Convergence of Message Passing Graph Neural Networks with Generic Aggregation On Large Random Graphs

Matthieu Cordonnier, Nicolas Keriven, Nicolas Tremblay, Samuel Vaiter

CONVERGENCE OF MESSAGE PASSING GRAPH NEURAL NETWORKS WITH GENERIC AGGREGATION ON LARGE RANDOM GRAPHS

MATTHIEU CORDONNIER[†], NICOLAS KERIVEN[‡], NICOLAS TREMBLAY[†], AND SAMUEL VAITER[§]

Abstract. We study the convergence of message passing graph neural networks on random graph models to their continuous counterpart as the number of nodes tends to infinity. Until now, this convergence was only known for architectures with aggregation functions in the form of degree-normalized means. We extend such results to a very large class of aggregation functions, that encompasses all classically used message passing graph neural networks, such as attention-based mesage passing or max convolutional message passing on top of (degree-normalized) convolutional message passing. Under mild assumptions, we give non asymptotic bounds with high probability to quantify this convergence. Our main result is based on the McDiarmid inequality. Interestingly, we treat the case where the aggregation is a coordinate-wise maximum separately, at it necessitates a very different proof technique and yields a qualitatively different convergence rate.

1. Introduction

Graph Neural Networks (GNNs) [1,2] are deep learning architectures largely inspired by Convolutional Neural Networks, that aim to extend convolutional methods to signal on graphs. GNNs are in practice of great interest as a large variety of real data live on an underlying graph structure. Examples of data for which GNNs have achieved state-of-the-art performance in the recent past include chemistry molecules, biological proteins and node clustering [3–5].

The duality of the convolutional product has led to two ways of defining GNNs. On the one hand, convolution as a pointwise product of frequencies in the Fourier domain has justified the design of so-called Spectral Graph Neural Networks [6] (SGNNs), in which one introduces a graph Fourier transform through a chosen graph shift operator [7] to legitimate the use of polynomial filters. On the other hand, the spatial interpretation sees the convolution as local aggregations of neighborhood information, leading to Message Passing Neural Networks (MPGNNs) [3, 8]. The message passing paradigm consists of iteratively updating each node via the aggregation of messages from each of its neighbors. This framework is often favored

[†]CNRS, Univ. Grenoble Alpes, Grenoble-INP, GIPSA-lab, Grenoble, France
[‡]CNRS, IRISA, Rennes, France
[§]CNRS, Univ. Côte d'Azur, Laboratoire J.A. Dieudonné, Nice, France

This work was partially supported by the French National Research Agency in the framework of the "France 2030" program (ANR-15-IDEX-0002), the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01), and the ANR grants GRANDMA (ANR-21-CE23-0006), GRANOLA (ANR-21-CE48-0009) and GRAVA (ANR-18-CE40-0005).

due to its inherent flexibility: messages and aggregation functions are unconstrained as long as they stay invariant to node reordering, *i.e.*, as long as they match on isomorphic graphs. Besides, SGNNs layers are mostly made of polynomials of graph shift operators which are a form of message passing, defined by a choice of graph shift operator and a polynomial degree. As such, SGNNs can be seen as a subcase of the more versatile message-passing framework.

From a theoretical perspective, the study of GNNs' expressivity, *i.e.*, the class of functions that GNNs can approximate, is an active research topic. While it is well known that multi-layer perceptrons are universal approximators [9, 10], the case is more involved for GNNs. Recent research has shown that their approximation power is equivalent to their ability to distinguish non isomorphic graphs [11], but cannot significantly outperform the so-called Weisfeiler-Lehman algorithm [12]. However, that point of view becomes questionable when dealing with very large graphs: it may turn irrelevant to focus on one-to-one correspondences between nodes and edges, especially given that real-world networks are rarely fully known and their structure may evolve quickly over time. In such cases, one should rather focus on global properties such as the degree distribution, the size of connected components, etc.

For that purpose, random graph models with growing number of nodes are privileged tools. Several papers have therefore focused on GNNs for large random graphs. In [13], it is shown that some class of GNN classifier will map a graph to a certain output with probability either zero or one, in the large graph limit. In [14–17], the authors show that SGNNs or MPGNNs with degree normalized aggregation, converge to a limit *continuous* architecture, which is called cGNN in [14, 15]. The discriminative power of those cGNNs have been studied in [18]. Most of them use dense random graph models where nodes are sampled from a compact metric space $(\mathcal{X}, P)$ and randomly connected through a kernel $W : \mathcal{X}^2 \to [0, 1]$ according to their latent positions. More precisely, they define a continuous equivalent of the discrete GNN as a deep architecture that propagates functions over the latent space and they show that, under mild regularity conditions, the discrete network converges to the continuous counterpart.

Contributions. In this paper, we study the convergence toward a continuous counterpart of MPGNNs with a generic aggregation function, whereas previous work [14–17] are restricted to SGNNs or MPGNNs with specific aggregations. Our main result, Theorem 5.7, states that for MPGNNs having a Lipschitz-type regularity, the discrete network on a large random graph is closed to is continuous counterpart with high probability. We quantify this convergence via a non asymptotic bound based on the well-known McDiarmid concentration inequality for multivariate functions of independent random variables. A special treatment is given to the case where the aggregation is a coordinate-wise maximum, for which Theorem 5.7 does not hold. In this case, we provide another non asymptotic proof of convergence towards a continuous counterpart, based on different concentration inequalities, in Theorem 5.12.

Related work. A classical method to gain insights on limits of discrete mathematical concepts is to embed them into a continuous framework. The Stirling formula or the Central Limit Theorem are two such examples. For graphs, a limit theory

has been initiated in the 2000's mainly by Lováz ans Szegedy [19]. They define a limit of graph as a kernel $W : \mathcal{X}^2 \to [0, 1]$ and call such a kernel a graphon. Further, they show that if one endows the space of graphons with a wisely chosen metric, one obtains a complete compact metric space. Moreover one can embed the set of all finite graphs in the space of graphons in such a way that makes this embedding dense. Conversely, a graphon on a probability space defines a random graph model. This type of random graphs model is very general and encompasses a lot of other models such as stochastic block models or Erdős–Rényi models. The idea of studying GNN limits from the point of view of graphons is not new: for instance, authors in [20] show that a SGNN trained on a random graph will perform well on another graph drawn from the same graphon. In [13], it is shown that GNN classifiers will match Erdős–Rényi random graphs to a particular output with probability either zero or one in the limit. Closer to this work are our previous articles [14, 15] in which we prove the convergence of SGNNs defined on a random graph model to continuous equivalents and then study their stability to deformation of the underlying graph model. Also, in [17], the authors prove convergence of MPGNNs in the particular case where the aggregation is a degree normalized mean, and use this result to further establish a generalization bound. The case where the aggregation function is a maximum is often mentioned as a straightforward possibility for MPGNNs, it is even implemented as a default option in popular python library such as PyTorch [21]. However, it is actually rarely considered in the literature: to our knowledge, [22] is the only work that makes significant use of it.

Outline In Section 2, we give some basic definitions. In Section 3 we define MPGNNs with a generic aggregation function as deep architectures that propagate a signal over a graph and that must be coherent to graph isomorphism. In Section 4 we introduce continuous-MPGNN (cMPGNN) which are the counterpart of discrete MPGNNs that propagate a function over a compact probability space, alongside a connectivity kernel. As a discrete MPGNN must be symmetric to graph isomorphism, we demand the cMPGNN to be symmetric to probability space isomorphism. In Section 5, we focus on MPGNNs when applied on random graphs and describe what class of cMPGNN would be their natural limit. Our main result is Theorem 5.7: it provides necessary conditions under which the discrete network converges to its continuous counterpart. We make use of the McDiarmid concentration inequality to derive a non asymptotic bound with high probability of the deviation between the outputs of the MPGNN and its limit cMPGNN. Overall, we conclude that a sufficient condition of convergence is for the aggregation to have sharp bounded differences. All along the paper, we illustrate our concepts on classical GNN examples from the basic Graph Convolutional Network to the more sophisticated Graph Attentional Network [23]. We give a particular treatment to the case of the maximum aggregation. Indeed, its behavior turns out to be significantly different than for the other examples and do not fit into the class of MPGNNs having sharp bounded differences. Nevertheless, in Theorem 5.12 we make use of other specific concentration bounds to prove another non asymptotic bound between max MPGNN and its limit cMPGNN.

2. Notations and Definitions

We start by expliciting the notations that will hold throughout the paper. The letter d (and its derived $d_0, d^{(0)}, \ldots$) will represent the dimension of a real vector space, the letter n will denote the number of nodes in a graph, and the letter L will refer to the total number of layers in a deep architecture. Whenever we need to index something relatively to vertices of a graph, we use a subscript indexation (*e.g.*, z_i) and in the case of layers, we employ a superscript (*e.g.*, $z^{(l)}$).

We fix a positive integer d and $(\mathbb{R}^d, \|\cdot\|_\infty, \mathbb{B}(\mathbb{R}^d))$ the d-dimensional real vector space endowed with the infinite norm $\|x\|_\infty = \max_i |x_i|$ as well as its Borel sigma algebra. Except when specified differently, any topological concept, such as balls, continuity, *etc.*, will be considered relatively to the norm $\|\cdot\|_\infty$. All along this paper, $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$ and $\mathcal{B}(\mathcal{X})$ its Borel sigma algebra defined as the sigma algebra generated by the $U \cap \mathcal{X}$, for the open sets U of $\mathbb{R}^d$.

The group of permutations of $\{1, \ldots, n\}$ is denoted as S_n. If $x = (x_1, \ldots, x_n)$ is a n-tuple and σ an element of S_n, we define the n-tuple $\sigma \cdot x$ as $\sigma \cdot x = (x_{\sigma^{-1}(1)}, \ldots, x_{\sigma^{-1}(n)})$.

The set of bijections ϕ of $\mathcal{X}$ such that both ϕ and ϕ^{-1} are measurable is a group for the composition of functions. We call this group the group of automorphisms of $\mathcal{X}$ and denote it as $\mathrm{Aut}(\mathcal{X})$. We denote as $\mathcal{P}(\mathcal{X})$ the set of probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. For a measure $P \in \mathcal{P}(\mathcal{X})$ and a bijection $\phi \in \mathrm{Aut}(\mathcal{X})$, the push forward measure of P through ϕ is defined as $\phi_\# P(A) = P(\phi^{-1}(A))$ for all A in $\mathcal{B}(\mathcal{X})$. Since this makes the group $\mathrm{Aut}(\mathcal{X})$ acting on the set of probability measures on $\mathcal{X}$, we also use the notation $\phi \cdot P = \phi_\# P$, which is standard for a (left) group action. For the same reason, we shall use the notation $\phi \cdot f = f \circ \phi^{-1}$ and $\phi \cdot W = W(\phi^{-1}(\cdot), \phi^{-1}(\cdot))$ whenever f is a measurable function on $\mathcal{X}$ and W is a bivariate measurable function on $\mathcal{X} \times \mathcal{X}$.

For $P \in \mathcal{P}(X)$, the space $L^\infty_P(\mathcal{X}, \mathbb{R}^p)$ is the space of essentially bounded (equivalence classes of) maps from $\mathcal{X}$ to $\mathbb{R}^p$ endowed with the norm $\|f\|_{P,\infty} = \operatorname*{ess\,sup}_{P,\, x \in \mathcal{X}} \|f(x)\|_\infty$. When there is no ambiguity on P, The norm $\|\cdot\|_{P,\infty}$ is noted $\|\cdot\|_\infty$. The space $\mathcal{C}(\mathcal{X}, \mathbb{R}^p)$ is made of the continuous functions from $\mathcal{X}$ to $\mathbb{R}^p$. Since $\mathcal{X}$ is compact, any continuous map is bounded thus essentially bounded, which makes $\mathcal{C}(\mathcal{X}, \mathbb{R}^p)$ a subspace of $L^\infty_P(\mathcal{X}, \mathbb{R}^p)$.

Sets are represented between braces $\{\cdot\}$, whereas multisets, that is, sets in which an element is allowed to appear twice or more, are represented by double braces $\{\!\{\cdot\}\!\}$. If m and m' are two multisets of same size, say n, containing elements from a metric space $(\mathcal{E}, \delta)$, we define their distance by:

$$\delta(m, m') = \min_{\sigma \in S_n} \max_{x_i \in m, x'_i \in m'} \delta(x_i, x'_{\sigma(i)}). \tag{2.1}$$

We define the sampling operator the following way. If $f : \mathcal{E}_0 \to \mathcal{E}_1$ and $X = (x_1, \ldots, x_n) \in \mathcal{E}_0^n$:

$$S_X f = (f(x_i), \ldots, f(x_n)) \in \mathcal{E}_1^n. \tag{2.2}$$

2.1. Graph-related definitions

In this subsection, we introduce the concepts of discrete graph, graph signal and graph isomorphism.

Graph. A non oriented weighted graph G with n vertices is defined by a triplet (V, E, w), where $V = \{v_1, \dots, v_n\}$ is a finite set called the set of vertices (or nodes) and E is the set of edges. The set of neighbors of a vertex v_i in G is referred to as $\mathcal{N}_G(v_i)$ or simply $\mathcal{N}(v_i)$ when the underlying graph is clear from context. The weight function w assigns a nonnegative number to each edge. It is often represented by a symmetric function $w : V^2 \to \mathbb{R}^+$ and the abbreviation $w_{i,j}$ is used to denote the weight $w(v_i, v_j) = w(v_j, v_i)$ where $\{v_i, v_j\} \in E$. In this paper, "graph" will always mean "undirected and weighted graph". The set of graphs defined on the vertex set V is denoted as $\mathcal{G}(V)$.

Graph signal. Given a graph $G \in \mathcal{G}(V)$, where $|V| = n$, a signal on G is a map from the set of vertices V to $\mathbb{R}^d$ that assigns a d-dimensional vector z_i to each vertex v_i. The images from all vertices are stacked into a tensor Z of size $n \times d$. Abusing notations, we may not distinguish between the map and its image Z, the latter being also named the signal.

Graph isomorphism. Two graphs G_1 and G_2 in $\mathcal{G}(V)$, where $|V| = n$, are said to be isomorphic if there is a permutation $\sigma \in S_n$ such that $E_2 := \{\{v_{\sigma^{-1}(i)}, v_{\sigma^{-1}(j)}\} | \{v_i, v_j\} \in E_1\}$ and $w_2(v_i, v_j) = w_1(v_{\sigma^{-1}(i)}, v_{\sigma^{-1}(j)})$. In this case we note $G_2 = \sigma \cdot G_1$. Moreover, if Z is a signal on G_1 and $\sigma \in S_n$, $\sigma \cdot Z$ is an isomorphic signal on the graph $\sigma \cdot G_1$.

2.2. Random Graph Models

Random Graph Model. A random graph model is a couple (W, P) where P is a Borel probability measure on $\mathcal{X}$ and $W : \mathcal{X} \times \mathcal{X} \mapsto [0, 1]$ is a kernel, *i.e.*, a symmetric measurable function. One can interpret W as a totally connected graph on the vertex set $\mathcal{X}$ and whose weight function is W.

Random Graph. We generate random graphs from a random graph model (W, P) as follows. Given a positive integer n, we first draw n independent and identically distributed random variables from the distribution P, represented by $X_1, \dots, X_n$, which form the vertex set of the graph. The random graph is fully connected and has weight function W :

$$X_1, \dots, X_n \overset{iid}{\sim} P, \quad w_{i,j} = w_{j,i} = W(X_i, X_j).$$

When convenient, we will use the short notation $X = (X_1, \dots, X_n)$ for the tuple of the vertices of a random graph. We call $\mathcal{G}_n(W, P)$ the distribution from which random graphs with n nodes are drawn. We bring the reader's attention to the fact that in the above definition, a random graph is always fully connected and edge may have a weight equal to zero. A common approach [14] is to add a Bernoulli distribution to the connectivity in order to model random graphs with prescribed expected sparsity, but it is not done here for the sake of simplicity.

Random Graph Model isomorphism. Two probability measures P_1 and P_2 on $\mathcal{X}$ are said isomorphic if there is some ϕ in $\mathrm{Aut}(\mathcal{X})$ such that $P_2 = \phi_{\#} P_1$. Similarly, two random graph models (W_1, P_1) and (W_2, P_2) on $\mathcal{X}$ are said to be isomorphic

if there is a ϕ in $\mathrm{Aut}(\mathcal{X})$ such that $(W_2, P_2) = (\phi \cdot W_1, \phi \cdot P_1)$, in this case, we will note $(W_2, P_2) = \phi \cdot (W_1, P_1)$.

3. Message Passing Graph Neural Networks (MPGNNs)

A multilayer MPGNN iteratively propagates a signal over a graph. At each step, the current representation of every node's neighbors are gathered, transformed, and combined to update the node's representation. Broadly speaking, a MPGNN can be defined as a collection of L applications $(F^{(l)})_{1 \leq l \leq L}$ that act as follows. Let $G \in \mathcal{G}(V)$ be a graph with n nodes, and $Z = Z^{(0)} \in \mathbb{R}^{n \times d_0}$ be a signal on it. At each layer, denoting $Z^{(l)}$ as the current state of the signal, $Z^{(l+1)}$ is computed node-wise by:

$$z_i^{(l+1)} = F^{(l+1)}\left(z_i^{(l)}, \left\{\!\!\left\{ \left(z_j^{(l)}, w_{i,j}\right) \right\}\!\!\right\}_{v_j \in \mathcal{N}(v_i)}\right) \in \mathbb{R}^{d_{l+1}}. \tag{3.1}$$

So $Z^{(l+1)}$ is a $n \times d_{l+1}$ tensor. In (3.1), the $F^{(l)}$ take as arguments a vector, which is the current node's representation, and a multiset of pairs. Each pair is composed of a node from the neighborhood of the aforementioned running node, along with the corresponding weight. In the literature, the $F^{(l)}$ are often referred to as *aggregations* [24]. Their major property is to ignore the order in which the neighborhood information is collected, which is handeled by the use of a multiset.

Depending on the context, the final output of the MPGNN may be a signal over the graph, or a single vector representation for the entire graph. We call these two versions respectively the *equivariant* and the *invariant* versions of the network. We denote $\Theta_G(Z)$ as the output in the first case and $\bar{\Theta}_G(Z)$ in the second case, where $\bar{\Theta}_G$ use an additional pooling operation over the nodes, $R : \mathbb{R}^{n \times d_L} \to \mathbb{R}^{d_L}$, called the *readout* [24] function :

$$\Theta_G(Z) = Z^{(L)} \in \mathbb{R}^{n \times d_L}, \qquad \bar{\Theta}_G(Z) = R\left(\left\{\!\!\left\{ z_1^{(L)}, \dots, z_n^{(L)} \right\}\!\!\right\}\right) \in \mathbb{R}^{d_L} \tag{3.2}$$

A fundamental requirement for graph neural networks is to be consistent with graph isomorphism. More precisely, relabeling the nodes of the input graph signal must be the same as relabeling the nodes of the output in the *equivariant* case, and must must leave the output unchanged in the *invariant* case. This exactly corresponds to the concepts of invariance and equivariance for group actions and follows naturally from the definition of MPGNNs, as stated in the following proposition.

Proposition 3.1 (Invariance and equivariance of MPGNNs)**.** *Let $G \in G(V)$ with $|V| = n$. Then, Θ and $\bar{\Theta}$ are respectively S_n-equivariant and S_n-invariant, in the sense that for all $\sigma \in S_n$, for all $Z \in \mathbb{R}^{n \times d_0}$, we have $\Theta_{\sigma \cdot G}(\sigma \cdot Z) = \sigma \cdot \Theta_G(Z)$ and $\bar{\Theta}_{\sigma \cdot G}(\sigma \cdot Z) = \bar{\Theta}_G(Z)$.*

Proof. We prove the equivariant case. Let us introduce $\Lambda_G^{(l)} : Z^{(l-1)} \mapsto Z^{(l)}$ the layer functions such that $\Theta_G = \Lambda_G^{(L)} \circ \cdots \circ \Lambda_G^{(1)}$ by construction. Let $Z \in \mathbb{R}^{n \times d_{l-1}}$ be a signal on G. On the one hand, $\Lambda_{\sigma \cdot G}^{(l)}(\sigma \cdot Z) = Y$ is the signal on $\sigma \cdot G$ such that

$$y_i = F^{(l)}\left(z_{\sigma^{-1}(i)}, \left\{\!\!\left\{ \left(z_{\sigma^{-1}(j)}, w_{\sigma^{-1}(i), \sigma^{-1}(j)}\right) \right\}\!\!\right\}_{v_{\sigma^{-1}(j)} \in \mathcal{N}_{\sigma \cdot G}(v_{\sigma^{-1}(i)})}\right),$$

by definition of $\sigma \cdot G$ and $\sigma \cdot Z$. On the other hand, $\sigma \cdot \Lambda_G^{(l)}(Z) = Y'$ is the signal on $\sigma \cdot G$ such that

$$\begin{aligned} y'_i &= F^{(l)}\left(z_{\sigma^{-1}(i)}, \left\{\!\!\left\{ \left(z_j, w_{\sigma^{-1}(i),j}\right) \right\}\!\!\right\}_{v_j \in \mathcal{N}_G(v_{\sigma^{-1}(i)})}\right) \\ &= F^{(l)}\left(z_{\sigma^{-1}(i)}, \left\{\!\!\left\{ \left(z_{\sigma^{-1}(j)}, w_{\sigma^{-1}(i),\sigma^{-1}(j)}\right) \right\}\!\!\right\}_{v_{\sigma^{-1}(j)} \in \mathcal{N}_{\sigma\cdot G}(v_{\sigma^{-1}(i)})}\right). \end{aligned}$$

So $Y = Y'$ which means that $\Lambda^{(l)}_{\sigma\cdot G}(\sigma \cdot Z) = \sigma \cdot \Lambda^{(l)}_G(Z)$ is equivariant for all l. Thereby $\Theta_{\sigma\cdot G}(\sigma \cdot Z) = \sigma \cdot \Theta_G(Z)$ by composition. For the invariant case, R is clearly S_n-invariant since it has a multiset as input. The fact that the composition of an equivariant map followed by an invariant map is invariant yields the result. □

The role of the functions $F^{(l)}$ in (3.1) is crucial and there is a wide range of designs for them [25]. Nevertheless, we can encompass a large class of those designs in the following description. Fix a layer $F^{(l+1)}$ and a node v_i. After being gathered, the signals on the neighbors of this node are transformed by a learnable operation which is usually a classical multilayer perceptron (MLP) denoted as $\psi^{(l+1)}$. Then the transformed neighbors $\psi^{(l)}(z_j)$ are combined along with some coefficients

$$c_{i,j}^{(l+1)} = c^{(l+1)}\left(z_i^{(l)}, z_j^{(l)}, w_{i,j}\right) \tag{3.3}$$

in a way that is invariant to node relabeling. It appears that a natural way of doing so is to perform a *mean*, in a broad sense: an arithmetic mean, a weighted mean, a maximum, *etc.* Thus, we have a *mean* operator $M^{(l+1)}$ such that (3.1) can be rewritten as

$$\begin{aligned} &F^{(l+1)}\left(z_i^{(l)}, \left\{\!\!\left\{ \left(z_j^{(l)}, w_{i,j}\right) \right\}\!\!\right\}_{v_j \in \mathcal{N}(v_i)}\right) \\ &= M^{(l+1)}\left(\left\{\!\!\left\{ \left(\psi^{(l+1)}(z_j)^{(l)}, c_{i,j}^{(l+1)}\right) \right\}\!\!\right\}_{v_j \in \mathcal{N}(v_i)}\right) \end{aligned} \tag{3.4}$$

In the sequel, we discuss four examples that follow (3.4), the first three of which are very popular in the literature.

Example 1 (Convolutional Message Passing [6, 8, 25])**.** *The $c_{i,j}$ are the graph weights $w_{i,j}$. Each neighbor representation is multiplied by its corresponding weight and we combine them with an arithmetic mean. Notice that this is equivalent to a Convolutional Graph Neural Network (GCN) with polynomial filters of degree one.*

$$z_i^{(l+1)} = \frac{1}{|\mathcal{N}(v_i)|} \sum_{v_j \in \mathcal{N}(v_i)} w_{i,j} \psi^{(l+1)}\left(z_j^{(l)}\right).$$

In the invariant case, the readout function is an arithmetic mean:

$$R\left(\left\{\!\!\left\{z_1^{(L)}, \ldots, z_n^{(L)}\right\}\!\!\right\}\right) = \frac{1}{n}\sum_{i=1}^{n} z_i^{(L)}.$$

Example 2 (Degree normalized convolution)**.** *The $c_{i,j}$ are still the graph weights $w_{i,j}$ but a weighted mean is performed [17].*

$$z_i^{(l+1)} = \sum_{j \in \mathcal{N}(v_i)} \frac{w_{i,j}}{\sum_{k \in \mathcal{N}(v_i)} w_{i,k}} \psi^{(l+1)}\left(z_j^{(l)}\right).$$

In the invariant case, the readout function is an arithmetic mean.

Example 3 (Attention based Message Passing)**.** *Unlike the two examples above, the attentional coefficients are learnable and depend on all the possible parameters mentioned in* (3.3) *[23]. A weighted mean is then used.*

$$z_i^{(l+1)} = \sum_{j \in \mathcal{N}(v_i)} \frac{c^{(l+1)}\left(z_i^{(l)}, z_j^{(l)}, w_{i,j}\right)}{\sum_{k \in \mathcal{N}(v_i)} c^{(l+1)}\left(z_i^{(l)}, z_k^{(l)}, w_{i,k}\right)} \psi^{(l+1)}\left(z_j^{(l)}\right).$$

In the invariant case, the readout function is an arithmetic mean.

Example 4 (Max Convolutional Message Passing)**.** *The aggregation maximum is often mentioned as a possibility in the literature but is rarely treated [22]. Here the $c_{i,j}$ are also the graph weights $w_{i,j}$ but an element-wise maximum is used to combine everything:*

$$z_i^{(l+1)} = \max_{v_j \in \mathcal{N}(v_i)} w_{i,j} \psi^{(l+1)}\left(z_j^{(l)}\right).$$

In the invariant case, the readout function is an element-wise maximum.

$$R\left(\left\{\!\!\left\{z_1^{(L)}, \dots, z_n^{(L)}\right\}\!\!\right\}\right) = \max_{i=1,\dots,n} z_i^{(L)}.$$

4. Continuous MPGNNs (cMPGNNs) on random graph models

We define the continuous counterpart of MPGNNs, that we call continuous MPGNNs (cMPGNNs). Analogously to the discrete case, a cMPGNN is defined to be L operators $(\mathcal{F}^{(l)})_{1 \leq l \leq L}$ that propagate a function on $\mathcal{X}$ relatively to a random graph model. Let (W, P) be a random graph model and $f = f^{(0)} \in L^\infty_P(\mathcal{X}, \mathbb{R}^{d_0})$, $f^{(l+1)}$ is recursively computed by:

$$\forall x \in \mathcal{X} \quad f^{(l+1)}(x) = \mathcal{F}_P^{(l+1)}\left(f^{(l)}(x), \left(f^{(l)}, W(x, \cdot)\right)\right) \in \mathbb{R}^{d_{l+1}}. \tag{4.1}$$

Notice that $\mathcal{F}^{(l+1)}$ depends on the measure P. Considering the functions $f^{(l)}$ as signals on the vertex set $\mathcal{X}$, the update $f^{(l+1)}(x)$ of a node $x \in \mathcal{X}$ is calculated from the knowledge of its current representation $f^{(l)}(x)$ and all its "weighted neighborhood" $(f^{(l)}, W(x, \cdot))$. The latter being a short notation for the map $y \mapsto (f^{(l)}(y), W(x, y))$ at x fixed, which is the continuum equivalent of the multiset of pairs of weighted neighbors $\{\!\{(z_j^{(l)}, w_{i,j})\}\!\}_{v_j \in \mathcal{N}(v_i)}$ from (3.1). We denote $\Theta_{W,P}(f) = f^{(L)}$ the output in the equivariant case and $\bar{\Theta}_{W,P}(f)$ in the invariant case.

$$\Theta_{W,P}(f) = f^{(L)} \in L^\infty_P(\mathcal{X}, \mathbb{R}^{d_L}), \qquad \bar{\Theta}_{W,P}(f) = \mathcal{R}_P(\Theta_{W,P}(f)) \in \mathbb{R}^{d_L} \tag{4.2}$$

Where $\bar{\Theta}_{W,P}$ involves an additional continuum readout operator $\mathcal{R} : \mathcal{P}(\mathcal{X}) \times L^\infty_P(\mathcal{X}, \mathbb{R}^{d_L}) \to \mathbb{R}^{d_L}$. Naturally, we also demand the equivariant and invariant versions of the cMPGNN to respectively be equivariant and invariant to random graph model isomorphisms. To that extent, we impose the following assumption on the operators $\mathcal{F}^{(l)}$ and on $\mathcal{R}$:

Assumption 4.1. *There is a subgroup $H \subset \mathrm{Aut}(\mathcal{X})$ such that $\forall\ 1 \leq l \leq L$, $\forall f \in L^\infty_P(\mathcal{X}, \mathbb{R}^{d_l}), \forall \phi \in H$:*

$$\mathcal{F}^{(l)}_{\phi \cdot P}(f(x), (\phi \cdot f, W(x, \phi^{-1}(\cdot)))) = \mathcal{F}^{(l)}_P(f(x), (f, W(x, \cdot))) \ \text{a.s.}$$

and

$$\mathcal{R}_{\phi\cdot P}(\phi\cdot f))=\mathcal{R}_P(f).$$

Assumption 4.1 is largely inspired by the classical change of variable formula by push forward measure in Lebesgue integration. This formula states that for any ϕ bijective and measurable (ϕ^{-1} need not to be measurable here) and any measurable map f,

$$\int f\circ\phi dP=\int f d\phi_{\#}P. \tag{4.3}$$

It is easy to check that if, for example, $\mathcal{R}_P(f)=\int f dP$, then (4.3) implies Assumption 4.1 when ϕ is an automorphism of $\mathcal{X}$. Contrary to the discrete case, where the symmetry is valid for the full group S_n, we require here a symmetry for a subgroup of $\mathrm{Aut}(\mathcal{X})$ only. Ideally, one would like Assumption 4.1 to hold for $H=\mathrm{Aut}(\mathcal{X})$. However, in the next section, we will interpret some cMPGNN as limits of discrete MPGNN such that the graph isomorphism symmetry becomes a random graph model isomorphism symmetry as the number of nodes tends to infinity. In this context, the example of maximum aggregation (Example 4-d in the next section) will highlight the fact that, for a matter of existence of such a limit, one may have to restrict to a subgroup of $\mathrm{Aut}(\mathcal{X})$.

Proposition 4.2 (Invariance and equivariance of cMPGNNs)**.** *Let (W,P) be a random graph model on $\mathcal{X}$. Then, under Assumption 4.1, Θ and $\bar{\Theta}$ are respectively H-equivariant and H-invariant. Meaning that for any f, for any $\phi\in H$, $\Theta_{\phi\cdot(W,P)}(\phi\cdot f)=\phi\cdot\Theta_{W,P}(f)$ and $\bar{\Theta}_{\phi\cdot(W,P)}(\phi\cdot f)=\bar{\Theta}_{W,P}(f)$.*

Proof. We start by the equivariant case and the invariant one follows directly by composition with $\mathcal{R}$. Let $\Lambda^{(l)}_{W,P}$ be the layer operators such that $\Theta^{(L)}_{W,P}=\Lambda^{(L)}_{W,P}\circ\cdots\circ\Lambda^{(1)}_{W,P}$. Let $f\in L^\infty_P(\mathcal{X},\mathbb{R}^{d_{l-1}})$. On the one hand, $\phi\cdot\Lambda^{(l)}_{W,P}(f)$ is the map

$$\phi\cdot\Lambda^{(l)}_{W,P}(f)(x)=\mathcal{F}^{(l)}_P\left(\phi\cdot f(x),\left(f,W(\phi^{-1}(x),\cdot)\right)\right).$$

On the other hand, $\Lambda^{(L)}_{\phi\cdot(W,P)}(\phi\cdot f)$ is the map

$$\begin{aligned}\Lambda^{(L)}_{\phi\cdot(W,P)}(\phi\cdot f)(x)&=\mathcal{F}^{(l)}_{\phi_{\#}P}\Big(\phi\cdot f(x),\Big(\phi\cdot f,W(\phi^{-1}(x),\phi^{-1}(\cdot))\Big)\Big)\\&=\mathcal{F}^{(l)}_P\left(\phi\cdot f(x),\left(f,W(\phi^{-1}(x),\cdot)\right)\right)\end{aligned}$$

by the assumption 4.1. So the Proposition is true on all the $\Lambda^{(l)}_{W,P}$, thus also true on $\Theta^{(L)}_{W,P}$ by composition. For the invariant case, it is clear from Assumption 4.1 that $\mathcal{R}$ is $\mathrm{Aut}(\mathcal{X})$-invariant. The fact that the composition of an equivariant map followed by an invariant map is invariant yields the result. □

In the following are some examples of cMPGNN. The reader will of course see the connection to the previous Examples 1, 2, 3 and 4. In the next section, we will precisely see in what sense Examples 1 to 4, when applied on random graphs and as n grows large, tend to the following cMPGGNs.

Example a (Convolutional Message Passing)**.** *The arithmetic mean becomes an integral over the probability space:*

$$f^{(l+1)}(x) = \int_{y \in \mathcal{X}} W(x,y) \psi^{(l+1)} \left(f^{(l)}(y) \right) dP(y)$$

and, in the invariant case, the continuous readout is :

$$\mathcal{R}_P \left(f^{(L)} \right) = \int_{\mathcal{X}} f^{(L)} dP.$$

Example b (Degree Normalized Convolutional Message Passing)**.** *The continuous counterpart is:*

$$f^{(l+1)}(x) = \int_{y \in \mathcal{X}} \frac{W(x,y)}{\int_{t \in \mathcal{X}} W(x,t) dP(t)} \psi^{(l+1)} \left(f^{(l)}(y) \right) dP(y).$$

In the invariant case, the readout is the integral relatively to P.

Example c (Attention based Message Passing)**.** *The continuous counterpart is:*

$$f^{(l+1)}(x) = \int_{y \in \mathcal{X}} \frac{c^{(l+1)} \left(f^{(l)}(x), f^{(l)}(y), W(x,y) \right)}{\int_{t \in \mathcal{X}} c^{(l+1)} \left(f^{(l)}(x), f^{(l)}(t), W(x,t) \right) dP(t)} \psi \left(f^{(l)}(y) \right) dP(y).$$

In the invariant case, the readout is the integral relatively to P.

Example d (Max Convolutional Message Passing)**.** *The maximum becomes a component-wise essential supremum according to the probability measure P:*

$$f^{(l+1)}(x) = \operatorname*{ess\,sup}_{y \in \mathcal{X},\, P} W(x,y) \psi^{(l+1)} \left(f^{(l)}(y) \right)$$

and, in the invariant case, the final readout is the component-wise :

$$\mathcal{R}_P \left(f^{(L)} \right) = \operatorname*{ess\,sup}_{y \in \mathcal{X},\, P} f^{(L)}(y).$$

Remark 4.3. *It can be easily verified that for all these examples, the underlying $\mathcal{F}^{(l)}$ functions satisfy Assumption 4.1. For the integral, it is ensured by the classical change of variable formula* (4.3)*.*

As for the essential supremum, a similar formula holds. Indeed, recall that for any measurable g, for any measurable bijection ϕ, one has

$$\operatorname*{ess\,sup}_{P} g \circ \phi = \inf\{M | P(g \circ \phi > M) = 0\}.$$

However, $(g \circ \phi > M) = \{x | g \circ \phi(x) > M\} = \{\phi^{-1}(y) | g(y) > M\} = \phi^{-1}(g > M)$, such that one finally has:

$$\inf\{M | P(g \circ \phi > M) = 0\} = \inf\{M | P(\phi^{-1}(g > M)) = 0\} = \operatorname*{ess\,sup}_{\phi_\# P} g.$$

5. cMPGNNs as limits of MPGNNs on large random graphs

This section contains the core of our contributions. We focus on MPGNNs when applied on random graphs G_n drawn from $\mathcal{G}_n(W, P)$. Specifically, given such a MPGNN, we are interested in its limit as n tends to infinity. We show that under mild regularity conditions, such a limit exists and is a cMPGNN. We further provide some non asymptotic bounds to control the deviation between a MPGNN and its limit cMPGNN with high probability.

This section is divided in two parts. In the first part (section 5.1), given a MPGNN, we define, when it exists, its associated canonical cMPGNN on (W, P) that we call **continuous counterpart**. The precise definition of this central concept is Definition 5.2: it states how that continuous counterpart is built out of the discrete network as a limit on random graphs $G_n \sim \mathcal{G}_n(W, P)$ of growing sizes. Then, we show that under mild regularity conditions, Examples a, b, c and d are indeed the continuous counterparts of Examples 1, 2, 3, and 4 according to our definition.

Note that all MPGNNs do not have a continuous counterpart in the sense of Definition 5.2: indeed, the definition is based on the existence of a limit (Eq. (5.8)). In addition, when the continuous counterpart exists, the MPGNN may not converge to its continuous counterpart as n tends to infinity. In the second part of this section (section 5.2), we study this convergence: we give sufficient conditions for this convergence to occur and provide convergence rates in the form of non asymptotic bounds with high probability.

Our main result, Theorem 5.7 in section 5.2.1, concerns a class of MPGNN that have a certain kind of Lipschitz continuity among other mild assumptions: in a few words, it states that such MPGNNs have a continuous counterpart to which they converge as n grows, with a controlled rate that we specify. The result we obtain is based on the so called McDiarmid inequality [26], that says that a multivariate function of independent random variable has a sub-Gaussian concentration around its mean if it satisfies the following notion of bounded differences.

Definition 5.1 (Bounded Differences Property)**.** *Let $f : \mathcal{E}^n \to \mathbb{R}$ be a function of n variables. We say that f has the bounded differences property if there exist n nonnegative constants $c_1, \dots, c_n$ such that for any $1 \leq i \leq n$:*

$$|f(x_1 \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n)| \leq c_i, \tag{5.1}$$

For any $x_1, \dots, x_n, x_i' \in \mathcal{E}$.

In plain terms, whenever one fixes all but one of the components of f, the variations should be bounded.

Our second result, Theorem 5.12 in section 5.2.2, is specific to the case of maximum aggregation (indeed, in this case, the bounded difference property is not verified and Theorem 5.7 is not applicable). It is based on another concentration inequality and leads to a bound with a dependence on the input dimension d (recall $\mathcal{X} \subset \mathbb{R}^d$), as opposed to the bounded differences method.

5.1. Limit of MPGNNs on large random graphs

Let (W, P) be a random graph model and $f \in L^\infty_P(\mathcal{X}, \mathbb{R}^d)$. In this subsection, we consider a single layer MPGNN applied on a random graph $G_n \sim \mathcal{G}_n(W, P)$ and input tensor $S_X f$ (recall the definition of Eq. (2.2)). We define a corresponding canonical cMPGNN layer on (W, P) with input map f. Since there is only one layer in this section, we drop the superscript indexation.

To motivate the next definition – that may seem overly technical at first sight– let us consider the simplest example, namely Examples 1 and a. Let us examine how Example a can be recovered from Example 1 at the limit.

Consider a one-layer convolutional cMPGNN from Example a, with input signal f, for which the update of $f(x)$ is given by

$$\int_{y\in\mathcal{X}} W(x,y)\psi\left(f(y)\right) dP(y). \tag{5.2}$$

It is fairly clear that, by the law of large numbers, this integral equals the limit of

$$\frac{1}{n}\sum_{i=1}^{n} W(x, X_i)\psi(f(X_i)) \tag{5.3}$$

for $X_1, \dots, X_n \overset{iid}{\sim} P$. Moreover, Eq. (5.3) is exactly the discrete message passing of Example 1 around a certain node on a certain graph. To be precise, let $G := G_n \cup \{x\}$ be the graph G_n to which a (deterministic) vertex x along with all its associated edges $\{x, X_i\}$ are added. Given the extended graph signal $(f(x), f(X_1), \dots, f(X_n))$, Eq. (5.3) is precisely an iteration of convolutional message passing from Example 1, around the vertex x, for the graph G, that gives the update of $f(x)$. We have thus obtained the cMPGNN of Example a via a limit of the MPGNN of Example 1 on random graphs.

Back to the general case, given an abstract discrete MPGNN F, we want to define a cMPGNN from the limit of the former on random graphs. Following the path of the above example, we look at the following limit :

$$\lim_{n\to\infty} F\left(f(x), \{\!\{(f(X_k), W(x, X_k))\}\!\}_{1\le k\le n}\right) \tag{5.4}$$

If it appears that this limit exists and that it defines the update of $f(x)$ via some cMPGNN, then we have found the continuous counterpart of F. Unfortunately, this existence is far from obvious and F must not always have a continuous counterpart. Actually, convergence of (5.4), as presented in the example of convolutional message passing, is an almost sure convergence of random variable. It is itself quite a strong requirement and we rather relax it to the convergence of

$$\mathbb{E}_{X_1,\dots,X_n}\left[F\left(f(x), \{\!\{(f(X_k), W(x, X_k))\}\!\}_{1\le k\le n}\right)\right]. \tag{5.5}$$

instead. In the first part of the upcoming definition, we say that if there is a $\mathcal{F}$ such as in Eq. (4.1) and such that the limit of Eq. (5.4) is

$$\mathcal{F}_P\left(f(x), (f, W(x,\cdot))\right) \tag{5.6}$$

then $\mathcal{F}$ is a good candidate to be the continuous counterpart of F. It is still only a candidate because Eq. (5.6) is not enough to define a cMPGNN. As we saw in Section 4, $\mathcal{F}$ must also be coherent to random graph model isomorphism, *i.e.,* must

verify Assumption 4.1. This is precisely the purpose of the second part of the following definition.

Definition 5.2 (Continuous counterpart)**.** *Let F be a MPGNN layer. For $f, g \in L^\infty_P(\mathcal{X}, \mathbb{R}^d)$, $W : \mathcal{X}^2 \to [0,1]$ and $P \in \mathcal{P}(\mathcal{X})$, define the sequence of functions in $L^\infty_P(\mathcal{X}, \mathbb{R}^{d'})$:*

$$F_{P,n}(f,(g,W))(x) \quad \hat{=} \quad \mathbb{E}_{X_1,\dots,X_n}\left[F\left(f(x), \{\!\{(g(X_k), W(x,X_k))\}\!\}_{1\le k\le n}\right)\right] \tag{5.7}$$

where the expected value is taken over all the $X_1, \dots, X_n \overset{iid}{\sim} P$.

Let $\mathcal{F}$ be an operator of the form (4.1) taking value in $L^\infty_P(\mathcal{X}, \mathbb{R}^{d'})$ and denote $F_P(f,(g,W))$ the function

$$F_P(f,(g,W))(x) = \mathcal{F}_P\left(f(x), (g, W(x,\cdot))\right).$$

Suppose that there exists H, a non trivial subgroup of $\mathrm{Aut}(\mathcal{X})$, such that for any $f \in L^\infty_P(\mathcal{X}, \mathbb{R}^d)$, for any $\phi \in H$, $F_{\phi\cdot P,n}(f,(\phi\cdot f, W(\cdot,\phi^{-1}(\cdot))))$ converges to $F_{\phi\cdot P}(f,(\phi\cdot f, W(\cdot,\phi^{-1}(\cdot))))$ in the $L^\infty_P(\mathcal{X}, \mathbb{R}^d)$ norm, i.e.:

$$F_{\phi\cdot P,n}(f,(\phi\cdot f, W(\cdot,\phi^{-1}(\cdot)))) \xrightarrow{L^\infty_P(\mathcal{X},\mathbb{R}^d)} F_{\phi\cdot P}(f,(\phi\cdot f, W(\cdot,\phi^{-1}(\cdot)))). \tag{5.8}$$

*Then we say that $\mathcal{F}$ is the **continuous counterpart** of F for H. When $H = \mathrm{Aut}(\mathcal{X})$, or when H is obvious from the context, we simply say that $\mathcal{F}$ is the **continuous counterpart** of F.*

Note that this definition does not say that the continuous counterpart of the MPGNN F is a cMPGNN. The reason being that a cMPGNN must be equivariant to the action of automorphisms of $\mathcal{X}$ which is not straightforward from the above definition. Nevertheless the stability condition (5.8) will ensure that Assumption 4.1 is satisfied by the continuous counterpart, as shown in the next proposition.

Proposition 5.3. *Let $\mathcal{F}$ be the continuous counterpart of F as defined in Definition 5.2. Then it satisfies Assumption 4.1 for any $\phi \in H$.*

Proof. Let $f \in L^\infty_P(\mathcal{X}, \mathbb{R}^d)$, $\phi \in H$, and $X_1, \dots, X_n \overset{iid}{\sim} P$, by (5.8) , we have for P-almost all x :

$$\begin{aligned}
&\mathcal{F}_P\left(f(x), (f, W(x,\cdot))\right)\\
&= \lim_{n\to\infty} \mathbb{E}_{X_1,\dots,X_n}\left[F\left(f(x), \{\!\{(f(X_k)), W(x,X_k))\}\!\}_{1\le k\le n}\right)\right]\\
&= \lim_{n\to\infty} \mathbb{E}_{X_1,\dots,X_n}\left[F\left(f(x), \{\!\{\left(\phi\cdot f(\phi(X_k))\right), W(x,\phi^{-1}(\phi(X_k)))\right)\}\!\}_{1\le k\le n}\right)\right]\\
&= \lim_{n\to\infty} \mathbb{E}_{Y_1,\dots,Y_n}\left[F\left(f(x), \{\!\{\left(\phi\cdot f(Y_k)\right), W(x,\phi^{-1}(Y_k))\right)\}\!\}_{1\le k\le n}\right)\right]\\
&= \lim_{n\to\infty} F_{\phi\cdot P,n}(f,(\phi\cdot f, W(\cdot,\phi^{-1}(\cdot))))
\end{aligned}$$

where $Y_i = \phi(X_i) \overset{iid}{\sim} \phi\cdot P$, then by (5.8) this is equal to

$$\mathcal{F}_{\phi_\# P}\left(f(x), \left(\phi\cdot f, W(x,\phi^{-1}(\cdot))\right)\right)$$

which concludes the proof. □

The same definition can be given for a readout layer.

Definition 5.4. *Let R be a MPGNN readout layer and $P \in \mathcal{P}(\mathcal{X})$. For $f \in L^\infty_P(\mathcal{X}, \mathbb{R}^d)$, we define the sequence of terms*

$$R_{P,n}(f) = \mathbb{E}_{X_1,\dots,X_n}\left[R\left(\{\!\{f(X_1),\dots,f(X_n)\}\!\}\right)\right] \in \mathbb{R}^{d'},$$

where the expected value is taken over all the $X_1,\dots,X_n \overset{iid}{\sim} P$.
Let $\mathcal{R}$ be a continuum readout operator of the form (4.2) taking values in $\mathbb{R}^{d'}$.
Suppose we have H a non trivial subgroup of $\mathrm{Aut}(\mathcal{X})$ such that for any $f \in L^\infty_P(\mathcal{X}, \mathbb{R}^d)$, for any $\phi \in H$, $R_{\phi\cdot P,n}(\phi \cdot f)$ converges to $\mathcal{R}_{\phi\cdot P}(\phi \cdot f)$ in the $\|\cdot\|_\infty$ norm of $\mathbb{R}^d$:

$$R_{\phi\cdot P,n}(\phi \cdot f) \to \mathcal{R}_{\phi\cdot P}(\phi \cdot f).$$

*Then we say that $\mathcal{R}$ is the **continuous counterpart** of R for H, unless $H = \mathrm{Aut}(\mathcal{X})$ or H is obvious from context, in which case we simply say that $\mathcal{R}$ is the continuous counterpart of R.*

Proposition 5.5. *Let $\mathcal{R}$ be the continuous counterpart of R as in definition 5.4. Then it satisfies Assumption 4.1 for any $\phi \in H$.*

Going back to our four examples of Sections 3 and 4, we now show that a, b and c are the continuous counterparts of 1, 2 and 3 for the full $\mathrm{Aut}(\mathcal{X})$ under a positivity condition for the coefficients in the degree normalized and GAT examples. The case 4- d is however more involved, as one has to be careful with the shape of $\mathcal{X}$ and the properties of P to avoid nullset issues at the boundary $\partial\mathcal{X}$. We show that if $\mathcal{X}$ contains no nonvoid open nullset, and if W, f are continuous, then d is the continuous counterpart of 4 for the subgroup H of $\mathrm{Aut}(\mathcal{X})$ consisting of all the homeomorphisms from $\mathcal{X}$ into itself.

Examples 1-a. *With no additional restriction on W, f, nor P, a is the continuous counterpart of 1 for the full* $\mathrm{Aut}(\mathcal{X})$.

Proof. By iid of the random variables an linearity of the expected value, the convergence in (5.8) is actually an equality for all integer n.

$$\mathbb{E}\left[\frac{1}{n}\sum_i W(x, X_i)\psi(f(X_i))\right] = \mathbb{E}\left[W(x, X_1)\psi(f(X_1))\right] = \int_{\mathcal{X}} W(x, y)\psi(y)dP(y).$$

Clearly this remains true replacing P by $\phi \cdot P$, f by $\phi \cdot f$ and W by $W(\cdot, \phi^{-1}(\cdot))$ for any $\phi \in \mathrm{Aut}(\mathcal{X})$. □

Examples 2-b. *Suppose that ψ is bounded and that there is a strictly positive α such that $W > \alpha$ almost surely. Then b is the continuous counterpart of 2 for the full* $\mathrm{Aut}(\mathcal{X})$.

Proof. For $X_1,\dots,X_n \overset{iid}{\sim} P$, we have $\forall x$

$$\int_{\mathcal{X}} \frac{W(x, y)\psi(f(y))}{\int_{\mathcal{X}} W(x, t)dP(t)}dP(y) = \frac{\mathbb{E}\left[W(x, X_1)\psi(f(X_1))\right]}{\mathbb{E}\left[W(x, X_1)\right]},$$

and

$$\mathbb{E}\left[\frac{\sum_i W(x, X_i)\psi(f(X_i))}{\sum_k W(x, X_k)}\right] = \mathbb{E}\left[\frac{\frac{1}{n}\sum_i W(x, X_i)\psi(f(X_i))}{\frac{1}{n}\sum_k W(x, X_k)}\right].$$

Then

$$
\begin{aligned}
&\left\|\mathbb{E}\left[\frac{\frac{1}{n}\sum_i W(x,X_i)\psi(f(X_i))}{\frac{1}{n}\sum_k W(x,X_k)}\right] - \frac{\mathbb{E}\left[W(x,X_1)\psi(f(X_1))\right]}{\mathbb{E}\left[W(x,X_1)\right]}\right\|_\infty \\
=&\left\|\mathbb{E}\left[\frac{\frac{1}{n}\sum_i W(x,X_i)\psi(f(X_i))}{\frac{1}{n}\sum_k W(x,X_k)} - \frac{\mathbb{E}\left[W(x,X_1)\psi(f(X_1))\right]}{\mathbb{E}\left[W(x,X_1)\right]}\right]\right\|_\infty \\
\leq&\mathbb{E}\left[\left\|\frac{\mathbb{E}\left[W(x,X_1)\right]\frac{1}{n}\sum_i W(x,X_i)\psi(f(X_i))}{\frac{1}{n}\sum_k W(x,X_k)\mathbb{E}\left[W(x,X_1)\right]}\right.\right. \\
&\qquad\left.\left. - \frac{\frac{1}{n}\sum_k W(x,X_k)\mathbb{E}\left[W(x,X_1)\psi(f(X_1))\right]}{\frac{1}{n}\sum_k W(x,X_k)\mathbb{E}\left[W(x,X_1)\right]}\right\|_\infty\right] \\
\leq&\frac{1}{\alpha^2}\mathbb{E}\left[\left\|\mathbb{E}\left[W(x,X_1)\right]\frac{1}{n}\sum_i W(x,X_i)\psi(f(X_i))\right.\right. \\
&\qquad\left.\left. - \frac{1}{n}\sum_k W(x,X_k)\mathbb{E}\left[W(x,X_1)\psi(f(X_1))\right]\right\|_\infty\right] \\
\leq&\frac{1}{\alpha^2}\mathbb{E}\left[\left\|\mathbb{E}\left[W(x,X_1)\right]\frac{1}{n}\sum_i W(x,X_i)\psi(f(X_i))\right.\right. \\
&\qquad\left.\left. - E\left[W(x,X_1)\right]E\left[W(x,X_1)\psi(f(X_1))\right]\right\|_\infty\right. \\
&\quad\left. + \left\|\mathbb{E}\left[W(x,X_1)\right]\mathbb{E}\left[W(x,X_1)\psi(f(X_1))\right]\right.\right. \\
&\qquad\left.\left. - \frac{1}{n}\sum_k W(x,X_k)\mathbb{E}\left[W(x,X_1)\psi(f(X_1))\right]\right\|_\infty\right] \\
\leq&\frac{1}{\alpha^2}\mathbb{E}\left[\left\|\frac{1}{n}\sum_i W(x,X_i)\psi(f(X_i)) - \mathbb{E}\left[W(x,X_1)\psi(f(X_1))\right]\right\|_\infty\right] \\
&\quad + \frac{\|\psi\|_\infty}{\alpha^2}\mathbb{E}\left[\left|\mathbb{E}\left[W(x,X_1)\right] - \frac{1}{n}\sum_k W(x,X_k)\right|\right].
\end{aligned}
$$

Using the formula $E(X) = \int_{t>0} P(X > t)dt$ for X nonnegative, we get that this last quantity is equal to

$$
\begin{aligned}
&\frac{1}{\alpha^2}\int_{t>0}\mathbb{P}\left(\left\|\frac{1}{n}\sum_i W(x,X_i)\psi(f(X_i)) - \mathbb{E}\left[W(x,X_1)\psi(f(X_1))\right]\right\|_\infty > t\right)dt \\
&\quad + \frac{\|\psi\|_\infty}{\alpha^2}\int_{t>0}\mathbb{P}\left(\left|\mathbb{E}\left[W(x,X_1)\right] - \frac{1}{n}\sum_k W(x,X_k)\right| > t\right)dt.
\end{aligned} \tag{5.9}
$$

Finally we use McDiarmid inequality (which turns out to be the same as Hoeffding inequality for a sum of independent random variables). It is easy to check that the concerned mutivariate maps have bounded differences of the form $c_i = K/n$ for all

i. Therefore, there is some positive constants K_1, K_s, K_3, K_4 independent of x such that (5.9) is bounded by

$$\frac{1}{\alpha^2}\int_{t>0} K_1 e^{-nK_1 t^2} dt + \frac{\|\psi\|_\infty}{\alpha^2}\int_{t>0} K_3 e^{-nK_4 t^2} dt = O(1/\sqrt{n}) \to 0.$$

This remains true replacing P by $\phi \cdot P$, f by $\phi \cdot f$ and W by $W(\cdot, \phi^{-1}(\cdot))$ for any $\phi \in \mathrm{Aut}(\mathcal{X})$. □

Examples 3-c. *Call $V(x,y) = c(f(x), f(y), W(x,y))$ and suppose that ψ is bounded and that there is two strictly positive constants $0 < \alpha < \beta$ such that $\alpha < V < \beta$ a. s. Then c is the continuous counterpart of 3 for the full $\mathrm{Aut}(\mathcal{X})$.*

Proof. We are brought to the previous example with V instead of W. □

Examples 4-d. *Suppose that W, ψ, and f are continuous and that the measure P is strictly positive on $\mathcal{X}$ i.e, any nonvoid relative open of $\mathcal{X}$ has a strictly positive measure by P. Then d is the continuous counterpart of 4 for $\mathrm{Hom}(\mathcal{X})$: the subgroup of $\mathrm{Aut}(\mathcal{X})$ made of the $\phi \in \mathrm{Aut}(\mathcal{X})$ that are homeomorphisms.*

Proof. We call $g(x,y) = W(x,y)\psi(f(y))$. We start by the case when g is real valued, since g is continuous and P is strictly positive, $\operatorname{ess\,sup}_P g(x,\cdot) = \sup g(x,\cdot) \forall x$ by Lemma D.4 in the Appendix. Let $\epsilon > 0$, By definition of the supremum and by independence of the X_i, we have that

$$\begin{aligned} &\mathbb{P}(|\max_i g(x, X_i) - \sup g(x,\cdot)| \geq \epsilon) \\ =&\mathbb{P}(\max_i g(x, X_i) \leq \sup g(x,\cdot) - \epsilon) \\ =&\mathbb{P}(g(x, X_1) \leq \sup g(x,\cdot) - \epsilon)^n \\ =&\mathbb{P}(|g(x, X_1) - \sup g(x,\cdot)| \geq \epsilon)^n. \end{aligned} \tag{5.10}$$

By continuity and compactness, there is $x^* \in \mathcal{X}$ such that $\sup g(x,\cdot) = g(x, x^*)$, so (5.10) is equal to

$$\begin{aligned} &\mathbb{P}(|g(x, X_1) - g(x, x^*)| \geq \epsilon)^n \\ =&\left(1 - \mathbb{P}(|g(x, X_1) - g(x, x^*)| < \epsilon)\right)^n. \end{aligned} \tag{5.11}$$

By continuity and compactness again, g is uniformly continuous so there is $\delta > 0$ such that $\|(x, X_1) - (x, x^*)\| = \|X_1 - x^*\| < \delta$ implies $|g(x, X_1) - g(x, x^*))| < \epsilon$. Thus (5.11) is bounded from above by

$$\left(1 - \mathbb{P}(\|X_1 - x^*\| < \delta)\right)^n = \left(1 - P(\mathcal{B}(x^*, \delta) \cap \mathcal{X})\right)^n \tag{5.12}$$

where $\mathcal{B}(x^*, \delta)$ is the open ball of center x^* and radius δ in $\mathbb{R}^d$. To finish let us justify that the measure of the $\mathcal{B}(x^*, \delta) \cap \mathcal{X}$ when x runs over $\mathcal{X}$ is bounded from below. Suppose this would not be the case, *i.e* that the measure of a ball of radius δ centered in $\mathcal{X}$ could be arbitrary small. By compactness, up to subsequence extraction, we can assume there is $(x_k) \in \mathcal{X}^{\mathbb{N}}$ such that $x_n \to x \in \mathcal{X}$ and $P(\mathcal{B}(x_k, \delta) \cap \mathcal{X}) \leq 1/2^k$. Call $U = \mathcal{B}(x, \delta/2) \cap \mathcal{X}$, there is rank k_0 such that $\forall k \geq k_0$, $x_k \in U$. Thus $U \subset \mathcal{B}(x_k, \delta) \cap \mathcal{X}$ $\forall k \geq k_0$ yielding $P(U) \leq 1/2^k$ $\forall k \geq k_0$ *i.e* $P(U) = 0$. Impossible since U is nonempty relative open of $\mathcal{X}$.

So there is $\eta > 0$ independent of x such that $P(\mathcal{B}(x^*, \delta) \cap \mathcal{X}) > \eta$ and, coming back to (5.12):

$$\mathbb{P}(|\max_i g(x, X_i) - \sup g(x, \cdot)| \geq \epsilon) \leq (1 - \eta)^n. \tag{5.13}$$

If g is vector valued, say in $\mathbb{R}^{d'}$, call $g_1, \dots, g_{d'}$ its components and η_k such that g_k satisfies (5.13) with $\eta = \eta_k$. Then by an union bound we have

$$\mathbb{P}(\|\max_i g(x, X_i) - \sup g(x, \cdot)\|_\infty \geq \epsilon) \leq \sum_{k=1}^{d'} (1 - \eta_k)^n. \tag{5.14}$$

A the end of the day, by letting $Z = \|\max_i g(x, X_i) - \sup g(x, \cdot)\|_\infty$, we have for any $\epsilon > 0$:

$$\begin{aligned} &\|\mathbb{E}(\max_i g(x, X_i)) - \sup g(x, \cdot)\|_\infty \\ \leq &\mathbb{E}(Z) \\ = &\mathbb{E}(Z \mathbb{1}_{Z \geq \epsilon}) + \mathbb{E}(Z \mathbb{1}_{Z < \epsilon}) \\ \leq &2\|g\|_\infty \sum_{k=1}^{d'} (1 - \eta_k)^n + \epsilon. \end{aligned} \tag{5.15}$$

This concludes the uniform convergence. To conclude the proof, we are left to check that the strict positiveness of P as well as the continuity of f and W are preserved by the action of homeomorphisms. It is clear for maps' continuity. Let $\phi \in \mathrm{Hom}(\mathcal{X})$ and $U \subset \mathcal{X}$ a relative non empty open of $\mathcal{X}$,

$$\phi \cdot P(U) = P(\phi^{-1}(U)) > 0$$

since $\phi^{-1}(U)$ is a non empty open of $\mathcal{X}$ as ϕ is continuous. □

5.2. Convergence of MPGNN on random graphs

Let (W, P) be a random graph model, and $(G_n)_{n \geq 1}$ be a sequence of random graphs drawn from $\mathcal{G}_n(W, P)$. We go back to the multi layer setup: consider a MPGNN $(F^{(l)})_{1 \leq l \leq L}$, a readout R and their continuous counterparts $(\mathcal{F}^{(l)})_{1 \leq l \leq L}$ and $\mathcal{R}$ in the sense of Definitions 5.2 and 5.4. For a $f \in L^\infty_P(\mathcal{X}, \mathbb{R}^{d_0})$, does the MPGNN on G_n with input signal $S_X f$ actually converge to the cMPGNN on (W, P) with input signal f? If yes, at which speed? In this section we provide non asymptotic bounds with high probability to quantify this convergence.

Our main theorems state that, under mild regularity condition and with high probability, $\Theta_{G_n}(S_X(f))$ is close to $\Theta_{W,P}(f)$ in the equivariant case and that $\bar{\Theta}_{G_n}(S_X(f))$ is close to $\bar{\Theta}_{W,P}(f)$ in the invariant case. For the latter, we can compare both the outputs directly since they belong to the same vector space. The comparison is however more involved in the equivariant case since $\Theta_{G_n}(S_X(f))$ is a tensor and $\Theta_{W,P}(f)$ is a function. In this case, we measure their deviation with the Maximum Absolute Error (MAE) defined by

$$\mathrm{MAE}_X(Z, f) = \max_{1 \leq i \leq n} \|z_i - f(X_i)\|_\infty.$$

Our first theorem is based on the Mcdiarmid inequality D.3. It encompasses a whole class of MPGNNs that includes Examples 1, 2 and 3 but not 4. For the latter, we

obtain a different bound based on other concentration inequalities similar to what has been done for the case 4-d in the previous section.

5.2.1. Bounded differences method

Since for all l, $\mathcal{F}^{(l)}$ is the continuous counterpart of $F^{(l)}$, and using the notations of Definition 5.2, we let $(a_n^{(l)})$ be a sequence of positive reals such that $a_n^{(l)} \to 0$ and (recall the definitions of $F_{P,n}$ and F_P from Definition 5.2):

$$\|F_{P,n}^{(l)}(f^{(l-1)}, (f^{(l-1)}, W)) - F_P^{(l)}(f^{(l-1)}, (f^{(l-1)}, W))\|_\infty \leq a_n^{(l)}. \tag{5.16}$$

for all n.

Similarly, we let (b_n) be another sequence of positive reals verifying $b_n \to 0$ and such that (recall the definitions of $R_{P,n}$ and $\mathcal{R}_P$ from Definition 5.4)

$$\|R_{P,n}(f^{(L)}) - \mathcal{R}_P(f^{(L)})\|_\infty \leq b_n \tag{5.17}$$

for all n.

For a fixed $x_1 \in \mathcal{X}$, we are interested in the bounded differences of

$$F^{(l)}(f^{(l-1)}(x_1), \{\!\{(f^{(l-1)}(x_k), W(x_1, x_k))\}\!\}_{k\geq 2}) \tag{5.18}$$

as a map of the $n-1$ variables $x_2, \ldots, x_n$. If $c_1(x_1) \ldots, c_n(x_1)$ satisfy (5.1), since (5.18) is invariant to the permutations of $x_2, \ldots, x_n$, they can be taken all equal. We call $D_n^{(l)}(x_1) = c_1(x_1) = \cdots = c_n(x_1)$. Moreover, since (5.18) belongs to $L_P^\infty(\mathcal{X}, \mathbb{R}^{d_l})$ as a function of x_1, it is P-essentially bounded by compactness. Define

$$D_n^{(l)} = \operatorname*{ess\,sup}_{P, x_1 \in \mathcal{X}} D_n^{(l)}(x_1) \tag{5.19}$$

Similarly, we call C_n the bounded difference of

$$(x_1, \ldots, x_n) \mapsto R(f^{(L)}(x_1), \ldots, f^{(L)}(x_n)).$$

Finally, we add a "Lipschitz-type" regularity assumption on the $F^{(l)}$. To sum up we suppose :

Assumption 5.6. (i) *The $\mathcal{F}^{(l)}$ as well as $\mathcal{R}$ are the continuous counterparts of the $F^{(l)}$, and R as defined is Definitions 5.2, and 5.4.*

(ii) *There exist some $D_n^{(l)}$ such as defined in* (5.19).

(iii) *There exist some $a_n^{(l)}$ and b_n such as defined in* (5.16) *and* (5.17).

(iv) *For all $1 \leq l \leq L$, we endow $\mathbb{R}^{d_{l-1}} \times [0,1]$ with the norm $\|(y,t)\|_1 = \|y\|_\infty + |t|$ and call δ_1 the corresponding distance on multisets as defined in* (2.1). *Let $x, x' \in \mathbb{R}^{d_{l-1}}$ and m, m' be two multisets of same cardinal n containing elements of $\mathbb{R}^{d_{l-1}} \times [0,1]$, then there exist two constants $\mu^{(l)} \geq 0$ and $\lambda_{F,n}^{(l)} > 0$ such that :*

$$\left\| F^{(l)}(x, m) - F^{(l)}(x', m') \right\|_\infty \leq \mu_F^{(l)} \|x - x'\|_\infty + \lambda_{F,n}^{(l)} \delta_1(m, m').$$

(v) *For m, m' being be two multisets of same cardinal n containing elements of $\mathbb{R}^{d_L}$, we define the distance $\delta_\infty(m, m')$ relatively to the $\|\cdot\|_\infty$ norm in $\mathbb{R}^{d_L}$ as in* (2.1)*. Then there exists $\lambda_{R,n} > 0$ such that*

$$\|R(m) - R(m')\|_\infty \leq \lambda_{R,n}\delta_\infty(m, m').$$

(vi) *The sequences $\left(\lambda_{F,n}^{(l)}\right)$ and $\left(\lambda_{R,n}^{(l)}\right)$ are bounded over n.*

Theorem 5.7 (MPGNN convergence towards cMPGNN)**.** *Under Assumption 5.6 for any $0 < \rho \leq 1$, with probability at least $1 - \rho$, the following assertions are verified:*

$$\mathrm{MAE}_X\left(\Theta_{G_n}(S_X(f)), \Theta_{W,P}(f)\right) \lesssim LD_n\sqrt{n\ln\left(\frac{n2^L d_{max}}{\rho}\right)} + La_{n-1}. \tag{5.20}$$

$$\begin{aligned}\left\|\bar{\Theta}_{G_n}(S_X(f)) - \bar{\Theta}_{W,P}(f)\right\|_\infty \lesssim\ & LD_n\sqrt{n\ln\left(\frac{n2^{L+1}d_{max}}{\rho}\right)} + C_n\sqrt{n\ln\left(\frac{4d_L}{\rho}\right)} \\ & + La_{n-1} + b_n.\end{aligned} \tag{5.21}$$

Where $D_n = \max_l D_n^{(l)}$, $d_{max} = \max_l d_l$, $a_n = \max_l a_n^{(l)}$ and $\lesssim$ hides some multiplicative constants which depend polynomially on $\lambda_{F,n}^{(1)}, \dots \lambda_{F,n}^{(L)}, \lambda_{R,n}, \mu_F^{(l)}$ and are bounded over n.

Sketch of proof. (See Appendix A for full proof) We prove the result by induction on the number of layers L. At each step, we bound $\|(S_X f)_i^{(L)} - f^{(L)}(X_i)\|$ for all i. This is done by conditioning over x_i and finding a bound of

$$\left\|F^{(L)}\left(f^{(L-1)}(x_i), \{\!\{\left(f^{(L-1)}(X_k), W(x_i, X_k)\right)\}\!\}_{k\neq i}\right) - f^{(L-1)}(x_1)\right\|_\infty$$

that does not depend on x_i, using a succession of triangular inequalities, the Lipschitz-type property from Assumption 5.6 *(v)* and McDiarmid's inequality. We then turn it into a bound for $\|(S_X f)_i^{(L)} - f^{(L)}(X_i)\|$ via the law of total probability and conclude with an union bound over i. □

The asymptotic behavior of (5.20) is determined by D_n : if it does not decrease fast enough, the inequality becomes meaningless. This suggests the following important corollary.

Corollary 5.8 (Sufficient condition for MPGNN convergence on a random graph)**.** *If $D_n = o\left(1/\sqrt{n\ln n}\right)$ then $\mathrm{MAE}_X(\Theta_{G_n}(S_X(f)), \Theta_{W,P}(f))$ converges in probability towards 0.*

This corollary provides a sufficient condition for a MPGNN on a random graph to converge to its continuous counterpart on the random graph model: in words, its aggregation function needs to have sharp enough bounded differences. We investigate whether our Examples 1, 2, 3 and 4 have such sharp bounded differences

Example	D_n	a_n	b_n	Convergence by Th. 5.7
1-a	$O(1/n)$	0	0	✓
3-c	$O(1/n)$	$O(1/\sqrt{n})$	0	✓
4-d	$O(1/n)$	$O(1/\sqrt{n})$	0	✓
4-d	$\Omega(1)$	–	–	✗

Table 1. *Table summing up the results of Proposition 5.9.*

and verify Assumption 5.6. Under mild regularity conditions this is the case for all examples but Example 4:

Proposition 5.9. *We present application of Theorem 5.7 on the Examples 1, 2 and 3 but not 4, the $\psi^{(l)}$ are supposed Lipschitz continuous and bounded. Additional regularity assumptions are needed for some examples.*

- **1-a** $D_n = O(1/n)$, $a_n = b_n = 0$.
- **2-b** *Suppose there is $\alpha > 0$ such that $W > \alpha$. Then $D_n = O(1/n)$, $a_n = O(1/\sqrt{n})$ and $b_n = 0$.*
- **3-c** *Suppose there is $\alpha, \beta > 0$ and $\lambda_c > 0$ such that $\alpha < c(x, y, t) < \beta$ and $|c(x, y, t) - c(x', y', t')| \leq \lambda_c(\|x - x'\|_\infty + \|y - y'\|_\infty + |t - t'|)$, $\forall x, x', y, y', t, t'$. Then $D_n = O(1/n)$, $a_n = O(1/\sqrt{n})$ and $b_n = 0$.*
- **4-d** *The bounded differences do not satisfy Corollary 5.8.*

Proof. Calculation and verification of the Theorem's assumptions are done in Appendix C. □

Table 1 sums up theses results. For a network with max aggregation, the bounded differences are not sharp enough for theorem 5.7 to conclude. We thus treat this case separately in the next section.

5.2.2. Convergence of max aggregation MPGNNs

We would like to follow the same line of proof used in the proof of 4-d, but when we reach Eq. (5.12), we need to be able to give an approximation of the measure of a ball in $\mathcal{X}$. To this end, we introduce the notion of **volume retaining property**.

Definition 5.10 (Volume retaining property)**.** *We say that the probability space $(\mathcal{X}, P)$ has the (r_0, κ)-volume retaining property if for any $r \leq r_0$ and for any $x \in \mathcal{X}$,*

$$P(\mathcal{B}(x, r) \cap \mathcal{X}) \geq \kappa m(\mathcal{B}(x, r)) \tag{5.22}$$

Where $\mathcal{B}(x, r)$ is the ball of center x and radius r and m is the classical d-dimensional Lebesgue measure in $\mathbb{R}^d$

Clearly, volume-retention implies strict positiveness of the measure. This property ensures that the measure of the intersection of small ball centered in a point of $\mathcal{X}$ with $\mathcal{X}$ is at least a portion of the volume of that ball in $\mathbb{R}^d$. That enables us to estimate from below the measure of balls anywhere in $\mathcal{X}$, due to translation

invariance of Lebesgue measure. To provide an example of volume retaining probability space, consider $\mathcal{X}$ the unit hypercube $[0,1]^d$ and P is the Lebesgue measure itself, it is easy to check that this has the $(1, 1/2^d)$-volume retaining property.

For a volume retaining probability space, we obtain the following new concentration inequality

Lemma 5.11 (Concentration inequality for volume retaining space)**.** *Let $g : \mathcal{X}^2 \to \mathbb{R}^{d'}$ be λ_g-Lipschitz and $(\mathcal{X}, P)$ have the (r, κ)-volume retaining property for some $r, \kappa > 0$. Then for any $\rho \geq e^{-n\kappa r_0^d 2^d}$, for any random variables $X_1, \ldots, X_n \overset{iid}{\sim} P$, with probability at least $1 - \rho$:*

$$\| \max_{1\leq i\leq n} g(x, X_i) - \sup g(x, \cdot)\|_\infty \leq \frac{\lambda_g}{2} \left(\frac{\ln(d'/\rho)}{n\kappa} \right)^{1/d}.$$

Proof. We write the proof assuming $d' = 1$, the case $d' \geq 1$ follows easily by an union bound. The proof is exactly the same than 4-d, with g having a single variable here, until (5.11) where we use Lipschitz continuity to get the bound

$$\left| \max_i g(x, X_i) - \sup g(x, \cdot) \right| \leq (1 - P(\mathcal{B}(x^*, \epsilon/\lambda_g) \cap \mathcal{X}))^n \tag{5.23}$$

By volume retention, for $\epsilon \leq r_0 \lambda_g$, (5.23) is bounded by

$$\left(1 - \kappa \left(\frac{2\epsilon}{\lambda_g} \right)^d \right)^n \leq e^{-n\kappa \frac{2^d \epsilon^d}{\lambda_g^d}}, \tag{5.24}$$

Which implies that for $\rho \geq e^{-n\kappa r_0^d 2^d}$ with probability at least $1 - \rho$:

$$| \max_{1\leq i\leq n} g(x, X_i) - \sup g(x, \cdot)| \leq \frac{\lambda_g}{2} \left(\frac{\ln(1/\rho)}{n\kappa} \right)^{1/d}.$$

□

Armed with this lemma, we are now ready to state the non asymptotic bound for a MPGNN with max aggregation.

Theorem 5.12 (Nonasymptotic convergence of max-MPGNN towards cMPGNN)**.** *Suppose, $(\mathcal{X}, P)$ has the (r_0, κ)-volume retaining property and that f, W and the $\psi^{(l)}$ are Lipschitz continuous. Let $\rho \geq 2^{L-1} n e^{-n\kappa r_0^d 2^d}$ and n large enough for $0 < \rho < 1$ to hold. Then with probability at least $1 - \rho$:*

$$\mathrm{MAE}_X(\Theta_{G_n}(S_X(f)), \Theta_{W,P}(f) \lesssim L \left(\frac{1}{n-1} \ln \left(\frac{2^{L-1} n d_{\max}}{\rho} \right) \right)^{1/d}, \tag{5.25}$$

and

$$\left\| \bar{\Theta}_{G_n}(S_X(f)) - \bar{\Theta}_{W,P}(f) \right\|_\infty \lesssim L \left(\frac{1}{n-1} \ln \left(\frac{2^L n d_{\max}}{\rho} \right) \right)^{1/d} + \left(\frac{1}{n} \ln \left(\frac{2 d_L}{\rho} \right) \right)^{1/d}, \tag{5.26}$$

where $d_{max} = \max_l d_l$.

Remark 5.13. *Since we made an assumption that involves the volume of a d-dimensional ball, the convergence rate for max convolution depends on the dimension of the latent space ($\mathcal{X} \subset \mathbb{R}^d$). Overall, it is slower than for the McDiarmid's method from Theorem 5.7 for high dimensional latent spaces.*

6. Conclusion

In this work, we have defined continuous counterparts of MPGNNs with very generic aggregation functions on a probability space with respect to a transition kernel. We then have shown that under certain conditions, cMPGNNs are limits of discrete MPGNNs on random graphs sampled from the corresponding random graph model. Until now, similar result were known for SGNNs, which are more restricted architectures, or for MPGNNs with a degree normalized mean aggregation. Our main contribution is to extend this to abstract MPGNNs with generic aggregation functions. All along this paper, a focus is given on examples based on mean or weighted mean aggregation (Examples 1, 2 and 3) and max aggregation (Example 4), but our theorems are not limited to these examples and is in fact verified for mild assumptions on the underlying model.

The techniques used for the three first examples are very different from the ones used for the "max example". As the max is the limit of L_p-means as p increases, future work could try to come up with a proof on L_p, which could perhaps bridge the two worlds and unify the two approaches described in this paper.

Appendices

A. Proof of Theorem 5.7

A.1. Equivariant case

We start with the equivariant case. We seek to bound:

$$\max_{1\leq i\leq n}\left\|(S_X f)_i^{(L)} - S_X(f^{(L)})_i\right\| = \max_{1\leq i\leq n}\left\|(S_X f)_i^{(L)} - f^{(L)}(X_i)\right\|.$$

We will prove the following sharper version of Theorem 5.7.

Theorem A.1. *Under same assumptions that Theorem 5.7 Let $\rho > 0$, then with probability at least $1-\rho$:*

(A.1)

$$\max_{1\leq i\leq n}\left\|(S_X f)_i^{(L)} - f^{(L)}(X_i)\right\| \leq \sum_{l=1}^{L} A_n^{(l,L)}\left[D_n^{(l)}\sqrt{\frac{1}{2}n\ln\left(\frac{2^{L+2-l}nd_l}{\rho}\right)} + a_{n-1}^{(l)}\right]$$

Where $A_n^{(l,L)} = \prod_{k=l+1}^{L}(\mu_F^{(k)} + \lambda_{F,n-1}^{(k)})$ with the conventions $\prod_{k=a}^{b}(\dots) = 1$ and $\sum_{k=a}^{b}(\dots) = 0$ if $a > b$.

Then Theorem 5.7 in the main text is actually the following corollary.

Corollary A.2 (Theorem 5.7 in the main text)**.** *With probability at least $1-\rho$:*

$$\max_i\left\|(S_X f)_i^{(L)} - f^{(L)}(X_i)\right\|_\infty \lesssim LD_n\sqrt{n\ln\left(\frac{n2^L d_{max}}{\rho}\right)} + La_n.$$

Proof. The $A_n^{(l,L)} = \prod_{k=l+1}^{L}(\mu_F^{(k)} + \lambda_{F,n-1}^{(k)})$ are bounded over n by assumption. Thus the corollary comes directly from Theorem A.1. □

Proof of Theorem A.1. We prove the result by induction on L. Let $\rho > 0$, until the end of this proof we denote by $H^{(L)}(\rho)$ the bound (A.1) :

$$H^{(L)}(\rho) = \sum_{l=1}^{L} A_n^{(l,L)}\left[D_n^{(l)}\sqrt{\frac{1}{2}n\ln\left(\frac{2^{L+2-l}nd_l}{\rho}\right)} + a_{n-1}^{(l)}\right].$$

We recall those notations from Definition 5.2

$$\begin{aligned}&F_{P,n}^{(l+1)}(f^{(l)},(f^{(l)},W))(x)\\ =&\mathbb{E}_{X_1,\dots,X_n}\left[F^{(l+1)}\left(f^{(l)}(x),\{\!\{\left(f^{(l)}(X_k),W(x,X_k)\right)\}\!\}_{1\leq k\leq n}\right)\right],\end{aligned}$$

and

$$F_P^{(l+1)}(f^{(l)},(f^{(l)},W))(x) = \mathcal{F}_P^{(l+1)}\left(f^{(l)}(x),\left(f^{(l)},W(x,\cdot)\right)\right) = f^{(l+1)}(x).$$

Suppose $L=1$, we shall find a quantity that bounds all the $\left\|(S_X f)_i^{(1)} - f^{(1)}(X_i)\right\|$ for $i=1,\dots,n$ with probability at least $1-\rho/n$. Thereby, by an union bound, this quantity will bound their maximum with probability at least $1-\rho$.

Choose $i \in \{1, \dots, n\}$ and $x_i \in \mathcal{X}$, consider

$$\left\| F^{(1)}\left(f^{(0)}(x_i), \{\!\{\left(f^{(0)}(X_k), W(x_i, X_k)\right)\}\!\}_{k \neq i}\right) - f^{(1)}(x_i) \right\|_\infty .$$

From a triangular inequality,

(A.2)
$$\begin{aligned}
&\left\| F^{(1)}\left(f^{0)}(x_i), \{\!\{\left(f^{(0)}(X_k), W(x_i, X_k)\right)\}\!\}_{k \neq i}\right) - f^{(1)}(x_i) \right\|_\infty \\
&\leq \left\| F^{(1)}\left(f^{(0)}(x_i), \{\!\{\left(f^{(0)}(X_k), W(x_i, X_k)\right)\}\!\}_{k \neq i}\right) - F^{(1)}_{P,n-1}(f^{(0)}, (f^{(0)}, W))(x_i) \right\|_\infty \\
&\quad + \left\| F^{(1)}_{P,n-1}(f^{(0)}, (f^{(0)}, W))(x_i) - f^{(1)}(x_1) \right\|_\infty \\
&\leq \left\| F^{(1)}\left(f^{(0)}(x_i), \{\!\{\left(f^{(0)}(X_k), W(x_i, X_k)\right)\}\!\}_{k \neq i}\right) - F^{(1)}_{P,n-1}(f^{(0)}, (f^{(0)}, W))(x_i) \right\|_\infty \\
&\quad + a^{(1)}_{n-1}.
\end{aligned}$$

by definition of a_n. Now we bound (A.2) with high probability using McDiarmid's inequality D.3 on

$$F^{(1)}\left(f^{(0)}(x_i), \{\!\{\left(f^{(0)}(x_k), W(x_i, x_k)\right)\}\!\}_{k \neq i}\right)$$

as a multivariate function of the $n-1$ variables $x_2, \dots, x_n$.

We obtain that for any x_1, with probability at least $1 - \rho/n$,

$$\begin{aligned}
&\left\| F^{(1)}\left(f^{(0)}(x_i), \{\!\{\left(f^{(0)}(X_k), W(x_i, X_k)\right)\}\!\}_{k \neq i}\right) - f^{(1)}(x_1) \right\|_\infty \\
&\leq D^{(1)}_n \sqrt{\frac{1}{2}(n-1) \ln\left(\frac{2d_1 n}{\rho}\right)} + a^{(1)}_{n-1} \leq D^{(1)}_n \sqrt{\frac{1}{2} n \ln\left(\frac{2d_1 n}{\rho}\right)} + a^{(1)}_{n-1}.
\end{aligned} \tag{A.3}$$

Hence, by conditioning over X_i and applying the law of total probability, (A.3) yields with probability at least $1 - \rho/n$:

$$\left\| (S_X f)^{(1)}_i - f^{(1)}(X_i) \right\|_\infty \leq D^{(1)}_n \sqrt{\frac{1}{2} n \ln\left(\frac{2d_1 n}{\rho}\right)} + a^{(1)}_{n-1}. \tag{A.4}$$

And by an union bound, we can conclude that with probability at least $1 - \rho$:

$$\max_i \left\| (S_X f)^{(1)}_i - f^{(1)}(X_i) \right\|_\infty \leq D^{(1)}_n \sqrt{\frac{1}{2} n \ln\left(\frac{2d_1 n}{\rho}\right)} + a^{(1)}_{n-1} \leq H^{(1)}(\rho). \tag{A.5}$$

Now suppose the theorem true for $L \geq 1$. For any node i,

$$
\begin{aligned}
&\left\|(S_X f)_i^{(L+1)} - f^{(L+1)}(X_i)\right\|_\infty \\
&\leq \left\|(S_X f)_i^{(L+1)} - F_{P,n-1}^{(L+1)}(f^{(L)}, (f^{(L)}, W))(X_i)\right\|_\infty \\
&\quad + \left\|F_{P,n-1}^{(L+1)}(f^{(L)}, (f^{(L)}, W))(X_i) + f^{(L+1)}(X_i)\right\|_\infty \\
&\leq \left\|(S_X f)_i^{(L+1)} - F_{P,n-1}^{(L+1)}(f^{(L)}, (f^{(L)}, W))(X_i)\right\|_\infty + a_{n-1}^{(L+1)} \\
&\leq \left\|(S_X f)_i^{(L+1)} - F^{(L+1)}\left(f^{(L)}(X_i), \{\!\{\left(f^{(L)}(X_k), W(X_i, X_k)\right)\}\!\}_{k\neq i}\right)\right\|_\infty \\
&\quad + \left\|F^{(L+1)}\left(f^{(L)}(X_i), \{\!\{\left(f^{(L)}(X_k), W(X_i, X_k)\right)\}\!\}_{k\neq i}\right) - F_{P,n-1}^{(L+1)}(f^{(L)}, (f^{(L)}, W))(X_i)\right\|_\infty \\
&\quad + a_{n-1}^{(L+1)} \\
&\leq \mu_F^{(L+1)} \left\|(S_X f)_i^{(L)} - f^{(L)}(X_i)\right\|_\infty + \lambda_{F,n-1}^{(L+1)} \max_{j\neq i} \left\|(S_X f)_j^{(L)} - f^{(L)}(X_j)\right\|_\infty \\
&\quad + \left\|F^{(L+1)}\left(f^{(L)}(X_i), \{\!\{\left(f^{(L)}(X_k), W(X_i, X_k)\right)\}\!\}_{k\neq i}\right) - F_{n-1}^{(L+1)}(f^{(L)}, (f^{(L)}, W))(X_i)\right\|_\infty \\
&\quad + a_{n-1}^{(L+i)}.
\end{aligned}
$$

Where the last inequality comes from the Lipschitz-like regularity Assumption 5.6 *(v)* on $F^{(L+1)}$. Now taking the maximum over the vertices :

$$
\begin{aligned}
&\max_i \left\|(S_X f)_i^{(L+1)} - f^{(L+1)}(X_i)\right\|_\infty \\
&\leq \mu_F^{(L+1)} \max_i \left\|(S_X f)_i^{(L)} - f^{(L)}(X_i)\right\| \\
&\quad + \lambda_{F,n-1}^{(L+1)} \max_i \max_{j\neq i} \left\|_\infty (S_X f)_j^{(L)} - f^{(L)}(X_j)\right\|_\infty \\
&\quad + \max_i \left\|F^{(L+1)}\left(f(X_i), \{\!\{(f(X_k), W(X_i, X_k))\}\!\}_{k\neq i}\right) - F_{P,n-1}^{(L+1)}(f^{(L)}, (f^{(L)}, W))(X_i)\right\|_\infty \\
&\quad + a_{n-1}^{(L+1)} \\
&\leq (\mu_F^{(L+1)} + \lambda_{F,n-1}^{(L+1)}) \max_i \left\|(S_X f)_i^{(L)} - f^{(L)}(X_i)\right\|_\infty \\
&\quad + \max_i \left\|F^{(L+1)}\left(f(X_i), \{\!\{(f(X_k), W(X_i, X_k))\}\!\}_{k\neq i}\right) - F_{P,n-1}^{(L+1)}(f^{(L)}, (f^{(L)}, W))(X_i)\right\|_\infty \\
&\quad + a_{n-1}^{(L+1)}
\end{aligned}
$$

(as $\max_i \max_{j\neq i} a_i = \max_i a_i$). Finally, we bound this last equation with high probability. The first term is handled by the induction hypothesis. For the second term, by conditioning over X_i, using McDiarmid and a union bound, the same way

we did in the case $L = 1$, we obtain with probability at least $1 - \rho$:

$$
\begin{aligned}
&\max_i \left\| (S_X f)_i^{(L+1)} - f^{(L+1)}(X_i) \right\|_\infty \\
&\leq (\mu_F^{(L+1)} + \lambda_{F,n-1}^{(L+1)}) H^{(L)}(\rho/2) + D_n^{(L+1)} \sqrt{\frac{1}{2} n \ln\left(\frac{4 d_{L+1} n}{\rho}\right)} + a_{n-1}^{(L+1)} \\
&= \sum_{l=2}^{L} (\mu_F^{(L+1)} + \lambda_{F,n-1}^{(L+1)}) A_n^{(l,L)} D_n^{(l)} \sqrt{\frac{1}{2} n \ln\left(\frac{2^{L+1+2-l} n d_l}{\rho}\right)} \\
&\quad + (\mu_F^{(L+1)} + \lambda_{F,n-1}^{(L+1)}) A_n^{(1,L)} D_n^{(1)} \sqrt{\frac{1}{2} n \ln\left(\frac{2^{L+1} n d_1}{\rho}\right)} \\
&\quad + (\mu_F^{(L+1)} + \lambda_{F,n-1}^{(L+1)}) \sum_{l=1}^{L} A_n^{(l,L)} a_{n-1}^{(l)} + D_n^{(L+1)} \sqrt{\frac{1}{2} n \ln\left(\frac{4 d_{L+1} n}{\rho}\right)} + a_{n-1}^{(L+1)} \\
&= \sum_{l=2}^{L} A_n^{(l,L+1)} D_n^{(l)} \sqrt{\frac{1}{2} n \ln\left(\frac{2^{L+1+2-l} n d_l}{\rho}\right)} \\
&\quad + A_n^{(1,L+1)} D_n^{(1)} \sqrt{\frac{1}{2} n \ln\left(\frac{2^{L+1} n d_1}{\rho}\right)} \\
&\quad + \sum_{l=1}^{L} A_n^{(l,L+1)} a_{n-1}^{(l)} + D_n^{(L+1)} \sqrt{\frac{1}{2} n \ln\left(\frac{4 d_{L+1} n}{\rho}\right)} + a_{n-1}^{(L+1)} \\
&= \sum_{l=2}^{L+1} A_n^{(l,L+1)} D_n^{(l)} \sqrt{\frac{1}{2} n \ln\left(\frac{2^{L+1+2-l} n d_l}{\rho}\right)} \\
&\quad + A_n^{(1,L+1)} D_n^{(1)} \sqrt{\frac{1}{2} n \ln\left(\frac{2^{L+1} n d_1}{\rho}\right)} + \sum_{l=1}^{L+1} A_n^{(l,L+1)} a_{n-1}^{(l)} \\
&\leq \sum_{l=1}^{L+1} A_n^{(l,L+1)} \left[D_n^{(l)} \sqrt{\frac{1}{2} n \ln\left(\frac{2^{L+1+2-l} n d_l}{\rho}\right)} + a_{n-1}^{(l)} \right] = H^{(L+1)}(\rho).
\end{aligned}
$$

□

A.2. Invariant case

For the invariant case, we use the bound of the equivariant case previously obtained, and we make an additional use of McDiarmid's concentration bound.

Proof.

$$
\begin{aligned}
&\left\|\bar{\Theta}_{G_n}(S_X(f)) - \bar{\Theta}_{W,P}(f)\right\|_\infty \\
\leq &\left\|R\left(\{\!\{(S_X f)_1^{(L)}, \dots, (S_X f)_n^{(L)}\}\!\}\right) - R\left(\{\!\{f^{(L)}(X_1), \dots, f^{(L)}(X_n)\}\!\}\right)\right\|_\infty \\
&+ \left\|R\left(\{\!\{f^{(L)}(X_1), \dots, f^{(L)}(X_n)\}\!\}\right) - R_{P,n}(f^{(L)})\right\|_\infty \\
&+ \left\|R_{P,n}(f^{(L)}) - \mathcal{R}_P(f^{(L)})\right\|_\infty \\
\leq &\lambda_{R,n} \max_i \left\|(S_X f)_i^{(L)} - f^{(L)}(X_i)\right\|_\infty \\
&+ \left\|R\left(\{\!\{f^{(L)}(X_1), \dots, f^{(L)}(X_n)\}\!\}\right) - R_{P,n}(f^{(L)})\right\|_\infty + b_n
\end{aligned} \tag{A.6}
$$

Using the bound of the equivariant case, McDiarmid's inequality and the fact that $(\lambda_{R,n})$ is bounded, we get that, with probability at least $1-\rho$:

$$
\begin{aligned}
\left\|\bar{\Theta}_{G_n}(S_X(f)) - \bar{\Theta}_{W,P}(f)\right\|_\infty \lesssim\ & L D_n \sqrt{n \ln\left(\frac{n 2^{L+1} d_{max}}{\rho}\right)} + C_n \sqrt{n \ln\left(\frac{4 d_L}{\rho}\right)} \\
&+ L a_{n-1} + b_n
\end{aligned}
$$

□

B. Proof of Theorem 5.12

We will need the following property.

Proposition B.1. *Under the hypothesis of Theorem 5.12, The functions $f^{(0)}, \dots, f^{(L)}$ are Lipschitz continuous. We denote by $\lambda_f = \lambda_{f^{(0)}}, \dots, \lambda_{f^{(L)}}$ their Lipschitz constants.*

Proof. It is already assumed for $l = 0$. Suppose it is true for $l \geq 1$, $f^{(l+1)}(x) = \sup_y W(x,y)\psi^{(l+1)}(f^{(l)}(y)) = \sup_y g(x,y)$ where g is $\lambda_W \|\psi^{(l+1)} \circ f^{(l)}\|_\infty + \lambda_{\psi^{(l)}} \lambda_{f^{(l)}}$ Lipschitz. Then from Lemma D.6 $f^{(l+1)}$ is also Lipschitz. □

B.1. Equivariant case

We will prove the following sharper version of Theorem 5.12

Theorem B.2. *Suppose, $(\mathcal{X}, P)$ has the (r_0, κ)-volume retaining property and that f, W and the $\psi^{(l)}$ are Lipschitz continuousLet $\rho \geq 2^{L-1} n e^{-n\kappa r_0^d 2^d}$ and n large enough for $0 < \rho < 1$ to hold. Then with probability at least $1-\rho$:*

$$
\begin{aligned}
&\max_{1 \leq i \leq n} \left\|(S_X f)_i^{(L)} - f^{(L)}(X_i)\right\|_\infty \\
&\leq \sum_{l=1}^{L} B^{(l,L)} \frac{\lambda_{f^{(l)}}}{2} \left(\frac{1}{n\kappa} \ln\left(\frac{2^{L+1-l} n d_l}{\rho}\right)\right)^{1/d}
\end{aligned} \tag{B.1}
$$

Where $B^{(l,L)} = \prod_{k=l+1}^{L} \lambda_{\psi^{(k)}}$ with the conventions $\prod_{k=a}^{b} = 1$ and $\sum_{k=a}^{b} = 0$ if $a > b$.

Corollary B.3 (Theorem 5.12 in the main text)**.** *Suppose, $(\mathcal{X}, P)$ has the (r_0, κ)-volume retaining property and that f, W and the $\psi^{(l)}$ are Lipschitz continuousLet*

$\rho \geq 2^{L-1} n e^{-n\kappa r_0^d 2^d}$ and n large enough for $0 < \rho < 1$ to hold. Then with probability at least $1-\rho$:

$$\max_{1\leq i\leq n} \left\| (S_X f)_i^{(L)} - f^{(L)}(X_i) \right\|_\infty \lesssim L \left(\frac{1}{n} \ln \left(\frac{2^{L-1} n d_{\max}}{\rho} \right) \right)^{1/d}$$

Proof of theorem B.2. Let $\rho > 0$. We will prove the theorem by induction on L. Until the end of the proof, we denote by $H^{(L)}(\rho)$ the bound (B.1) :

$$H^{(L)}(\rho) = \sum_{l=1}^{L} B^{(l,L)} \frac{\lambda_{f^{(l)}}}{2} \left(\frac{1}{n\kappa} \ln \left(\frac{2^{L+1-l} n d_l}{\rho} \right) \right)^{1/d}.$$

For $L = 1$, let us note $g(x,y) = W(x,y)\psi^{(1)}(f^{(0)}(y))$. The map g is $\lambda_{f^{(1)}} = \lambda_{\psi^{(1)}}\lambda_{f^{(0)}} + \|\psi^{(1)} \circ f^{(0)}\|_\infty \lambda_W$ Lipschitz continuous from Property B.1. Fix $i \in \{1, \dots, n\}$ and $x_i \in \mathcal{X}$, by lemma 5.11, for $\rho \geq n e^{-n\kappa r_0^d 2^d}$, with probability at least $1 - \rho/n$, we have

$$\left\| \max_{j\neq i} g(x, X_j) - \sup_{y\in\mathcal{X}} g(x,y) \right\|_\infty \leq \frac{\lambda_{f^{(1)}}}{2} \left(\frac{1}{(n-1)\kappa} \ln \left(\frac{n d_0}{\rho} \right) \right)^{1/d} \qquad \forall x \in \mathcal{X}.$$

Thus, by the law of total probability, with probability at least $1 - \rho/n$,

$$\left\| \max_{j\neq i} g(X_i, X_j) - \sup_{y\in\mathcal{X}} g(X_i,y) \right\|_\infty \leq \frac{\lambda_{f^{(1)}}}{2} \left(\frac{1}{(n-1)\kappa} \ln \left(\frac{n d_0}{\rho} \right) \right)^{1/d}.$$

And by maximizing over i and doing an union bound, for $\rho \geq n e^{-n\kappa r_0^d 2^d}$, with probability at least $1 - \rho$:

$$\max_i \left\| (S_X f)_i^{(1)} - f^{(1)}(X_i) \right\|_\infty \leq \frac{\lambda_{f^{(1)}}}{2} \left(\frac{1}{(n-1)\kappa} \ln \left(\frac{n d_0}{\rho} \right) \right)^{1/d} \leq H^{(1)}(\rho).$$

That concludes the case $L = 1$. Now let $L \geq 1$, fix $i \in \{1, \dots, n\}$:

(B.2)

$$\begin{aligned}
&\left\| (S_X f)_i^{(L+1)} - f^{(L+1)}(X_i) \right\|_\infty \\
=& \left\| \max_{j\neq i} W(X_i, X_j) \psi^{(L+1)}((S_X f)_j^{(L)}) - f^{(L+1)}(X_j) \right\|_\infty \\
\leq& \left\| \max_{j\neq i} W(X_i, X_j) \psi^{(L+1)}((S_X f)_j^{(L)}) - \max_{j\neq i} W(X_i, X_j) \psi^{(L+1)}(f^{(L)}(X_j)) \right\|_\infty \\
&\quad + \left\| \max_{j\neq i} W(X_i, X_j) \psi^{(L+1)}(f^{(L)}(X_j)) - \sup_{y\in\mathcal{X}} W(X_i, y) \psi^{(L+1)}(f^{(L)}(y)) \right\|_\infty \\
\leq& \lambda_{\psi^{(L+1)}} \max_{j\neq i} \left\| (S_X f)_j^{(L)} - f^{(L)}(X_j) \right\|_\infty \\
&\quad + \left\| \max_{j\neq i} W(X_i, X_j) \psi^{(L+1)}(f^{(L)}(X_j)) - \sup_{y\in\mathcal{X}} W(X_i, y) \psi^{(L+1)}(f^{(L)}(y)) \right\|_\infty
\end{aligned}$$

Where the last inequality uses Lemma D.5, $|W| \leq 1$, and Lipschitz continuity of $\psi^{(L+1)}$. Thus taking the maximum over i :

(B.3)

$$\begin{aligned}
&\max_i \left\| (S_X f)_i^{(L+1)} - f^{(L+1)}(X_i) \right\|_\infty \\
&\leq \lambda_{\psi^{(L+1)}} \max_i \left\| (S_X f)_j^{(L)} - f^{(L)}(X_j) \right\|_\infty \\
&\quad + \max_i \left\| \max_{j \neq i} W(X_i, X_j) \psi^{(L+1)}(f^{(L)}(X_j)) - \sup_{y \in \mathcal{X}} W(X_i, y) \psi^{(L+1)}(f^{(L)}(y)) \right\|_\infty
\end{aligned}$$

Now we bound (B.3) with high probability. We use the induction hypothesis for the first term. For the second term, we set $g(x, y) = W(x, y)\psi^{(L+1)}(f^{(L)}(y))$ and use lemma 5.11 on g which is $\lambda_{f^{(L+1)}} = \lambda_{\psi^{(L+1)}} \lambda_{f^{(L)}} + \|\psi^{(L+1)} \circ f^{(L)}\|_\infty \lambda_W$ Lipschitz. The method is the same than in the case $L = 1$ and by conditioning over X_i followed by an union bound. We obtain that for $\rho \geq 2^L n e^{-n\kappa r_0^d 2^d}$, with probability at least $1 - \rho$:

$$\begin{aligned}
&\max_i \left\| (S_X f)_i^{(L+1)} - f^{(L+1)}(X_i) \right\|_\infty \\
&\leq \lambda_{\psi^{(L+1)}} H^{(L)}(\rho/2) + \frac{\lambda_{f^{(L+1)}}}{2} \left(\frac{1}{(n-1)\kappa} \ln\left(\frac{2nd_L}{\rho} \right) \right)^{1/d} \\
&= \sum_{l=2}^{L} \lambda_{\psi^{(L+1)}} B^{(l,L)} \frac{\lambda_{f^{(l)}}}{2} \left(\frac{1}{(n-1)\kappa} \ln\left(\frac{2^{L+2-l} n d_l}{\rho} \right) \right)^{1/d} \\
&\quad + \lambda_{\psi^{(L+1)}} B^{(1,L)} \frac{\lambda_{f^{(1)}}}{2} \left(\frac{1}{(n-1)\kappa} \ln\left(\frac{2^L n d_0}{\rho} \right) \right)^{1/d} \\
&\quad + \frac{\lambda_{f^{(L+1)}}}{2} \left(\frac{1}{(n-1)\kappa} \ln\left(\frac{2nd_L}{\rho} \right) \right)^{1/d} \\
&= \sum_{l=2}^{L} B^{(l,L+1)} \frac{\lambda_{f^{(l)}}}{2} \left(\frac{1}{(n-1)\kappa} \ln\left(\frac{2^{L+2-l} n d_l}{\rho} \right) \right)^{1/d} \\
&\quad + B^{(1,L+1)} \frac{\lambda_{f^{(1)}}}{2} \left(\frac{1}{(n-1)\kappa} \ln\left(\frac{2^L n d_0}{\rho} \right) \right)^{1/d} \\
&\quad + \frac{\lambda_{f^{(L+1)}}}{2} \left(\frac{1}{(n-1)\kappa} \ln\left(\frac{2nd_L}{\rho} \right) \right)^{1/d} \\
&= \sum_{l=2}^{L+1} B^{(l,L+1)} \frac{\lambda_{f^{(l)}}}{2} \left(\frac{1}{(n-1)\kappa} \ln\left(\frac{2^{L+2-l} n d_l}{\rho} \right) \right)^{1/d} \\
&\quad + B^{(1,L+1)} \frac{\lambda_{f^{(1)}}}{2} \left(\frac{1}{(n-1)\kappa} \ln\left(\frac{2^L n d_0}{\rho} \right) \right)^{1/d} \\
&\leq H^{(L+1)}(\rho).
\end{aligned}$$

□

B.2. Invariant case

Theorem B.4. *Suppose, $(\mathcal{X}, P)$ has the (r_0, κ)-volume retaining property and that f, W and the $\psi^{(l)}$ are Lipschitz continuousLet $\rho \geq 2^{L-1} n e^{-n\kappa r_0^d 2^d}$ and n large enough for $0 < \rho < 1$ to hold. Then with probability at least $1 - \rho$:*

$$\left\|\bar{\Theta}_{G_n}(S_X(f)) - \bar{\Theta}_{W,P}(f)\right\|_\infty \lesssim L \left(\frac{1}{n-1} \ln\left(\frac{2^L n d_{\max}}{\rho}\right)\right)^{1/d} + \left(\frac{1}{n} \ln\left(\frac{2 d_L}{\rho}\right)\right)^{1/d}.$$

Proof.

$$\begin{aligned}
\left\|\bar{\Theta}_{G_n}(S_X(f)) - \bar{\Theta}_{W,P}(f)\right\|_\infty &= \left\|\max_i (S_X f)_i^{(L)} - \sup f^{(L)}\right\|_\infty \\
&\leq \left\|\max_i (S_X f)_i^{(L)} - \max_i f^{(L)}(X_i)\right\|_\infty + \left\|\max_i f^{(L)}(X_i) - \sup f^{(L)}\right\|_\infty \\
&\leq \max_i \left\|(S_X f)_i^{(L)} - f^{(L)}(X_i)\right\|_\infty + \left\|\max_i f^{(L)}(X_i) - \sup f^{(L)}\right\|_\infty.
\end{aligned}$$

Using the bound for the equivariant case and Lemma 5.11 on $f^{(L)}$, we obtain the result. □

C. Examples

For notational convenience, we drop any subscript or superscript referring to layers. Recal that ψ is supposed Lipschitz and bounded, we denote λ_ψ its Lipschitz constant and $\|\psi\|_\infty = \sup_x \|\psi(x)\|_\infty$. For Examples 1, 2 and 3,

- We check Assumptions 5.6.
- We compute the a_n from (5.16).
- We compute the bounded differences

For Example 4, we show that the bounded differences are not sharp enough.

C.1. Examples 1 and a: Convolutional message passing with mean aggregation

Check of Assumptions 5.6. Let $x, x' \in \mathbb{R}^{d_{l-1}}$ and $m = \{\!\{(y_i, t_i),\}\!\}_{1 \leq i \leq n}, m' = \{\!\{(y'_i, t'_i)\}\!\}_{1 \leq i \leq n} \subset \mathbb{R}^{d_{l-1}} \times [0, 1]$.

$$\begin{aligned}
\|F(x, m) - F(x', m')\|_\infty &\leq \frac{1}{n} \sum_{1 \leq i \leq n} \|t_i \psi(y_i) - t'_i \psi(y'_i)\|_\infty \\
&\leq \frac{1}{n} \sum_{1 \leq i \leq n} \|t_i \psi(y_i) - t_i \psi(y'_i)\|_\infty + \|t_i \psi(y'_i) - t'_i \psi(y'_i)\|_\infty \\
&\leq \frac{1}{n} \sum_{1 \leq i \leq n} \lambda_\psi \|y_i - y'_i\| + \|\psi\|_\infty |t_i - t'_i| \leq \max(\lambda_\psi, \|\psi\|_\infty) \max_{1 \leq i \leq n} \|y_i - y'_i\|_\infty + |t_i - t'_i|.
\end{aligned}$$

This inequality does not depend on any ordering of the (y'_i, t'_i) so that

$$\|F(x, m) - F(x', m')\|_\infty \leq \max(\lambda_\psi, \|\psi\|_\infty) \max_{1 \leq i \leq n} \|y_i - y'_{\sigma(i)}\|_\infty + |t_i - t'_{\sigma(i)}|$$

for any permutation σ. Taking the minimum over S_n we get the Assumption with $\mu_F = 0$ and $\lambda_{F,n} = \max(\lambda_\psi, \|\psi\|_\infty)$ which is bounded over n.

Calculation of a_n By linearity of the expected value, it is clear that any $X_1, \dots, X_n \overset{iid}{\sim} P$ and any f,

$$\mathbb{E}\left[\frac{1}{n}\sum_i W(x, X_i)\psi(X_i)\right] = \mathbb{E}\left[W(x, X_1)\psi(X_1)\right] = \int_{\mathcal{X}} W(x, y)\psi(y)dP(y) \quad \forall x$$

So $a_n = 0$.

Calculation of bounded differences $\mathbf{D_n}$. Let x_1, x_n and $x'_2, \dots, x'_n$ be such that $x_i = x'_i$ except at $i = 2$

$$\begin{aligned}
&\|F\left(f(x_1), \{\!\{(f(x_k), W(x_1, x_k))\}\!\}_{2\le k\le n}\right) - F\left(f(x_1), \{\!\{(f(x'_k), W(x_1, x'_k))\}\!\}_{2\le k\le n}\right)\|_\infty \\
&= \frac{1}{n-1}\left\|W(x_1, x_2)\psi(f(x_2)) - W(x_1, x'_2)\psi^{(l)}(f(x'_2))\right\| \\
&= O(1/n).
\end{aligned}$$

Since $\psi \circ f$ is bounded by continuity and compactness.

C.2. Example 2 and b : Degree normalized convolutional message passing with sum aggregation

We make the additional assumption that there exists $\alpha > 0$ such that $W(x, y) > \alpha,\ \forall x, y$.

Check of Assumption 5.6. Let $x, x' \in \mathbb{R}^d$ and $m = \{\!\{(y_i, t_i), \}\!\}_{1\le i\le n}, m' = \{\!\{(y'_i, t'_i)\}\!\}_{1\le i\le n} \subset \mathbb{R}^{d_{l-1}} \times [\alpha, 1]$.

$$\begin{aligned}
&\|F(x, m) - F(x', m')\|_\infty \le \sum_{1\le i\le n}\left\|\frac{t_i\psi(y_i)}{\sum_k t_k} - \frac{t'_i\psi(y_i)}{\sum_k t'_k}\right\|_\infty \\
&= \sum_{1\le i\le n}\left\|\frac{\sum_k t'_k t_i\psi(y_i) - \sum_k t_k t'_i\psi(y'_i)}{\sum_k t_k \sum_k t'_k}\right\|_\infty \\
&\le \frac{1}{n^2\alpha^2}\sum_{1\le i,k\le n}\|t'_k t_i\psi(y_i) - t_k t'_i\psi(y'_i)\|_\infty \\
&\le \frac{1}{n^2\alpha^2}\sum_{1\le i,k\le n}\|t'_k t_i\psi(y_i) - t'_k t'_i\psi(y'_i)\|_\infty + \|t'_k t'_i\psi(y'_i) - t_k t'_i\psi(y'_i)\|_\infty \\
&\le \frac{1}{n^2\alpha^2}\sum_{1\le i,k\le n}\|t_i\psi(y_i) - t'_i\psi(y'_i)\|_\infty + \|t'_k\psi(y'_i) - t_k\psi(y'_i)\|_\infty \\
&\le \frac{1}{n^2\alpha^2}\sum_{1\le i,k\le n}\|t_i\psi(y_i) - t_i\psi(y'_i)\|_\infty + \|t_i\psi(y'_i) - t'_i\psi(y'_i)\|_\infty + \|\psi\|_\infty|t_k - t'_k| \\
&\le \frac{1}{n^2\alpha^2}\sum_{1\le i,k\le n}\lambda_\psi\|y_i - y'_i\|_\infty + \|\psi\|_\infty|t_i - t'_i| + \|\psi\|_\infty|t_k - t'_k| \\
&\le \frac{1}{\alpha^2}\left(\lambda_\psi \max_i\|y_i - y'_i\|_\infty + 2\|\psi\|_\infty \max_i|t_i - t'_i|\right) \\
&= \frac{\max\left(\lambda_\psi, 2\|\psi\|_\infty\right)}{\alpha^2}\max_{1\le i\le n}\|y_i - y'_i\|_\infty + |t_i - t'_i|.
\end{aligned}$$

This inequality does not depend on any ordering of the (y'_i, t'_i) so that

$$\|F(x,m) - F(x',m')\|_\infty \leq \frac{\max(\lambda_\psi, 2\|\psi\|_\infty)}{\alpha^2} \max_{1\leq i\leq n} \|y_i - y'_{\sigma(i)}\|_\infty + |t_i - t'_{\sigma(i)}|$$

for any permutation σ. Taking the minimum over S_n we get the Assumption with $\mu_F = 0$ and $\lambda_{F,n} = \frac{\max(\lambda_\psi, 2\|\psi\|_\infty)}{\alpha^2}$ which is bounded over n.

Calculation of a_n For $X_1, \ldots, X_n \overset{iid}{\sim} P$, we have $\forall x$

$$\int_{\mathcal{X}} \frac{W(x,y)\psi(f(y))}{\int_{\mathcal{X}} W(x,t)dP(t)} dP(y) = \frac{\mathbb{E}\left[W(x,X_1)\psi(f(X_1))\right]}{\mathbb{E}\left[W(x,X_1)\right]},$$

and

$$\mathbb{E}\left[\frac{\sum_i W(x,X_i)\psi(f(X_i))}{\sum_k W(x,X_k)}\right] = \mathbb{E}\left[\frac{\frac{1}{n}\sum_i W(x,X_i)\psi(f(X_i))}{\frac{1}{n}\sum_k W(x,X_k)}\right].$$

$$\begin{aligned}
&\left\|\mathbb{E}\left[\frac{\frac{1}{n}\sum_i W(x,X_i)\psi(f(X_i))}{\frac{1}{n}\sum_k W(x,X_k)}\right] - \frac{\mathbb{E}\left[W(x,X_1)\psi(f(X_1))\right]}{\mathbb{E}\left[W(x,X_1)\right]}\right\|_\infty \\
&= \left\|\mathbb{E}\left[\frac{\frac{1}{n}\sum_i W(x,X_i)\psi(f(X_i))}{\frac{1}{n}\sum_k W(x,X_k)} - \frac{\mathbb{E}\left[W(x,X_1)\psi(f(X_1))\right]}{\mathbb{E}\left[W(x,X_1)\right]}\right]\right\|_\infty \\
&\leq \mathbb{E}\left[\left\|\frac{\mathbb{E}\left[W(x,X_1)\right]\frac{1}{n}\sum_i W(x,X_i)\psi(f(X_i)) - \frac{1}{n}\sum_k W(x,X_k)\mathbb{E}\left[W(x,X_1)\psi(f(X_1))\right]}{\frac{1}{n}\sum_k W(x,X_k)\mathbb{E}\left[W(x,X_1)\right]}\right\|_\infty\right] \\
&\leq \frac{1}{\alpha^2}\mathbb{E}\left[\left\|\mathbb{E}\left[W(x,X_1)\right]\frac{1}{n}\sum_i W(x,X_i)\psi(f(X_i)) - \frac{1}{n}\sum_k W(x,X_k)\mathbb{E}\left[W(x,X_1)\psi(f(X_1))\right]\right\|_\infty\right] \\
&\leq \frac{1}{\alpha^2}\mathbb{E}\left[\left\|\mathbb{E}\left[W(x,X_1)\right]\frac{1}{n}\sum_i W(x,X_i)\psi(f(X_i)) - E\left[W(x,X_1)\right]E\left[W(x,X_1)\psi(f(X_1))\right]\right\|_\infty\right. \\
&\qquad \left. + \left\|\mathbb{E}\left[W(x,X_1)\right]\mathbb{E}\left[W(x,X_1)\psi(f(X_1))\right] - \frac{1}{n}\sum_k W(x,X_k)\mathbb{E}\left[W(x,X_1)\psi(f(X_1))\right]\right\|_\infty\right] \\
&\leq \frac{1}{\alpha^2}\mathbb{E}\left[\left\|\frac{1}{n}\sum_i W(x,X_i)\psi(f(X_i)) - \mathbb{E}\left[W(x,X_1)\psi(f(X_1))\right]\right\|_\infty\right] \\
&\qquad + \frac{\|\psi\|_\infty}{\alpha^2}\mathbb{E}\left[\left|\mathbb{E}\left[W(x,X_1)\right] - \frac{1}{n}\sum_k W(x,X_k)\right|\right]
\end{aligned}$$

Using the formula $E(X) = \int_{t>0} P(X > t)dt$ for X nonnegative,

$$\begin{aligned}
&= \frac{1}{\alpha^2}\int_{t>0} \mathbb{P}\left(\left\|\frac{1}{n}\sum_i W(x,X_i)\psi(f(X_i)) - \mathbb{E}\left[W(x,X_1)\psi(f(X_1))\right]\right\|_\infty > t\right) dt \\
&\qquad + \frac{\|\psi\|_\infty}{\alpha^2}\int_{t>0} \mathbb{P}\left(\left|\mathbb{E}\left[W(x,X_1)\right] - \frac{1}{n}\sum_k W(x,X_k)\right| > t\right) dt
\end{aligned}$$

Using McDiarmid (or Hoeffding) inequality, there is some positive constants K_1, K_s, K_3, K_4 independent of x such that this is bounded by

$$\frac{1}{\alpha^2}\int_{t>0} K_1 e^{-nK_1 t^2} dt + \frac{\|\psi\|_\infty}{\alpha^2}\int_{t>0} K_3 e^{-nK_4 t^2} dt = O(1/\sqrt{n}) = a_n.$$

Calculation of bounded differences $\mathbf{D_n}$. Let x_1, x_n and $x'_2, \dots, x'_n$ be such that $x_i = x'_i$ except at $i = 2$. This is the same calculation than the previous paragraph where $x = x' = f(x_1)$, $y_i = f(x_i)$, $y'_i = f(x'_i)$, $t_i = W(x_1, x_i)$ and $t'_i = W(x_1, x'_i)$ for $i \geq 2$. We get

$$\begin{aligned}
&\|F\left(f(x_1), \{\!\{(f(x_k), W(x_1, x_k))\}\!\}_{2\leq k\leq n}\right) - F\left(f(x_1), \{\!\{(f(x'_k), W(x_1, x'_k))\}\!\}_{2\leq k\leq n}\right)\|_\infty \\
&\leq \frac{1}{(n-1)^2\alpha^2} \sum_{1\leq i,k\leq n-1} \lambda_\psi \left\|f(x_i) - f(x'_i)\right\|_\infty + \|\psi\|_\infty |W(x_1, x_i) - W(x_1, x'_i)| + \|\psi\|_\infty |W(x_1, x_k) - W(x_1, x'_k)| \\
&= \frac{1}{(n-1)^2\alpha^2}((n-1)\lambda_\psi \left\|f(x_2) - f(x'_2)\right\|_\infty + (n-1)\|\psi\|_\infty |W(x_1, x_2) - W(x_1, x'_2)| \\
&\qquad\qquad + (n-1)\|\psi\|_\infty |W(x_1, x_2) - W(x_1, x'_2)|) \\
&\leq \frac{2\lambda_\psi \|f\|_\infty + 4\|\psi\|_\infty}{(n-1)\alpha^2} \\
&= O(1/n) \\
&= D_n.
\end{aligned}$$

So $D_n = O(1/n)$.

C.3. Example 3 and c : Attentional message passing with sum aggregation

We make the additional assumption that there exists $\alpha, \beta > 0$ and $\lambda_c > 0$ such that $\alpha < c(x, y, t) < \beta$ and $|c(x, y, t) - c(x', y', t')| \leq \lambda_c(\|x - x'\|_\infty + \|y - y'\|_\infty + |t - t'|)$, $\forall x, x', y, y', t, t'$. As a consequence, c is bounded on any compact set.

Check of Assumption 5.6. Let $x, x' \in \mathbb{R}^d$ and $m = \{\!\{(y_i, t_i)\}\!\}_{1\leq i\leq n}, m' = \{\!\{(y'_i, t'_i)\}\!\}_{1\leq i\leq n} \subset \mathbb{R}^{d_{l-1}} \times [0, 1]$. Let us shorten $c(x, y_i, t_i)$ and $c(x, y'_i, t'_i)$ as c_i and

c'_i.

$$
\begin{aligned}
\|F(x,m) - F(x',m')\|_\infty &\leq \sum_{1\leq i\leq n} \left\| \frac{c_i\psi(y_i)}{\sum_k c_k} - \frac{c'_i\psi(y_i)}{\sum_k c'_k} \right\|_\infty \\
&= \sum_{1\leq i\leq n} \left\| \frac{\sum_k c'_k t_i \psi(y_i) - \sum_k c_k c'_i \psi(y'_i)}{\sum_k c_k \sum_k c'_k} \right\|_\infty \\
&\leq \frac{1}{n^2\alpha^2} \sum_{1\leq i,k\leq n} \|c'_k c_i \psi(y_i) - c_k c'_i \psi(y'_i)\|_\infty \\
&\leq \frac{1}{n^2\alpha^2} \sum_{1\leq i,k\leq n} \|c'_k c_i \psi(y_i) - c'_k c'_i \psi(y'_i)\|_\infty + \|c'_k c'_i \psi(y'_i) - c_k c'_i \psi(y'_i)\|_\infty \\
&\leq \frac{\beta}{n^2\alpha^2} \sum_{1\leq i,k\leq n} \|c_i \psi(y_i) - c'_i \psi(y'_i)\|_\infty + \|c'_k \psi(y'_i) - c_k \psi(y'_i)\|_\infty \\
&\leq \frac{\beta}{n^2\alpha^2} \sum_{1\leq i,k\leq n} \|c_i \psi(y_i) - c_i \psi(y'_i)\|_\infty + \|c_i \psi(y'_i) - c'_i \psi(y'_i)\|_\infty + \|\psi\|_\infty |c_k - c'_k| \\
&\leq \frac{\beta}{n^2\alpha^2} \sum_{1\leq i,k\leq n} \lambda_\psi \beta \, \|y_i - y'_i\|_\infty + \|\psi\|_\infty |c_i - c'_i| + \|\psi\|_\infty |c_k - c'_k| \\
&\leq \frac{\beta}{n^2\alpha^2} \sum_{1\leq i,k\leq n} \lambda_\psi \beta \, \|y_i - y'_i\|_\infty + \|\psi\|_\infty \lambda_c (\|x - x'\|_\infty + \|y_i - y'_i\|_\infty + |t_i - t'_i|) \\
&\qquad\qquad + \|\psi\|_\infty \lambda_c (\|x - x'\|_\infty + \|y_k - y'_k\|_\infty + |t_k - t'_k|) \\
&\leq \frac{2\beta\|\psi\|_\infty \lambda_c}{\alpha^2} \|x - x'\|_\infty + \frac{\beta \max(\beta\lambda_\psi, 2\|\psi\|_\infty \lambda_c)}{\alpha^2} \max_{1\leq i\leq n} \|y_i - y'_i\|_\infty + |t_i - t'_i|.
\end{aligned}
$$

This inequality does not depend on any ordering of the (y'_i, t'_i) so that

$$
\|F(x,m) - F(x',m')\|_\infty \leq \frac{2\beta\|\psi\|_\infty \lambda_c}{\alpha^2} \|x - x'\|_\infty + \frac{\beta \max(\beta\lambda_\psi, 2\|\psi\|_\infty \lambda_c)}{\alpha^2} \max_{1\leq i\leq n} \|y_i - y'_{\sigma(i)}\|_\infty + |t_i - t'_{\sigma(i)}|
$$

for any permutation σ. Taking the minimum over S_n we get the Assumption with $\mu_F = \frac{2\beta\|\psi\|_\infty \lambda_c}{\alpha^2}$ and $\lambda_{F,n} = \frac{\beta \max(\beta\lambda_\psi, 2\|\psi\|_\infty \lambda_c)}{\alpha^2}$ which are bounded over n.

Calculation of a_n Let us denote $V(x,y) = c(f(x), f(y), W(x,y))$. For $X_1, \ldots, X_n \overset{iid}{\sim} P$, we have $\forall x$

$$
\int_{\mathcal{X}} \frac{c(f(x), f(y), W(x,y))\psi(f(y))}{\int_{\mathcal{X}} c(f(x), f(y), W(x,y)) dP(t)} dP(y) = \int_{\mathcal{X}} \frac{V(x,y)\psi(f(y))}{\int_{\mathcal{X}} V(x,t) dP(t)} dP(y) = \frac{\mathbb{E}\left[V(x, X_1)\psi(f(X_1))\right]}{\mathbb{E}\left[V(x, X_1)\right]},
$$

and

$$
\mathbb{E}\left[\frac{\sum_i V(x, X_i)\psi(f(X_i))}{\sum_k W(x, X_k)} \right] = \mathbb{E}\left[\frac{\frac{1}{n}\sum_i V(x, X_i)\psi(f(X_i))}{\frac{1}{n}\sum_k V(x, X_k)} \right].
$$

This is the same setup than in example 2, with V instead of W and $\alpha < V < \beta$ instead of $\alpha < W < 1$. So the same calculation gives the result with $a_n = O(1/\sqrt{n})$

Calculation of bounded differences $\mathbf{D_n}$. Let x_1, x_n and $x'_2, \ldots, x'_n$ be such that $x_i = x'_i$ except at $i = 2$. Using again the notation $V(x,y) = c(f(x), f(y), W(x,y))$ and the fact that V is bounded by β, we end up performing the same calculation

than in the case of example 3-c. We get

$$
\begin{aligned}
&\|F\left(f(x_1), \{\!\{(f(x_k), W(x_1,x_k))\}\!\}_{2\le k\le n}\right) - F\left(f(x_1), \{\!\{(f(x'_k), W(x_1,x'_k))\}\!\}_{2\le k\le n}\right)\|_\infty \\
&\le \frac{1}{(n-1)^2\alpha^2} \sum_{1\le i,k\le n-1} \lambda_\psi\beta \left\|f(x_i) - f(x'_i)\right\|_\infty + \|\psi\|_\infty |V(x_1,x_i) - V(x_1,x'_i)| + \|\psi\|_\infty |V(x_1,x_k) - V(x_1,x'_k)| \\
&= \frac{1}{(n-1)^2\alpha^2} ((n-1)\lambda_\psi\beta \left\|f(x_2) - f(x'_2)\right\|_\infty + (n-1)\|\psi\|_\infty |V(x_1,x_2) - V(x_1,x'_2)| \\
&\qquad\qquad + (n-1)\|\psi\|_\infty |V(x_1,x_2) - V(x_1,x'_2)|) \\
&\le \frac{2\lambda_\psi \|f\|_\infty \beta + 4\|\psi\|_\infty\beta}{(n-1)\alpha^2} \\
&= O(1/n) \\
&= D_n.
\end{aligned}
$$

So $D_n = O(1/n)$.

C.4. Example 2 and b Convolutional Message Passing with max aggregation

Here we check the bounded differences are not sharp. Recall the hypothesis

Bounded differences are not sharp. We show that $(x_1,\dots,x_n) \mapsto \max_i W(x,x_i)\psi(f(x_i))$ has no bounded differences in $o(1/\sqrt{n\ln n})$. Call $g(x,y) = W(x,y)f^{(l-1)}(y)$, and $(g_1,\dots,g_{d_l})$ its components which are real functions. We suppose g not constant, so there is k such that g_k is not constant, say $k=1$. By compactness and continuity of g_1 there is x^* such that $g(x,x^*) = sup_y g(x,y)$. Since g_1 is not constant, for any n, there exist $x_1,\dots,x_n$ such that $g(x,x_1),\dots,g(x,x_n)$ are all strictly smaller that $g(x,x^*)$. Up to reordering them we suppose $g(x,x_1) = \max_{2\le i\le n} g(x,x_i)$ and call $\alpha = |g_1(x,x^*) - g_1(x,x_1)| > 0$.

$$
\begin{aligned}
&\|\max\{g(x,x^*), g(x,x_2),\dots,g(x,x_n)\} - \max\{g(x,x_1), g(x,x_2),\dots,g(x,x_n)\}\|_\infty \\
&\ge |\max\{g_1(x,x^*), g_1(x,x_2),\dots,g_1(x,x_n)\} - \max\{g_1(x,x_1), g_1(x,x_2),\dots,g_1(x,x_n)\}| \\
&= |g_1(x,x^*) - g_1(x,x_1)| \\
&> \alpha.
\end{aligned}
$$

So for any n

$$\alpha < \sup \|\max\{g(x,x_1),\dots,g(x,x_n)\} - \max\{g(x,x'_1),\dots,\dots,g(x,x'_n)\}\|_\infty$$

where the supremum is taken over $x, x_2,\dots,x_n,x'_2,\dots,x'_n \in \mathcal{X}$ such that $(x_2,\dots,x_n)$ and $(x'_2,\dots,x'_n)$ differ from only one component. That proves that the bounded differences are not $o(1/\sqrt{n\ln n})$.

D. Useful results

The following remark can be useful,

Proposition D.1. *f has the bounded difference property if and only if f is bounded.*

Proof. If f is bounded, clearly $c_1 = \cdots = c_n = \|f\|_\infty$ are valid bounded differences. Conversely, if $c_1,\dots,c_n$ are bounded differences of f, pick $a \in \mathcal{E}$ and for any $x \in \mathcal{E}^n$,

introduce $x^k = (x_1, \dots, x_k, a \dots, a), 0 \leq k \leq n$. Then

$$|f(x)| \leq |f(a)| + |f(x) - f(a)| = |f(a)| + \left| \sum_{k=0}^{n-1} f(x^{k+1}) - f(x^k) \right| \leq |f(a)| + \sum c_i.$$

□

Theorem D.2 (McDiarmid inequality [27])**.** *Suppose $\mathcal{E}$ is a probability space and let $f : \mathcal{E}^n \to \mathbb{R}$ be a function of n variables. Suppose that f satisfies the bounded differences property with the n nonnegative constants $c_1, \dots, c_n$. Then for any independent random variables $X_1, \dots, X_n$ in $\mathcal{E}$, for any $\epsilon > 0$:*

$$\mathbb{P}(|f(X_1, \dots, X_n) - E(f(X_1, \dots, X_n))| > \epsilon) \leq 2e^{-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}}.$$

Notice that the X_i are not required to be identically distributed. By an union bound and reformulating Proposition D.2 to a bound with high probability one can obtain the following result for function taking multidimensional values.

Corollary D.3 (Multi dimensional McDiarmid inequality)**.** *Suppose that $f : \mathcal{E}^n \to \mathbb{R}^d$ satisfies a vectorial version on the bounded difference : $\|f(x) - f(x')\|_\infty \leq c_i$ whenever x and x' differ only from the i-th component. Then for any independent random variables $X_1, \dots, X_n$ in $\mathcal{E}$, for any $\rho > 0$:*

$$\|f(X_1, \dots, X_n) - E(f(X_1, \dots, X_n))\|_\infty \leq \sqrt{\frac{1}{2} \sum_{i=1}^n c_i^2 \ln\left(\frac{2d}{\rho}\right)}$$

holds with probability at least $1 - \rho$.

Lemma D.4. *Suppose P is strictly positive i.e : for all $U \subset \mathbb{R}^d$, $P(U \cap \mathcal{X}) > 0$ if and only if $U \cup \mathcal{X}$ is nonvoid. Then for any continuous map $f : \mathcal{X} \to \mathbb{R}$,*

$$\operatorname*{ess\,sup}_P f = \sup f < +\infty.$$

Proof. Clearly $\operatorname*{ess\,sup}_P f \leq \sup f$ and $\sup f < +\infty$ by continuity and compactness. Suppose that $\operatorname*{ess\,sup}_P f < \sup f$ then there is M such that $\operatorname*{ess\,sup}_P f < M < \sup f$. By definition of $\sup f$, the set $(f > M) = f^{-1}(]M; +\infty[)$ is nonempty, it is also a relative open of $\mathcal{X}$ since it is the inverse image of an open by a continuous map. Thus this set has strictly positive measure, which yields a contradiction with the fact that $\operatorname*{ess\,sup}_P f < M$. □

Lemma D.5. *Let $(a_i)_{i \in I}$ and $(b_i)_{i \in I}$ be two finite families of vectors in $\mathbb{R}^m$. We have the following properties :*

(i) $\|\max_i a_i\|_\infty \leq \max_i \|a_i\|_\infty$.

(ii) $\|\max_i a_i - \max_i b_i\|_\infty \leq \max_i \|a_i - b_i\|_\infty$.

Proof. (i)For $m = 1$, $a_i \leq |a_i| \implies \max_i a_i \leq \max_i |a_i| \implies |\max_i a_i| \leq |\max_i |a_i|| = \max_i |a_i|$. For $m \geq 1$, $\|\max_i a_i\|_\infty = \max_k |\max_i a_i^{(k)}| \leq \max_k \max_i |a_i^{(k)}| = \max_i \max_k |a_i^{(k)}| = \max_i \|a_i\|_\infty$.

(ii) For $m = 1$. Let i_a (resp. i_b) be an index that realizes $\max_i a_i$ (resp. $\max_i b_i$). We have

$$\max_i a_i - \max_i b_i = a_{i_a} - b_{i_b} = a_{i_a} - b_{i_a} + \underbrace{b_{i_a} - b_{i_b}}_{\leq 0}$$
$$\leq a_{i_a} - b_{i_a} \leq \max_i a_i - b_i \leq |\max_i a_i - b_i| \leq \max_i |a_i - b_i|.$$

Then analogously $\max_i b_i - \max_i a_i \leq \max_i |b_i - a_i| = \max_i |a_i - b_i|$.

For $m \geq 1$, $\|\max_i a_i - \max_i b_i\|_\infty = \max_k |\max_i a_i^{(k)} - \max_i b_i^{(k)}| \leq \max_k \max_i |a_i^{(k)} - b_i^{(k)}| = \max_i \|a_i - b_i\|_\infty$. □

Lemma D.6. *Let $g : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^m$ be λ_g-Lipschitz continuous. Then $f : x \mapsto \sup_{y \in \mathcal{X}} g(x, y)$ is also λ_g-Lipschitz continuous.*

Proof. For $m = 1$. Let $x, x' \in \mathcal{X}$,by continuity on a compact, $\exists\, x^*, x'^*$ such that $f(x) = g(x, x^*)$ and $f(x') = g(x', x'^*)$. Then $f(x) - f(x') = g(x, x^*) - g(x', x^*) + \underbrace{g(x', x^*) - g(x', x'^*)}_{\leq 0} \leq \lambda_g \|x - x'\|_\infty$, and permuting x and x' we obtain the Lipschitz condition.

For $m \geq 1$, $\|f(x) - f(x')\|_\infty = \max_i |\sup_y g_i(x, y) - \sup_y g_i(x', y)| = \max_i |g_i(x, x^*) - g_i(x', x'^*)| \leq \max_i \lambda_g \|x - x'\|_\infty = \lambda_g \|x - x'\|_\infty$. □

. References

[1] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The Graph Neural Network Model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, Jan. 2009, conference Name: IEEE Transactions on Neural Networks.

[2] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, vol. 2, Jul. 2005, pp. 729–734 vol. 2, iSSN: 2161-4407.

[3] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 1263–1272. [Online]. Available: https://proceedings.mlr.press/v70/gilmer17a.html

[4] Z. Chen, L. Li, and J. Bruna, "Supervised community detection with line graph neural networks," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=H1g0Z3A9Fm

[5] A. Fout, J. Byrd, B. Shariat, and A. Ben-Hur, "Protein interface prediction using graph convolutional networks," in *NIPS*, 2017.

[6] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/04df4d434d481c5bb723be1b6df1ee65-Paper.pdf

[7] N. Tremblay, P. Gonçalves, and P. Borgnat, "Design of graph filters and filterbanks," in *Cooperative and Graph Signal Processing*, P. M. Djurić and C. Richard, Eds. Academic Press, Jun. 2018, pp. 299–324. [Online]. Available: https://inria.hal.science/hal-01675375

[8] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *Proceedings of the 5th International Conference on Learning Representations*, ser. ICLR '17, 2017. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl

[9] Kurt and Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, no. 2, pp. 251–257, 1991. [Online]. Available: http://www.sciencedirect.com/science/article/pii/089360809190009T

[10] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems (MCSS)*, vol. 2, no. 4, pp. 303–314, Dec. 1989. [Online]. Available: http://dx.doi.org/10.1007/BF02551274

[11] Z. Chen, S. Villar, L. Chen, and J. Bruna, "On the equivalence between graph isomorphism testing and function approximation with gnns," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/71ee911dd06428a96c143a0b135041a4-Paper.pdf

[12] C. Morris, G. Rattan, and P. Mutzel, "Weisfeiler and leman go sparse: Towards scalable higher-order graph embeddings," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 21 824–21 840. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/f81dee42585b3814de199b2e88757f5c-Paper.pdf

[13] S. Adam-Day, T. M. Iliant, and I. I. Ceylan, "Zero-one laws of graph neural networks," Preprint submitted to the International Conference on Machine Learning, 2023, preprint submitted to the International Conference on Machine Learning.

[14] N. Keriven, A. Bietti, and S. Vaiter, "Convergence and Stability of Graph Convolutional Networks on Large Random Graphs," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 21 512–21 523. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/f5a14d4963acf488e3a24780a84ac96c-Paper.pdf

[15] ——, "On the universality of graph neural networks on large random graphs," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 6960–6971. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/38181d991caac98be8fb2ecb8bd0f166-Paper.pdf

[16] L. Ruiz, L. F. O. Chamon, and A. Ribeiro, "Transferability Properties of Graph Neural Networks," *arXiv:2112.04629 [cs, eess]*, Dec. 2021, arXiv: 2112.04629. [Online]. Available: http://arxiv.org/abs/2112.04629

[17] S. Maskey, R. Levie, Y. Lee, and G. Kutyniok, "Generalization analysis of message passing neural networks on large random graphs," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 4805–4817. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/1eeaae7c89d9484926db6974b6ece564-Paper-Conference.pdf

[18] A. Magner, M. Baranwal, and A. O. Hero, "The power of graph convolutional networks to distinguish random graph models," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 2664–2669.

[19] L. Lovász, *Large Networks and Graph Limits*, ser. Colloquium Publications. Providence, Rhode Island: American Mathematical Society, Dec. 2012, vol. 60. [Online]. Available: http://www.ams.org/coll/060

[20] L. Ruiz, L. Chamon, and A. Ribeiro, "Graphon Neural Networks and the Transferability of Graph Neural Networks," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1702–1712. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/12bcd658ef0a540cabc36cdf2b1046fd-Paper.pdf

[21] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," in *ICLR 2019 Workshop on Representation Learning on Graphs and Manifolds*, 2019. [Online]. Available: https://arxiv.org/abs/1903.02428

[22] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates,

Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf

[23] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," *6th International Conference on Learning Representations*, 2017.

[24] S. Jegelka, "Theory of Graph Neural Networks: Representation and Learning," Apr. 2022, arXiv:2204.07697 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2204.07697

[25] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A Comprehensive Survey on Graph Neural Networks," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021, arXiv: 1901.00596. [Online]. Available: http://arxiv.org/abs/1901.00596

[26] C. McDiarmid, *On the method of bounded differences*, ser. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989, p. 148–188.

[27] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: a nonasymptotic theory of independence*. Oxford: Oxford university press, 2013.