{(l')}.
\end{aligned} \tag{44}$$

By Lemma B.7, we have

$$||f^{(l)}||_\infty \leq B_1^{(l)} + B_2^{(l)} ||f||_\infty, \tag{45}$$

where $B_1^{(l)}$, $B_2^{(l)}$ are independent of f. Furthermore, we have

$$L_{f^{(l)}} \leq Z_1^{(l)} + Z_2^{(l)} \|f\|_\infty + Z_3^{(l)} L_f, \tag{46}$$

where $Z_1^{(l)}$, $Z_2^{(l)}$ and $Z_3^{(l)}$ are independent of f, and defined in (31). We plug the bound of $L_{f^{(l-1)}}$ from (46) into (43)

$$\begin{aligned}
&\mathrm{dist}\big(\Theta_G(f(X)), \Theta_W(f)\big) \\
&\leq \sum_{l=1}^{T} L_{\Psi^{(l)}} \Bigg(4\frac{\varepsilon_1}{\sqrt{N}\mathrm{d}_{min}^2} \|W\|_\infty (L_{\Phi^{(l)}} \|f^{(l-1)}\|_\infty + \|\Phi^{(l)}(0,0)\|_\infty) + N^{-\frac{1}{2(D_\chi+1)}} \\
&\cdot \Big(\frac{2\|W\|_\infty}{\mathrm{d}_{min}} L_{\Phi^{(l)}} (Z_1^{(l-1)} + Z_2^{(l-1)} \|f\|_\infty + Z_3^{(l-1)} L_f) + \frac{2L_W}{\mathrm{d}_{min}} (L_{\Phi^{(l)}} \|f^{(l-1)}\|_\infty + \|\Phi^{(l)}(0,0)\|_\infty) \\
&+ \frac{C_\chi}{\sqrt{2}} \big(\frac{\|W\|_\infty}{\mathrm{d}_{min}} (L_{\Phi^{(l)}} \|f^{(l-1)}\|_\infty + \|\Phi^{(l)}(0,0)\|_\infty) \big) \\
&\cdot \sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)} \log(N) + \log(2/p)} \Big) \Bigg) \prod_{l'=l+1}^{T} K^{(l')}.
\end{aligned}$$

We insert the bound of $\|f^{(l-1)}\|_\infty$ from (45) in the above expression, to get

$$\begin{aligned}
&\leq \sum_{l=1}^{T} L_{\Psi^{(l)}} \Bigg(4\frac{\varepsilon_1}{\sqrt{N}\mathrm{d}_{min}^2} \|W\|_\infty \big(L_{\Phi^{(l)}} (B_1^{(l-1)} + B_2^{(l-1)} ||f||_\infty) + \|\Phi^{(l)}(0,0)\|_\infty\big) + N^{-\frac{1}{2(D_\chi+1)}} \\
&\cdot \Big(\frac{2\|W\|_\infty}{\mathrm{d}_{min}} L_{\Phi^{(l)}} (Z_1^{(l-1)} + Z_2^{(l-1)} \|f\|_\infty + Z_3^{(l-1)} L_f) + \frac{2L_W}{\mathrm{d}_{min}} \big(B_1^{(l-1)} + B_2^{(l-1)} ||f||_\infty\big) \\
&+ \frac{C_\chi}{\sqrt{2}} \big(\frac{\|W\|_\infty}{\mathrm{d}_{min}} \big(L_{\Phi^{(l)}} (B_1^{(l-1)} + B_2^{(l-1)} ||f||_\infty) + \|\Phi^{(l)}(0,0)\|_\infty\big) \big) \\
&\sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)} \log(N) + \log(2/p)} \Big) \Bigg) \prod_{l'=l+1}^{T} K^{(l')}.
\end{aligned} \tag{47}$$

We insert the bound for ε_1, defined in (22) as

$$\varepsilon_1 = \zeta\Big(L_W \big(\sqrt{\log(C_\chi)} + \sqrt{D_\chi}\big) + \big(\sqrt{2}\|W\|_\infty + L_W\big) \sqrt{\log 2/p} \Big),$$

into (47) to get

$$
\begin{aligned}
&\leq \sum_{l=1}^{T} L_{\Psi^{(l)}} \Bigg(4 \frac{\zeta\Big(L_W\big(\sqrt{\log(C_\chi)}+\sqrt{D_\chi}\big)+\big(\sqrt{2}\|W\|_\infty + L_W\big)\sqrt{\log 2/p}\Big)}{\sqrt{N}\mathrm{d}_{min}^2} \\
&\cdot \|W\|_\infty \big(L_{\Phi^{(l)}}(B_1^{(l-1)}+B_2^{(l-1)}||f||_\infty)+\|\Phi^{(l)}(0,0)\|_\infty\big) \\
&+ N^{-\frac{1}{2(D_\chi+1)}} \Big(\frac{2\|W\|_\infty}{\mathrm{d}_{min}} L_{\Phi^{(l)}}(Z_1^{(l-1)}+Z_2^{(l-1)}\|f\|_\infty + Z_3^{(l-1)}L_f) \\
&+ \frac{2L_W}{\mathrm{d}_{min}}\big(B_1^{(l-1)}+B_2^{(l-1)}||f||_\infty\big) \\
&+ \frac{C_\chi}{\sqrt{2}}\big(\frac{\|W\|_\infty}{\mathrm{d}_{min}}\big(L_{\Phi^{(l)}}(B_1^{(l-1)}+B_2^{(l-1)}||f||_\infty)+\|\Phi^{(l)}(0,0)\|_\infty\big)\big) \\
&\cdot \sqrt{\log(C_\chi)+\frac{D_\chi}{2(D_\chi+1)}\log(N)+\log(2/p)}\Big)\Bigg) \prod_{l'=l+1}^{T} K^{(l')}.
\end{aligned}
$$

Then, rearranging the terms yields

$$
\begin{aligned}
&= \sum_{l=1}^{T} L_{\Psi^{(l)}} 4 \frac{\zeta L_W\big(\sqrt{\log(C_\chi)}+\sqrt{D_\chi}\big)}{\sqrt{N}\mathrm{d}_{min}^2}\|W\|_\infty\big(L_{\Phi^{(l)}}B_1^{(l-1)}+\|\Phi^{(l)}(0,0)\|_\infty\big)\prod_{l'=l+1}^{T} K^{(l')} \\
&+ \sum_{l=1}^{T} L_{\Psi^{(l)}} 4 \frac{\zeta(\sqrt{2}\|W\|_\infty + L_W)\sqrt{\log 2/p}}{\sqrt{N}\mathrm{d}_{min}^2}\|W\|_\infty\big(L_{\Phi^{(l)}}B_1^{(l-1)}+\|\Phi^{(l)}(0,0)\|_\infty\big)\prod_{l'=l+1}^{T} K^{(l')} \\
&+ \sum_{l=1}^{T} L_{\Psi^{(l)}} 4 \frac{\zeta L_W\big(\sqrt{\log(C_\chi)}+\sqrt{D_\chi}\big)}{\sqrt{N}\mathrm{d}_{min}^2}\|W\|_\infty\big(L_{\Phi^{(l)}}B_2^{(l-1)}\|f\|_\infty\big)\prod_{l'=l+1}^{T} K^{(l')} \\
&+ \sum_{l=1}^{T} L_{\Psi^{(l)}} 4 \frac{\zeta(\sqrt{2}\|W\|_\infty + L_W)\sqrt{\log 2/p}}{\sqrt{N}\mathrm{d}_{min}^2}\|W\|_\infty\big(L_{\Phi^{(l)}}B_2^{(l-1)}\|f\|_\infty\big)\prod_{l'=l+1}^{T} K^{(l')} \\
&+ \sum_{l=1}^{T} L_{\Psi^{(l)}} N^{-\frac{1}{2(D_\chi+1)}}\Big(\frac{2\|W\|_\infty}{\mathrm{d}_{min}}L_{\Phi^{(l)}}Z_1^{(l-1)}+\frac{2L_W}{\mathrm{d}_{min}}B_1^{(l-1)}\Big)\prod_{l'=l+1}^{T} K^{(l')} \\
&+ \sum_{l=1}^{T} L_{\Psi^{(l)}} N^{-\frac{1}{2(D_\chi+1)}}\Big(\frac{2\|W\|_\infty}{\mathrm{d}_{min}}L_{\Phi^{(l)}}Z_2^{(l-1)}\|f\|_\infty+\frac{2L_W}{\mathrm{d}_{min}}B_2^{(l-1)}\|f\|_\infty\Big)\prod_{l'=l+1}^{T} K^{(l')} \\
&+ \sum_{l=1}^{T} L_{\Psi^{(l)}} N^{-\frac{1}{2(D_\chi+1)}} 2\frac{\|W\|_\infty}{\mathrm{d}_{min}}L_{\Phi^{(l)}}Z_3^{(l-1)}L_f\prod_{l'=l+1}^{T} K^{(l')} \\
&+ \sum_{l=1}^{T} L_{\Psi^{(l)}} N^{-\frac{1}{2(D_\chi+1)}}\frac{C_\chi}{\sqrt{2}}\frac{\|W\|_\infty}{\mathrm{d}_{min}}(L_{\Phi^{(l)}}B_1^{(l-1)}+\|\Phi^{(l)}(0,0)\|_\infty) \\
&\cdot \sqrt{\log(C_\chi)+\frac{D_\chi}{2(D_\chi+1)}\log(N)+\log(2/p)}\Big)\prod_{l'=l+1}^{T} K^{(l')} \\
&+ \sum_{l=1}^{T} L_{\Psi^{(l)}} N^{-\frac{1}{2(D_\chi+1)}}\frac{C_\chi}{\sqrt{2}}\frac{\|W\|_\infty}{\mathrm{d}_{min}}L_{\Phi^{(l)}}B_2^{(l-1)}\|f\|_\infty \\
&\cdot \sqrt{\log(C_\chi)+\frac{D_\chi}{2(D_\chi+1)}\log(N)+\log(2/p)}\Big)\prod_{l'=l+1}^{T} K^{(l')}
\end{aligned}
$$

$$
\begin{aligned}
&=: \Omega_1 \frac{1}{\sqrt{N}} + \Omega_2 \frac{\log(2/p)}{\sqrt{N}} + \Omega_3 \frac{\|f\|_\infty}{\sqrt{N}} + \Omega_4 \frac{\|f\|_\infty \log(2/p)}{\sqrt{N}} \\
&+ N^{-\frac{1}{2(D_\chi+1)}} \left(\Omega_5 + \Omega_6 \|f\|_\infty + \Omega_7 L_f\right) \\
&+ N^{-\frac{1}{2(D_\chi+1)}} \sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)} \log(N) + \log(2/p)} \cdot (\Omega_8 + \Omega_9 \|f\|_\infty),
\end{aligned}
$$

where we define

$$
\begin{aligned}
\Omega_1 &= \sum_{l=1}^{T} L_{\Psi^{(l)}} 4 \frac{\zeta L_W \left(\sqrt{\log(C_\chi)} + \sqrt{D_\chi}\right)}{\mathrm{d}_{min}^2} \|W\|_\infty \left(L_{\Phi^{(l)}} B_1^{(l-1)} + \|\Phi^{(l)}(0,0)\|_\infty\right) \prod_{l'=l+1}^{T} K^{(l')} \\
\Omega_2 &= \sum_{l=1}^{T} L_{\Psi^{(l)}} 4 \frac{\zeta(\sqrt{2}\|W\|_\infty + L_W)}{\mathrm{d}_{min}^2} \|W\|_\infty \left(L_{\Phi^{(l)}} B_1^{(l-1)} + \|\Phi^{(l)}(0,0)\|_\infty\right) \prod_{l'=l+1}^{T} K^{(l')} \\
\Omega_3 &= \sum_{l=1}^{T} L_{\Psi^{(l)}} 4 \frac{\zeta L_W \left(\sqrt{\log(C_\chi)} + \sqrt{D_\chi}\right)}{\mathrm{d}_{min}^2} \|W\|_\infty \left(L_{\Phi^{(l)}} B_2^{(l-1)}\right) \prod_{l'=l+1}^{T} K^{(l')} \\
\Omega_4 &= \sum_{l=1}^{T} L_{\Psi^{(l)}} 4 \frac{\zeta(\sqrt{2}\|W\|_\infty + L_W)\sqrt{\log 2/p}}{\mathrm{d}_{min}^2} \|W\|_\infty \left(L_{\Phi^{(l)}} B_2^{(l-1)}\right) \prod_{l'=l+1}^{T} K^{(l')} \\
\Omega_5 &= \sum_{l=1}^{T} L_{\Psi^{(l)}} \left(\frac{2\|W\|_\infty}{\mathrm{d}_{min}} L_{\Phi^{(l)}} Z_1^{(l-1)} + \frac{2L_W}{\mathrm{d}_{min}} B_1^{(l-1)}\right) \prod_{l'=l+1}^{T} K^{(l')} \\
\Omega_6 &= \sum_{l=1}^{T} L_{\Psi^{(l)}} \left(\frac{2\|W\|_\infty}{\mathrm{d}_{min}} L_{\Phi^{(l)}} Z_2^{(l-1)} + \frac{2L_W}{\mathrm{d}_{min}} B_2^{(l-1)}\right) \prod_{l'=l+1}^{T} K^{(l')} \\
\Omega_7 &= \sum_{l=1}^{T} L_{\Psi^{(l)}} 2 \frac{\|W\|_\infty}{\mathrm{d}_{min}} L_{\Phi^{(l)}} Z_3^{(l-1)} \prod_{l'=l+1}^{T} K^{(l')} \\
\Omega_8 &= \sum_{l=1}^{T} L_{\Psi^{(l)}} \frac{C_\chi}{\sqrt{2}} \frac{\|W\|_\infty}{\mathrm{d}_{min}} \left(L_{\Phi^{(l)}} B_1^{(l-1)} + \|\Phi^{(l)}(0,0)\|_\infty\right) \prod_{l'=l+1}^{T} K^{(l')} \\
\Omega_9 &= \sum_{l=1}^{T} L_{\Psi^{(l)}} \frac{C_\chi}{\sqrt{2}} \frac{\|W\|_\infty}{\mathrm{d}_{min}} L_{\Phi^{(l)}} B_2^{(l-1)} \prod_{l'=l+1}^{T} K^{(l')},
\end{aligned}
\tag{48}
$$

where $Z_1^{(l-1)}, Z_2^{(l-1)}, Z_3^{(l-1)}$ are defined in (31), $B_1^{(l-1)}$ and $B_2^{(l-1)}$ are defined in (27) and (28), and

$$
K^{(l')} = \sqrt{(L_{\Psi^{(l')}})^2 + \frac{8\|W\|_\infty^2}{\mathrm{d}_{\min}^2} (L_{\Phi^{(l')}})^2 (L_{\Psi^{(l')}})^2}.
$$

□

Next we study the convergence of MPNNs after global pooling. We give the following lemma.

Lemma B.16. *Let (χ, d, μ) be a metric-measure space and W be a kernel s.t. Assumptions A.10.1-6. are satisfied. Suppose that $X_1, \ldots, X_N$ are drawn i.i.d. from μ on χ such that $(X_1, \ldots, X_N) \in \mathcal{E}_{\mathrm{Lip}}^p$, where the event $\mathcal{E}_{\mathrm{Lip}}^p$ is defined in Lemma B.3. Then, for every MPNN Θ satisfying Assumption A.10.7*

and $f : \chi \to \mathbb{R}^F$ with Lipschitz constant L_f,

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{i=1}^{N} \big(S^X \Theta_W(f)\big)(X_i) - \int_\chi \Theta_W(f)(y) d\mu(y) \right\|_\infty \\
&\leq N^{-\frac{1}{2(D_\chi+1)}} \Bigg(2(Z_1^{(T)} + Z_2^{(T)} \|f\|_\infty + Z_3^{(T)} L_f) + \frac{C_\chi}{\sqrt{2}} (B_1^{(T)} + B_2^{(T)} \|f\|_\infty) \\
&\cdot \sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)} \log(N) + \log(2/p)} \Bigg).
\end{aligned} \tag{49}$$

Here, $Z_1^{(T)}, Z_2^{(T)}, Z_3^{(T)}$ and $B_1^{(T)}, B_2^{(T)}$ are defined in (45) and (46).

Proof. By Lemma B.7, we have

$$\|\Theta_W^{(T)}(f)\|_\infty \leq B_1^{(T)} + \|f\|_\infty B_2^{(T)}$$

and, by Corollary B.8, we have

$$L_{\Theta_W^{(T)}(f)} \leq Z_1^{(T)} + Z_2^{(T)} \|f\|_\infty + Z_3^{(T)} L_f$$

for all MPNNs Θ and metric-space signals f considered. Hence, by Lemma B.3, equation (49) holds. □

Corollary B.17. *Let (χ, d, μ) be a metric-measure space and W be a kernel s.t. Assumptions A.10.1-6. are satisfied. Consider a graph $(G, \mathbf{f}) \sim (W, f)$ with N nodes and corresponding graph features, where N satisfies (14). If the event $\mathcal{F}_{\text{Lip}}^p$ from Lemma B.5 occurs, then for every MPNN Θ satisfying Assumption A.10.7 and every $f : \chi \to \mathbb{R}^F$ with Lipschitz constant L_f,*

$$\begin{aligned}
\left\| \Theta_G^P(\mathbf{f}) - \Theta_W^P(f) \right\|_\infty^2 \leq \frac{S_1 + S_2 \|f\|_\infty^2}{N} + \frac{R_1 + R_2 \|f\|_\infty^2 + R_3 L_f^2}{N^{\frac{1}{D_\chi+1}}} + \frac{T_1 + T_2 \|f\|_\infty^2}{N^{\frac{1}{D_\chi+1}}} \log(N) \\
+ \frac{S_3 + S_4 \|f\|_\infty^2}{N} \log^2(2/p) + \frac{R_4 + R_5 \|f\|_\infty^2}{N^{\frac{1}{D_\chi+1}}} \log(2/p),
\end{aligned}$$

where the constants are defined in (51) below.

Proof. We have

$$
\begin{aligned}
&\left\|\Theta_G^P(\mathbf{f}) - \Theta_W^P(f)\right\|_\infty \\
&= \|\frac{1}{N}\sum_{i=1}^N \Theta_G(\mathbf{f})(X_i) - \int_\chi \Theta_W(f)(y)d\mu(y)\|_\infty \\
&\le \|\frac{1}{N}\sum_{i=1}^N \Theta_G(\mathbf{f})(X_i) - \frac{1}{N}\sum_{i=1}^N \big(S^X\Theta_W(f)\big)(X_i)\|_\infty \\
&+ \|\frac{1}{N}\sum_{i=1}^N \big(S^X\Theta_W(f)\big)(X_i) - \int_\chi \Theta_W(f)(y)d\mu(y)\|_\infty \\
&\le \frac{1}{N}\sum_{i=1}^N \|\Theta_G(\mathbf{f})(X_i) - \big(S^X\Theta_W(f)\big)(X_i)\|_\infty \\
&+ \|\frac{1}{N}\sum_{i=1}^N \big(S^X\Theta_W(f)\big)(X_i) - \int_\chi \Theta_W(f)(y)d\mu(y)\|_\infty \\
&\le \frac{1}{N}\sum_{i=1}^N \|\Theta_G(\mathbf{f})(X_i) - \big(S^X\Theta_W(f)\big)(X_i)\|_\infty \\
&+ N^{-\frac{1}{2(D_\chi+1)}}\Bigg(2(Z_1^{(T)} + Z_2^{(T)}\|f\|_\infty + Z_3^{(T)}L_f) + \frac{C_\chi}{\sqrt{2}}(B_1^{(T)} + B_2^{(T)}\|f\|_\infty) \\
&\cdot \sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)}\log(N) + \log(2/p)}\Bigg) \\
&= \mathrm{dist}\big(\Theta_G(\mathbf{f}), \Theta_W(f)\big) \\
&+ N^{-\frac{1}{2(D_\chi+1)}}\Bigg(2(Z_1^{(T)} + Z_2^{(T)}\|f\|_\infty + Z_3^{(T)}L_f) + \frac{C_\chi}{\sqrt{2}}(B_1^{(T)} + B_2^{(T)}\|f\|_\infty) \\
&\cdot \sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)}\log(N) + \log(2/p)}\Bigg),
\end{aligned}
$$

where the last inequality holds by Lemma B.16. Together with Theorem B.15, we get

$$
\begin{aligned}
&\left\|\Theta_G^P(\mathbf{f}) - \Theta_W^P(f)\right\|_\infty \\
&\le \frac{\Omega_1 + \Omega_2\log(2/p) + \Omega_3\|f\|_\infty + \Omega_4\|f\|_\infty\log(2/p)}{N^{\frac{1}{2}}} \\
&+ \frac{\Omega_5 + \Omega_6\|f\|_\infty + \Omega_7 L_f}{N^{\frac{1}{2(D_\chi+1)}}} \\
&+ \frac{\Omega_8 + \Omega_9\|f\|_\infty}{N^{\frac{1}{2(D_\chi+1)}}}\sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)}\log(N) + \log(2/p)} \\
&+ N^{-\frac{1}{2(D_\chi+1)}}\Bigg(2(Z_1^{(T)} + Z_2^{(T)}\|f\|_\infty + Z_3^{(T)}L_f) + \frac{C_\chi}{\sqrt{2}}(B_1^{(T)} + B_2^{(T)}\|f\|_\infty) \\
&\cdot \sqrt{\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)}\log(N) + \log(2/p)}\Bigg).
\end{aligned} \tag{50}
$$

Now we use the inequality

$$
\left(\sum_{i=1}^n a_i\right)^2 \le n\sum_{i=1}^n a_i^2
$$

for any $a_i \in \mathbb{R}_+$, $i = 1, \dots, N$, and square both sides of (50) to get

$$
\begin{aligned}
&\left\|\Theta_G^P(\mathbf{f}) - \Theta_W^P(f)\right\|_\infty^2 \\
&\leq 14\frac{\Omega_1^2 + \Omega_3^2\|f\|_\infty^2}{N} + 14\frac{\Omega_5^2 + \Omega_6^2\|f\|_\infty^2 + \Omega_7^2 L_f^2}{N^{\frac{1}{D_\chi+1}}} \\
&+ 14\frac{\Omega_8^2 + \Omega_9^2\|f\|_\infty^2}{N^{\frac{1}{D_\chi+1}}}\left(\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)}\log(N)\right) \\
&+ 56\frac{(Z_1^{(T)})^2 + (Z_2^{(T)})^2\|f\|_\infty^2 + (Z_3^{(T)})^2 L_f^2}{N^{\frac{1}{D_\chi+1}}} \\
&+ 7\frac{\left(C_\chi^2(B_1^{(T)})^2 + C_\chi^2(B_2^{(T)})^2\|f\|_\infty^2\right)\left(\log(C_\chi) + \frac{D_\chi}{2(D_\chi+1)}\log(N)\right)}{N^{\frac{1}{D_\chi+1}}} \\
&+ 14\frac{\left(\Omega_2^2 + \Omega_4^2\|f\|_\infty^2\right)\log^2(2/p)}{N} + 14\frac{\Omega_8^2 + \Omega_9^2\|f\|_\infty^2}{N^{\frac{1}{D_\chi+1}}}\log(2/p) \\
&+ 7\frac{\left(C_\chi^2(B_1^{(T)})^2 + C_\chi^2(B_2^{(T)})^2\|f\|_\infty^2\right)\log(2/p)}{N^{\frac{1}{D_\chi+1}}} \\
&=: \frac{S_1 + S_2\|f\|_\infty^2}{N} + \frac{R_1 + R_2\|f\|_\infty^2 + R_3 L_f^2}{N^{\frac{1}{D_\chi+1}}} + \frac{T_1 + T_2\|f\|_\infty^2}{N^{\frac{1}{D_\chi+1}}}\log(N) \\
&+ \frac{S_3 + S_4\|f\|_\infty^2}{N}\log^2(2/p) + \frac{R_4 + R_5\|f\|_\infty^2}{N^{\frac{1}{D_\chi+1}}}\log(2/p),
\end{aligned}
$$

where

$$
\begin{aligned}
S_1 &= 14\Omega_1^2 \\
S_2 &= 14\Omega_3^2 \\
S_3 &= 14\Omega_2^2 \\
S_4 &= 14\Omega_4^2 \\
R_1 &= 14\Omega_5^2 + 14\Omega_8^2\log(C_\chi) + 56(Z_1^{(T)})^2 + 7C_\chi^2(B_1^{(T)})^2\log(C_\chi) \\
R_2 &= 14\Omega_6^2 + 14\Omega_9^2\log(C_\chi) + 56(Z_2^{(T)})^2 + 7C_\chi^2(B_2^{(T)})^2\log(C_\chi) \\
R_3 &= 14\Omega_7^2 + 56(Z_3^{(T)})^2 \\
R_4 &= 14\Omega_8^2 + 7C_\chi^2(B_1^{(T)})^2 \\
R_5 &= 14\Omega_9^2 + 7C_\chi^2(B_2^{(T)})^2 \\
T_1 &= 14\Omega_8^2\frac{D_\chi}{2(D_\chi+1)} + 7C_\chi^2(B_1^{(T)})^2\frac{D_\chi}{2(D_\chi+1)} \\
T_2 &= 14\Omega_9^2\frac{D_\chi}{2(D_\chi+1)} + 7C_\chi^2(B_2^{(T)})^2\frac{D_\chi}{2(D_\chi+1)},
\end{aligned}
\tag{51}
$$

and $\Omega_1, \dots, \Omega_9$ are defined in (48), and $B_1^{(T)}$ and $B_2^{(T)}$ are defined in (27) and (28). □

We now write a version of Theorem 3.1 (about the convergence error of MPNNs) with detailed constants, and prove it.

Theorem B.18. *Let (χ, d, μ) be a metric-measure space and W be a kernel s.t. Assumptions A.10.1-6. and Assumptions A.10.8 are satisfied. Consider a graph $(G, \mathbf{f}) \sim (W, f)$ with N nodes and*

corresponding graph features. Then, for every $f : \chi \to \mathbb{R}^F$ with Lipschitz constant L_f,

$$\mathbb{E}_{X_1,\dots,X_N\sim\mu^N}\left[\sup_{\Theta\in\mathrm{Lip}_{L,B}}\left\|\Theta_G^P(\mathbf{f})-\Theta_W^P(f)\right\|_\infty^2\right]$$
$$\leq 6\sqrt{\pi}\left(\frac{S_1+S_3+(S_2+S_4)\|f\|_\infty^2}{N}+\frac{R_1+R_4+(R_2+R_5)\|f\|_\infty^2+R_3L_f^2}{N^{\frac{1}{D_\chi+1}}}\right.$$
$$\left.+\frac{\left(T_1+T_2\|f\|_\infty^2\right)\log(N)}{N^{\frac{1}{D_\chi+1}}}\right)+\mathcal{O}\left(\exp(-N)N^{\frac{3}{2}T-\frac{3}{2}}\right),$$

where the constants are defined in (51).

Proof. For any $p>0$, we have with probability at least $1-2p$ for every $\Theta\in\mathrm{Lip}_{L,B}$, by Corollary B.17, that

$$\left\|\Theta_G^P(\mathbf{f})-\Theta_W^P(f)\right\|_\infty^2\leq H_1+H_2\log(2/p)+H_3\log^2(2/p)$$

if (14) holds, where

$$H_1=\frac{S_1+S_2\|f\|_\infty^2}{N}+\frac{R_1+R_2\|f\|_\infty^2+R_3L_f^2}{N^{\frac{1}{D_\chi+1}}}+\frac{T_1+T_2\|f\|_\infty^2}{N^{\frac{1}{D_\chi+1}}}\log(N),$$
$$H_2=\frac{R_4+R_5\|f\|_\infty^2}{N^{\frac{1}{D_\chi+1}}}\text{ and }H_3=\frac{S_3+S_4\|f\|_\infty^2}{N}.$$

Further, for every $p\in(0,1/2)$, we consider $k>0$ such that $p=2\exp(-k^2)$. This means, if p respectively k satisfies (14), we have with probability at least $1-4\exp(-k^2)$ for every $\Theta\in\mathrm{Lip}_{L,B}$,

$$\left\|\Theta_G^P(\mathbf{f})-\Theta_W^P(f)\right\|_\infty^2\leq H_1+H_2k+H_3k^2.$$

If k does not satisfy (14), we get

$$k>N_0=D_1+D_2\sqrt{N},$$

where $D_1\in\mathbb{R}$ and $D_2>0$ are the matching constants in (14). By Lemma B.10 and Lemma B.7, we get in this case

$$\begin{aligned}\left\|\Theta_G^P(\mathbf{f})-\Theta_W^P(f)\right\|_\infty^2&=\left\|\frac{1}{N}\sum_{i=1}^N\Theta_G(\mathbf{f})_i-\int_\chi\Theta_W(f)(y)d\mu(y)\right\|_\infty^2\\&\leq\frac{4}{N}\sum_{i=1}^N\|\Theta_G(\mathbf{f})_i\|_\infty^2+2\left\|\int_\chi\Phi_W(f)(y)d\mu(y)\right\|_\infty^2\\&\leq\frac{4}{N}\|\Theta_G(\mathbf{f})\|_{2;\infty}^2+2\|\Theta_W(f)\|_\infty^2\\&\leq\frac{4}{N}N^T(A'+A''\|f\|_\infty^2)+2(B_1^{(T)}+\|f\|_\infty B_2^{(T)})^2=:q(N),\end{aligned}\tag{52}$$

where the first inequality holds by applying the triangle inequality and Cauchy-Schwarz.

We then calculate the expected value by partitioning the integral over the event space into the following sum.

$$\begin{aligned}&\mathbb{E}_{X_1,\dots,X_N\sim\mu^N}\left[\sup_{\Theta\in\mathrm{Lip}_{L,B}}\left\|\Theta_G^P(\mathbf{f})-\Theta_W^P(f)\right\|_\infty^2\right]\\&\leq\sum_{k=0}^{N_0}\mathbb{P}\big(H_1+H_2k+H_3k^2\leq\sup_{\Theta\in\mathrm{Lip}_{L,B}}\left\|\Theta_G^P(\mathbf{f})-\Theta_W^P(f)\right\|_\infty^2<H_1+H_2(k+1)+H_3(k+1)^2\big)\\&\quad\cdot\big(H_1+H_2(k+1)+H_3(k+1)^2\big)\\&+\sum_{k=N_0}^{\infty}\mathbb{P}\big(H_1+H_2k+H_3k^2\leq\sup_{\Theta\in\mathrm{Lip}_{L,B}}\left\|\Theta_G^P(\mathbf{f})-\Theta_W^P(f)\right\|_\infty^2<H_1+H_2(k+1)+H_3(k+1)^2\big)\\&\quad\cdot q(N)\end{aligned}\tag{53}$$

To bound the second sum, note that it is a finite sum, since $\left\|\Theta_G^P(\mathbf{f}) - \Theta_W^P(f)\right\|_\infty^2$ is bounded by $q(N)$, which is defined in (52). The summands are zero if $H_1 + H_2 k + H_3 k^2 > q(N)$, which holds for $k > \sqrt{\frac{q(N)}{H_3}}$. Hence, we calculate with the right-hand-side of (53) by

$$\begin{aligned}
&\leq 2\sum_{k=0}^{N_0} 2\exp(-k^2)\cdot\left(H_1 + H_2(k+1) + H_3(k+1)^2\right) \\
&+ \sum_{k=N_0}^{\left\lceil\sqrt{\frac{q(N)}{H_3}}\right\rceil} 4\exp(-N_0^2)\cdot q(N) \\
&\leq 2\int_0^\infty 2\exp(-k^2)\cdot\left(H_1 + H_2(k+1) + H_3(k+1)^2\right) \\
&+ 4\exp(-N_0^2)q(N)\left\lceil\sqrt{\frac{q(N)}{H_3}}\right\rceil,
\end{aligned} \tag{54}$$

where $q(N) = O(N^{T-1})$ is a polynomial in N as defined above. The first term on the right-hand-side is bounded by using

$$\int_0^\infty 2(t+1)^2 e^{-t^2}dt,\ \int_0^\infty 2(t+1)e^{-t^2}dt,\ \int_0^\infty 2e^{-t^2}dt \leq 3\sqrt{\pi}.$$

For the second term we remember that $N_0 = D_1 + D_2\sqrt{N}$. Hence,

$$\begin{aligned}
&\mathbb{E}_{X_1,\ldots,X_N\sim\mu^N}\left[\sup_{\Theta\in\mathrm{Lip}_{L,B}}\left\|\Theta_G^P(\mathbf{f}) - \Theta_W^P(f)\right\|_\infty^2\right] \\
&\leq 6\sqrt{\pi}(H_1 + H_2 + H_3) + \mathcal{O}(\exp(-N)N^{\frac{3}{2}T-\frac{3}{2}}).
\end{aligned}$$

□

C Generalization Analysis

In this section, we provide details on our generalization analysis of MPNNs. In Subsection C.1, we detail the data distribution from the graph classification task, which was introduced in Subsection 2.4. In Subsection C.2, we provide a detailed version and a proof for Theorem 3.3 (about the generalization bound of MPNNs). This is followed by a derivation of the asymptotics of our generalization bound in Subsection C.3 and a comparison of the asymptotics of our generalization bound with other related generalization bounds in Subsection C.4.

C.1 The Probability Space of the Dataset

Recall that the measure on the space χ^j is denoted by μ^j. Given a class j and $N \in \mathbb{N}$, the space of graphs with N nodes from class j is defined to be $(\chi^j)^N$. The measure on $(\chi^j)^N$ is defined to be $(\mu^j)^N$, namely, the direct product of the measure μ^j with itself N times. The space $\mathcal{G}_j$ of graphs of any size, which are sampled from class j, is defined to be

$$\mathcal{G}_j := \bigcup_{n\in\mathbb{N}} (\chi^j)^N.$$

The measure on $\mathcal{G}_j$ is denoted by $\mu_{\mathcal{G}_j}$, and defined as follows.

Definition C.1. *A set of graphs $S \subset \mathcal{G}_j$ is called measurable, if for each $N \in \mathbb{N}$, the restriction*

$$S_N := \{G \in S \mid G \text{ has } N \text{ nodes}\} \subset (\chi^j)^N$$

is measurable with respect to $(\mu^j)^N$. The measure of a measurable set $S \subset \mathcal{G}_j$ is defined to be

$$\mu_{\mathcal{G}_j}(S) := \sum_{N=1}^\infty \nu(N)(\mu^j)^N(S_N),$$

where $\nu(N)$ is the probability of choosing a graph with N nodes (see Subsection 2.4).

The space of graphs of either of the classes $j = 1, \ldots, \Gamma$ is defined to be

$$\mathcal{G} := \bigcup_{j=1}^{\Gamma} \mathcal{G}_j.$$

The measure on $\mathcal{G}$ is denoted by $\mu_{\mathcal{G}}$, and defined as follows.

Definition C.2. *A set of graphs $S \subset \mathcal{G}$ is called measurable, if for each $j = 1, \ldots, \Gamma$, the restriction*

$$S_j := \{G \in S \mid G \text{ is sampled from class } j\} \subset \mathcal{G}_j$$

is measurable with respect to $\mu_{\mathcal{G}_j}$. The measure of a measurable $S \subset \mathcal{G}$ is defined to be

$$\mu_{\mathcal{G}}(S) = \sum_{j=1}^{\Gamma} \gamma_j \mu_{\mathcal{G}_j}(S_j),$$

where γ_j is the probability of choosing class j (see Subsection 2.4).

With these notations, the space of graph datasets of size m is defined to be $\mathcal{G}^m$ with the direct product measure $\mu_{\mathcal{G}}^m$. We denote a random graph sampled from the space of graphs by $(G, \mathbf{f}, y) \sim \mu_{\mathcal{G}}$. Here, y denotes the class of the graph, namely, the value y such that $(G, \mathbf{f})$ is sampled from class y.

The next lemma is direct, and given without proof.

Lemma C.3. *The spaces $\{\mathcal{G}, \mu_{\mathcal{G}}\}$ and $\{\mathcal{G}_j, \mu_{\mathcal{G}_j}\}$, $j = 1, \ldots, \Gamma$, are measure spaces, and $\mu_{\mathcal{G}}$ and $\mu_{\mathcal{G}_j}$, $j = 1, \ldots, \Gamma$, are probability measures.*

Let us next derive a re-parameterization of the space of datasets $\mathcal{G}^m$. Given $\mathcal{T} \sim \mu_{\mathcal{G}}^m$, for every $j = 1, \ldots, \Gamma$, let m_j denote the number of graphs in $\mathcal{T}$ that fall into the class j. Note that $\mathbf{m} = (m_1, \ldots, m_\Gamma)$ has a multinomial distribution with parameters m and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_\Gamma)$, which we denote by $\mathrm{MN}_{m,\boldsymbol{\gamma}}$. Conditioning the choice of the graphs on the choice of $\mathbf{m}$, we can formulate the data sampling procedure as first sampling $\mathbf{m}$ from $\mathrm{MN}_{m,\boldsymbol{\gamma}}$, and then sampling $\{G_i^j, \mathbf{f}_i^j\}_{i=1}^{m_j} \sim (\mu_{\mathcal{G}_j})^{m_j}$, $j = 1 \ldots, \Gamma$ independently of each other. Now, the measure $\mu_{\mathcal{G}}^m$ of the space of datasets can be parameterized as follows.

First, we define the following measure space. Let $\mathbf{m} = (m_1, \ldots, m_\Gamma)$ satisfy $\sum_{j=1}^{\Gamma} m_j = m$. We define the space

$$\mathcal{G}^{\mathbf{m}} := \prod_{j=1}^{\Gamma} \mathcal{G}_j^{m_j},$$

with the measure

$$\mu_{\mathcal{G}^{\mathbf{m}}} := \prod_{j=1}^{\Gamma} \mu_{\mathcal{G}_j}^{m_j}. \tag{55}$$

The space $\mathcal{G}^{\mathbf{m}}$ is interpreted as the space of datasets with exactly m_j samples in each class j.

We can now show the following parametrization of the measure space $\mathcal{G}^m$ of datasets of size m. The lemma is direct, and given without proof.

Lemma C.4. *A set of datasets $S \subset \mathcal{G}^m$ is measurable, if and only if for every $\mathbf{m} = (m_1, \ldots, m_\Gamma)$ with $\sum_{j=1}^{\Gamma} m_j = m$, the restriction*

$$S_{\mathbf{m}} = \{\mathcal{T} \in S \mid \forall 1 \leq j \leq \Gamma, \ \ \mathcal{T} \text{ contains } m_j \text{ graphs from class } j\} \subset \mathcal{G}^{\mathbf{m}}$$

is measurable with respect to $\mu_{\mathcal{G}^{\mathbf{m}}}$.

With these notations, μ_{G}^m is decomposed as follows: $\mathcal{G}^m = \bigcup_{\mathbf{m}} \mathcal{G}^{\mathbf{m}}$, and for every measurable set of datasets $S \subset \mathcal{G}^m$,

$$\mu_{\mathrm{G}}^m(S) = \sum_{\mathbf{m}:\ m_1 + \ldots + m_\Gamma = m} \mu_{\mathrm{MN}_{m,\boldsymbol{\gamma}}}(\mathbf{m}) \sum_{j=1}^{\Gamma} \sum_{i=1}^{m_j} \mu_{\mathrm{G}_j}(S_{\mathbf{m}}).$$

C.2 Proof of Theorem 3.3

The following corollary computes the expected robustness of a random graph, of arbitrary size, sampled from $\mu_{\mathcal{G}_j}$, and is a direct result of Definition C.1 and Theorem B.18.

Corollary C.5. *Let $\{(W^j, f^j)\}$ be a RGM on the corresponding metric-measure space (χ^j, d^j, μ^j) that satisfies Assumptions A.10.1.-6. and A.10.8. Let $\mu_{\mathcal{G}_j}$ be the distribution from Definition C.1. Then,*

$$\begin{aligned}
&\mathbb{E}_{(G^j,\mathbf{f}^j)\sim\mu_{\mathcal{G}_j}}\big[\sup_{\Theta\in\mathrm{Lip}_{L,B}}\left\|\Theta^P_{G^j}(\mathbf{f}^j)-\Theta^P_{W^j}(f^j)\right\|^2_\infty\big]\\
&\leq 6\sqrt{\pi}\Bigg(\big(S_1^{(j)}+S_3^{(j)}+(S_2^{(j)}+S_4^{(j)})\|f^j\|^2_\infty\big)\mathbb{E}_{N\sim\nu}\left[N^{-1}\right]\\
&+\big(R_1^{(j)}+R_4^{(j)}+(R_2^{(j)}+R_5^{(j)})\|f^j\|^2_\infty+R_3^{(j)}L^2_{f^j}\big)\mathbb{E}_{N\sim\nu}\left[N^{-\frac{1}{D_{\chi^j}+1}}\right]\\
&+\big(T_1^{(j)}+T_2^{(j)}\|f^j\|^2_\infty\big)\mathbb{E}_{N\sim\nu}\left[\log(N)N^{-\frac{1}{D_{\chi^j}+1}}\right]\Bigg)+\mathcal{O}\left(\mathbb{E}_{N\sim\nu}\left[\exp(-N)N^{\frac{3T}{2}-\frac{3}{2}}\right]\right),
\end{aligned}$$

where $S_l^{(j)}, R_l^{(j)}, T_l^{(j)}$ are the according constants from Theorem B.18 for each class j and are defined in (51).

When sampling a dataset $\mathcal{T}\sim p^m$, the numbers of samples m_j that fall in class χ^j, for $j=1,\dots,\Gamma$, are distributed multinomially. We hence recall a concentration of measure result for multinomial variables.

Lemma C.6 (Proposition A.6 in [Vaart and Wellner, 1996], Bretagnolle-Huber-Carol inequality)**.** *If the random vector $(m_1,\dots m_\Gamma)$ is multinomially distributed with parameters m and $\gamma_1,\dots,\gamma_\Gamma$, then*

$$\mathbb{P}\left(\sum_{i=1}^{\Gamma}|m_i-m\gamma_i|\geq 2\sqrt{m}\lambda\right)\leq 2^\Gamma\exp(-2\lambda^2)$$

for any $\lambda>0$.

We now write a version of Theorem 3.3 (about the generalization error of MPNNs) with detailed constants, and prove it.

Theorem C.7. *Let $\{(W^j,f^j)\}_{j=1}^\Gamma$ be a collection of RGMs on corresponding metric-measure spaces $\{(\chi^j,d^j,\mu^j)\}_{j=1}^\Gamma$ such that each one satisfies Assumptions A.10.1.-6. and A.10.8. Let $\mu_{\mathcal{G}}$ denote the data distribution from Definition C.2. Let $\mathcal{T}=\big((G_1,\mathbf{f}_1,y_1),\dots,(G_m,\mathbf{f}_m,y_m)\big)\sim\mu_{\mathcal{G}}^m$ be a dataset of graphs. Then,*

$$\begin{aligned}
&\mathbb{E}_{\mathcal{T}\sim\mu_{\mathcal{G}}^m}\left[\sup_{\Theta\in\mathrm{Lip}_{L,B}}\left(\frac{1}{m}\sum_{i=1}^m\mathcal{L}(\Theta^P_{G_i}(\mathbf{f}_i),y_i)-\mathbb{E}_{(G,\mathbf{f},y)\sim\mu_{\mathcal{G}}}\left[\mathcal{L}(\Theta^P_G(\mathbf{f}),y)\right]\right)^2\right]\\
&\leq 2^\Gamma\frac{8\|\mathcal{L}\|^2_\infty}{m}\pi+\frac{6\sqrt{\pi}}{m}2^\Gamma\Gamma\sum_{j=1}^\Gamma\gamma_jL^2_{\mathcal{L}}\Bigg(\sqrt{\pi}\Big(\big(S_1^{(j)}+S_3^{(j)}+(S_2^{(j)}+S_4^{(j)})\|f^j\|^2_\infty\big)\mathbb{E}_{N\sim\nu}\left[N^{-1}\right]\\
&+\big(R_1^{(j)}+R_4^{(j)}+(R_2^{(j)}+R_5^{(j)})\|f^j\|^2_\infty+R_3^{(j)}L^2_{f^j}\big)\mathbb{E}_{N\sim\nu}\left[N^{-\frac{1}{D_{\chi^j}+1}}\right]\\
&+\big(T_1^{(j)}+T_2^{(j)}\|f^j\|^2_\infty\big)\mathbb{E}_{N\sim\nu}\left[\log(N)N^{-\frac{1}{D_{\chi^j}+1}}\right]\Big)+\mathcal{O}\left(\mathbb{E}_{N\sim\nu}\left[\exp(-N)N^{\frac{3}{2}T-\frac{3}{2}}\right]\right)\Bigg),
\end{aligned}$$

where $S_l^{(j)}, R_l^{(j)}, T_l^{(j)}$ are the according constants from Theorem B.18 for each class j and are defined in (51).

Proof. Given $\mathbf{m}=(m_1,\dots,m_\Gamma)$ with $\sum_{j=1}^\Gamma m_j=m$, recall that $\mathcal{G}^{\mathbf{m}}$ is the space of datasets with fixed number of samples m_j from each class $j=1,\dots,\Gamma$. The probability measure on $\mathcal{G}^{\mathbf{m}}$ is given

by $\mu_{\mathcal{G}^{\mathbf{m}}}$ (see (55)). Similarly to the notation of Lemma C.4, we denote the conditional choice of the dataset on the choice of $\mathbf{m}$ by

$$\mathcal{T}_{\mathbf{m}} := \left\{\{G_i^j, \mathbf{f}_i^j\}_{i=1}^{m_j}\right\}_{j=1}^{\Gamma} \sim \mu_{\mathcal{G}^{\mathbf{m}}}.$$

Given $k \in \mathbb{Z}$, denote by $\mathcal{M}_k$ the set of all $\mathbf{m} = (m_1, \ldots, m_\Gamma) \in \mathbb{N}_0^\Gamma$ with $\sum_{j=1}^{\Gamma} m_j = m$, such that $2\sqrt{m}k \leq \sum_{j=1}^{\Gamma} |m_j - m\gamma_j| < 2\sqrt{m}(k+1)$. Using these notations, we decompose the expected generalization error as follows.

$$\begin{aligned}
&\mathbb{E}_{\mathcal{T}\sim\mu_{\mathcal{G}}^m}\left[\sup_{\Theta\in\mathrm{Lip}_{L,B}}\left(\frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(\Theta_{G_i}^P(\mathbf{f}_i), y_i) - \mathbb{E}_{(G,\mathbf{f},y)\sim\mu_{\mathcal{G}}}\left[\mathcal{L}(\Theta_G^P(\mathbf{f}), y)\right]\right)^2\right] \\
&= \mathbb{E}_{\mathcal{T}\sim\mu_{\mathcal{G}}^m}\left[\sup_{\Theta\in\mathrm{Lip}_{L,B}}\left(\frac{1}{m}\sum_{j=1}^{\Gamma}\sum_{i=1}^{m_j}\mathcal{L}(\Theta_{G_i^j}^P(\mathbf{f}_i^j), y_j) - \mathbb{E}_{(G,\mathbf{f},y)\sim\mu_{\mathcal{G}}}\left[\mathcal{L}(\Theta_G^P(\mathbf{f}), y)\right]\right)^2\right] \\
&= \mathbb{E}_{\mathcal{T}\sim\mu_{\mathcal{G}}^m}\left[\sup_{\Theta\in\mathrm{Lip}_{L,B}}\left(\sum_{j=1}^{\Gamma}\left(\frac{1}{m}\sum_{i=1}^{m_j}\mathcal{L}(\Theta_{G_i^j}^P(\mathbf{f}_i^j), y_j) - \gamma_j\mathbb{E}_{(G^j,\mathbf{f}^j)\sim\mu_{\mathcal{G}_j}}\left[\mathcal{L}(\Theta_{G^j}^P(\mathbf{f}^j), y_j)\right]\right)\right)^2\right] \\
&\leq \sum_k \mathbb{P}\big(\mathbf{m}\in\mathcal{M}_k\big) \times \sup_{\mathbf{m}\in\mathcal{M}_k}\mathbb{E}_{\mathcal{T}_{\mathbf{m}}\sim\mu_{\mathcal{G}^{\mathbf{m}}}}\left[\sup_{\Theta\in\mathrm{Lip}_{L,B}}\left(\sum_{j=1}^{\Gamma}\left(\frac{1}{m}\sum_{i=1}^{m_j}\mathcal{L}(\Theta_{G_i^j}^P(\mathbf{f}_i^j), y_j)\right.\right.\right. \\
&\qquad\qquad\left.\left.\left. - \frac{1}{m}\sum_{i=1}^{m\gamma_j}\mathbb{E}_{(G^j,\mathbf{f}^j)\sim\mu_{\mathcal{G}_j}}\left[\mathcal{L}(\Theta_{G^j}^P(\mathbf{f}^j), y_j)\right]\right)\right)^2\right]
\end{aligned} \tag{56}$$

We bound the last term of (56) as follows. For $j = 1, \ldots, \Gamma$, if $m_j \leq m\gamma_j$, we add "ghost samples", i.e., we add additional i.i.d. sampled graphs $(G_{m_j}^j, \mathbf{f}_{m_j}^j), \ldots, (G_{m\gamma_j}^j, \mathbf{f}_{m\gamma_j}^j) \sim (W^j, f^j)$. By convention, for any two $l, q \in \mathbb{N}_0$ with $l < q$, we define

$$\sum_{j=q}^{l} c_j = -\sum_{j=l}^{q} c_j$$

for any sequence c_j of reals, and define $\sum_{j=q}^{q} c_j = 0$. With these notations, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{T}_{\mathbf{m}}\sim\mu_{\mathcal{G}^{\mathbf{m}}}}\Bigg[\sup_{\Theta\in\mathrm{Lip}_{L,B}}\Bigg(\sum_{j=1}^{\Gamma}\Bigg(\frac{1}{m}\sum_{i=1}^{m_j}\mathcal{L}(\Theta^P_{G_i^j}(\mathbf{f}_i^j),y_j)\\
&\qquad -\frac{1}{m}\sum_{i=1}^{m\gamma_j}\mathbb{E}_{(G^j,\mathbf{f}^j)\sim\mu_{\mathcal{G}_j}}\left[\mathcal{L}(\Theta^P_{G^j}(\mathbf{f}^j),y_j)\right]\Bigg)\Bigg)^2\Bigg]\\
&=\mathbb{E}_{\mathcal{T}_{\mathbf{m}}\sim\mu_{\mathcal{G}^{\mathbf{m}}}}\Bigg[\sup_{\Theta\in\mathrm{Lip}_{L,B}}\Bigg(\sum_{j=1}^{\Gamma}\Bigg(\frac{1}{m}\sum_{i=1}^{m\gamma_j}\mathcal{L}(\Theta^P_{G_i^j}(\mathbf{f}_i^j),y_j)+\frac{1}{m}\sum_{i=m\gamma_j}^{m_j}\mathcal{L}(\Theta^P_{G_i^j}(\mathbf{f}_i^j),y_j)\\
&-\frac{1}{m}\sum_{i=1}^{m\gamma_j}\mathbb{E}_{(G^j,\mathbf{f}^j)\sim\mu_{\mathcal{G}_j}}\left[\mathcal{L}(\Theta^P_{G^j}(\mathbf{f}^j),y_j)\right]\Bigg)\Bigg)^2\Bigg] \qquad (57)\\
&\le\mathbb{E}_{\mathcal{T}_{\mathbf{m}}\sim\mu_{\mathcal{G}^{\mathbf{m}}}}\Bigg[\sup_{\Theta\in\mathrm{Lip}_{L,B}}2\Bigg(\sum_{j=1}^{\Gamma}\Bigg(\frac{1}{m}\sum_{i=1}^{m\gamma_j}\mathcal{L}(\Theta^P_{G_i^j}(\mathbf{f}_i^j),y_j)\\
&\qquad -\frac{1}{m}\sum_{i=1}^{m\gamma_j}\mathbb{E}_{(G^j,\mathbf{f}^j)\sim\mu_{\mathcal{G}_j}}\left[\mathcal{L}(\Theta^P_{G^j}(\mathbf{f}^j),y_j)\right]\Bigg)\Bigg)^2\Bigg]\\
&+\mathbb{E}_{\mathcal{T}_{\mathbf{m}}\sim\mu_{\mathcal{G}^{\mathbf{m}}}}\left[2\left(\sum_{j=1}^{\Gamma}\left(\frac{1}{m}|m\gamma_j-m_j|\|\mathcal{L}\|_\infty\right)\right)^2\right].
\end{aligned}
$$

Let us first bound the last term of the above bound. Since any $\mathbf{m}\in\mathcal{M}_{\mathbf{k}}$ satisfies $\sum_{j=1}^{\Gamma}|m_j-m\gamma_j|<2\sqrt{m}(k+1)$, we have

$$
\begin{aligned}
\mathbb{E}_{\mathcal{T}_{\mathbf{m}}\sim\mu_{\mathcal{G}^{\mathbf{m}}}}\left[2\left(\sum_{j=1}^{\Gamma}\left(\frac{1}{m}|m\gamma_j-m_j|\|\mathcal{L}\|_\infty\right)\right)^2\right]&\le\frac{2}{m^2}\|\mathcal{L}\|_\infty^2\left(\sum_{j=1}^{\Gamma}|m\gamma_j-m_j|\right)^2\\
&\le\frac{2}{m^2}\|\mathcal{L}\|_\infty^2 4m(k+1)^2=\frac{8\|\mathcal{L}\|_\infty^2}{m}(k+1)^2.
\end{aligned}
$$

Hence, by Lemma C.6,

$$
\begin{aligned}
&\sum_k\mathbb{P}\big(\mathbf{m}\in\mathcal{M}_k\big)\times\sup_{\mathbf{m}\in\mathcal{M}_k}\mathbb{E}_{\mathcal{T}_{\mathbf{m}}\sim\mu_{\mathcal{G}^{\mathbf{m}}}}\left[2\left(\sum_{j=1}^{\Gamma}\left(\frac{1}{m}|m\gamma_j-m_j|\|\mathcal{L}\|_\infty\right)\right)^2\right]\\
&\le\sum_k\mathbb{P}\big(\mathbf{m}\in\mathcal{M}_k\big)\times\frac{8\|\mathcal{L}\|_\infty^2}{m}(k+1)^2\\
&\le\sum_k 2^\Gamma\exp(-2k^2)\frac{8\|\mathcal{L}\|_\infty^2}{m}(k+1)^2\\
&\le\int_0^\infty 2^\Gamma\exp(-2k^2)\frac{8\|\mathcal{L}\|_\infty^2}{m}(k+1)^2dk\\
&=2^\Gamma\frac{8\|\mathcal{L}\|_\infty^2}{m}\int_0^\infty\exp(-2k^2)(k+1)^2dk\\
&\le 2^\Gamma\frac{8\|\mathcal{L}\|_\infty^2}{m}\pi.
\end{aligned}
$$

To bound the first term of the right-hand-side of (57), we have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{T}_{\mathbf{m}}\sim\mu_{\mathcal{G}^{\mathbf{m}}}}\Bigg[\sup_{\Theta\in\mathrm{Lip}_{L,B}}\Bigg(\sum_{j=1}^{\Gamma}\Bigg(\frac{1}{m}\sum_{i=1}^{m\gamma_j}\mathcal{L}(\Theta^P_{G_i^j}(\mathbf{f}_i^j),y_j)\\
&\qquad\qquad -\frac{1}{m}\sum_{i=1}^{m\gamma_j}\mathbb{E}_{(G^j,\mathbf{f}^j)\sim\mu_{\mathcal{G}_j}}\Bigg[\sup_{\Theta\in\mathrm{Lip}_{L,B}}\mathcal{L}(\Theta^P_{G^j}(\mathbf{f}^j),y_j)\Bigg]\Bigg)\Bigg)^2\Bigg]\\
\leq&\Gamma\sum_{j=1}^{\Gamma}\mathbb{E}_{\mathcal{T}_{\mathbf{m}}\sim\mu_{\mathcal{G}^{\mathbf{m}}}}\Bigg[\sup_{\Theta\in\mathrm{Lip}_{L,B}}\Bigg(\frac{1}{m}\sum_{i=1}^{m\gamma_j}\mathcal{L}(\Theta^P_{G_i^j}(\mathbf{f}_i^j),y_j)\\
&\qquad\qquad -\frac{1}{m}\sum_{i=1}^{m\gamma_j}\mathbb{E}_{(G^j,\mathbf{f}^j)\sim\mu_{\mathcal{G}_j}}\Bigg[\sup_{\Theta\in\mathrm{Lip}_{L,B}}\mathcal{L}(\Theta^P_{G^j}(\mathbf{f}^j),y_j)\Bigg]\Bigg)^2\Bigg]\\
=&\Gamma\sum_{j=1}^{\Gamma}\mathrm{Var}_{(G^j,\mathbf{f}^j)\sim\mu_{\mathcal{G}_j}}\Bigg[\sup_{\Theta\in\mathrm{Lip}_{L,B}}\frac{1}{m}\sum_{i=1}^{\gamma_j\cdot m}\mathcal{L}(\Theta^P_{G^j}(\mathbf{f}^j),y_j)\Bigg]\\
=&\Gamma\sum_{j=1}^{\Gamma}\frac{\gamma_j}{m}\mathrm{Var}_{(G^j,\mathbf{f}^j)\sim\mu_{\mathcal{G}_j}}\Bigg[\sup_{\Theta\in\mathrm{Lip}_{L,B}}\mathcal{L}(\Theta^P_{G^j}(\mathbf{f}^j),y_j)\Bigg]\\
\leq&\Gamma\sum_{j=1}^{\Gamma}\frac{\gamma_j}{m}\mathbb{E}_{(G^j,\mathbf{f}^j)\sim\mu_{\mathcal{G}_j}}\Bigg[\sup_{\Theta\in\mathrm{Lip}_{L,B}}\left|\mathcal{L}(\Theta^P_{G^j}(\mathbf{f}^j),y_j)-\mathcal{L}(\Theta^P_{W^j}(f^j),y_j)\right|^2\Bigg]\\
\leq&\Gamma\sum_{j=1}^{\Gamma}\frac{\gamma_j}{m}\mathbb{E}_{(G^j,\mathbf{f}^j)\sim\mu_{\mathcal{G}_j}}\Bigg[\sup_{\Theta\in\mathrm{Lip}_{L,B}}L_{\mathcal{L}}^2\|\Theta^P_{G^j}(\mathbf{f}^j)-\Theta^P_{W^j}(f^j)\|_\infty^2\Bigg].
\end{aligned}
$$

We now apply Corollary C.5 to get

$$
\begin{aligned}
\leq\Gamma\sum_{j=1}^{\Gamma}\frac{\gamma_j}{m}L_{\mathcal{L}}^2\Bigg(6\sqrt{\pi}\Bigg(&\big(S_1+S_3+(S_2+S_4)\|f^j\|_\infty^2\big)\mathbb{E}_{N\sim\nu}\left[N^{-1}\right]\\
&+\big(R_1+R_4+(R_2+R_5)\|f^j\|_\infty^2+R_3L_{f^j}^2\big)\mathbb{E}_{N\sim\nu}\left[N^{-\frac{1}{D_{\chi^j}+1}}\right]\\
&+\big(T_1+T_2\|f^j\|_\infty^2\big)\mathbb{E}_{N\sim\nu}\left[\frac{\log(N)}{N^{\frac{1}{D_\chi^j+1}}}\right]\Bigg)+\mathcal{O}\left(\mathbb{E}_{N\sim\nu}\left[\exp(-N)N^{\frac{3}{2}T-\frac{3}{2}}\right]\right)\Bigg).
\end{aligned}
$$

Hence, by Lemma C.6,

$$\sum_k \mathbb{P}\big(\mathbf{m} \in \mathcal{M}_k\big) \times \sup_{\mathbf{m}\in\mathcal{M}_k} \mathbb{E}_{\mathcal{T}_{\mathbf{m}}\sim\mu_{\mathcal{G}^{\mathbf{m}}}} \Bigg[\sup_{\Theta\in\mathrm{Lip}_{L,B}} \Bigg(\sum_{j=1}^{\Gamma} \Bigg(\frac{1}{m}\sum_{i=1}^{m\gamma_j} \mathcal{L}(\Theta^P_{G_i^j}(\mathbf{f}_i^j), y_j) - \frac{1}{m}\sum_{i=1}^{m\gamma_j} \mathbb{E}_{(G^j,\mathbf{f}^j)\sim\mu_{\mathcal{G}_j}} \left[\mathcal{L}(\Theta^P_{G^j}(\mathbf{f}^j), y_j)\right] \Bigg)\Bigg)^2 \Bigg]$$

$$\leq \frac{\sqrt{\pi}}{2} 2^{\Gamma} \sum_{j=1}^{\Gamma} \frac{\gamma_j}{m} \mathbb{E}_{\mathcal{T}_{\mathbf{m}}\sim\mu_{\mathcal{G}^{\mathbf{m}}}} \Bigg[\sup_{\Theta\in\mathrm{Lip}_{L,B}} \Bigg(\sum_{j=1}^{\Gamma} \Bigg(\frac{1}{m}\sum_{i=1}^{m\gamma_j} \mathcal{L}(\Theta^P_{G_i^j}(\mathbf{f}_i^j), y_j) - \frac{1}{m}\sum_{i=1}^{m\gamma_j} \mathbb{E}_{(G^j,\mathbf{f}^j)\sim\mu_{\mathcal{G}_j}} \left[\mathcal{L}(\Theta^P_{G^j}(\mathbf{f}^j), y_j)\right] \Bigg)\Bigg)^2 \Bigg]$$

$$\leq \frac{\sqrt{\pi}}{2} 2^{\Gamma}\Gamma \sum_{j=1}^{\Gamma} \frac{\gamma_j}{m} L_{\mathcal{L}}^2 \Bigg(6\sqrt{\pi}\Bigg(\big(S_1^{(j)} + S_3^{(j)} + (S_2^{(j)} + S_4^{(j)})\|f^j\|_\infty^2\big)\mathbb{E}_{N\sim\nu}\left[N^{-1}\right]$$
$$+ \big(R_1^{(j)} + R_4^{(j)} + (R_2^{(j)} + R_5^{(j)})\|f^j\|_\infty^2 + R_3^{(j)} L_{f^j}^2\big)\mathbb{E}_{N\sim\nu}\left[N^{-\frac{1}{D_{\chi^j}+1}}\right]$$
$$+ \big(T_1^{(j)} + T_2^{(j)}\|f^j\|_\infty^2\big)\mathbb{E}_{N\sim\nu}\left[\frac{\log(N)}{N^{\frac{1}{D_{\chi}^j+1}}}\right]\Bigg) + \mathcal{O}\left(\mathbb{E}_{N\sim\nu}\left[\exp(-N)N^{\frac{3}{2}T-\frac{3}{2}}\right]\right)\Bigg),$$

where $S_l^{(j)}, R_l^{(j)}, T_l^{(j)}$ are the according constants from Theorem B.18 for each class j and are defined in (51). All in all, we get

$$\mathbb{E}_{\mathcal{T}\sim\mu_{\mathcal{G}}^m} \Bigg[\sup_{\Theta\in\mathrm{Lip}_{L,B}} \Bigg(\frac{1}{m}\sum_{j=1}^{\Gamma}\sum_{i=1}^{m_j} \mathcal{L}(\Theta^P_{G_i^j}(\mathbf{f}_i^j), y_j) - \mathbb{E}_{(G,\mathbf{f},y)\sim\mu_{\mathcal{G}}}\left[\mathcal{L}(\Theta^P_G(\mathbf{f}), y)\right] \Bigg)^2 \Bigg]$$
$$\leq 2^{\Gamma}\frac{8\|\mathcal{L}\|_\infty^2}{m}\pi + \frac{\sqrt{\pi}}{m} 2^{\Gamma}\Gamma \sum_{j=1}^{\Gamma} \gamma_j L_{\mathcal{L}}^2 \Bigg(6\sqrt{\pi}\Big(\big(S_1^{(j)} + S_3^{(j)} + (S_2^{(j)} + S_4^{(j)})\|f^j\|_\infty^2\big)\mathbb{E}_{N\sim\nu}\left[N^{-1}\right]$$
$$+ \big(R_1^{(j)} + R_4^{(j)} + (R_2^{(j)} + R_5^{(j)})\|f^j\|_\infty^2 + R_3^{(j)} L_{f^j}^2\big)\mathbb{E}_{N\sim\nu}\left[N^{-\frac{1}{D_{\chi^j}+1}}\right]$$
$$+ \big(T_1^{(j)} + T_2^{(j)}\|f^j\|_\infty^2\big)\mathbb{E}_{N\sim\nu}\left[\frac{\log(N)}{N^{\frac{1}{D_{\chi}^j+1}}}\right]\Big) + \mathcal{O}\left(\mathbb{E}_{N\sim\nu}\left[\exp(-N)N^{\frac{3}{2}T-\frac{3}{2}}\right]\right)\Bigg).$$

We define

$$C = 6\sqrt{\pi} \max_{j=1,\dots,\Gamma} \left(\sum_{i=1}^{4} S_i^{(j)} + \sum_{i=1}^{5} R_i^{(j)} + \sum_{i=1}^{2} T_i^{(j)} \right), \tag{58}$$

leading to

$$\mathbb{E}_{\mathcal{T}\sim\mu_{\mathcal{G}}^m}\Big[\sup_{\Theta\in\mathrm{Lip}_{L,B}} \Big(R_{emp}(\Theta^P) - R_{exp}(\Theta^P)\Big)^2\Big]$$
$$\leq \frac{2^{\Gamma}8\|\mathcal{L}\|_\infty^2\pi}{m} + \frac{2^{\Gamma}\Gamma L_{\mathcal{L}}^2 C}{m}\sum_j \gamma_j\big(1 + \|f^j\|_\infty^2 + L_{f^j}^2\big) \cdot \left(\mathbb{E}_{N\sim\nu}\left[\frac{1}{N} + \frac{1+\log(N)}{N^{1/D_{\chi^j}+1}} + \mathcal{O}\left(\exp(-N)N^{\frac{3}{2}T-\frac{3}{2}}\right)\right]\right).$$

□

C.3 Asymptotics of the Generalization Bound

In this subsection, we derive the asymptotic dependency of our generalization bound in Theorem 3.3 with respect to the uniform Lipschitz bound L of the message and update function, the depth T, the maximal hidden dimension h and the average graph size, that we denote in this section by abuse of notation N. Since we bound the expected square generalization error, and most other related generalization bounds are formulated in high probability, we transform our bound in expectation to a bound in high probability, using, e.g., Markov's Inequality (and then taking the square root of the square error). By this, the comparison with other generalization bounds formulated in high probability are valid. Hence, we focus on the constant $\sqrt{C}$, where C is the constant from Theorem 3.3. We reformulated Theorem 3.3 as Theorem C.7, where we observed that $C \leq 6\sqrt{\pi}\max_{j=1,\dots,\Gamma}\left(\sum_{i=1}^{4} S_i^{(j)} + \sum_{i=1}^{5} R_i^{(j)} + \sum_{i=1}^{2} T_i^{(j)}\right)$, where $S_l^{(j)}, R_l^{(j)}, T_l^{(j)}$ are the according constants from Theorem B.18 for each class j and are defined in (51). For a better presentation, we drop the class-superscript by setting $S_l = \max_j S_l^{(j)}$, for $l = 1, \dots, 4$, $R_l = \max_j R_l^{(j)}$, for $l = 1, \dots, 5$ and $T_l = \max_j T_l^{(j)}$, for $l = 1, 2$. Further, denote $C_\chi = \max_j C_{\chi^j}$, $D_\chi = \max_j D_{\chi^j}$, $L_W = \max_j L_{W^j}$ and $\|W\|_\infty = \max_j \|W^j\|_\infty$.

The constants R_i, S_i and T_i are bounded by a polynomial of order 2 in Ω_j, for $j = 1, \dots, 9$, defined in (48). The constants Ω_j, $j = 1, \dots, 9$, depend on a polynomial of degree one in $Z_1^{(l)}, Z_2^{(l)}, Z_3^{(l)}$, $B_1^{(l)}, B_2^{(l)}$ and on a polynomial of degree at most $T-1$ in $K^{(l)}$ for $l = 1, \dots, T-1$. Here, $Z_1^{(l)}, Z_2^{(l)}, Z_3^{(l)}$ are defined in (31), $B_1^{(l)}$ and $B_2^{(l)}$ are defined in (27) and (28), and

$$K^{(l')} = \sqrt{(L_{\Psi^{(l')}})^2 + \frac{8\|W\|_\infty^2}{\mathrm{d}_{\min}^2}(L_{\Phi^{(l')}})^2(L_{\Psi^{(l')}})^2}.$$

Hence, our strategy is as follows. We first work out the asymptotic behaviour of $Z_1^{(l)}, Z_2^{(l)}, Z_3^{(l)}$, $B_1^{(l)}, B_2^{(l)}$ and $K^{(l)}$ for $l = 1, \dots, T-1$ with respect to the parameters. Then, we derive the asymptotics of Ω_j, $j = 1, \dots, 9$. These already agree with the asymptotic of $\sqrt{C}$. For this, we write$A \lesssim x^k$ if A is bounded by a polynomial of order k in x.

We begin with observing that $K^{(l')} \lesssim L^2 \frac{\|W\|_\infty}{\mathrm{d}_{\min}}$. Since we only consider MPNNs $\Theta \in \mathrm{Lip}_{L,B}$, we have for $l = 1, \dots, T$,

$$\begin{aligned} B_1^{(l)} &\leq \sum_{k=1}^{l} \left(L_{\Psi^{(k)}} \frac{\|W\|_\infty}{\mathrm{d}_{\min}} \|\Phi^{(k)}(0,0)\|_\infty + \|\Psi^{(k)}(0,0)\|_\infty\right) \prod_{l'=k+1}^{l} L_{\Psi^{(l')}} \left(1 + \frac{\|W\|_\infty}{\mathrm{d}_{\min}} L_{\Phi^{(l')}}\right) \\ &\lesssim \sum_{k=1}^{l} LB \frac{\|W\|_\infty}{\mathrm{d}_{\min}} \left(\frac{\|W\|_\infty}{\mathrm{d}_{\min}} L^2\right)^{l-k} \lesssim \frac{\|W\|_\infty^l}{\mathrm{d}_{\min}^l} L^{2l} B. \end{aligned}$$

and

$$\begin{aligned} B_2^{(l)} &\leq \prod_{k=1}^{l} L_{\Psi^{(k)}} \left(1 + \frac{\|W\|_\infty}{\mathrm{d}_{\min}} L_{\Phi^{(k)}}\right) \\ &\lesssim \frac{\|W\|_\infty^l}{\mathrm{d}_{\min}^l} L^{2l}. \end{aligned}$$

For $l = 1, \dots, T$, the constants $Z_1^{(l)}$, are defined in (31). We have

$$
\begin{aligned}
Z_1^{(l)} &\leq \sum_{k=1}^{l} \Bigg(\Big(L_{\Psi^{(k)}} \frac{L_W}{\mathrm{d}_{\min}} \|\Phi^{(k)}(0,0)\|_\infty + L_{\Psi^{(k)}} \|W\|_\infty \|\Phi^{(k)}(0,0)\|_\infty \frac{L_W}{\mathrm{d}_{\min}^2} \Big) \\
&\quad + B_1^{(k-1)} \Big(L_{\Psi^{(k)}} \frac{L_W}{\mathrm{d}_{\min}} L_{\Phi^{(k)}} + L_{\Psi^{(k)}} \|W\|_\infty L_{\Phi^{(k)}} \frac{L_W}{\mathrm{d}_{\min}^2} \Big) \Bigg) \prod_{l'=k+1}^{l} L_{\Psi^{(l')}} \Big(1 + \frac{\|W\|_\infty}{\mathrm{d}_{\min}} L_{\Phi^{(l')}} \Big) \\
&\lesssim \sum_{k=1}^{l} B_1^{(k-1)} L^2 \frac{L_W}{\mathrm{d}_{\min}^2} \left(\frac{\|W\|_\infty}{\mathrm{d}_{\min}} L^2 \right)^{l-k} \\
&\lesssim \sum_{k=1}^{l} B \|W\|_\infty \frac{\|W\|_\infty^{k-1}}{\mathrm{d}_{\min}^{k-1}} (L^2)^{k-1} L^2 \frac{L_W}{\mathrm{d}_{\min}^2} \left(\frac{\|W\|_\infty}{\mathrm{d}_{\min}} L^2 \right)^{l-k} \\
&\lesssim B \frac{\|W\|_\infty^l L_W}{\mathrm{d}_{\min}^{l+1}} L^{2l}.
\end{aligned}
$$

We have

$$
\begin{aligned}
Z_2^{(l)} &\leq \sum_{k=1}^{l} B_2^{(k)} \Big(L_{\Psi^{(k)}} \frac{L_W}{\mathrm{d}_{\min}} L_{\Phi^{(k)}} + L_{\Psi^{(k)}} \|W\|_\infty L_{\Phi^{(k)}} \frac{L_W}{\mathrm{d}_{\min}^2} \Big) \prod_{l'=k+1}^{l} L_{\Psi^{(l')}} \Big(1 + \frac{\|W\|_\infty}{\mathrm{d}_{\min}} L_{\Phi^{(l')}} \Big) \\
&\lesssim \sum_{k=1}^{l} B_2^{(k-1)} L^2 \frac{L_W}{\mathrm{d}_{\min}} \|W\|_\infty \left(\frac{\|W\|_\infty}{\mathrm{d}_{\min}} L^2 \right)^{l-k} \\
&\lesssim \sum_{k=1}^{l} \frac{\|W\|_\infty^{k-1}}{\mathrm{d}_{\min}^{k-1}} (L^2)^{k-1} L^2 \frac{L_W}{\mathrm{d}_{\min}} \|W\|_\infty \left(\frac{\|W\|_\infty}{\mathrm{d}_{\min}} L^2 \right)^{l-k} \\
&\lesssim \frac{L_W \|W\|_\infty^l}{\mathrm{d}_{\min}^l} L^{2l}.
\end{aligned}
$$

We have

$$
\begin{aligned}
Z_3^{(l)} &\leq \prod_{k=1}^{l} L_{\Psi^{(k)}} \Big(1 + \frac{\|W\|_\infty}{\mathrm{d}_{\min}} L_{\Phi^{(k)}} \Big) \\
&\lesssim \frac{\|W\|_\infty^l}{\mathrm{d}_{\min}^l} L^{2l}.
\end{aligned}
$$

For $i = 1, \dots, 9$, the constant Ω_i depends on $K^{(l)}$ for which we have

$$
K^{(l')} \leq \sqrt{(L_{\Psi^{(l')}})^2 + \frac{8\|W\|_\infty^2}{\mathrm{d}_{\min}^2} (L_{\Phi^{(l')}})^2 (L_{\Psi^{(l')}})^2} \lesssim \frac{\|W\|_\infty}{\mathrm{d}_{\min}} L^2
$$

For Ω_1, we calculate

$$
\begin{aligned}
\Omega_1 &\leq \sum_{l=1}^{T} L_{\Psi^{(l)}} 4 \frac{\zeta L_W \big(\sqrt{\log(C_\chi)} + \sqrt{D_\chi} \big)}{\mathrm{d}_{min}^2} \|W\|_\infty \big(L_{\Phi^{(l)}} B_1^{(l-1)} + \|\Phi^{(l)}(0,0)\|_\infty \big) \prod_{l'=l+1}^{T} K^{(l')} \\
&\lesssim \big(\sqrt{\log(C_\chi)} + \sqrt{D_\chi} \big) \sum_{l=1}^{T} L^2 B_1^{(l-1)} \frac{L_W \|W\|_\infty}{\mathrm{d}_{\min}^2} (L^2)^{T-l} \\
&\lesssim \big(\sqrt{\log(C_\chi)} + \sqrt{D_\chi} \big) B L^{2T} \frac{L_W \|W\|_\infty^T}{\mathrm{d}_{\min}^{T+1}}
\end{aligned}
$$

Similar calculations lead to

$$
\Omega_i \lesssim \big(\sqrt{\log(C_\chi)} + \sqrt{D_\chi} \big) B (L^2)^T \frac{L_W \|W\|_\infty^T}{\mathrm{d}_{\min}^{T+1}}.
$$

Hence,

$$GE \lesssim \frac{2^{\Gamma/2}}{\sqrt{m}} + \frac{2^{\Gamma/2}\left(\sqrt{\log(C_\chi)} + \sqrt{D_\chi}\right) B L^{2T} L_W \|W\|_\infty^T}{\sqrt{m}\mathrm{d}_{\min}^{T+1}} \mathbb{E}_{N\sim\nu}\left[\frac{\sqrt{\log(N)}}{N^{\frac{1}{2(D_\chi+1)}}}\right]. \tag{59}$$

C.4 Generalization Bound Comparison

In this subsection, we compare our generalization bound, especially the asymptotics derived in the previous subsection, with other related generalization bounds. Since related work does neither consider the same network architecture, nor the same data distribution as our work, we emphasize the setting of each of the cited results. We then write the asymptotics of the cited bounds in terms of the maximal hidden dimension h, depth T, Lipschitz bound L of the message and update functions, maximal node d degree and graph size N. We recall (59), where we derived the asymptotics of our generalization bound from Theorem 3.3 with respect to T, L and N as

$$\mathcal{O}\left(\mathbb{E}_{N\sim\nu}\left[\frac{\sqrt{\log(N)}}{N^{\frac{1}{2(D_\chi+1)}}}\right]\right), \qquad \mathcal{O}\left(L^{2T}\right)$$

and $\mathcal{O}(1)$ with respect to h.

C.4.1 PAC-Bayesian Approach based Bound

The generalization analysis of [Liao et al., 2021] considers MPNNs with sum aggregation for a K-class graph classification setting. The authors differentiate between the input node feature vectors $\mathbf{x}_v$, which is an unchanged input for every layer, and the node embedding/representation in the l-th layer $\mathbf{f}^{(l)}$, where they take $\mathbf{f}^{(0)} = 0$. More formally, the MPNNs takes the following form.

Definition C.8. *Let G be a graph with graph features $\mathbf{x}$. A MPNN (in [Liao et al., 2021]) with T layers is defined by taking the input feature representation $\mathbf{f}^{(0)} = 0$, and mapping it to the features $\mathbf{f}^{(l)}$ in the l-th layer, which are defined recursively by*

$$\mathbf{f}_v^{(l)} = \Psi\left(W_1\mathbf{x}_v + W_2\rho\left(\sum_{u\in\mathcal{N}(v)} \Phi(\mathbf{f}_u^{(l-1)})\right)\right), \tag{60}$$

where ρ, Ψ and Φ are nonlinear transformations, and W_1 and W_2 are linear transformations. This is followed by a global pooling layer, which takes as an input $\mathbf{f}^{(T-1)} \in \mathbb{R}^{N\times K}$, and returns the vector

$$\frac{1}{N}\mathbf{1}_N\mathbf{f}^{(T-1)}W_T \in \mathbb{R}^{1\times K},$$

where W_T is a linear transformation. Here $\mathbf{1}_N$ denotes the vector $(1,\dots,1) \in \mathbb{R}^{1\times N}$, where N is the number of nodes in the graph.

The message and update functions in Definition C.8 are the same in every layer. It is assumed that Ψ, ρ and Φ have Lipschitz constants L_Ψ, L_ρ and L_Φ. Furthermore it is assumed that W_1, W_2 and W_T have bounded norms, i.e., $\|W_1\|_2 \le B_1, \|W_2\|_2 \le B_2$ and $\|W_T\|_2 \le B_T$.

The expected multiclass margin loss is then defined as

$$R_{\mathcal{D},\gamma}(\Theta) = \mathbb{P}_{(G,\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left(\left(\Theta_G^P(\mathbf{x})\right)_{\mathbf{y}} \le \gamma + \max_{j\neq\mathbf{y}}\left(\Theta_G^P(\mathbf{x})\right)_j\right),$$

where $\mathcal{D}$ is the unknown data distribution, $\gamma > 0$ and Θ_G^P is the MPNN after pooling. Accordingly, the empirical loss is defined as

$$R_{\mathcal{T},\gamma}(\Theta) = \frac{1}{m}\sum_{(G_i,\mathbf{x}_i,\mathbf{y}_i)\in\mathcal{T}} \mathbb{1}\left(\left(\Theta_{G_i}^P(\mathbf{x}_i)\right)_{\mathbf{y}_i} \le \gamma + \max_{j\neq\mathbf{y}_i}\left(\Theta_{G_i}^P(\mathbf{x}_i)\right)_j\right),$$

where the summand $\mathbb{1}\left(\left(\Theta_{G_i}^P(\mathbf{x}_i)\right)_{\mathbf{y}_i} \le \gamma + \max_{j\neq\mathbf{y}_i}\left(\Theta_{G_i}^P(\mathbf{x}_i)\right)_j\right)$ is equal to 1 if $\left(\Theta_{G_i}^P(\mathbf{x}_i)\right)_{\mathbf{y}_i} \le \gamma + \max_{j\neq\mathbf{y}_i}\left(\Theta_{G_i}^P(\mathbf{x}_i)\right)_j$ and otherwise 0.

Furthermore, the following assumptions hold for the training set and the considered MPNNs

Assumption C.9.

1. *The training set $\mathcal{T} = \{(G_1, \mathbf{x}_1, \mathbf{y}_1), \ldots, (G_m, \mathbf{x}_m, \mathbf{y}_m)\}$ is drawn i.i.d. from some distribution $\mathcal{D}$, where all graphs are simple and have node degrees at most $d-1$.*

2. *The maximum hidden dimension across all layers is h.*

3. *The node features are drawn in an l^2-ball with radius B from the node feature space $\mathcal{X}$.*

The generalization bound is formulated in terms of the following constants: $\zeta = \min(\|W_1\|_2, \|W_2\|_2, \|W_T\|_2)$, $|w|_2^2 = \|W_1\|_F^2 + \|W_2\|_F^2 + \|W_T\|_F^2$, $\lambda = \|W_1\|_2 \|W_T\|_2$, $\xi = L_\Psi \frac{(d\mathcal{C})^{l-1}-1}{d\mathcal{C}-1}$, and the percolation complexity $\mathcal{C} = L_\Psi L_\rho L_\Phi \|W_2\|_2$. We summarize the main result [Liao et al., 2021, Theorem 3.4] as follows.

Theorem. *Let $T > 1$. Then for any $\delta, \gamma > 0$, with probability at least $1-\delta$ over the choice of the training set $\mathcal{T} \sim \mathcal{D}^m$ of m graphs, for any T-layered MPNN Θ, we have,*

1. *If $d\mathcal{C} = 1$, then*

$$R_{\mathcal{D},0}(\Theta) \leq R_{\mathcal{T},\gamma}(\Theta) + \mathcal{O}\left(\sqrt{\frac{B^2 \max\left(\zeta^{-6}, \lambda^3 L_\Psi^3\right)(T+1)^4 h \log(Th)|w|_2^2 + \log\frac{m}{\delta}}{\gamma^2 m}}\right).$$

2. *If $d\mathcal{C} \neq 1$, then*

$$R_{\mathcal{D},0}(\Theta) \leq R_{\mathcal{T},\gamma}(\Theta) + \mathcal{O}\left(\sqrt{\frac{B^2 \left(\max\left(\zeta^{-(T+1)}, (\lambda\xi)^{(T+1)/T}\right)\right)^2 T^2 h \log(Th)|w|_2^2 + \log\frac{m(T+1)}{\delta}}{\gamma^2 m}}\right).$$

We only consider the non-degenerative case $d\mathcal{C} \neq 1$, as it is the generic case, which can again be split into two cases. As the authors in [Liao et al., 2021] mention, these two cases correspond to $\max(\zeta^{-1}, (\lambda\xi)^{\frac{1}{T}}) = \zeta^{-1}$ (case A) and $\max(\zeta^{-1}, (\lambda\xi)^{\frac{1}{T}}) = (\lambda\xi)^{\frac{1}{T}}$ (case B). In practice case B occurs more often, where the generalization bound depends on the parameters with orders $\mathcal{O}\left(d^{\frac{(T+1)(T-2)}{T}}\right)$, $\mathcal{O}\left(\sqrt{h \log h}\right)$ and $\mathcal{O}\left(\lambda^{1+\frac{1}{T}} \xi^{1+\frac{1}{T}} \sqrt{\|W_1\|_F^2 + \|W_2\|_F^2 + \|W_l\|_F^2}\right)$. In case A, the generalization bound depends on the parameters with orders $\mathcal{O}\left(\sqrt{h \log h}\right)$ and $\mathcal{O}\left(\zeta^{-(T+1)} \sqrt{\|W_1\|_F^2 + \|W_2\|_F^2 + \|W_l\|_F^2}\right)$.

We now describe the architecture in Definition C.8 in terms of the message passing framework from (1). For $l = 1, \ldots, T$, we denote by $\mathbf{m}_i^{(l)}$ and $\mathbf{f}_i^{(l)}$ the message and graph feature of node i in the l-th layer, respectively. Given a simple graph G with node features $(\mathbf{x}_i)_i$, we set $\mathbf{f}_i = \mathbf{x}_i$ as the input for the MPNN. Then the message function in the first layer is given by $\Phi^{(1)}(\mathbf{f}_i, \mathbf{f}_j) = \mathbf{f}_i$. We recall that the message in MPNNs with sum aggregation is calculated as $\mathbf{m}_i^{(1)} = \sum_{j \in \mathcal{N}(i)} \Phi^{(1)}(\mathbf{f}_i, \mathbf{f}_j)$. The update function in the first layer is given by $\Psi^{(1)}(\mathbf{f}_i, \mathbf{m}_i^{(1)}) = \left(\Phi(W_1 \mathbf{f}_i), \mathbf{f}_i\right)$. For $l = 2, \ldots, T-1$, the message functions are defined as

$$\Phi^{(l)}(\mathbf{f}_i^{(l-1)}, \mathbf{f}_j^{(l-1)}) = \Phi(\mathbf{f}_i^{(l-1)})$$

and the update functions are defined as

$$\Psi^{(l)}(\mathbf{f}_i^{(l-1)}, \mathbf{m}_i^{(l)}) = \Psi\left(W_1(\mathbf{f}_i^{(l-1)})_2 + W_2 \rho\left(\mathbf{m}_i^{(l)}\right), (\mathbf{f}_i^{(l-1)})_2\right),$$

where $(\mathbf{f}_i^{(l-1)})_2$ stays unchanged through all layers, and is equal to the input graph features $\mathbf{x}_i$. The aggregation scheme is given by sum aggregation. Finally, the pooling in Definition C.8 can be described by a graph MPNN layer with update function W_T followed by average pooling. With this construction of message and update functions the MPNN $\Theta = \left((\Psi^{(l)})_{l=1}^T, (\Phi^{(l)})_{l=1}^T\right)$ with sum aggregation matches the architecture in Definition C.8.

We summarize the Lipschitz bounds for the message and update functions by $L_{\Phi^{(1)}} = 1, L_{\Phi^{(T)}} = 1, L_{\Psi^{(1)}} = L_\Phi, L_{\Phi^{(T)}} = \|W_T\|_2$ and $L_{\Phi^{(l)}} = L_\Phi, L_{\Psi^{(l)}} = L_\Psi\big(\|W_1\|_2 + \|W_2\|_2 L_\rho\big)$ for $l = 2, \dots, T-1$. For deriving our generalization bound in Theorem 3.1, we assume that there exists a uniform Lipschitz bound for the message and update functions, denoted by L. Hence, we assume that $L_\Phi \leq L$, $L_\Psi \|W_1\|_2 + L_\Psi \|W_2\|_2 L_\rho \leq L$ and $\|W_T\|_2 \leq L$.

For simplicity and better comparison with our generalization bound, we make use of the following upper bounds,

$$\begin{aligned}
&\mathcal{C} = L_\Psi L_\rho L_\Phi \|W_2\| \leq L^2, \\
&\xi = L_\Psi \frac{(d\mathcal{C})^{T-1} - 1}{d\mathcal{C} - 1} \leq L(L^2)^{T-2}, \\
&\zeta = \min(\|W_1\|_2, \|W_2\|_2, \|W_l\|_2) \leq L \text{ and} \\
&\lambda = \|W_1\|_2 \|W_l\|_2 \leq L.
\end{aligned} \tag{61}$$

This leads to

$$\begin{aligned}
\mathcal{O}\left(\lambda^{1+\frac{1}{l}} \xi^{1+\frac{1}{l}} \sqrt{\|W_1\|_F^2 + \|W_2\|_F^2 + \|W_l\|_F^2}\right) &= \mathcal{O}\left(L^{1+\frac{1}{T}} (L(L^2)^{T-2})^{1+\frac{1}{T}} L\right) \\
&= \mathcal{O}\left(L^{2T-2/T+1}\right).
\end{aligned}$$

Hence, the asympotics of the generalization bound in [Liao et al., 2021] with respect to the maximal hidden dimension h, the Lipschitz bound L, the depth T and the maximum node degree d can be summarized respectively as

$$\mathcal{O}\left(d^{\frac{(T+1)(T-2)}{T}}\right), \ \mathcal{O}\left(\sqrt{h \log h}\right) \text{ and } \mathcal{O}\left(L^{2T-2/T+1}\right).$$

C.4.2 Rademacher Complexity based Bound

We next analyze the bound derived in [Garg et al., 2020]. Since [Garg et al., 2020] consider the same architecture, defined in Definition C.8, as [Liao et al., 2021], we adopt the notation from Subsection C.4.1. The authors in [Garg et al., 2020] consider a binary graph classification task with the same Assumptions C.9 on the training set and the MPNN as in [Liao et al., 2021]. The main result can be summarized as follows.

Theorem. *Let $T > 1$. Then for any $\delta, \gamma > 0$, with probability at least $1 - \delta$ over the choice of the training set $\mathcal{T} \sim \mathcal{D}^m$ of m graphs, for any T-layered MPNN Θ, we have,*

$$R_{\mathcal{D},0}(\Theta) \leq R_{\mathcal{T},\gamma}(\Theta)$$
$$+ \mathcal{O}\left(\frac{1}{\gamma m} + h\|W_T\|_2 Z \sqrt{\frac{\log\left(\|W_T\|_2 \sqrt{m} \max\left(Z, \xi\sqrt{h} \max\left(B\|W_1\|_2, \bar{R}\|W_2\|_2\right)\right)\right)}{\gamma^2 m}} + \sqrt{\frac{\frac{1}{\delta}}{m}}\right),$$

where $\bar{R}$ is a constant specified in [Garg et al., 2020] that satisfies $\bar{R} \leq L_\rho L_\Phi d B \|W_1\|_2 \xi$, and $Z = B\|W_1\|_2 \|W_T\|_2$.

We only consider the case $\max\left(Z, \xi\sqrt{h} \max\left(BB_1, \bar{R}B_2\right)\right) = \xi\sqrt{h}\bar{R}B_2$, which is the generic case (see [Liao et al., 2021, Subsection A.5.2] for the other cases). Thus the generalization bound from [Garg et al., 2020] depends on the parameters with orders $\mathcal{O}\left(d^{T-1}\sqrt{\log(d^{2T-3})}\right), \mathcal{O}\left(h\sqrt{\log\sqrt{h}}\right)$ and $\mathcal{O}\left(\lambda\mathcal{C}\xi\sqrt{\log(\|W_2\|_2 \lambda \xi^2)}\right)$.

Similarly to Subsection C.4.1, we consider a uniform Lipschitz bound L for the message and update functions. We thus consider the upper bounds on ξ, λ and $\mathcal{C}$, summarized in (61), which leads to

$$\mathcal{O}\left(\lambda\mathcal{C}\xi\sqrt{\log(\|W_2\|_2 \lambda \xi^2))}\right) = \mathcal{O}\left(L^{2T}\sqrt{\log(L^{4T-4})}\right).$$

Hence, the asympotics of the Rademacher based generalization bound in [Garg et al., 2020] with respect to the maximal hidden dimension h, the Lipschitz bound L, the depth T and the maximum node degree d can be summarized as

$$\mathcal{O}\left(d^{T-1}\sqrt{\log(d^{2T-3})}\right), \ \mathcal{O}\left(h\sqrt{\log\sqrt{h}}\right) \text{ and } \mathcal{O}\left(L^{2T}\sqrt{\log(L^{4T-4})}\right).$$

VC-Dimension Based Bound [Scarselli et al., 2018] The work by [Scarselli et al., 2018] considers graph neural networks in supervised classification or regression tasks, where the input is a graph G with graph feature map $\mathbf{x}$ and one node of interest v in which we want to produce a prediction. They apply a recurrent graph neural network on the graph G with graph feature $\mathbf{x}$, and then evaluate the output graph feature map $\mathbf{f}$ only in v. They then calculate the loss between $\mathbf{f}(v)$ and its given desired target $\mathbf{y}$. More formally, the training dataset $\mathcal{T}$ is defined as $\mathcal{T} = \{(G^i, \mathbf{x}^i, v^i, \mathbf{y}^i) \mid 1 \leq i \leq m\}$, where m is the number of graphs and each tuple $(G^i, \mathbf{x}^i, v^i, \mathbf{y}^i)$ denotes a graph G^i with graph features $\mathbf{x}^i$, the supervised node v^i, and the desired target $\mathbf{y}^i$ for that node.

Given a graph $G = (V, E)$ with graph features $\mathbf{x}$ the graph neural network architecture is defined implicitly, as a method that solves a system of equations, and the solution is the output of the network. The equation is given by

$$\mathbf{f}_i = \sum_{j \in \mathcal{N}(i)} \Phi(\mathbf{x}_i, \mathbf{f}_j, \mathbf{x}_j), \ \forall i \in V \tag{62}$$

where Φ is a multi-layer-perceptron with input $[\mathbf{x}_i, \mathbf{f}_j, \mathbf{x}_j]$, and the solution $\mathbf{f}$ to (62) is defined as the output of this part of the network. The output of the network $\mathbf{o}_i \in \mathbb{R}$ for the node i is then defined by

$$\mathbf{o}_i = g(\mathbf{x}_i, \mathbf{f}_i), \tag{63}$$

where g is a multi-layer-perceptron. Given the training data set $\mathcal{T}$, the empirical loss R_{emp} is then defined by the sum of the squared errors, i.e.,

$$R_{\text{emp}} = \sum_{i=1}^{m} (\mathbf{y}^i - \mathbf{o}_{v^i})^2.$$

One way to solve the fixed point problem (62) is by a fixed point iteration, which means that we can interpret the architecture as a recurrent message passing network (theoretically with infinite depth), where all message functions in all layers are equal to Φ.

Scarselli et al. [2018] derive VC-dimension bounds for the mapping that takes as an input a tuple $(G, \mathbf{x}, v)$ of a graph G with features $\mathbf{x}$ and node of interest v and outputs $\mathbf{o}_v$ as defined in (62) and (63). The VC-dimension bound depends on the total number of parameters p of the network and a predefined maximum graph size N. Furthermore, the bound for the VC-dimension depends on the choice of the activation function in the MLPs Φ and g. If the activation is given by tanh and logistic sigmoid activations the VC-dimension scales as $\mathcal{O}(p^4 N^2)$. Since p can be related to the maximum hidden dimension h by $p \in \mathcal{O}(h^2)$, the VC-dimension scales as $\mathcal{O}(h^8 N^2)$. Consequently, the asymptotics of the generalization bounds in [Scarselli et al., 2018] with respect to h and N can be summarized as

$$\mathcal{O}(h^4) \text{ and } \mathcal{O}(N).$$

For piecewise polynomial activations the VC-dimension scales as $\mathcal{O}(h^4 \log(N) N)$, hence the generalization bound scales in this case as

$$\mathcal{O}(h^2) \text{ and } \mathcal{O}(\sqrt{\log(N) N})$$

with respect to h and N.

D Details on Numerical Experiments and Additional Experiments

In this section we report additional experiments and write all details corresponding to Section 4. We First give an example that illustrate our convergence theorem (Theorem 3.1), and then introduce a comparison between our generalization bound and the Rademacher complexity [Garg et al., 2020] and PAC-Bayesian [Liao et al., 2021] bounds, evaluated on synthetic datasets.

D.1 Convergence Experiments

In this section, we show simple numerical experiments on the convergence of sampled MPNNs from a random geometric graph model, on toy data. We consider random geometric graphs [Penrose, 2003], which can be described by using RGMs with the kernel $W(x, y) = \mathbb{1}_{B_r(x)}(y)$ on $[0, 1]^2$, equipped with the uniform distribution and the standard Euclidean norm. Here $\mathbb{1}_{B_r(x)}$ is the indicator function of the ball around x with radius r. Even though $\mathbb{1}_{B_r(x)}(y)$ is not Lipschitz continuous, and

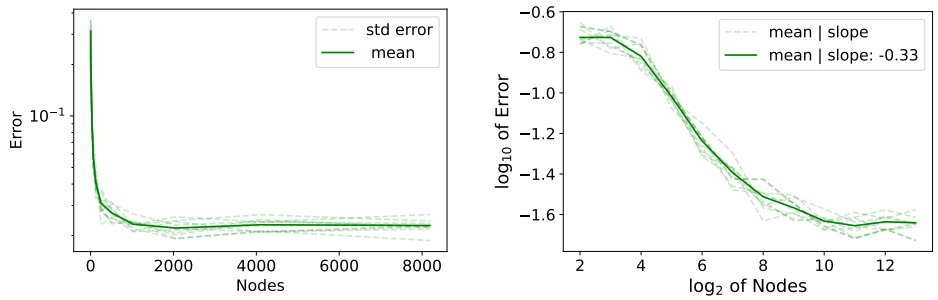


Figure 2: The average worst-case error between MPNNs realized on graphs and on the limit RGM, with varying number of nodes, drawn from the RGM $W(x, y) = \mathbb{1}_{B_r(x)}(y)$ (where $\mathbb{1}_{B_r(x)}$ is the indicator function of the ball around x with radius $r = 0.2$ in the space $([0, 1]^2, \|\cdot\|_{\mathbb{R}^2}, \mathcal{L})$), and a random low frequency signal. Left: graph sizes on the x-Axis and error on logarithmic y-Axis. Right: $\log_2$ of the graph sizes on the x-Axis and $\log_{10}$ of the error on the y-Axis. The slope of the curve represents the exponential dependency of the error on N.

hence does not satisfy the conditions of Theorems 3.1, $\mathbb{1}_{B_r(x)}(y)$ can be approximated by a Lipschitz continuous function. As the metric-space signal we consider a random low frequency signal (see Figure 2).

For our network, we choose untrained MPNNs with random weights, where each layer is defined using EdgeConv [Bronstein et al., 2017] with mean aggregation, and is implemented using Pytorch Geometric [Fey and Lenssen, 2019]. More precisely, we consider MPNNs with 2 layers. The message function in the first layer is defined as $\Phi^{(1)}(\mathbf{f}_i, \mathbf{f_j}) = h^{(1)}(\mathbf{f}_i, \mathbf{f}_j - \mathbf{f}_i)$, where $h^{(1)}$ is a 1-layered MLP with ReLU activation, input dimension 2 and output dimension 3. The message function in the second layer is defined as $\Phi^{(2)}(\mathbf{f}_i^{(1)}, \mathbf{f_j}^{(1)}) = h^{(2)}(\mathbf{f}_i^{(1)}, \mathbf{f}_j^{(1)} - \mathbf{f}_i^{(1)})$, where $h^{(2)}$ is a 1-layered MLP with ReLU activation, input dimension 6 and output dimension 1. The update functions are given by $\Psi(\mathbf{f}_i^{(1)}, \mathbf{m}_i^{(2)}) = \mathbf{m}_i^{(2)}$. This is a followed by an average pooling layer.

We ran the experiments that depend on random variables 10 times and report the average results with error bars that indicate the standard error. One run consists of the following steps. We consider 10 different graph sequences, where each graph sequence contains randomly sampled graphs of 2^i nodes, with $i = 1, \ldots, 13$. We then consider 50 (different) randomly initialized MPNNs, and compute for each graph sequence the worst-case error between the output of the cMPNN to its sampled graphs, i.e., for every graph size N, we pick the MPNN with the highest error. We then average the resulting 10 errors over the 10 different graph sequences, to approximate the expected error over the choice of the graph. In Figure 2, we plot the average error over the 10 runs on the logarithmic y-axis and the number of nodes on the x-Axis. We also provide a log-log-graph of this relation. Recall that in a log-log-graph a function of the form $f(x) = x^c$ appears as a line with slope c. We observe that in this toy example the worst-case error, which corresponds roughly to the uniform convergence result in Theorem 3.1, decays faster than our theoretical worst-case error bound $-1/6$. This suggests that, at least for band limited signals on random geometric graphs, our convergence bounds are not tight.

Computing the exact cMPNN would involve computing integrals. To approximate this integral, we sampled a large graph from the RGM. For the largest graph, we choose 2^{14} nodes. Our smaller graphs consist of 2^i nodes, with $i = 1, \ldots, 13$, and are sampled directly from the RGM. As the metric-space signal we consider a discrete random band-limited signal of resolution 256x256, defined as $f = \mathcal{F}^{-1}(v)$, where v consists of randomly chosen Fourier coefficients in the low positive frequency band 20x20 such that the coefficients in the lowest positive frequency band 8x8 are amplified by a factor of 10, and $\mathcal{F}^{-1}$ is the inverse Finite Fourier Transform.

D.2 Generalization Experiments

In this subsection, we provide details for the numerical experiments from Section 4 and report additional generalization experiments.

Table 2: Readout of the constants depending on the datasets. Each column represents the value of the respective dataset parameter used for the calculations of our generalization bounds.

	$\|W\|_\infty$	L_W	$\|f\|_\infty$	f_L	N	m	$d_{\min}$
ER-SBM	0.41	0.5	0.5	0.5	50	100K	0.25
ER-EXP	0.5	1	0.5	0.5	50	100K	0.373
EXP-SBM	0.5	1	0.5	0.5	50	100K	0.25

D.2.1 Dataset

We create three different synthetic datasets of random graphs from different random graph models. The domains of the graphons (the metric space), is taken as the Euclidean space $[0, 1]$. First, we consider Erdös-Rényi graphs with edge probably 0.4 with constant signal, represented by (W_1, f_1) with $W_1(x, y) = 0.4$ and $f_1(x) = 0.5$. We also consider a smooth version of a stochastic block model, represented by (W_2, f_2) with $W_2(x, y) = \sin(2\pi x)\sin(2\pi y)/2\pi + 0.25$ and $f_2(x) = \sin(x)/2$. Last, we consider an exponential radial graphon, represented by (W_3, f_3) with $W_3(x, y) = \exp(-|x - y|^2)/2$ and $f_3(x) = 0.5x$. For each graphon, we create 50K graphs of size 50. We call the Erdös-Renyi dataset **ER**, the stochstic block model dataset **SBM**, and the exponential radial dataset **EXP**. We then consider all possible pairs, i.e., **ER-SBM**, **ER-EXP** and **SBM-EXP**, and train a binary classifier for each pair. We split each dataset to 90% training examples and 10% test.

D.2.2 MPNN Details

For our network, we choose MPNNs intialized with random weights, where each layer is defined using GraphSage [Hamilton et al., 2017], and is implemented with Pytorch Geometric [Fey and Lenssen, 2019]. We consider MPNNs with 1,2 and 3 layers. The message functions are defined by

$$\Phi^{(l)}(\mathbf{f}_i^{(l-1)}, \mathbf{f}_j^{(l-1)}) = \mathbf{f}_j^{(l-1)}.$$

The update functions are given by

$$\Psi^{(l)}(\mathbf{f}_i^{(l-1)}, \mathbf{m}_i^{(l-1)}) = \rho\big(W_1^{(l)}\mathbf{f}_i^{(l-1)} + W_2^{(l)}\mathbf{m}_i^{(l-1)}\big),$$

where $W_1^{(1)} \in \mathbb{R}^{128\times 1}$, $W_2^{(1)} \in \mathbb{R}^{128\times 1}$ and $W_1^{(2)}, W_1^{(3)} \in \mathbb{R}^{128\times 128}$, $W_2^{(3)}, W_2^{(3)} \in \mathbb{R}^{128\times 128}$. We then consider a global mean pooling layer, and apply a last linear layer Q (including bias) with input dimension 128 and output dimension 2. This last linear layer is seen as part of the loss function in the analysis, and contributes to the generalization bound via the Lipschitz constant and infinity norm of the loss, as seen in Theorem 3.3.

D.2.3 Experimental Setup

The loss is given by soft-max composed with cross-entropy (composed on the last MLP). We consider Adam with learning rate $lr = 0.01$. For experiments with weight decay, we use an l^2-regularization on the weights with factors $0.27, 0.15$ and 0.05 for the **ER-SBM** dataset. For the **ER-EXP** dataset we consider weight decay factors $0.37, 0.15$ and 0.05. For the **SBM-EXP** dataset we consider $0.28, 0.05$ and 0.05. We train for 1 epoch. The batch size is 64. We consider 1, 2 and 3 layers.

D.2.4 Details on Computations of Our Bound

We compute our generalization bound according to the full formula given in Theorem C.7. The terms depending on the dataset are: the size of the training dataset m, the average graph size N, the minimum degree d, the largest infinity norm of the graphons $\|W\|_\infty$, largest Lipschitz norm of the graphons L_W, the largest infinity norm of the metric-space signal $\|f\|_\infty$, the largest Lipschitz norm of the metric-spaces signals L_f and the number of classes is $\Gamma = 2$. For every dataset, we summarize these terms depending on the dataset in Table 2.

Our bound depend also on the Lipschitz constants of the trained GraphSage MPNN, i.e., on the Lipschitz norms $L_{\Psi^{(l)}}$ and $L_{\Phi^{(l)}}$ of the update function $\Psi^{(l)}$ and message function $\Phi^{(l)}$, given in

Subsection D.2.2. We have $L_{\Phi^{(l)}} = \left\|[W_1^{(l)}, W_2^{(l)}]\right\|_\infty$ and $L_{\Phi^{(l)}} = 1$. We readout the norms $\left\|[W_1^{(l)}, W_2^{(l)}]\right\|_\infty$ for every layer, and plug it into our bound. The bound also depends on the infinity norm and Lipschitz constant of the loss. We compute these constants in the next subsection.

D.2.5 Computation of the Infinity Norm and Lipschitz Constant of the Loss

Next we bound the Lipschitz constant and infinity norm of the loss. Namely, we derive properties of softmax composed on cross-entropy. Softmax composed with the cross-entropy loss in the case of binary classes take the form

$$\mathcal{L}_{\mathrm{CE}}(\mathbf{x};\mathbf{y}) = -y_1 \log\left(\frac{e^{x_1}}{e^{x_1}+e^{x_2}}\right) - y_2 \log\left(\frac{e^{x_2}}{e^{x_1}+e^{x_2}}\right),$$

where $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$ and $(y_1, y_2) \in \{e_1, e_2\}$ depends on the target label, where $e_1 = (1, 0)$ and $e_2 = (0, 1)$. When the target label is fixed, we write in short $\mathcal{L}_{\mathrm{CE}}(\mathbf{x}) := \mathcal{L}_{\mathrm{CE}}(\mathbf{x};\mathbf{y})$.

Lemma D.1. *The loss $\mathcal{L}_{\mathrm{CE}}$ is Lipschitz continuous with Lipschitz constant 1. Additionally, $\mathcal{L}_{\mathrm{CE}}$ is locally bounded in the following sense:*

$$\|\mathcal{L}_{\mathrm{CE}}\|_{L^\infty([-K,K]^2)} \leq \log(1+e^{2K}),$$

where $\|\mathcal{L}_{\mathrm{CE}}\|_{L^\infty([-K,K]^2)} = \max_{\mathbf{x}\in[-K,K]^2} \|\mathcal{L}_{\mathrm{CE}}(\mathbf{x})\|$.

Proof. We compute

$$\begin{aligned}\frac{\partial}{\partial x_1}\mathcal{L}_{\mathrm{CE}}(x_1,x_2) =& -y_1\left(1-\frac{e^{x_1}}{e^{x_1}+e^{x_2}}\right) + y_2\frac{e^{x_1}}{e^{x_1}+e^{x_2}}\\ =&(y_1+y_2)\frac{e^{x_1}}{e^{x_1}+e^{x_2}} - y_1\\ =&\frac{e^{x_1}}{e^{x_1}+e^{x_2}} - y_1\end{aligned}$$

Since $\frac{e^{x_1}}{e^{x_1}+e^{x_2}} \in [0,1]$ and $y_1 \in \{0,1\}$, this implies

$$\left|\frac{\partial}{\partial x_1}\mathcal{L}_{\mathrm{CE}}(x_1,x_2)\right| \leq 1.$$

By symmetry we conclude that $\mathcal{L}_{\mathrm{CE}}$ is Lipschitz continuous with constant 1.

Last, let $(x_1, x_2) \in [-K, K]^2$ and without loss of generality $y_1 = 1$ and $y_2 = 0$. We have

$$\begin{aligned}|\mathcal{L}_{\mathrm{CE}}(x_1,x_2)| =& -\log\left(\frac{e^{x_1}}{e^{x_1}+e^{x_2}}\right)\\ =&\log\left(1+\frac{e^{x_2}}{e^{x_1}}\right)\\ \leq&\log\left(1+e^{2K}\right).\end{aligned}$$

□

The above lemma tells us that in order to bound the infinity norm of the loss we must bound the domain of the loss - the output of the MPNN.

Lemma D.2. *Let $\Theta = \left((\Phi^{(l)})_{l=1}^T, (\Psi^{(l)})_{l=1}^T\right)$ be a MPNN s.t. Assumption 7. is satisfied. Consider a graph with N nodes and a graph feature map $\mathbf{f} \in \mathbb{R}^{N\times F}$. Then,*

$$\|\Theta_G^P(\mathbf{f})\|_\infty \leq A' + A''\|\mathbf{f}\|_{\infty;\infty},$$

where

$$A' = \sum_{l=1}^T \left(L_{\Psi^{(l)}}\|\Phi^{(l)}(0,0)\|_\infty + \|\Psi^{(l)}(0,0)\|_\infty\right)\prod_{l'=l+1}^T L_{\Psi^{(l')}}\left(L_{\Phi^{(l')}}+1\right)$$

and

$$A'' = \prod_{l=1}^T L_{\Psi^{(l)}}\left(1+L_{\Phi^{(l)}}\right).$$

Proof. Let G be a graph with weight matrix $\mathbf{W} = (W_{i,j})_{i,j=1\ldots,N}$. Let $l = 0, \ldots, T-1$. Then, for $k = 0, \ldots, l$, we have

$$\begin{aligned}
\|\mathbf{f}_i^{(k+1)}\|_\infty &= \left\|\Psi^{(k+1)}\left(\mathbf{f}_i^{(k)}, \mathbf{m}_i^{(k+1)}\right)\right\|_\infty \\
&\leq \left\|\Psi^{(k+1)}\left(\mathbf{f}_i^{(k)}, \mathbf{m}_i^{(k+1)}\right) - \Psi^{(k+1)}(0,0)\right\|_\infty + \|\Psi^{(k+1)}(0,0)\|_\infty \\
&\leq L_{\Psi^{(k+1)}}\left(\|\mathbf{f}_i^{(k)}\|_\infty + \left\|\mathbf{m}_i^{(k+1)}\right\|_\infty\right) + \|\Psi^{(k+1)}(0,0)\|_\infty,
\end{aligned} \tag{64}$$

where $\mathbf{m}_i^{(k+1)} = \frac{1}{\sum_{j=1}^N W_{i,j}} \sum_{j=1}^N W_{i,j} \Phi^{(k+1)}\left(\mathbf{f}_i^{(k)}, \mathbf{f}_j^{(k)}\right)$. For this message term, we have

$$\begin{aligned}
\|\mathbf{m}_i^{(k+1)}\|_\infty &= \left\|\frac{1}{\sum_{j=1}^N W_{i,j}} \sum_{j=1}^N W_{i,j} \Phi^{(k+1)}\left(\mathbf{f}_i^{(k)}, \mathbf{f}_j^{(k)}\right)\right\|_\infty \\
&\leq \left\|\max_{j=1,\ldots,N} \Phi^{(k+1)}\left(\mathbf{f}_i^{(k)}, \mathbf{f}_j^{(k)}\right)\right\|_\infty \\
&\leq \max_{j=1,\ldots,N} L_{\Phi^{(k+1)}} \|\mathbf{f}_j^{(k)}\|_\infty + \|\Phi^{(k+1)}(0,0)\|_\infty.
\end{aligned} \tag{65}$$

Denote $\|\mathbf{f}\|_{\infty;\infty} = \max_{i=1,\ldots,N} \max_{j=1,\ldots,F} |\mathbf{f}_{i,j}|$ for $\mathbf{f} \in \mathbb{R}^{N \times F}$. We have as a result of (64) and (65)

$$\begin{aligned}
&\|\mathbf{f}^{(k+1)}\|_{\infty;\infty} \\
&\leq L_{\Psi^{(k+1)}}\left(\|\mathbf{f}^{(k)}\|_{\infty;\infty} + \left(L_{\Phi^{(k+1)}} \|\mathbf{f}^{(k)}\|_{\infty;\infty} + \|\Phi^{(k+1)}(0,0)\|_\infty\right)\right) + \|\Psi^{(k+1)}(0,0)\|_\infty
\end{aligned}$$

which we can write as

$$\begin{aligned}
&\|\mathbf{f}^{(k+1)}\|_{\infty;\infty} \\
&\leq L_{\Psi^{(k+1)}}\left(1 + L_{\Phi^{(k+1)}}\right) \|\mathbf{f}^{(k)}\|_{\infty;\infty} + L_{\Psi^{(k+1)}} \|\Phi^{(k+1)}(0,0)\|_\infty + \|\Psi^{(k+1)}(0,0)\|_\infty.
\end{aligned}$$

We apply Lemma B.11 to solve this recurrence relation, to get

$$\begin{aligned}
\|\mathbf{f}^{(k)}\|_{\infty;\infty} \leq &\sum_{l=1}^k \left(L_{\Psi^{(l)}} \|\Phi^{(l)}(0,0)\|_\infty + \|\Psi^{(l)}(0,0)\|_\infty\right) \prod_{l'=l+1}^k L_{\Psi^{(l')}}(1 + L_{\Phi^{(l')}}) \\
&+ \|\mathbf{f}^{(0)}\|_{\infty;\infty} \prod_{l=1}^k L_{\Psi^{(l)}}(1 + L_{\Phi^{(l)}})
\end{aligned}$$

Now, since for general bounded functions $F : \chi \to \mathbb{R}^n$ and $x_1, \ldots x_N \in \chi$

$$\left\|\frac{1}{N} \sum_{i=1}^N F(x_i)\right\|_\infty \leq \|F\|_{\infty,\infty},$$

the proof is done. □

Note that using our analysis, for the MPNN architecture presented in Section D.2.2, the loss is not just $\mathcal{L}_{\text{CE}}$, but the composition of $\mathcal{L}_{\text{CE}}$ on the last linear layer of the network. We denote this total loss by $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} \circ Q$. Hence, in our analysis the Lipschitz constant of the total loss is bounded by

$$L_{\mathcal{L}_{\text{total}}} = \|Q\|_\infty,$$

where $\|Q\|_\infty$ is the induced infinity norm of the matrix Q. The infinity norm bound of the total loss is bounded by

$$\|\mathcal{L}_{\text{total}}\|_\infty \leq \log(1 + e^{2(\|Q\|_\infty K + b)}),$$

where K is the infinity norm of the MPNN.

D.2.6 Details on the Computation of Bounds from Other Papers

The papers [Liao et al., 2021] and [Garg et al., 2020] do not provide generalization bounds for general MPNNs, but only for a specific architecture – GNNs with mean field updates, as defined in Definition C.8, namely

$$\mathbf{f}_i^{(l)} = \Psi\left(W_1\mathbf{x}_i + W_2\rho\left(\sum_{u\in\mathcal{N}(v)}\Phi(\mathbf{f}_j^{(l-1)})\right)\right),$$

where ρ, Ψ and Φ are nonlinear transformations, and W_1 and W_2 are linear transformations. This is followed by a global pooling layer, which takes as an input $\mathbf{f}^{(T-1)} \in \mathbb{R}^{N\times K}$, and returns the vector

$$\frac{1}{N}\mathbf{1}_N\mathbf{f}^{(T-1)}W_T \in \mathbb{R}^{1\times K},$$

where W_T is a linear transformation. Here $\mathbf{1}_N$ denotes the vector $(1,\dots,1) \in \mathbb{R}^{1\times N}$, where N is the number of nodes in the graph. As described in Subsection C.4.1, GNNs with mean field updates are a special case of MPNNs.

The generalization bounds in [Liao et al., 2021] and [Garg et al., 2020] are formulated in terms of the following constants: $\zeta = \min\left(\|W_1\|_2, \|W_2\|_2, \|W_T\|_2\right)$, $|w|_2^2 = \|W_1\|_F^2 + \|W_2\|_F^2 + \|W_T\|_F^2$, $\lambda = \|W_1\|_2\|W_T\|_2$, $\xi = L_\Psi\frac{(d\mathcal{C})^{l-1}-1}{d\mathcal{C}-1}$, and the percolation complexity $\mathcal{C} = L_\Psi L_\rho L_\Phi\|W_2\|_2$, where L_Ψ, L_ρ and L_Φ are the Lipschitz constants of Ψ, ρ and Φ. For the calculation of the generalization bounds, we use the fully non-asymptotic generalizations bounds, provided in [Liao et al., 2021, Subsection A.7]. There, the PAC-Bayesian based bound is given by

$$\sqrt{\frac{42^2B^2\left(\max\left(\eta^{-(T+1)}, (\lambda\zeta)^{\frac{T+1}{T}}\right)\right)^2T^2h\log(4Th)|w|_2^2}{\gamma^2m}}. \tag{66}$$

The Rademacher based bound is given by

$$48h\|W_T\|_2Z\sqrt{\frac{3\log\left(24\|W_T\|_2\sqrt{m}\max\left(Z, M\sqrt{h}\max(B\|W_1\|_2, \bar{R}\|W_2\|_2)\right)\right)}{\gamma^2m}}. \tag{67}$$

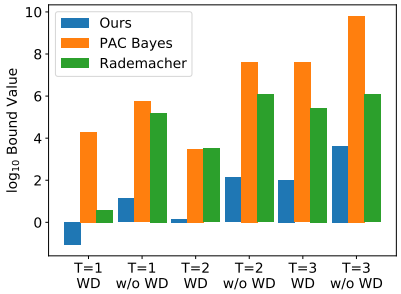
Note that GraphSage cannot be described in terms of mean field update networks, and vice versa. In order to still report some comparison between the generalization bounds, we offer some conversion between the constants of the two methods, and then apply the PAC-Bayes and Rademacher bounds on the converted bounds. It should be noted that the comparison is a bit like "comparing apples to oranges," but still gives insight into the respective bounds, their asymptotics, and their usefulness in practical situations.

Since the transformation by $\Psi(W_1(\cdot) + W_2 \circ \rho(\cdot))$ can be seen as an update function, similarly to the one in GraphSage, we set in the PAC-Bayes bound $\|W_1\|_2 = \|W_2\|_2 = 1/T\sum_{l=1}^{T}L_{\Psi^{(l)}}$, where $L_{\Psi^{(l)}}$ is the Lipschitz constant of the update function of GraphSage in the l-th layer. The message function in GraphSage is the identity, which corresponds to Φ. We thus convert this to $L_\Phi = 1$ in the PAC-Bayes generalization bound. Finally, we give a lower bound for the maximum degree over all graphs in the datasets by setting $d = Nd_{\min}$ (note that the PAC-Bayes and Rademacher complexity based bounds increase with increasing maximum degree).

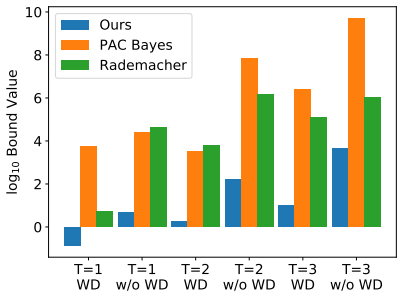
D.2.7 Generalization Comparison Results

The results are reported in Figure 3. The different experimental setting are given on the x-Axis. We report experiments for MPNNs with depth $T = 1, 2, 3$ with weight decay (WD) and without weight decay (w/o WD). The bound values are reported in a logarithmic y-Axis to improve comparability. In addition to the figures, we also provide numerical values of the bound calculations in Table 3.

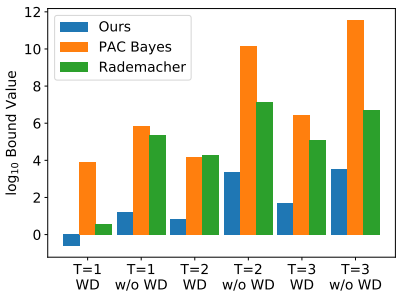
Our generalization bound is tighter than the PAC-Bayes bound and the Rademacher bound under all settings, i.e., for all datasets, for all depths, with weight decay and also without weight decay.



i) Generalization bounds on **ER-SBM**



ii) Generalization bounds on **ER-EXP**



iii) Generalization bounds on **EXP-SBM**

Figure 3: The generalization bounds on all datasets, i.e., i) **ER-SBM**, ii) **ER-EXP** and iii) **EXP-SBM**, for different number of layers $T = 1, 2, 3$ with weight decay (WD) and without weight decay (w/o WD). Bounds are given in a $\log_{10}$-scale.

D.3 Additional Comparison of the Generalization Bounds

In this subsection, we present additional plots of the generalization bounds which showcase the dependency on the average graph sizes in the dataset. The parameters in these plots are set not for a specific dataset and trained network. The plots can be interpreted as the bounds corresponding to training with certain constraints or regularization terms leading to the respective constants (Lipschitz bounds and infinity norms).

We consider a theoretical setting in which we assume that the following parameters are given: The dataset has 50K graphs, randomly sampled from RGMs with graphons that have maximum infinity norm $\|W\|_\infty = 0.4$ and Lipschitz norm $L_W = 0.5$. We assume that the metric-space signal are bounded by 0.5 and have Lipschitz constants of maximum 0.5. Furthermore, we assume there is a linear layer after pooling such that the norms of weight matrix and of the bias are upper bounded by 0.5 and 0.1, respectively. The infinity and Lipschitz norms of the loss function are assumed to be bounded by 1.

We then consider different datasets with graphs of average size $N = 2^4, 2^5, \ldots, 2^{25}$. Since the PAC-Bayes and Rademacher generalization bounds scale with the maximum node degree d of the graphs, we estimate the degree by setting $d = N \cdot d_{\min}$, where $d_{\min}$ is the graphon degree. We report our generalization bound with respect to the graph size in Figure 4. The comparison with other generalization bounds is given in Figure 5. As expected by our theoretical results, our generalization bound decays with respect to the average graph size. In contrast, we see that both the PAC Bayes based bound and the Rademacher based bound increase with respect to the increasing graph size.

In Figure 6 we showcase the dependency of our generalization bound on the Lipschitz constant of the graphons. For this, we fix the graph sizes in the dataset to 1000, and compute the resulting bounds for increasing Lipschitz norms. The rest of the parameters are as specified above. We plot the generalization bound for MPNNs with depth $1, 2$ and 3 in Figure 6.

Table 3: Bound comparisons on all synthetic datasets.

T = 1 WD	ER - SBM	ER - EXP	SBM - EXP
Rademacher	3.9597×10^{0}	5.5278×10^{0}	3.6869×10^{0}
PAC-Bayesian	1.9597×10^{4}	5.8187×10^{3}	7.8245×10^{3}
Ours	$\mathbf{8.8373 \times 10^{-2}}$	$\mathbf{1.325 \times 10^{-1}}$	$\mathbf{2.4495 \times 10^{-1}}$
T = 1 w/o WD			
Rademacher	1.7329×10^{5}	4.3686×10^{4}	2.2311×10^{5}
PAC-Bayesian	6.0161×10^{5}	2.6146×10^{4}	7.2969×10^{5}
Ours	$\mathbf{1.3534 \times 10^{1}}$	$\mathbf{4.4827 \times 10^{0}}$	$\mathbf{1.7109 \times 10^{1}}$
T = 2 WD			
Rademacher	3.2439×10^{3}	6.3963×10^{3}	1.9428×10^{4}
PAC-Bayesian	3.0695×10^{3}	3.2992×10^{3}	1.4299×10^{4}
Ours	$\mathbf{1.2817 \times 10^{0}}$	$\mathbf{1.7047 \times 10^{0}}$	$\mathbf{5.4857 \times 10^{0}}$
T = 2 w/o WD			
Rademacher	1.1943×10^{6}	1.4238×10^{6}	1.3619×10^{7}
PAC-Bayesian	4.0392×10^{7}	6.9262×10^{7}	1.3817×10^{10}
Ours	$\mathbf{1.1875 \times 10^{2}}$	$\mathbf{1.478 \times 10^{2}}$	$\mathbf{1.9353 \times 10^{3}}$
T = 3 WD			
Rademacher	2.7221×10^{5}	1.2286×10^{5}	1.2529×10^{5}
PAC-Bayesian	3.9963×10^{7}	2.6016×10^{6}	2.6141×10^{6}
Ours	$\mathbf{8.1312 \times 10^{1}}$	$\mathbf{8.1904 \times 10^{0}}$	$\mathbf{2.7574 \times 10^{1}}$
T = 3 w/o WD			
Rademacher	1.1762×10^{6}	1.0872×10^{6}	4.8271×10^{6}
PAC-Bayesian	6.225×10^{9}	5.1689×10^{9}	3.7028×10^{11}
Ours	$\mathbf{3.1606 \times 10^{3}}$	$\mathbf{3.5428 \times 10^{3}}$	$\mathbf{1.8596 \times 10^{3}}$

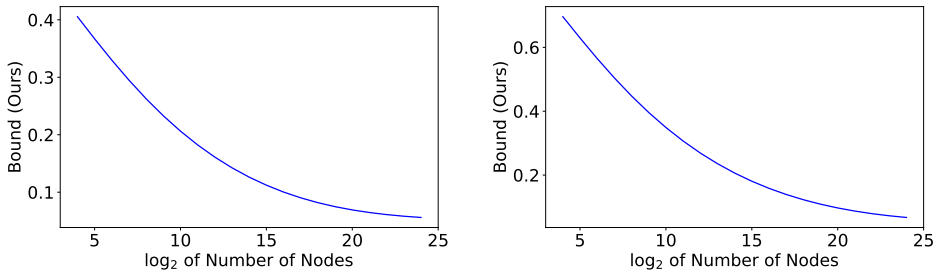


Figure 4: Our generalization bounds with respect to increasing average graph sizes. On the x-axis we give the average number of nodes in the dataset in $\log_2$-scale. On the y-axis, we give our generalization bound.

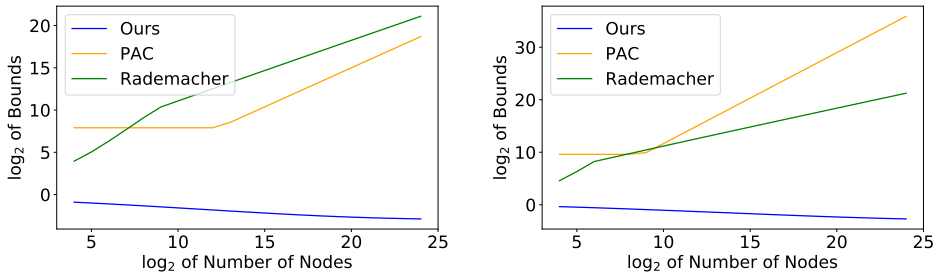


Figure 5: The generalization bounds with respect to increasing average graph sizes. On the x-axis, we give the average number of nodes in the dataset in $\log_2$-scale. On the y-Axis, we give our generalization bound, the PAC-Bayes based bound and the Rademacher complexity based bound for MPNNs with depth 2 (left) and depth 3 (right) also in $\log_2$-scale.

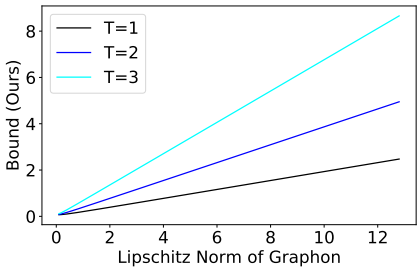


Figure 6: Our generalization bounds with respect to increasing average Lipschitz norm of the graphon. On the x-axis we give the maximal Lipschitz norm of the graphons from which we sampl the dataset. On the y-axis, we give our generalization bound. The rest of the parameters are equal to the parameters in the setting of Figure 5 (see Subsection D.3) with fixed graph size $N = 1000$.

E Background in Random Processes

In this section, we provide background information in probability theory, and focus on random processes and concentration of measure inequalities.

Definition E.1 (Definition 7.1.1. in [Vershynin, 2018]). *A* random process *is a collection of random variables $(Y_t)_{t\in T}$ on the same probability space, which are indexed by the elements t of some set T.*

The following lemma provides an upper bound on the probability that the sum of bounded independent random variables deviates from its expected value by more than a certain amount.

Theorem E.2 (Hoeffding's Inequality). *Let $Y_1, \dots, Y_N$ be independent random variables such that $a \leq Y_i \leq b$ almost surely. Then, for every $k > 0$,*

$$\mathbb{P}\Big(\Big|\frac{1}{N}\sum_{i=1}^{N}(Y_i - \mathbb{E}[Y_i])\Big| \geq k\Big) \leq 2\exp\Big(-\frac{2k^2N}{(b-a)^2}\Big).$$

Definition E.3 (Definition 2.5.6 in [Vershynin, 2018]). *A random variable Y is called a* sub-Gaussian random variable *if there exists a constant $K \in \mathbb{R}$ such that $\mathbb{E}\big[\exp\big(Y^2/K^2\big)\big] \leq 2$. The* sub-Gaussian norm *of a sub-Gaussian random variable X is defined as*

$$\|Y\|_{\psi_2} = \inf\Big\{t > 0 : \mathbb{E}\big[\exp\big(Y^2/t^2\big)\big] \leq 2\Big\}.$$

Lemma E.4 (Example 2.5.8 in [Vershynin, 2018]). *Any bounded random variable Y is sub-Gaussian with*

$$\|Y\|_{\psi_2} \leq \frac{1}{\sqrt{\ln(2)}}\|Y\|_\infty.$$

Definition E.5 (Sub-Gaussian increments, Definition 8.1.1 in [Vershynin, 2018]). *Consider a random process $(Y_x)_{x\in\chi}$ on a metric space (χ, d). We say that the process has* sub-Gaussian increments *if there exists a constant $K \geq 0$ such that*

$$\|Y_x - Y_{x'}\|_{\psi_2} \leq K d(x, x')$$

for all $x, x' \in \chi$. We call $(\|Y_x - Y_{x'}\|_{\psi_2})_{x,x'\in\chi}$ the sub-Gaussian increments of $(Y_x)_{x\in\chi}$.

Lemma E.6 (Centering of sub-Gaussian random variables, Lemma 2.6.8 in [Vershynin, 2018]). *If Y is a sub-Gaussian random variable, then so is $Y - \mathbb{E}[Y]$, and*

$$\|Y - \mathbb{E}[Y]\|_{\psi_2} \leq \Big(\frac{2}{\ln(2)} + 1\Big)\|Y\|_{\psi_2}.$$

Lemma E.7 (Proposition 2.6.1 in [Vershynin, 2018]). *Let $Y_1, \ldots, Y_N$ be independent mean-zero sub-Gaussian random variables. Then, $\sum_{i=1}^N Y_i$ is also a sub-Gaussian random variable, and*

$$\|\sum_{i=1}^N Y_i\|_{\psi_2}^2 \leq \frac{2}{\sqrt{2}} e \sum_{i=1}^N \|Y_i\|_{\psi_2}^2.$$

Theorem E.8 (Dudley's Inequality, Theorem 8.1.6 in [Vershynin, 2018]). *Let $(Y_x)_x$ be a random process on a metric space (χ, d) with sub-Gaussian increments, i.e., there exists a $K \geq 0$ such that $\|Y_x - Y_{x'}\|_{\psi_2} \leq K d(x, x')$ for all $x, x' \in \chi$. Then, for every $u \geq 0$, the event*

$$\sup_{x,x'\in\chi} |Y_x - Y_{x'}| \leq CK\Big(\int_0^\infty \sqrt{\log \mathcal{C}(\chi, \varepsilon, d)}d\varepsilon + u\mathrm{diam}(\chi)\Big)$$

holds with probability at least $1 - 2\exp(-u^2)$, where $\mathcal{C}(\chi, \varepsilon, d)$ is defined in Definition A.1 and C is a universal constant, specified in [Vershynin, 2018, Chapter 8].