Journal of Machine Learning Research 25 (2024) 1-49 Submitted 7/23; Revised 8/24; Published 11/24

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

Matthieu Cordonnier MATTHIEU.CORDONNIER@GIPSA-LAB.FR
GIPSA-lab, Université Grenoble Alpes, Grenoble

Nicolas Keriven NICOLAS.KERIVEN@CNRS.FR
CNRS, IRISA, Rennes

Nicolas Tremblay NICOLAS.TREMBLAY@CNRS.FR
CNRS, GIPSA-lab, Grenoble

Samuel Vaiter SAMUEL.VAITER@CNRS.FR
CNRS, LJAD, Nice

Editor: Joan Bruna

Abstract

We study the convergence of message-passing graph neural networks on random graph models toward their continuous counterparts as the number of nodes tends to infinity. Until now, this convergence was only known for architectures with aggregation functions in the form of normalized means, or, equivalently, of an application of classical operators like the adjacency matrix or the graph Laplacian. We extend such results to a large class of aggregation functions, that encompasses all classically used message-passing graph neural networks, such as attention-based message-passing, max convolutional message-passing, (degree-normalized) convolutional message-passing, or moment-based aggregation 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, this result does not apply to the case where the aggregation is a coordinate-wise maximum. We treat this case separately and obtain a different convergence rate.

Keywords: Graph Neural Networks, Random Graphs, Message-Passing, Large Graphs, Aggregation Function, Concentration.

1. Introduction

Graph Neural Networks (GNNs) (Scarselli et al., 2008; Gori et al., 2005) are deep learning architectures largely inspired by Convolutional Neural Networks, that aim to extend convolutional methods to signals on graphs. Indeed, in many domains, the measured data live on a graph structure: examples for which GNNs have achieved state-of-the-art performance include molecules, proteins, and node clustering (Gilmer et al., 2017; Chen et al., 2019; Fout et al., 2017). Nevertheless, it has been observed that GNNs have limitations, both in practice (Wu et al., 2019; Hu et al., 2020) and in their theoretical understanding (Morris et al., 2024). Hence, the design of more reliable and powerful architectures is a current active and fast evolving area of research.

From a theoretical perspective, a large part of the literature has focused on the *expressive power* of GNNs, *i.e.,* what class of functions can GNNs approximate. This notion is fundamental in classical Deep Learning and is related to the so-called *Universal Approximation Theorem* (Hornik, 1991; Cybenko, 1989). Studying the expressive power of GNNs is however more involved, as they are by definition designed to be invariant or equivariant to the relabeling of nodes in a graph (see Section 3). Hence, in Xu et al. (2019) the authors relate their expressivity to the *graph isomorphism problem*, that is, deciding if two graphs are isomorphic of one another, a long-standing combinatorial problem in graph theory. The main avenue to analyze the expressive power of GNNs compares their performances to the traditional Weisfeiler-Lehman algorithm (WL) (Weisfeiler and Leman, 1968), which process is very similar to the message-passing paradigm at the core of GNNs. Hence, by construction, basic GNNs are at most as powerful as WL (Xu et al., 2019). From this point, a lot of effort has been made to design innovative GNN architectures to outperform the classical WL (Maron et al., 2019b,a; Keriven and Peyré, 2019; Vignac et al., 2020; Papp and Wattenhofer, 2022; Morris et al., 2019).

Nevertheless, while this combinatorial approach is worth considering for reasonably small graphs, its relevance in the context of large graphs is somewhat limited. Two real large graphs may share similar patterns, but will never be isomorphic, one main simple reason being that they most likely do not even share the same number of nodes. Large graphs are better described by some global properties such as edge density or number of communities. To that extent, the privileged mathematical tools are random graph models (Crane, 2018; Goldenberg et al., 2010). A generic family of models of interest to study GNNs on large graphs is the class of Latent Position Models (Keriven et al., 2020, 2021; Ruiz et al., 2020; Levie et al., 2021; Maskey et al., 2022). In such a model, the nodes of a random graph are first sampled as latent random variables (the latent positions) in a probability space $(\mathcal{X}, P)$, and then, the adjacency is decided via the sampling of a connectivity kernel $W : \mathcal{X}^2 \to [0, 1]$ at the random latent positions. This encompasses models like stochastic block models (Lei and Rinaldo, 2015) (SBM) or graphon models (Lovász, 2012), depending on how exactly we define the edge appearance procedure.

The key idea in studying GNNs on large random graphs is to embed the discrete problem into a continuous setting for which we expect to understand their properties with more ease. For instance, some authors derived new properties of geometric stability on large random graphs (Keriven et al., 2020; Levie et al., 2021; Ruiz et al., 2021), which are known to be key in other deep models (Mallat, 2012) but cannot easily be characterized on fixed deterministic graphs. Large random graphs also shed light on the expressive power of GNNs (Keriven et al., 2021) in a different manner than the discrete WL test, and some models are indeed proved to be more powerful than others (Keriven and Vaiter, 2023). To study GNNs on large random graphs, we match the GNN to a "continuous" counterpart, referred to as a continuous-GNN (cGNN) (Keriven et al., 2020; Ruiz et al., 2020). While the discrete GNN propagates a signal over the nodes of the graph, the cGNN propagates *a mapping over the latent space* $\mathcal{X}$. Such a map can be interpreted as a signal over the graph where the "continuum" of nodes would be all the points of $\mathcal{X}$. Then, as the random graph grows large, the GNN must behave similarly to its cGNN counterpart. To justify this, it is necessary to describe the cGNN as a limit of GNNs on random graphs and to ensure that the GNN converges to the cGNN as the number of nodes increases (Keriven et al., 2020; Maskey et al., 2022). This convergence problem is precisely the focus of the present work.

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 (Defferrard et al., 2016) (SGNNs), in which one introduces a graph Fourier transform through a chosen graph shift operator (Tremblay et al., 2018) 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) (Gilmer et al., 2017; Kipf and Welling, 2017; Corso et al., 2020). 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 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.

Contributions. In this paper, we study the convergence toward a continuous counterpart of MPGNNs with a **generic aggregation function** (Corso et al., 2020), whereas previous work (Keriven et al., 2020, 2021; Ruiz et al., 2021; Maskey et al., 2022) are restricted to SGNNs or MPGNNs with specific aggregations. We use a simple version of the Latent Space Model where random graphs are totally connected and weighted accordingly the sampling of the kernel W at the latent positions. Our main result, Theorem 15, states that for MPGNNs having a Lipschitz-type regularity, the discrete network on a large random graph is close to its continuous counterpart with high probability. We quantify this convergence via a non-asymptotic bound based on the 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* (Fey and Lenssen, 2019). For that particular case, Theorem 15 does not hold. Thus, we provide another proof of convergence based on a specific concentration inequality, in Theorem 22. This results in a significantly different theoretical convergence rate.

Related work. The closest related works to ours are the results from Keriven et al. (2020, 2021), where they establish convergence of SGNNs on Latent Position random graphs. We also mention Maskey et al. (2022) who study a particular case of MPGNN on large random graphs, where the aggregation is defined to be a mean normalized by the degree of the node, which is akin to an SGNN using the degree-normalized Laplacian matrix. The present paper can be considered as a direct extension of both these works in the setting of MPGNNs with generic aggregation.

Further, the concept of limit of a SGNN on large random graphs has shown fruitful to tackle several problems. For instance, multiple works from different authors, among which Keriven et al. (2020); Maskey et al. (2023); Levie et al. (2021); Ruiz et al. (2021); Cerviño et al. (2023), have focused on stability to deformation or transferability. The idea is that, since the same GNN can be applied to any graph, no matter its size or structure, we expect the outputs to be close on similar graphs, which is particularly relevant for large random graphs drawn from the same (or almost the same) model. Concerning the expressive power on large graphs, Keriven et al. (2020, 2021) exploit their convergence theorems to propose a description of the function space that SGNNs on random graphs can approximate in Keriven and Vaiter (2023) and derive certain properties of universality. About other topics related to the learning procedure such as generalization as well as oversmoothing, the authors in Maskey et al. (2022) derive a generalization bound that gets tighter for large graphs, while the results described in Keriven (2022) make use of Latent Position random graphs to search a threshold between beneficial finite smoothing and oversmoothing.

Beyond random graphs, large (dense) graphs can be described through the theory of graphons (Lovász, 2012), and several works aim to characterize the convergence of GNNs on large graphs with these mathematical tools. In Ruiz et al. (2020); Maskey et al. (2023), the authors define limits of graph polynomial filters of SGNNs designed from graph shift operators as integral operators with regard to the underlying graphon, and make use of the theory of self-adjoint operators and Hilbert spaces to study them. More recently, authors in Böker et al. (2024) consider a continuous version of the WL test via graphon estimation to study expressive power and the paper Levie (2023) is devoted to extending the concept of sampling graphs from graphon to sampling graph signals from graphon signals.

Outline. In Section 2, we give some basic definitions. In Section 3 we define MPGNNs with a generic aggregation function, that is, any function on sets used to gather and combine neighborhood information in the message-passing paradigm. In Section 4 we introduce continuous-MPGNNs (cMPGNNs) which are the counterparts of discrete MPGNNs that propagate a function over a compact probability space, alongside a connectivity kernel. As a discrete MPGNN must be coherent with graph isomorphism, we give mild conditions under which the cMPGNN is coherent with respect to some notion of 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 15: 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. Along the paper, we illustrate our concepts on classical GNN examples from the basic Graph Convolutional Network to the more sophisticated Graph Attentional Network (Veličković et al., 2017). We give a particular treatment to the case of *maximum* aggregation. Indeed, its behavior turns out to be significantly different than the other examples and does not fit into the class of MPGNNs having sharp bounded differences. Nevertheless, in Theorem 22 we make use of other specific concentration bounds to prove another non asymptotic bound between max MPGNN and its limit cMPGNN, with a significantly different convergence rate.

2. Notations and Definitions

We start by introducing the notations that will hold throughout the paper. The letter d (and its derivatives $d_0, \ldots$) will represent the dimension of a real vector space, the letter n will denote the number of nodes in a graph, and L will refer to the total number of layers in a deep architecture. Whenever we need to index something relatively to the vertices of a graph, we use a subscript indexation (*e.g.,* z_i) and in the case of layers, we employ a parenthesized 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 $\|x\|_\infty \stackrel{\text{def.}}{=} \max_i |x_i|$ 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 $\mathbb{B}(\mathcal{X})$ 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 $\{1, \dots, n\}$ is denoted as $\mathfrak{S}_n$. If $x = (x_1, \dots, x_n)$ is an n-tuple and σ an element of $\mathfrak{S}_n$, we define the n-tuple $\sigma \cdot x$ as $\sigma \cdot x \stackrel{\text{def.}}{=} (x_{\sigma^{-1}(1)}, \dots, x_{\sigma^{-1}(n)})$.

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 it 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 $(\mathcal{X}, \mathbb{B}(\mathcal{X}))$. 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 $\phi_\# P(A) \stackrel{\text{def.}}{=} P(\phi^{-1}(A))$ for all A in $\mathbb{B}(\mathcal{X})$. Since this makes the group $\mathrm{Aut}(\mathcal{X})$ acting on the set of probability measures on $\mathcal{X}$, we also use the notation $\phi \cdot P \stackrel{\text{def.}}{=} \phi_\# P$, which is standard for a (left) group action. For the same reason, we shall use the notations $\phi \cdot f \stackrel{\text{def.}}{=} f \circ \phi^{-1}$ and $\phi \cdot W \stackrel{\text{def.}}{=} W(\phi^{-1}(\cdot), \phi^{-1}(\cdot))$ 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 $\mathrm{L}^\infty_P(\mathcal{X}, \mathbb{R}^p)$ is the space of essentially bounded (equivalence classes of) maps from $\mathcal{X}$ to $\mathbb{R}^p$ endowed with the norm $\|f\|_{P,\infty} \stackrel{\text{def.}}{=} \operatorname{ess\,sup}_{x \in \mathcal{X},\, P} \|f(x)\|_\infty$. 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 $\mathrm{L}^\infty_P(\mathcal{X}, \mathbb{R}^p)$.

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 $\{\!\{\cdot\}\!\}$. We define the sampling operator in the following way. For any $f : \mathcal{E}_0 \to \mathcal{E}_1$ and $X = (x_1, \dots, x_n) \in \mathcal{E}_0^n$, the sampling of f at X, denoted as $\iota_X f$, is defined as

$$\iota_X f \stackrel{\text{def.}}{=} (f(x_i), \dots, f(x_n)) \in \mathcal{E}_1^n\,. \tag{1}$$

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 is defined by a triplet (V, E, w), corresponding to a set of *nodes* (or *vertices*) V, a set of *edges* E, and a *weight function* $w : V^2 \to \mathbb{R}^+$. The cardinality $|V| = n$ is the *size* of the graph. Formally, the set $V = \{v_1, \dots, v_n\}$ may contain arbitrary elements with an arbitrary numbering. However, for simplicity of the presentation, whenever the exact nature of the vertices does not matter,[1] we assume that $V = \{1, \dots, n\}$. The set of edges E contains 2-element subsets of V. The set of neighbors of a vertex i in G is referred to as $\mathcal{N}_G(i)$ or simply $\mathcal{N}(i)$ when the underlying graph is clear from the context. The weight function w assigns a nonnegative number to each edge. It is represented by a

1. It will matter later when considering random graphs whose nodes are latent positions in a metric space.

symmetric function $w : V^2 \to \mathbb{R}^+$ and the abbreviation $w_{i,j}$ is used to denote the weight $w(i,j) = w(j,i)$ where $\{i,j\} \in E$. In this paper, "graph" will always mean "undirected and weighted graph".

Graph signal. A *graph signal* is a map from the set of vertices V of a graph to a d-dimensional vector space $\mathbb{R}^d$. In other words, it is the data of n vectors in $\mathbb{R}^d$, one assigned to each node. Hence, we represent a graph signal by a n times d matrix $Z \in \mathbb{R}^{n\times d}$.

Graph isomorphism. Two graphs $G_1 = (V, E_1, w_1)$ and $G_2 = (V, E_2, w_2)$ are said to be isomorphic if there is a permutation $\sigma \in \mathfrak{S}_n$ such that $E_2 = \left\{\{i,j\} \,\middle|\, \{\sigma^{-1}(i), \sigma^{-1}(j)\} \in E_1\right\}$ and $w_2(i,j) = w_1\left(\sigma^{-1}(i), \sigma^{-1}(j)\right)$. In this case, we note $G_2 = \sigma \cdot G_1$. Moreover, if Z is a signal on G_1 and $\sigma \in \mathfrak{S}_n$, then $\sigma \cdot Z$ is an isomorphic signal on the graph $\sigma \cdot G_1$.

2.2 Random Graph Models

In this subsection, we define our random graph model of interest, as well as the concept of random graph model isomorphism.

Random graph model and random graphs. A *random graph model* is a couple (W, P) where P is a Borel probability measure on the compact set $\mathcal{X} \subset \mathbb{R}^d$ and $W : \mathcal{X} \times \mathcal{X} \mapsto [0,1]$ is a symmetric measurable function called *connectivity kernel*. A random graph model is used to generate random graphs 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 then fully connected and has weight function W:

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

Often, W is decreasing with the distance between the X_i, such that nodes with similar latent variables have stronger connections; classical examples include the Gaussian kernel $W(X,X') = e^{-\frac{\|X-X'\|^2}{2\sigma^2}}$ or the so-called ε-graphs $W(X,X') = 1\!\!1_{\|X-X'\|\leqslant\varepsilon}$. 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 *but* edges may have a weight equal to zero (e.g., for ε-graphs). Also note that, in the rest of the paper, we generally do not put any assumptions on the model (W,P), except when using max-aggregation (Exemples 5-e and Sec. 5.2.2), where we assume that W is Lipschitz-continuous in addition to some conditions on the probability space $(\mathcal{X}, P)$ (see Definition 19).

Another common model (Keriven et al., 2020; Lei and Rinaldo, 2015) is to add a Bernoulli distribution on edges, similarly to SBM models, in order to model unweighted random graphs, potentially with prescribed expected sparsity. It is not done here for the sake of simplicity. In the literature, weighted random graphs without Bernoulli edges are routinely used to analyze machine learning algorithms (Von Luxburg et al., 2008; Maskey et al., 2022), as they essentially model the underlying phenomena of interest in many cases.

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

A multilayer Message-Passing Neural Network (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. We synthesize these three operations into a single one for the sake of readability. Broadly speaking, an MPGNN can be defined as a collection of L applications $(F^{(l)})_{1 \leqslant l \leqslant L}$ that act as follows. Let G 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:

$$z_i^{(l+1)} \stackrel{\text{def.}}{=} F^{(l+1)} \left(z_i^{(l)}, \left\{\!\!\left\{ \left(z_j^{(l)}, w_{i,j} \right) \;\middle|\; j \in \mathcal{N}(i) \right\}\!\!\right\} \right) \in \mathbb{R}^{d_{l+1}} . \tag{2}$$

So $Z^{(l+1)}$ is a matrix in $\mathbb{R}^{n \times d_{l+1}}$. In Equation (2), $F^{(l)}$ takes 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 running node, along with the corresponding edge weight. In the literature, the $F^{(l)}$ are often referred to as the *aggregations* (Jegelka, 2022). Their main property is to ignore the order in which the neighborhood information is collected, through 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. Following the literature, 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 $\overline{\Theta}_G(Z)$ in the second case, where $\overline{\Theta}_G$ use an additional pooling operation over the nodes, $R : \mathbb{R}^{n \times d_L} \to \mathbb{R}^{d_L}$, called the *readout* (Jegelka, 2022) function:

$$\Theta_G(Z) \stackrel{\text{def.}}{=} Z^{(L)} \in \mathbb{R}^{n \times d_L}, \quad \text{and} \quad \overline{\Theta}_G(Z) \stackrel{\text{def.}}{=} R\left(\left\{\!\!\left\{ z_1^{(L)}, \dots, z_n^{(L)} \right\}\!\!\right\} \right) \in \mathbb{R}^{d_L} . \tag{3}$$

We know that a fundamental property of GNNs is that they are 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 leave the output unchanged in the *invariant* case. This is stated the proposition below.

Proposition 1 (Invariance and equivariance of MPGNNs)**.** *With the definition of the message-passing from Equations* (2) *and* (3)*,* Θ *and* $\overline{\Theta}$ *are respectively* $\mathfrak{S}_n$*-equivariant and* $\mathfrak{S}_n$*-invariant, in the sense that for all* $\sigma \in \mathfrak{S}_n$*, 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* $\overline{\Theta}_{\sigma \cdot G}(\sigma \cdot Z) = \overline{\Theta}_G(Z)$*.*

Proof We prove the equivariant case. Let us introduce the layer functions $\Lambda_G^{(l)} : Z^{(l-1)} \mapsto Z^{(l)}$, 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 $\sigma \cdot G$, obtained from the message-passing of $\sigma \cdot Z$ with regard to the graph $\sigma \cdot G$ and the weight function $\sigma \cdot w$. Locally, for all i, y_i is the result of the message-passing at the node labeled i in $\sigma \cdot G$. Recalling that the ith row of

$\sigma \cdot Z$ is $z_{\sigma^{-1}(i)}$, and that the (i,j)-entry of $\sigma \cdot w$ is $w_{\sigma^{-1}(i),\sigma^{-1}(j)}$, we get that,

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

Then, we do the change of variable $j = \sigma(j')$ in the multiset indexing, we obtain

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

Now, by definition of $\sigma \cdot G$, we have the equivalence $\sigma(j) \in \mathcal{N}_{\sigma\cdot G}(i)$ if and only if $j \in \mathcal{N}_G(\sigma^{-1}(i))$. This implies that the multiset $\{\!\!\{\left(z_j, w_{\sigma^{-1}(i),j}\right) \mid \sigma(j) \in \mathcal{N}_{\sigma\cdot G}(i)\}\!\!\}$ is the same as the multiset $\{\!\!\{\left(z_j, w_{\sigma^{-1}(i),j}\right) \mid j \in \mathcal{N}_G\left(\sigma^{-1}(i)\right)\}\!\!\}$. Thus,

$$y_i = F^{(l)}\left(z_{\sigma^{-1}(i)}, \{\!\!\{\left(z_j, w_{\sigma^{-1}(i),j}\right) \mid j \in \mathcal{N}_G\left(\sigma^{-1}(i)\right)\}\!\!\}\right).$$

On the other hand, $\sigma \cdot \Lambda_G^{(l)}(Z) = Y'$ is the signal $\Lambda_G^{(l)}(Z)$ to which the rows have been permuted a posteriori, so

$$y'_i = F^{(l)}\left(z_{\sigma^{-1}(i)}, \{\!\!\{\left(z_j, w_{\sigma^{-1}(i),j}\right) \mid j \in \mathcal{N}_G(\sigma^{-1}(i))\}\!\!\}\right).$$

Hence $Y = Y'$, which means that $\Lambda_{\sigma\cdot G}^{(l)}(\sigma \cdot Z) = \sigma \cdot \Lambda_G^{(l)}(Z)$, and that $\Lambda^{(l)}$ 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 $\mathfrak{S}_n$-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 Equation (2) is crucial and there is a wide range of designs for them (Wu et al., 2021). Nevertheless, they mostly take the following "message-then-combine" form. At layer $l+1$, the signals of the neighbors of a node are transformed by a learnable message operation $m^{(l+1)}$. Then these *messages* $m^{(l+1)}(z_j^{(l)})$ are *aggregated* along with some optional *weight coefficients*, whose expressions are very general here,

$$c_{i,j}^{(l+1)} = c^{(l+1)}\left(z_i^{(l)}, z_j^{(l)}, w_{i,j}\right), \tag{4}$$

in a way that is invariant to node relabeling. It appears that a natural way of doing the aggregation step is to perform a *mean*, in a broad sense: an arithmetic mean, a weighted mean, a maximum, and so on (Bullen, 2013). Thus, we have a *mean* operator $M^{(l+1)}$ such that Equation (2) is expressed as

$$F^{(l+1)}\left(z_i^{(l)}, \left\{\!\!\left\{\left(z_j^{(l)}, w_{i,j}\right) \,\middle|\, j \in \mathcal{N}(i)\right\}\!\!\right\}\right) = M^{(l+1)}\left(\left\{\!\!\left\{\left(m^{(l+1)}(z_j^{(l)}), c_{i,j}^{(l+1)}\right) \,\middle|\, j \in \mathcal{N}(i)\right\}\!\!\right\}\right). \tag{5}$$

To our knowledge, Equation (5) encompasses most of the existing popular MPGNN architectures of the literature. We note that it is essentially a more verbose reformulation of Equation (2), the two different expressions mostly provide a different level of intuition on the message-passing process. In the sequel, we discuss five examples that follow Equation (5). For each example, we also give the corresponding readout function that will be used in our results, for the invariant case.

Example 1 (Convolutional Message-Passing)**.** *The $c_{i,j}$ are the graph weights $w_{i,j}$ (Kipf and Welling, 2017; Defferrard et al., 2016; Gilmer et al., 2017). Each neighbor representation is multiplied by its corresponding weight and we combine them with an arithmetic mean. Notice that this is equivalent to a SGNN with polynomial filters of degree one.*

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

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 (Maskey et al., 2022).*

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

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

Example 3 (Attention based Message-Passing)**.** *Unlike the two examples above, the* attention *coefficients are learnable and depend on all the possible parameters mentioned in Equation* (4) *(Veličković et al., 2017). A weighted mean is then used.*

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

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

Example 4 (Generalized mean)**.** *Similar to Example 1, but with additional functions used to compute "generalized" or semi-arithmetic means (Kortvelesy et al., 2023; de Carvalho, 2016):*

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

for some invertible function h (possibly defined on a bounded domain). In this case, following the popular formalism from Kolmogorov and Castelnuovo (1930), in order to legitimately be considered as a generalized mean, some regularity conditions are required on h, typically, h must be a strictly increasing continuous function. For instance, taking $h^{-1} = x \mapsto \ln(x)$ (assuming positivity of the inputs) and $h = x \mapsto e^x$ yields the geometric mean, while taking $h^{-1} = x \mapsto x^p$ and $h = x \mapsto x^{1/p}$ (again, assuming nonnegativity of the message if necessary) yields a power mean as employed in the moment-based aggregations from Corso et al. (2020) (up to centering, which we omit for simplicity).

In the invariant case, the readout function is again an arithmetic mean (for simplicity).

Example 5 (Max Convolutional Message-Passing)**.** *The aggregation maximum is often mentioned as a possibility in the literature (Hamilton et al., 2017), but we note that it is less common in practice. Here the $c_{i,j}$ are also the graph weights $w_{i,j}$ but a* coordinate-wise *maximum is used to combine the messages:*

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

In the invariant case, the readout function is a coordinate-wise maximum.

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

Notice that, assuming positivity of the inputs, this aggregation is also obtained as the limit, for $p \to \infty$ of the generalized power mean from Example 4 with $h(x) = x^{1/p}$.

4. Continuous Message-Passing GNNs on Random Graph Models

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

$$f^{(l+1)}(x) \stackrel{\text{def.}}{=} \mathcal{F}_P^{(l+1)} \left(f^{(l)}(x), \left(f^{(l)}, W(x, \cdot) \right) \right) \in \mathbb{R}^{d_{l+1}} , \tag{6}$$

for all $x \in \mathcal{X}$. 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" $\left(f^{(l)}, W(x, \cdot) \right)$. The latter is a short notation for the map $y \mapsto \left(f^{(l)}(y), W(x,y) \right)$ at x fixed, which is the continuum equivalent of the multiset of pairs of weighted neighbors $\left\{\!\!\left\{ (z_j^{(l)}, w_{i,j}) \;\middle|\; j \in \mathcal{N}(i) \right\}\!\!\right\}$ from Equation (2).

For conciseness, we will often overload Equation (6) with the short notation:

$$f^{(l+1)} \stackrel{\text{def.}}{=} \mathcal{F}_P^{(l+1)} \left(f^{(l)}, W \right) . \tag{7}$$

We denote $\Theta_{W,P}(f)$ the output in the equivariant case and $\overline{\Theta}_{W,P}(f)$ in the invariant case:

$$\Theta_{W,P}(f) \stackrel{\text{def.}}{=} f^{(L)} \in \mathrm{L}_P^\infty(\mathcal{X}, \mathbb{R}^{d_L}), \quad \text{and} \quad \overline{\Theta}_{W,P}(f) \stackrel{\text{def.}}{=} \mathcal{R}_P(\Theta_{W,P}(f)) \in \mathbb{R}^{d_L} , \tag{8}$$

where $\overline{\Theta}_{W,P}$ involves an additional continuum readout operator

$$\mathcal{R} : \mathcal{P}(\mathcal{X}) \times \mathrm{L}_P^\infty(\mathcal{X}, \mathbb{R}^{d_L}) \to \mathbb{R}^{d_L} .$$

Naturally, we also demand the equivariant and invariant versions of the cMPGNN to respectively be equivariant and invariant to random graph model isomorphism. To that extent, we impose the following assumption on the operators $\mathcal{F}^{(l)}$ and on $\mathcal{R}$.

Assumption 2. *There is a subgroup $H \subset \mathrm{Aut}(\mathcal{X})$ such that $\forall\, 1 \leqslant l \leqslant L$, $\forall f \in \mathrm{L}^\infty_P(\mathcal{X}, \mathbb{R}^{d_l})$, $\forall \phi \in H$,*

$$\mathcal{F}^{(l)}_{\phi \cdot P}(\phi \cdot f, \phi \cdot W) = \phi \cdot \mathcal{F}^{(l)}_P(f, W)\,,$$

and,

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

Assumption 2 is largely inspired by the classical change of variable formula by push forward measure in Lebesgue integration. This formula states that for any $\phi \in \mathrm{Aut}(\mathcal{X})$ and any measurable map f,

$$\int f \mathrm{d}P = \int \phi \cdot f \mathrm{d}(\phi \cdot P)\,. \tag{9}$$

It is not difficult to verify that if, for example, $\mathcal{F}_P(f, W) = \int f(y)W(x, y)\mathrm{d}P(y)$ and $\mathcal{R}_P(f) = \int f \mathrm{d}P$, then Equation (9) implies Assumption 2 with $H = \mathrm{Aut}(\mathcal{X})$.

Contrary to the discrete case, where the symmetry is valid for the full group $\mathfrak{S}_n$, we require here a symmetry for a subgroup of $\mathrm{Aut}(\mathcal{X})$ only. Ideally, one would like Assumption 2 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 (Examples 5 and e in the next section) will highlight the fact that, for a matter of *existence* of such a limit, one may have to impose some conditions on P, and thus restrict to a subgroup of $\mathrm{Aut}(\mathcal{X})$.

Proposition 3 (Invariance and equivariance of cMPGNNs)**.** *Let (W, P) be a random graph model on $\mathcal{X}$. Then, under Assumption 2, Θ and $\overline{\Theta}$ are respectively H-equivariant and H-invariant. In other words, for any f, for any $\phi \in H$,*

$$\Theta_{\phi \cdot (W,P)}(\phi \cdot f) = \phi \cdot \Theta_{W,P}(f)\,,$$

and,

$$\overline{\Theta}_{\phi \cdot (W,P)}(\phi \cdot f) = \overline{\Theta}_{W,P}(f)\,.$$

Proof We start with the equivariant case, the invariant one will follow immediately by composition with $\mathcal{R}$. Let the $\mathcal{L}^{(l)}_{W,P} \colon \mathbb{R}^{n \times d_{l-1}} \to \mathbb{R}^{n \times d_l}$ be the layer operators such that $\Theta^{(L)}_{W,P} = \mathcal{L}^{(L)}_{W,P} \circ \cdots \circ \mathcal{L}^{(1)}_{W,P}$. Let $f \in \mathrm{L}^\infty_P(\mathcal{X}, \mathbb{R}^{d_{l-1}})$, using Assumption 2 we obtain

$$\phi \cdot \mathcal{L}^{(l)}_{W,P}(f) = \phi \cdot \mathcal{F}^{(l)}_P\,(f, W) = \mathcal{F}^{(l)}_{\phi \cdot P}\,(\phi \cdot f, \phi \cdot W) = \mathcal{L}^{(L)}_{\phi \cdot (W,P)}(\phi \cdot f)\,.$$

So the Proposition is true on all the $\mathcal{L}^{(l)}_{W,P}$, thus also true on $\Theta^{(L)}_{W,P}$ by composition. For the invariant case, it is clear from Assumption 2 that $\mathcal{R}$ is H-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 intuitive connection to the previous Examples 1 to 5 In the next section, we will precisely see in what sense Examples 1 to 5, when applied on random graphs and as n grows large, converge to the following cMPGNNs.

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

$$f^{(l+1)}(x) = \int_{y \in \mathcal{X}} W(x,y) m^{(l+1)} \left(f^{(l)}(y) \right) \mathrm{d}P(y)$$

and, in the invariant case, the continuous readout is

$$\mathcal{R}_P \left(f^{(L)} \right) = \int_{\mathcal{X}} f^{(L)} \mathrm{d}P \,.$$

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

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

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

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

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

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

Example d (Generalized mean). *We simply add the function h to the mean example*

$$f^{(l+1)}(x) = h \left(\int_{y \in \mathcal{X}} h^{-1} \left(W(x,y) m^{(l+1)} \left(f^{(l)}(y) \right) \right) \mathrm{d}P(y) \right) \,.$$

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

Example e (Max Convolutional Message-Passing). *The maximum becomes a coordinate-wise essential supremum according to the probability measure P (that is, a supremum P-almost everywhere)*

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

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

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

Remark that we use the P-essential supremum and not the classical supremum here, since this will represent the limit maximum of variables that are P-distributed. Nevertheless, if the measure P is strictly positive, notice that the essential supremum of a continuous function is nothing more than the usual supremum, see Lemma 27 in Appendix D

Remark 4. *It can be easily verified that for all these examples, the underlying $\mathcal{F}^{(l)}$ functions satisfy Assumption 2 with $H = \mathrm{Aut}(\mathcal{X})$. For the integral, it is ensured by the classical change of variable formula Equation* (9).

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

$$\operatorname*{ess\,sup}_P g \circ \phi = \inf\{M \mid P\left(g \circ \phi > M\right) = 0\}\,,$$

where $(g \circ \phi > M)$ is the probabilistic notation for the set $\{x \mid g \circ \phi(x) > M\}$. Hence, since by the bijectivity of ϕ,

$$\{x \mid g \circ \phi(x) > M\} = \left\{\phi^{-1}(y) \,\middle|\, g(y) > M\right\},$$

one has $(g \circ \phi > M) = \phi^{-1}(g > M)$, which finally yields

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

5. Continuous Message-Passing GNNs as Limits of Discrete Message-Passing GNNs 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 an 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. Furthermore, we provide some non-asymptotic bounds to control the deviation between an MPGNN and its limit cMPGNN with high probability.

This section is divided in two parts. In the first part (Section 5.1), given an 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: 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 to d are indeed the continuous counterparts of Examples 1 to 4 according to our definition.

In the second part of this section (Section 5.2), we study the convergence of MPGNNs towards their continuous counterpart, when it exists: we give sufficient conditions for this convergence to occur and provide convergence rates in the form of non-asymptotic bounds with high probability. Our first main result, Theorem 15 in Section 5.2.1, concerns a class of MPGNNs 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. As we will see, this applies to all examples *but* max aggregation. Our second result, Theorem 22 in Section 5.2.2, is specific to the case of maximum aggregation, as in this case the bounded difference property is not verified and Theorem 15 is not applicable. It is based on another concentration inequality and leads to a convergence rate with a dependence on the input dimension d (recall $\mathcal{X} \subset \mathbb{R}^d$), contrary to the bounded differences method.

5.1 Definition of the Limit Message-Passing GNN

Let (W, P) be a random graph model and $f \in \mathrm{L}^\infty_P(\mathcal{X}, \mathbb{R}^d)$. The main purpose of this subsection is to *define* the natural limit of discrete MPGNNs to cMPGNNs, before examining more precisely the rate of convergence from one to the other in the next subsection.

To this end, we consider a single-layer MPGNN applied on a random graph $G_n \sim \mathcal{G}_n(W, P)$ and input node features $Z = \iota_X f$ as a sampling of some function (recall the definition of the sampling operator in Equation (1)). We define a corresponding 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. The update of $f(x)$ is given by

$$\mathcal{F}_P(f, W)(x) = \int_{y \in \mathcal{X}} W(x, y) m\left(f(y)\right) \mathrm{d}P(y)\,. \tag{10}$$

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

$$\frac{1}{n} \sum_{i=1}^{n} W(x, X_i) m(f(X_i))\,, \tag{11}$$

for $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P$. Coincidentally, Equation (11) is exactly the discrete message-passing of Example 1 around node x of a random graph with node latent variables $\{x, X_1, \dots, X_n\}$, signal $Z = \{f(x), f(X_1), \dots, f(X_n)\}$ and kernel W. We have thus naturally obtained the cMPGNN from Example a via a limit of the MPGNN of Example 1 on random graphs.

Back to the general case, given an abstract discrete (single-layer) MPGNN with aggregation F, we want to define a cMPGNN from its limit on random graphs. Following the path of the above example, we could look at the almost sure limit of the message-passing equation

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

However, the existence of this limit is far from obvious in the general case. As we will see in the next definition, we will rather relax it to the convergence of the expectation instead:

$$\mathbb{E}_{X_1, \dots, X_n}\left[F\left(f(x), \{\!\{(f(X_i), W(x, X_i))\}\!\}_{1 \leqslant i \leqslant n}\right)\right]\,. \tag{12}$$

Up to some details related to isomorphism invariance, the existence of this limit is how we *define* the continuous counterpart of an MPGNN.

Definition 5 (Continuous counterpart)**.** *Let F be an MPGNN layer. For any $f \in \mathrm{L}^\infty_P(\mathcal{X}, \mathbb{R}^d)$, $W : \mathcal{X}^2 \to [0, 1]$ and $P \in \mathcal{P}(\mathcal{X})$, define the sequence of functions in $\mathrm{L}^\infty_P(\mathcal{X}, \mathbb{R}^{d'})$ by*

$$F_{P,n}(f, W) : x \mapsto \mathbb{E}\left[F\left(f(x), \{\!\{(g(X_i), W(x, X_i))\}\!\}_{2 \leqslant i \leqslant n}\right)\right]\,. \tag{13}$$

where the expected value is taken over all the $X_2, \dots, X_n \overset{\text{i.i.d.}}{\sim} P$.

Let $\mathcal{F}$ be an operator of the form Equation (7) taking value in $\mathrm{L}^\infty_P(\mathcal{X}, \mathbb{R}^{d'})$, and suppose that there exists H, a non-trivial subgroup of $\mathrm{Aut}(\mathcal{X})$, such that for any $f \in \mathrm{L}^\infty_P(\mathcal{X}, \mathbb{R}^d)$ and any $\phi \in H$, the operator $\mathcal{F}_{\phi \cdot P}\left(\phi \cdot f, \phi \cdot W\right)$ arises as the limit

$$\mathcal{F}_{\phi \cdot P}\left(\phi \cdot f, \phi \cdot W\right) = \lim_{n \to \infty} F_{\phi \cdot P, n}\left(\phi \cdot f, \phi \cdot W\right)\,, \tag{14}$$

*where the convergence occurs with respect to the norm of the space $\mathrm{L}^\infty_{\phi \cdot P}(\mathcal{X}, \mathbb{R}^d)$. 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 we consider only $n-1$ random variables for convenience with later definitions (in an n-sized graph, each node has potentially $n-1$ neighbors). In this definition of limit GNNs, we have *not* explicitly assumed that the $\mathcal{F}$ defined above satisfies Assumption 2 related to continuous isomorphisms and cMPGNN. Fortunately, this assumption is in fact automatically verified, for the natural subgroup H involved in the definition. In other words, every continuous counterpart of MPGNN as defined above is indeed a valid cMPGNN.

Proposition 6. *Let $\mathcal{F}$ be the continuous counterpart of F for H as defined in Definition 5. Then it satisfies Assumption 2 for any $\phi \in H$.*

Proof Let $f \in \mathrm{L}^\infty_P(\mathcal{X}, \mathbb{R}^d)$, $\phi \in H$, and $X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$, we have for P-almost all x,

$$\begin{aligned}
\phi \cdot F_{P,n}(f, W)(x) &= \mathbb{E}_{X_2,\ldots,X_n}\left[F\left(f(\phi^{-1}(x)), \{\!\!\{\left(f(X_i), W(\phi^{-1}(x), X_i)\right)\}\!\!\}_{2 \leqslant i \leqslant n}\right)\right] \\
&= \mathbb{E}_{Y_2,\ldots,Y_n}\left[F\left(\phi \cdot f(x), \{\!\!\{(\phi \cdot f(Y_i), \phi \cdot W(x, Y_i))\}\!\!\}_{2 \leqslant i \leqslant n}\right)\right] \qquad (15) \\
&= F_{\phi \cdot P,n}\left(\phi \cdot f, \phi \cdot W\right)(x)\,.
\end{aligned}$$

Where in Equation (15), we have set $Y_i = \phi(X_i)$ for all i, which are identically distributed random variable with common law $\phi \cdot P$, and used the classical change of variable formula Equation (9). Thus, by taking the limit, Equation (14) implies

$$\mathcal{F}_{\phi \cdot P}(\phi \cdot f, \phi \cdot W) = \phi \cdot \mathcal{F}_P(f, W)\,,$$

which is the desired result. ■

The same definitions and propositions as above can also be given for a readout final layer.

Definition 7. *Let R be an MPGNN readout layer and $P \in \mathcal{P}(\mathcal{X})$. For $f \in \mathrm{L}^\infty_P(\mathcal{X}, \mathbb{R}^d)$, we define the sequence of functions*

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

where the expected value is taken over all the $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$.

Let $\mathcal{R}$ be a continuum readout operator of the form Equation (8) taking values in $\mathbb{R}^{d'}$. Suppose we have H a non-trivial subgroup of $\mathrm{Aut}(\mathcal{X})$ such that for any $f \in \mathrm{L}^\infty_P(\mathcal{X}, \mathbb{R}^d)$, 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.*

Naturally, Definition 7 also implies invariance with respect to the proper subgroup.

Proposition 8. *Let $\mathcal{R}$ be the continuous counterpart of R as in Definition 7. Then it satisfies Assumption 2 for any $\phi \in H$.*

Going back to our four examples of Sections 3 and 4, we now show that Examples a to d are the continuous counterparts of Examples 1 to 4 for the full $\mathrm{Aut}(\mathcal{X})$. We will assume some mild positivity conditions for the coefficients in the degree normalized and the GAT examples. These are satisfied for most instances of kernels (*e.g.,* Gaussian) or attention coefficients (*e.g.,* the typical exponentials of LeakyReLUs from Veličković et al. (2017)), as the data live in a bounded domain $\mathcal{X}$. For the generalized mean of Examples 4 and d, we will rely on an Hölder type regularity assumption.

Examples 5 and e are however more involved, as one has to be careful with the shape of $\mathcal{X}$ and the properties of P to avoid null set issues at the boundary of $\mathcal{X}$. We will show that if $\mathcal{X}$ contains no nonempty open null set, and if W and f are continuous, then Example e is the continuous counterpart of Example 5 for the subgroup H of $\mathrm{Aut}(\mathcal{X})$ consisting of all the *homeomorphisms* from $\mathcal{X}$ into itself.

The proof of the next proposition is given in Appendix C.

Proposition 9 (Continuous counterpart of Examples)**.** *The following holds:*

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

- ***Examples 2 and b.*** *Suppose that m is bounded and that there is a strictly positive a such that $0 < a \leqslant W$. Then Example b is the continuous counterpart of Example 2 for the full* $\mathrm{Aut}(\mathcal{X})$*.*

- ***Examples 3 and c.*** *Suppose that m is bounded and that there is two positive constants $0 < a < b$ such that $a \leqslant c(f(x), f(y), W(x, y)) \leqslant b$. Then Example c is the continuous counterpart of Example 3 for the full* $\mathrm{Aut}(\mathcal{X})$*.*

- ***Examples 4 and d.*** *Suppose that h^{-1} is bounded and that h is α_h-Hölder on the range of h^{-1} with $0 < \alpha_h \leqslant 1$, that is, $\|h(x) - h(y)\|_\infty \leqslant K_h \|x - y\|_\infty^{\alpha_h}$. Then Example d is the continuous counterpart of Example 4 for the full* $\mathrm{Aut}(\mathcal{X})$*.*

- ***Examples 5 and e.*** *Suppose that W, m, and f are continuous and that the measure P is strictly positive on $\mathcal{X}$,* i.e., *any nonempty relative open of $\mathcal{X}$ has a strictly positive measure by P. Then Example e is the continuous counterpart of Example 5 for* $\mathrm{Hom}(\mathcal{X})$*: the subgroup of* $\mathrm{Aut}(\mathcal{X})$ *made of the $\phi \in \mathrm{Aut}(\mathcal{X})$ that are homeomorphisms.*

5.2 Convergence of Message-Passing GNNs to their Continuous Counterpart

Let (W, P) be a random graph model, and $(G_n)_{n\geqslant 1}$ be a sequence of random graphs drawn from $\mathcal{G}_n(W, P)$. We go back to the multi-layer setup: consider an MPGNN $(F^{(l)})_{1\leqslant l\leqslant L}$, a readout R and assume that their continuous counterparts $(\mathcal{F}^{(l)})_{1\leqslant l\leqslant L}$ and $\mathcal{R}$ in the sense of Definitions 5 and 7 exist. For an $f \in \mathrm{L}^\infty_P(\mathcal{X}, \mathbb{R}^{d_0})$, does the MPGNN on G_n with input signal $\iota_X f$ 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 conditions and with high probability, $\Theta_{G_n}(\iota_X(f))$ is close to $\Theta_{W,P}(f)$ in the equivariant case and that $\overline{\Theta}_{G_n}(\iota_X(f))$ is close to $\overline{\Theta}_{W,P}(f)$ in the invariant case. For the latter, we can compare both outputs directly since they belong to the same vector space. The comparison is however more involved in the equivariant case since $\Theta_{G_n}(\iota_X(f))$ is a matrix of size $n \times d_L$ and $\Theta_{W,P}(f)$ is a function from $\mathcal{X}$ to $\mathbb{R}^{d_L}$. In this case, we choose to sample the function, which results in measuring the deviation with the Maximum Absolute Error (MAE) defined by

$$\mathrm{MAE}_X(Z, f) = \max_{1 \leqslant i \leqslant n} \|z_i - f(X_i)\|_\infty . \tag{16}$$

Applied on an MPGNN and its cMPGNN counterpart, it can be written as

$$\mathrm{MAE}_X\left(\Theta_{G_n}\left(\iota_X f\right), \Theta_{W,P}(f)\right) = \left\| (\iota_X f)_i^{(L)} - \iota_X\left(f^{(L)}\right)_i \right\|_\infty ,$$

which corresponds to the following process. Start with $f \in \mathrm{L}_P^\infty\left(\mathcal{X}, \mathbb{R}^{d_0}\right)$. On the one hand, sample f to get a graph signal on G_n, and then pass it through the discrete MPGNN Θ_{G_n}. On the other hand, first pass f through the continuous counterpart cMPGNN $\Theta_{W,P}$, and then sample the result to get a graph signal on G_n. Compare these two graph signals.

Our first theorem is based on the bounded differences method and the McDiarmid inequality (Corollary 26). It encompasses numerous examples of MPGNNs that include Examples 1 to 4. For Example 5 however, we obtain a different bound based on other concentration inequalities.

5.2.1 The bounded differences method

Let us detail the assumptions that we make on the MPGNN. Our result is based on the so-called McDiarmid inequality (McDiarmid, 1989) (Theorem 24 and Corollary 26), which says that a multivariate function of independent random variable has a sub-Gaussian concentration around its mean, provided that it satisfies the following notion of bounded differences.

Definition 10 (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 $c_1, \ldots, c_n$ such that for any $1 \leqslant i \leqslant n$*

$$\left| f(x_1 \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) \right| \leqslant c_i , \tag{17}$$

for any $x_1, \ldots, x_n, x_i' \in \mathcal{E}$.

In plain terms, whenever one fixes all but one of the components of f, the variations should be bounded. For a fixed $x_1 \in \mathcal{X}$, we are then interested in the bounded differences of the layers of our MPGNN

$$(x_2, \ldots, x_n) \mapsto F^{(l)}\left(f^{(l-1)}(x_1), \left\{\!\!\left\{ \left(f^{(l-1)}(x_i), W(x_1, x_i) \right) \right\}\!\!\right\}_{i \geqslant 2} \right) , \tag{18}$$

as a map of the $n-1$ variables $x_2, \ldots, x_n$. These bounded differences depend on x_1. Remark that, if we call them $c_2(x_1) \ldots, c_n(x_1)$, since Equation (18) is invariant to the permutations of $x_2, \ldots, x_n$, they can be taken all equal $c_2(x_1) = \cdots = c_n(x_1)$. To handle some examples, we allow to take a transform ψ of the layers before computing the bounded difference, which, as we will see, may influence the final rate of convergence. For instance, for Example 4 with $h = (\cdot)^{1/p}$, one should lift back to the power p *before* computing the bounded differences to get the proper rate (*i.e.,* take $\psi = h$ in the assumption below). The precise assumption that we make is the following.

Assumption 11 (Bounded differences of layers)**.** *There is a function ψ, assumed invertible on the proper domain such that for all layers l:*

(i) We have the bounded difference inequality

$$\operatorname*{ess\,sup}_{x_1\in\mathcal{X},\,P}\Big\|\psi^{-1}\left(F^{(l)}\left(f^{(l-1)}(x_1),\{\!\!\{(f^{(l-1)}(x_i),W(x_1,x_i))\}\!\!\}_{i\geqslant 2}\right)\right) \\ -\psi^{-1}\left(F^{(l)}\left(f^{(l-1)}(x_1),\{\!\!\{(f^{(l-1)}(x_i'),W(x_1,x_i'))\}\!\!\}_{i\geqslant 2}\right)\right)\Big\|_\infty \leqslant D_n^{(l)}\,,$$

for all x_i, x_i', $i\geqslant 2$, and $x_i = x_i'$ except for $i=2$, where $D_n^{(l)}$ is some rate.

(ii) We have

$$\operatorname*{ess\,sup}_{x_1\in\mathcal{X},\,P}\Big\|\psi\left(\mathbb{E}\psi^{-1}\left(F^{(l)}\left(f^{(l-1)}(x_1),\{\!\!\{(f^{(l-1)}(X_i),W(x_1,X_i))\}\!\!\}_{i\geqslant 2}\right)\right)\right) \\ -\mathbb{E}F^{(l)}\left(f^{(l-1)}(x_1),\{\!\!\{(f^{(l-1)}(X_i),W(x_1,X_i))\}\!\!\}_{i\geqslant 2}\right)\Big\|_\infty \leqslant \widetilde{D}_n^{(l)}\,,$$

where the expectation is over $X_2,\dots,X_n\sim P$ and $\widetilde{D}_n^{(l)}$ is some rate.

(iii) The function ψ is α_ψ-Hölder on its domain with $0<\alpha_\psi\leqslant 1$:

$$\big\|\psi(y)-\psi(y')\big\|_\infty \leqslant K_\psi\big\|y-y'\big\|_\infty^{\alpha_\psi}\,.$$

In several examples, it will be enough to consider $\psi = \mathrm{id}$, such that the layers themselves satisfy the bounded difference property. In that case, Items (i) and (ii) from Assumption 11 above are automatically satisfied with exponent $\alpha_\psi = 1$. However, we allow for a transform ψ^{-1} before computing the bounded differences, to handle more general cases (namely, the generalized mean Examples 4 and d) through the modified McDiarmid inequality Corollary 26 in Appendix D. The next assumption relates to the existence of the continuous counterpart of the layers as defined by Definition 5. We recall that this is defined by the convergence of the expectation of the layers, here we just denote the rate of convergence by s_n.

Assumption 12 (Continuous counterpart of layers)**.** *For any layer l, $\mathcal{F}^{(l)}$ is the continuous counterpart of $F^{(l)}$, and using the notations of Definition 5, we let $\left(s_n^{(l)}\right)$ be a sequence of positive reals such that*

$$\left\|F_{P,n}^{(l)}\left(f^{(l-1)},W\right)-\mathcal{F}_P^{(l)}\left(f^{(l-1)},W\right)\right\|_\infty \leqslant s_n^{(l)} \to 0. \tag{19}$$

for all n.

Finally, we state a final assumption on the Hölder stability of the layers, which is usually easy to verify.

Assumption 13 (Hölder property of layers)**.** *There exists an exponent $0<\alpha_F\leqslant 1$, and, for any $1\leqslant l\leqslant L$, there exist a constant $K_{F,n}^{(l)}>0$ such that the aggregations $F^{(l)}$ satisfy*

$$\left\|F^{(l)}\left(z_1,\{\!\!\{(z_i,w_i)\}\!\!\}_{i\geqslant 2}\right)-F^{(l)}\left(z_1',\{\!\!\{(z_i',w_i)\}\!\!\}_{i\geqslant 2}\right)\right\|_\infty \leqslant K_{F,n}^{(l)}\max_{1\leqslant i\leqslant n}\big\|z_i-z_i'\big\|_\infty^{\alpha_F}$$

for all $z_i, z_i'\in\mathbb{R}^{d_{l-1}}$ and $w_i\in[0,1]$. Moreover, the $K_{F,n}^{(l)}$ are bounded over n.

Note that the "for all" statement implies that the z_i' can be permuted, and the bound must stay valid. We finish by stating the same assumptions for the readout function in the invariant case.

Assumption 14 (Readout function). *The readout function satisfies the following*

(i) The readout function R has a continuous counterpart $\mathcal{R}$, and we let r_n be such that (recall the definitions of $R_{P,n}$ and $\mathcal{R}_P$ from Definition 7)

$$\left\| R_{P,n}(f^{(L)}) - \mathcal{R}_P(f^{(L)}) \right\|_\infty \leqslant r_n \,. \tag{20}$$

(ii) The readout function has bounded differences(Definition 10). Since the n bounded differences can be taken all equal due to permutation invariance, we call C_n the common bounded difference of

$$(x_1, \ldots, x_n) \mapsto R\left(\{\!\!\{ f^{(L)}(x_1), \ldots, f^{(L)}(x_n) \}\!\!\}\right)$$

at each coordinate.

(iii) There exists $K_{R,n} > 0$ such that

$$\left\| R\left(\{\!\!\{ z_i \}\!\!\}_{i\leqslant 1\leqslant n}\right) - R\left(\{\!\!\{ z_i' \}\!\!\}_{i\leqslant 1\leqslant n}\right) \right\|_\infty \leqslant K_{R,n} \max_{1\leqslant i\leqslant n} \left\| z_i - z_i' \right\|_\infty ,$$

for all $z_i, z_i' \in \mathbb{R}^L$. Moreover, the $K_{R,n}$ are bounded over n.

Again, the "for all" statement implies that the bound remains under any permutation of the z_i'. Note that, for simplicity, we do not involve any "Holder" exponent in the readout function, as all our examples can be treated without. Our main result is then the following.

Theorem 15 (MPGNN convergence towards cMPGNN). *Under Assumptions 11, 12 and 13, for any $0 < \rho \leqslant 1$, the following assertions are verified.*

- ***Equivariant case.*** *With probability at least $1-\rho$, it holds*

$$\begin{aligned} &\mathrm{MAE}_X\left(\Theta_{G_n}(\iota_X(f)), \Theta_{W,P}(f)\right) \\ &\leqslant \sum_{l=1}^{L} A_n^{(l,L)} \left[\left(\frac{1}{2}\left(D_n^{(l)}\right)^2 n \ln\left(\frac{2^{L+2-l} d_l n}{\rho}\right) \right)^{\alpha_\psi/2} + \widetilde{D}_n^{(l)} + s_n^{(l)} \right]^{\alpha_F^{L-l}} , \end{aligned} \tag{21}$$

- ***Invariant case.*** *With additionally Assumption 14, it holds with probability at least $1-\rho$*

$$\begin{aligned} &\left\| \overline{\Theta}_{G_n}(\iota_X(f)) - \overline{\Theta}_{W,P}(f) \right\|_\infty \\ &\leqslant K_{R,n} \sum_{l=1}^{L} A_n^{(l,L)} \left[\left(\frac{1}{2}\left(D_n^{(l)}\right)^2 n \ln\left(\frac{2^{L+3-l} d_l n}{\rho}\right) \right)^{\alpha_\psi/2} + \widetilde{D}_n^{(l)} + s_n^{(l)} \right]^{\alpha_F^{L-l}} \\ &\quad + C_n \sqrt{n \ln\left(\frac{4 d_L}{\rho}\right)} + r_n \,. \end{aligned} \tag{22}$$

In both cases, $A_n^{(l,L)} = \prod_{k=l+1}^{L} \left(K_{F,n}^{(k)}\right)^{\alpha_F^{L-k}}$ *for* $1 \leqslant l \leqslant L$*, with the usual convention that a product indexed by the empty set always equals* 1*.*

Recall that $D_n^{(l)}, \widetilde{D}_n^{(l)}$ *and* α_ψ *are respectively the bounded differences and the Hölder exponent from Assumption 11;* $s_n^{(l)}$ *is from Assumption 12; the* $K_{F,n}^{(l)}$ *and* α_F *are from Assumption 13; and* $r_n, K_{R,n}$ *and* C_n *are from Assumption 14.*

Proof [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 $\|z_i^{(L)} - f^{(L)}(X_i)\|$ for all i. This is done by conditioning over x_i and finding a bound of

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

that does not depend on x_i, using a succession of triangular inequalities, the Hölder-type property of Assumption 13 and McDiarmid's inequality. We then turn it into a bound for $\|z_i^{(L)} - f^{(L)}(X_i)\|_\infty$ via the law of total probability and conclude with a union bound over i. ■

We present some corollaries and their consequences on our pool of examples. We can get a more explicit rate by disregarding multiplicative constants, as in the following corollary.

Corollary 16. *Under the assumptions of Theorem 15, for* n *large enough, and for* $0 < \rho < 1$*, the following holds with probability at least* $1 - \rho$*.*

$$\mathrm{MAE}_X \left(\Theta_{G_n}(\iota_X(f)), \Theta_{W,P}(f) \right) \lesssim \left(D_n^2 n \ln \left(\frac{n}{\rho} \right) \right)^{\alpha_F^{L-1} \alpha_\psi / 2} + \widetilde{D}_n^{\alpha_F^{L-1}} + s_n^{\alpha_F^{L-1}}, \tag{23}$$

Where $D_n = \max_l D_n^{(l)}$*,* $\widetilde{D}_n = \max_l \widetilde{D}_n^{(l)}$*,* $s_n = \max_l s_n^{(l)}$ *and the symbol* $\lesssim$ *hides some multiplicative constants which depend on* $K_{F,n}^{(1)}, \dots K_{F,n}^{(L)}, K_{R,n}$ *and are bounded over* n*.*

Proof Knowing that (D_n) and (s_n) both tend to zero, there is an integer n_0 such that, for any $n \geqslant n_0$, for each term of the sum in the right-hand side of Inequality (21), the base under the exponent α_{L-l} is smaller than 1. After, roughly majorizing each term of the sum, we disregard the multiplicative constants that are bounded over n under the symbol $\lesssim$. Finally, we use the inequality $(x + y)^a \leqslant x^a + y^a$ for $a \leqslant 1$. Recall that the $K_{F,n}^{(l)}$ and $K_{R,n}$ are assumed bounded from Assumptions 13 and 14. ■

Notice that Inequality (21) from Theorem 15 was non-asymptotic, that is, valid for any integer n. It is no longer in Corollary 16 above, Inequality (23) only holds for n large enough where the order of magnitude of the "large enough" will depend on the example.

The asymptotic behavior of Inequality (23) is generally determined by D_n. If the latter does not decrease fast enough, the inequality becomes meaningless, as it does not yield to convergence. In particular, we have the following important corollary.

Corollary 17. *If* $D_n = o\left(\frac{1}{\sqrt{n \ln n}}\right)$ *then* $\mathrm{MAE}_X(\Theta_{G_n}(\iota_X(f)), \Theta_{W,P}(f))$ *converges in probability towards* 0*. Moreover, if* $D_n \sim n^{-\beta}$ *with* $\beta > 1/2$*, then the convergence is almost sure.*

Proof From Inequality (23), the first part is immediate. For the second part, denote $Y_n = \mathrm{MAE}_X(\Theta_{G_n}(\iota_X(f)), \Theta_{W,P}(f))$. There is a constant C such that for any $\rho > 0$,

$$Y_n \leqslant C\left(\left(D_n^2 n \ln\left(\frac{n}{\rho}\right)\right)^{\alpha_F^{L-1}\alpha_\psi/2} + \widetilde{D}_n^{\alpha_F^{L-1}} + s_n^{\alpha_F^{L-1}}\right),$$

holds with probability as least $1-\rho$. This is equivalent to

$$\mathbb{P}\left(Y_n \geqslant \varepsilon\right) \leqslant n e^{-\frac{1}{nD_n^2}\left(\frac{\varepsilon}{C} - \widetilde{D}_n^{\alpha_F^{L-1}} - s_n^{\alpha_F^{L-1}}\right)^{-\alpha_F^{L-1}\alpha_\psi/2}}. \tag{24}$$

For any $\varepsilon > 0$. We claim that $\mathbb{P}\left(Y_n > \varepsilon\right)$ is summable for any ε. Indeed, since $\widetilde{D}_n$ and s_n tend to zero,

$$\left(\frac{\varepsilon}{C} - \widetilde{D}_n^{\alpha_F^{L-1}} - s_n^{\alpha_F^{L-1}}\right)^{-\alpha_F^{L-1}\alpha_\psi/2} = C_\varepsilon + o(1)\,,$$

where $C_\varepsilon = \left(\frac{\varepsilon}{C}\right)^{-\alpha_F^{L-1}\alpha_\psi/2}$. Therefore, since $\frac{1}{nD_n^2} \sim n^{2\beta-1}$, and $\beta > 1/2$, Inequality (24) gives,

$$\begin{aligned}
\mathbb{P}\left(Y_n > \varepsilon\right) &\leqslant n e^{-\left(\frac{C_\varepsilon}{nD_n^2} + o\left(\frac{1}{nD_n^2}\right)\right)} = n e^{-\left(\frac{C_\varepsilon}{nD_n^2} + o\left(n^{2\beta-1}\right)\right)} \\
&= n e^{-\left(C_\varepsilon\left(n^{2\beta-1} + o\left(n^{2\beta-1}\right)\right) + o\left(n^{2\beta-1}\right)\right)} = n e^{-\left(C_\varepsilon n^{2\beta-1} + O\left(n^{2\beta-1}\right)\right)} \\
&O\left(n e^{-C_\varepsilon n^{2\beta-1}}\right) = o\left(\frac{1}{n^2}\right),
\end{aligned}$$

which proves the summability of $\mathbb{P}\left(Y_n > \varepsilon\right)$. We conclude by Lemma 30 of Borel-Cantelli. ■

This corollary provides a sufficient condition for an 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. Below we investigate whether our examples have such sharp bounded differences. Under mild regularity conditions, this is the case for all examples but Example 5. The proof is in Appendix C.

Proposition 18. *We present the application of Theorem 15 on the Examples. In all cases, the message functions $m^{(l)}$ are supposed Lipschitz continuous and bounded. Additional regularity assumptions are needed for some examples.*

- ***Examples 1 and a.*** *$D_n = O(1/n)$, $\widetilde{D}_n = 0$, $s_n = r_n = 0$, and $\alpha_\psi = \alpha_F = 1$. The final rate is therefore $O\left(\sqrt{\ln n/n}\right)$.*
- ***Examples 2 and b.*** *Suppose that W is bounded away from zero, that is, there is $a > 0$ such that $W \geqslant a$. Then $D_n = O(1/n)$, $\widetilde{D}_n = 0$, $s_n = O(1/\sqrt{n})$, and $\alpha_\psi = \alpha_F = 1$. The final rate is therefore $O\left(\sqrt{\ln n/n}\right)$.*
- ***Examples 3 and c.*** *Suppose there is $a, b > 0$ and $K_c > 0$ such that $a < c(x,y,t) < b$ and $|c(x,y,t) - c(x',y',t)| \leqslant K_c(\|x-x'\|_\infty + \|y-y'\|_\infty)$, $\forall x, x', y, y', t$. Then $D_n = O(1/n)$, $\widetilde{D}_n = 0$, $s_n = O\left(1/\sqrt{n}\right)$, and $\alpha_\psi = \alpha_F = 1$. The final rate is therefore $O\left(\sqrt{\ln n/n}\right)$.*

- ***Examples 4 and d.*** *Suppose that h is α_h-Holder and that h^{-1} is bounded, and that Assumption 13 is satisfied for some exponent $0 < \alpha_F \leqslant 1$ (see examples below). Then $D_n = O(1/n)$, $\widetilde{D}_n = O(1/n^{\alpha_h/2})$, $s_n = O(1/n^{\alpha_h/2})$, and $\alpha_\psi = \alpha_h$. The final rate is therefore $O(\ln n/n)^{\alpha_h \alpha_F^L/2})$. For the exponents α_h, α_F, we can give several usual cases:*
  - *If $h = x \mapsto x^{1/p}$ and $h^{-1} = x \mapsto x^p$ (moment-based aggregation from Corso et al. (2020)), then $\alpha_h = 1/p$ and $\alpha_F = 1$.*
  - *If h^{-1} is Lipschitz with regard to, x (and recall that h is α_h-Holder), then $\alpha_F = \alpha_h$.*
  - *If the data domain is bounded and h^{-1} and h are both Lipschitz (e.g. for geometric mean on a bounded domain), then $\alpha_h = \alpha_F = 1$.*
- ***Examples 5 and e.*** *The bounded differences do not converge to zero, therefore cannot satisfy Corollary 17.*

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

Table 1 sums up these results. For a network with max aggregation, the bounded differences are not sharp enough for Theorem 15 to conclude. We thus treat this case separately in the next section.

Example	D_n	$\widetilde{D}_n$	s_n	Convergence by Corollary 16
1-a	$O(1/n)$	0	0	$O\left(\sqrt{\ln n/n}\right)$
2-b	$O(1/n)$	0	$O\left(1/\sqrt{n}\right)$	$O\left(\sqrt{\ln n/n}\right)$
3-c	$O(1/n)$	0	$O\left(1/\sqrt{n}\right)$	$O\left(\sqrt{\ln n/n}\right)$
4-d	$O(1/n)$	$O\left(n^{-\alpha_h/2}\right)$	$O\left(n^{-\alpha_h/2}\right)$	$O\left(\left(\ln n/n\right)^{\alpha_h \alpha_F^L/2}\right)$
5-e	$\Omega(1)$	–	–	✗

Table 1: Table summing up the convergence rates of this section. See Proposition 18 and Corollary 16 for details.

5.2.2 Convergence of *max* Aggregation Message-Passing GNNs

In this subsection, we specifically treat the example of max aggregation. Since the bounded differences method fails, we need another method to estimate the deviation between a *maximum* message-passing on a large random graph and its continuous counterpart.

We shall start by observing a simple example where everything is real-valued and smooth. Let f be a feature map on the latent space $\mathcal{X}$ and m be a message function which we assume to be real-valued. We call

$$g(x, y) = W(x, y)m(f(y)) \,,$$

such that the *maximum* message-passing around a node i is $\max_{j\neq i} g(X_i, X_j)$. Moreover, we suppose that $g\colon \mathcal{X}^2 \to \mathbb{R}$ is K_g-Lipschitz continuous and that the measure P is strictly positive. Therefore, in virtue of Lemma 27, the continuous counterpart of the *maximum* message-passing around any point $x \in \mathcal{X}$ is

$$\operatorname{ess\,sup}_P g(x,\cdot) = \sup g(x,\cdot)\,.$$

Our goal is to estimate

$$\mathbb{P}(|\max_i g(x, X_i) - \sup g(x,\cdot)| \geqslant \varepsilon)\,,$$

for $\varepsilon > 0$ and $x \in \mathcal{X}$. By definition of the supremum and by independence of the X_i, we have that

$$\begin{aligned}
\mathbb{P}(|\max_i g(x, X_i) - \sup g(x,\cdot)| \geqslant \varepsilon) &= \mathbb{P}(\max_i g(x, X_i) \leqslant \sup g(x,\cdot) - \varepsilon) \\
&= \mathbb{P}(g(x, X_1) \leqslant \sup g(x,\cdot) - \varepsilon)^n \\
&= \mathbb{P}(|g(x, X_1) - \sup g(x,\cdot)| \geqslant \varepsilon)^n \\
&= \big(1 - \mathbb{P}(|g(x, X_1) - \sup g(x,\cdot)| < \varepsilon)\big)^n\,.
\end{aligned}$$

By continuity and compactness, there is $x^* \in \mathcal{X}$ such that $\sup g(x,\cdot) = g(x, x^*)$, and, by Lipschitz continuity of g, for any $x \in \mathcal{X}$,

$$\|(x, X_1) - (x, x^*)\| = \|X_1 - x^*\| < \varepsilon/K_g \implies |g(x, X_1) - g(x, x^*))| < \varepsilon\,.$$

Thus, we obtain the bound

$$\begin{aligned}
\mathbb{P}(|\max_i g(x, X_i) - \sup g(x,\cdot)| \geqslant \varepsilon) &= \big(1 - \mathbb{P}(|g(x, X_1) - g(x, x^*)| < \varepsilon)\big)^n \\
&\leqslant (1 - \mathbb{P}(|X_1 - x^*| < \varepsilon/K_g))^n \\
&= \big(1 - P(B(x^*, \varepsilon/K_g) \cap \mathcal{X})\big)^n\,. \qquad (25)
\end{aligned}$$

where $B(x^*, \varepsilon/K_g)$ is the open ball, for the infinite norm, of center x^* and radius ε/K_g in $\mathbb{R}^d$.

We have now reached a point where, if we seek to go further, 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 "retention" of the Lebesgue measure which we call the *volume retaining property.* The purpose is to estimate from below the measure of a ball centered anywhere in $\mathcal{X}$.

Definition 19 (Volume retaining property)**.** *We say that the probability space $(\mathcal{X}, P)$ has the (r_0, κ)-volume retaining property if for any $r \leqslant r_0$ and for any $x \in \mathcal{X}$,*

$$P(B(x, r) \cap \mathcal{X}) \geqslant \kappa \lambda_d(B(x, r))\,. \qquad (26)$$

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

Notice that the volume retention property implies that the measure must be strictly positive.

The condition from Definition 19 is simultaneously a condition on the probability measure P as well as the geometry of $\mathcal{X}$. In the particular case where P itself is the Lebesgue measure, this becomes a purely geometrical condition on the shape of $\mathcal{X}$ at the boundary. For instance, it is no difficulty to see that the unit hypercube $[0,1]^d$ has the $(1, 1/2^d)$-volume retaining property. On the other hand, a typical shape that makes this property fail is an arbitrarily sharp peak. For example, the peak at the contact point of the complementary of two tangent open disks.

This hypothesis is also standard in other related contexts in which estimating the measure of balls is required in order to obtain some convergence rates. In Set Estimation, it is a cornerstone assumption. The goal of this branch of statistics is to estimate the compact support of a probability distribution on a metric space from samples, which is typically achieved by considering the union of balls centered at points drawn from that distribution. To this end, the property from Definition 19 was introduced by (Cuevas, 1990; Cuevas and Fraiman, 1997; Cuevas and Rodríguez-Casal, 2004). More contemporary, in Topological Data Analysis, it is used to measure the convergence rate of persistence diagrams, when the data is assumed to be drawn from a probability distribution supported on a compact metric space (Chazal et al., 2015a,b, 2016).

For a volume retaining probability space, we prove the following concentration inequality.

Lemma 20 (Concentration inequality for volume retaining space). *Let $g : \mathcal{X}^2 \to \mathbb{R}^q$ be K_g-Lipschitz and $(\mathcal{X}, P)$ have the (r_0, κ)-volume retaining property for some $r_0, \kappa > 0$. Recall that $\mathcal{X} \subset R^d$, then for any $\rho \geqslant e^{-n\kappa r_0^d 2^d}$, for any random variables $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$, with probability at least $1-\rho$, it holds*

$$\|\max_{1 \leqslant i \leqslant n} g(x, X_i) - \sup g(x, \cdot)\|_\infty \leqslant \frac{K_g}{2} \left(\frac{\ln(q/\rho)}{n\kappa}\right)^{1/d} . \tag{27}$$

Proof We write the proof assuming $q = 1$, the case $q \geqslant 1$ follows easily by a union bound. Clearly, volume-retention implies strict positiveness of the measure. The calculation is exactly the same as conducted in the introductory part of this Section 5.2.2, until we reach Inequality (25), where we must estimate

$$(1 - P(B(x^*, \varepsilon/K_g) \cap \mathcal{X}))^n . \tag{28}$$

Recall that we consider balls with regard to the infinity norm. Thus, by volume retention, for $\varepsilon \leqslant r_0 K_g$, Equation (28) is bounded by

$$(1 - P(B(x^*, \varepsilon/K_g) \cap \mathcal{X}))^n \leqslant \left(1 - \kappa \left(\frac{2\varepsilon}{K_g}\right)^d\right)^n \leqslant e^{-n\kappa \left(\frac{2\varepsilon}{K_g}\right)^d} . \tag{29}$$

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

$$|\max_{1 \leqslant i \leqslant n} g(x, X_i) - \sup g(x, \cdot)| \leqslant \frac{K_g}{2} \left(\frac{\ln(1/\rho)}{n\kappa}\right)^{1/d} , \tag{30}$$

Finally, Inequality (30) combined with a union bound yields to the desired result for $q \geqslant 1$. ■

Convergence in the case of maximum aggregation relies on the smoothness of the feature map f. Therefore, we will need the following regularity property about the intermediate layers.

Proposition 21. *Assume that W, $f = f^{(0)}$, as well as the $m^{(l)}$ are all Lipschitz continuous, then the functions $f^{(0)}, \dots, f^{(L)}$ are Lipschitz continuous too. We denote by $K_f = K_{f^{(0)}}, \dots, K_{f^{(L)}}$ their Lipschitz constants.*

Proof It is already assumed for $l = 0$. Suppose it is true for $l \geqslant 1$, we have

$$f^{(l+1)}(x) = \sup_y W(x,y) m^{(l+1)}(f^{(l)}(y)) = \sup_y g(x,y)\,,$$

where g is $K_W \|m^{(l+1)} \circ f^{(l)}\|_\infty + K_{m^{(l)}} K_{f^{(l)}}$-Lipschitz. Then from Lemma 29 $f^{(l+1)}$ is also Lipschitz continuous. ■

We are now ready to state the non-asymptotic bound for an MPGNN with maximum aggregation.

Theorem 22 (Non-asymptotic convergence of max-MPGNN towards cMPGNN). *Suppose that, $(\mathcal{X}, P)$ has the (r_0, κ)-volume retaining property and that f, W and the $m^{(l)}$ are Lipschitz continuous. Let $\rho \geqslant 2ne^{-n\kappa r_0^d 2^d}$ and n large enough for $0 < \rho < 1$ to hold, the following inequalities hold.*

- ***Equivariant case.*** *With probability at least $1-\rho$,*

$$\mathrm{MAE}_X\left(\Theta_{G_n}\left(\iota_X(f)\right), \Theta_{W,P}(f)\right) \leqslant \sum_{l=1}^{L} B^{(l,L)} \frac{K_{f^{(l)}}}{2} \left(\frac{1}{n\kappa} \ln\left(\frac{2^{L+1-l} n d_l}{\rho}\right)\right)^{1/d}, \quad (31)$$

- ***Invariant case.*** *With probability at least $1-\rho$,*

$$\begin{aligned}\left\|\overline{\Theta}_{G_n}(\iota_X(f)) - \overline{\Theta}_{W,P}(f)\right\|_\infty \leqslant &\sum_{l=1}^{L} B^{(l,L)} \frac{K_{f^{(l)}}}{2} \left(\frac{1}{n\kappa} \ln\left(\frac{2^{L+2-l} n d_l}{\rho}\right)\right)^{1/d} \\ &+ \left(\frac{1}{n} \ln\left(\frac{2d_L}{\rho}\right)\right)^{1/d}.\end{aligned} \quad (32)$$

Where $B^{(l,L)} = \prod_{k=l+1}^{L} K_{m^{(k)}}$ with the usual convention that a product indexed by the empty set always equals 1.

By grossly majoring each term of the sum in Inequality (31), and disregarding the constants, we get the following corollary. We only write the statement for the equivariant case as the bound for the invariant case would be similar.

Corollary 23. *Under the assumptions of Theorem 22, let $\rho \geqslant 2ne^{-n\kappa r_0^d 2^d}$ and n large enough for $0 < \rho < 1$ to hold. Then with probability at least $1-\rho$:*

$$\mathrm{MAE}_X\left(\Theta_{G_n}\left(\iota_X(f)\right), \Theta_{W,P}(f)\right) \lesssim \left(\frac{1}{n} \ln\left(\frac{n}{\rho}\right)\right)^{1/d},$$

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 $\mathcal{X} \subset \mathbb{R}^d$, where it is roughly equal to $O\left(n^{-1/d}\right)$, as opposed to the generally faster rate $O\left(n^{-1/2}\right)$ obtained with the McDiarmid's method from Theorem 15. Intuitively, this is to be expected, as the fast rate is akin to the central limit theorem, while the rate for max convolution follows from the number of balls necessary to cover the latent space (covering numbers), which scales exponentially in its dimension (Vershynin, 2018).

5.3 Experimental illustrations

We illustrate the convergence rates from both Theorems 15 and 22 on toy examples.

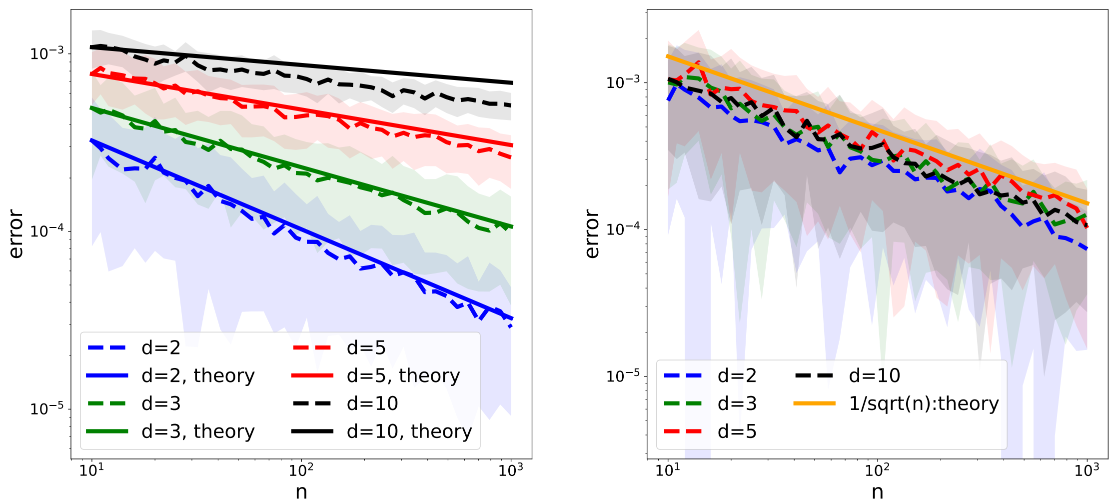

Figure 1: Numerical experiences for observing the trends of the rates of convergence arising from Theorems 15 and 22. The plots are in logarithmic scale. Left: max aggregation. Right: mean aggregation. For both figures, the dashed lines represent the experimental error as the graph size increases, while the full lines represent the theoretical rates arising from Theorems 15 and 22. This experiment has been conducted for various values of the latent space dimension d. The theoretical rates are $1/\sqrt{n}$ for a mean aggregation and $1/n^{1/d}$ for a *max* aggregation.

One difficulty in illustrating our convergence results is that the limit cGNN cannot be computed explicitly in most cases. Moreover, since our model of random graphs always produces dense (or even complete weighted) graphs, there are quite strong computational limits to testing very large n to approximate the cGNN. Nevertheless, we found that performing several rounds of Monte-Carlo simulation (50 is our experiments) yields a reasonable approximation of the limit for the mean example (Example 1). Additionally, there is a trick for the max example (Example 5). When all parameters are *nonnegative*, the non-linearity is increasing and nonnegative on $\mathbb{R}^+$ (*e.g.,* sigmoid), and, say, the input signal lives in $[0, 1]^d$, the cGNN can be computed explicitly: it is obtained when all points are $x_i = (1, \ldots, 1)$. Hence we limit our experiments to Examples 1 and 5, and our aim is to highlight the influence of the input dimension d on the convergence rate from Theorems 15 and 22: in theory, $O(1/\sqrt{n})$ for Example 1 and $O\left(1/n^{1/d}\right)$ for Example 5. We leave other examples for future work (in particular the generalized mean Example 4, which we found difficult to observe in practice).

For both experiments, the MPGNN has four layers. Each layer uses a single layer MLP with sigmoid activation function, and random weights in $[0, 1]$. A *mean* and a *max* aggregation are respectively used on the left and the right sides of Figure 1. The input signal is a dot product with a random vector in $[0, 1]^d$, and the latent variables are uniformly distributed in $[0, 1]^d$. Each experiment is run for $d = 2, 3, 5$ and 10. The MAE error is averaged over 50 experiments, and the standard deviation is also reported in Figure 1.

We indeed observe that the convergence rate is in $O(1/\sqrt{n})$ for a *mean* aggregation, no matter the dimension of the latent space. Whereas for a *max* aggregation, the speed follows $O\left(1/n^{1/d}\right)$, with d being the latent space's dimension. The standard deviation evolves as the error (note that the scale is logarithmic), we do not observe additional concentration phenomenon.

The reproducible code can be found in the following Github repository `https://github.com/Matthieu-Cordonnier/convergence-mpgnn/`.

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 to 4) and max aggregation (Example 5), but our theorems are not limited to these examples and is in fact verified for mild assumptions on the underlying model.

Throughout this paper, we have emphasized the fact that *mean* and *max* aggregation behave differently. Albeit, a link between the two still exists: as mentioned before, the generalized mean (Example 4) with moment $h = x \mapsto x^{1/p}$ naturally converges to the maximum aggregation when $p \to \infty$. However the bounded difference proof gives a rate in $n^{-1/2p}$, while for the max aggregation, our specific proof based on covering numbers for Theorem 22 yields $n^{-1/d}$. This is intuitively the "worst" rate possible in dimension d, but is better when $p > d/2$. Hence future work could try to unify the two proofs, to obtain a smooth transition between the different rates.

Acknowledgments

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 GRANOLA (ANR-21-CE48-0009), GRANDMA (ANR-21-CE23-0006) and GRAVA (ANR-18-CE40-0005).

Appendix A. Proof of Theorem 15

We proceed to the proof of the Theorem 15. In all this proof, we denote by $H^{(L)}(\rho)$ the upper bound from Inequality (21):

$$H^{(L)}(\rho) = \sum_{l=1}^{L} A_n^{(l,L)} \left[\left(\frac{1}{2} \left(D_n^{(l)} \right)^2 n \ln \left(\frac{2^{L+2-l} d_l n}{\rho} \right) \right)^{\alpha_\psi/2} + \widetilde{D}_n^{(l)} + s_n^{(l)} \right]^{\alpha_F^{L-l}} .$$

We will prove the results by induction on the depth L, we will detail the demonstration for the equivariant case, and the invariant case will easily follow.

Proof [Proof of the equivariant case, Inequality (21) of Theorem 15.] We start with the equivariant case. Calling $z_i^{(l)} = (\iota_X f)^{(l)}$ the signal at node i of the intermediate layer l of the GNN, we seek to bound

$$\begin{aligned} \mathrm{MAE}_X \left(\Theta_{G_n}(\iota_X f), \Theta_{W,P}(f) \right) &= \max_{1 \leqslant i \leqslant n} \left\| z_i^{(L)} - \iota_X (f^{(L)})_i \right\| \\ &= \max_{1 \leqslant i \leqslant n} \left\| z_i^{(L)} - f^{(L)}(X_i) \right\| . \end{aligned}$$

We prove the result by induction on L. Let $\rho > 0$, we recall some notations from Definition 5:

$$F_{P,n}^{(l+1)} \left(f^{(l)}, W \right)(x) = \mathbb{E}_{X_2,\ldots,X_n} \left[F^{(l+1)} \left(f^{(l)}(x), \left\{\!\!\left\{ \left(f^{(l)}(X_k), W(x, X_k) \right) \right\}\!\!\right\}_{2 \leqslant k \leqslant n} \right) \right] ,$$

and,

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

Suppose $L = 1$, we shall find a quantity that bounds all the $\left\| z_i^{(1)} - f^{(1)}(X_i) \right\|$, for i ranging from 1 to n, with probability at least $1 - \rho/n$. Thereby, by a union bound, this quantity will bound their maximum with probability at least $1 - \rho$. Choose an index $i \in \{1, \ldots, n\}$ and let $x_i \in \mathcal{X}$, consider

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

From a triangular inequality, and by definition of s_n (Inequality (19)), we get the majoration

$$\begin{aligned} &\left\| F^{(1)} \left(f^{(0)}(x_i), \left\{\!\!\left\{ \left(f^{(0)}(X_k), W(x_i, X_k) \right) \right\}\!\!\right\}_{k \neq i} \right) - f^{(1)}(x_i) \right\|_\infty \\ &\leqslant \left\| F^{(1)} \left(f^{(0)}(x_i), \left\{\!\!\left\{ \left(f^{(0)}(X_k), W(x_i, X_k) \right) \right\}\!\!\right\}_{k \neq i} \right) - F_{P,n}^{(1)} \left(f^{(0)}, W \right)(x_i) \right\|_\infty \\ &\quad + \left\| F_{P,n}^{(1)} \left(f^{(0)}, W \right)(x_i) - f^{(1)}(x_1) \right\|_\infty \\ &\leqslant \left\| F^{(1)} \left(f^{(0)}(x_i), \left\{\!\!\left\{ \left(f^{(0)}(X_k), W(x_i, X_k) \right) \right\}\!\!\right\}_{k \neq i} \right) - F_{P,n}^{(1)} \left(f^{(0)}, W \right)(x_i) \right\|_\infty + s_n^{(1)} . \end{aligned} \tag{33}$$

Now let us bound Inequality (33) from above with high probability using our Hölder version of McDiarmid's inequality, Corollary 26, on

$$(x_1,\dots,x_{i-1},x_{i+1},\dots,x_n)\mapsto F^{(1)}\left(f^{(0)}(x_i),\left\{\!\!\left\{\left(f^{(0)}(x_k),W(x_i,x_k)\right)\right\}\!\!\right\}_{k\neq i}\right)\,,$$

as a multivariate function of the $n-1$ variables x_k for $k\neq i$. We obtain that for any $x_i\in\mathcal{X}$, with probability at least $1-\rho/n$,

$$\begin{aligned}&\left\|F^{(1)}\left(f^{(0)}(x_i),\left\{\!\!\left\{\left(f^{(0)}(X_k),W(x_i,X_k)\right)\right\}\!\!\right\}_{k\neq i}\right)-f^{(1)}(x_i)\right\|_\infty\\&\qquad\leqslant\left(\frac{1}{2}\left(D_n^{(1)}\right)^2 n\ln\left(\frac{2d_1 n}{\rho}\right)\right)^{\alpha_\psi/2}+\widetilde{D}_n^{(1)}+s_n^{(1)}\,.\end{aligned}\tag{34}$$

Hence, by conditioning over X_i and applying the law of total probability, Inequality (34) yields with probability at least $1-\rho/n$

$$\left\|z_i^{(1)}-f^{(1)}(X_i)\right\|_\infty\leqslant\left(\frac{1}{2}\left(D_n^{(1)}\right)^2 n\ln\left(\frac{2d_1 n}{\rho}\right)\right)^{\alpha_\psi/2}+\widetilde{D}_n^{(1)}+s_n^{(1)}\,.$$

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

$$\max_i\left\|z_i^{(1)}-f^{(1)}(X_i)\right\|_\infty\leqslant\left(\frac{1}{2}\left(D_n^{(1)}\right)^2 n\ln\left(\frac{2d_1 n}{\rho}\right)\right)^{\alpha_\psi/2}+\widetilde{D}_n^{(1)}+s_n^{(1)}\leqslant H^{(1)}(\rho)\,,$$

which proves the case $L=1$.

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

$$\begin{aligned}&\left\|z_i^{(L+1)}-f^{(L+1)}(X_i)\right\|_\infty\\&\leqslant\left\|z_i^{(L+1)}-F_{P,n}^{(L+1)}(f^{(L)},W)(X_i)\right\|_\infty\\&\quad+\left\|F_{P,n}^{(L+1)}(f^{(L)},W)(X_i)+f^{(L+1)}(X_i)\right\|_\infty\\&\leqslant\left\|z_i^{(L+1)}-F_{P,n}^{(L+1)}(f^{(L)},W)(X_i)\right\|_\infty+s_n^{(L+1)}\\&\leqslant\left\|z_i^{(L+1)}-F^{(L+1)}\left(f^{(L)}(X_i),\left\{\!\!\left\{\left(f^{(L)}(X_k),W(X_i,X_k)\right)\right\}\!\!\right\}_{k\neq i}\right)\right\|_\infty\\&\quad+\left\|F^{(L+1)}\left(f^{(L)}(X_i),\left\{\!\!\left\{\left(f^{(L)}(X_k),W(X_i,X_k)\right)\right\}\!\!\right\}_{k\neq i}\right)-F_{P,n}^{(L+1)}(f^{(L)},W)(X_i)\right\|_\infty+s_n^{(L+1)}\\&\leqslant K_{F,n}^{(L+1)}\max_i\left\|z_i^{(L)}-f^{(L)}(X_i)\right\|_\infty^{\alpha_F}\\&\quad+\left\|F^{(L+1)}\left(f^{(L)}(X_i),\left\{\!\!\left\{\left(f^{(L)}(X_k),W(X_i,X_k)\right)\right\}\!\!\right\}_{k\neq i}\right)-F_{P,n}^{(L+1)}(f^{(L)},W)(X_i)\right\|_\infty+s_n^{(L+1)}\,,\end{aligned}\tag{35}$$

where the last Inequality (35) comes from the Hölder-like regularity Assumption 13 on $F^{(L+1)}$. Taking the maximum over the vertices, Inequality (35) yields

$$\begin{aligned}
&\max_i \left\| z_i^{(L+1)} - f^{(L+1)}(X_i) \right\|_\infty \\
&\leqslant K_{F,n}^{(L+1)} \max_i \left\| z_i^{(L)} - f^{(L)}(X_i) \right\|_\infty^{\alpha_F} \\
&\quad + \max_i \left\| F^{(L+1)}\left(f(X_i), \{\!\!\{ (f(X_k), W(X_i, X_k)) \}\!\!\}_{k \neq i} \right) - F_{P,n}^{(L+1)}(f^{(L)}, W)(X_i) \right\|_\infty + s_n^{(L+1)} .
\end{aligned} \tag{36}$$

Finally, we bound Inequality (36) from above with high probability. The first term is handled by the induction hypothesis. For the second term, we employ the same technique as we did in the case $L = 1$. By conditioning over X_i, using the Hölder version McDiarmid's inequality, and a union bound, we obtain with probability at least $1 - \rho$,

$$\begin{aligned}
&\max_i \left\| z_i^{(L+1)} - f^{(L+1)}(X_i) \right\|_\infty \\
&\leqslant K_{F,n}^{(L+1)} H^{(L)} \left(\frac{\rho}{2}\right)^{\alpha_F} + \left(\frac{1}{2} \left(D_n^{(L+1)} \right)^2 n \ln \left(\frac{4 d_{L+1} n}{\rho} \right) \right)^{\alpha_\psi/2} + \widetilde{D}_n^{(L+1)} + s_n^{(L+1)} \\
&\leqslant K_{F,n}^{(L+1)} \left(\sum_{l=1}^{L} A_n^{(l,L)} \left[\left(\frac{1}{2} \left(D_n^{(l)} \right)^2 n \ln \left(\frac{2^{L+3-l} d_l n}{\rho} \right) \right)^{\alpha_\psi/2} + \widetilde{D}_n^{(l)} + s_n^{(l)} \right]^{\alpha_F^{L-l}} \right)^{\alpha_F} \\
&\quad + \left(\frac{1}{2} \left(D_n^{(L+1)} \right)^2 n \ln \left(\frac{4 d_{L+1} n}{\rho} \right) \right)^{\alpha_\psi/2} + \widetilde{D}_n^{(L+1)} + s_n^{(L+1)} \\
&\leqslant \sum_{l=1}^{L} K_{F,n}^{(L+1)} (A_n^{(l,L)})^{\alpha_F} \left[\left(\frac{1}{2} \left(D_n^{(l)} \right)^2 n \ln \left(\frac{2^{L+3-l} d_l n}{\rho} \right) \right)^{\alpha_\psi/2} + \widetilde{D}_n^{(l)} + s_n^{(l)} \right]^{\alpha_F^{L+1-l}} \\
&\quad + \left(\frac{1}{2} \left(D_n^{(L+1)} \right)^2 n \ln \left(\frac{4 d_{L+1} n}{\rho} \right) \right)^{\alpha_\psi/2} + \overline{D}_n^{(L+1)} + s_n^{(L+1)} \\
&= H^{(L+1)}(\rho).
\end{aligned}$$

where we have used that $(x + y)^a \leqslant x^a + y^a$ for $a \leqslant 1$, as well as the recursive expression

$$\begin{cases} A^{(l,l)} = 1 \\ A^{(l,L+1)} = K_{F,n}^{(L+1)} (A^{(l,L)})^{\alpha_F} , \end{cases}$$

which yields the result. ■

We now turn to the invariant case, which follows as a corollary of Inequality (21). We will use Inequality (21), and make additional use of McDiarmid's concentration bound.

Proof [Proof of the invariant case, Inequality (22) of Theorem 15.] Under Assumptions 11 to 13 and 14, we obtain the following inequality.

$$\begin{aligned}
&\left\|\overline{\Theta}_{G_n}\left(\iota_X(f)\right)-\overline{\Theta}_{W,P}(f)\right\|_\infty \\
&\leqslant \left\|R\left(\{\!\!\{z_1^{(L)},\dots,z_n^{(L)}\}\!\!\}\right)-R\left(\{\!\!\{f^{(L)}(X_1),\dots,f^{(L)}(X_n)\}\!\!\}\right)\right\|_\infty \\
&\quad+\left\|R\left(\{\!\!\{f^{(L)}(X_1),\dots,f^{(L)}(X_n)\}\!\!\}\right)-R_{P,n}\left(f^{(L)}\right)\right\|_\infty \\
&\quad+\left\|R_{P,n}(f^{(L)})-\mathcal{R}_P\left(f^{(L)}\right)\right\|_\infty \\
&\leqslant K_{R,n}\max_i\left\|z_i^{(L)}-f^{(L)}(X_i)\right\|_\infty \\
&\quad+\left\|R\left(\{\!\!\{f^{(L)}(X_1),\dots,f^{(L)}(X_n)\}\!\!\}\right)-R_{P,n}\left(f^{(L)}\right)\right\|_\infty+r_n\,.
\end{aligned}$$

Using a union bound, Inequality (21) and McDiarmid's inequality, we get that, with probability at least $1-\rho$:

$$\left\|\overline{\Theta}_{G_n}(\iota_X(f))-\overline{\Theta}_{W,P}(f)\right\|_\infty \leqslant K_{R,n}H^{(L)}(\rho/2)+C_n\sqrt{n\ln\left(\frac{4d_L}{\rho}\right)}+r_n\,,$$

which concludes the proof. ■

Appendix B. Proof of Theorem 22

We prove Theorem 22. Until the end of the proof, we denote by $H^{(L)}(\rho)$ the left-hand side of Inequality (31),

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

Proof [Proof of the equivariant case, Inequality (31) of Theorem 22] Let $\rho>0$. We will prove the theorem by induction on L. For $L=1$, let us note $g^{(1)}(x,y)=W(x,y)m^{(1)}(f^{(0)}(y))$, such that $f^{(1)}(x)=\sup_y g^{(1)}(x,y)$. The map $f^{(1)}$ is $K_{f^{(1)}}=K_{m^{(1)}}K_{f^{(0)}}+\|m^{(1)}\circ f^{(0)}\|_\infty K_W$ Lipschitz continuous from Proposition 21.

Fix a node $i\in\{1,\dots,n\}$ and let $x_i\in\mathcal{X}$, by Lemma 20, for $\rho\geqslant ne^{-n\kappa r_0^d 2^d}$, with probability at least $1-\rho/n$, we have

$$\|\max_{j\neq i}g^{(1)}(x,X_j)-\sup_{y\in\mathcal{X}}g^{(1)}(x,y)\|_\infty\leqslant\frac{K_{f^{(1)}}}{2}\left(\frac{\ln(nd_1/\rho)}{n\kappa}\right)^{1/d}.$$

Thus, by conditioning over X_i and using the law of total probability, with probability at least $1-\rho/n$, it holds

$$\|\max_{j\neq i}g^{(1)}(X_i,X_j)-\sup_{y\in\mathcal{X}}g^{(1)}(X_i,y)\|_\infty\leqslant\frac{K_{f^{(1)}}}{2}\left(\frac{\ln(nd_1/\rho)}{n\kappa}\right)^{1/d}.$$

Since $\max_{j\neq i} g^{(1)}(X_i, X_j) = z_i^{(1)}$ and $\sup_{y\in\mathcal{X}} g^{(1)}(X_i, y) = f^{(1)}(X_i)$, by maximizing over i and doing a union bound, we obtain that with probability at least $1-\rho$, it holds

$$\max_i \left\| z_i^{(1)} - f^{(1)}(X_i) \right\|_\infty \leqslant \frac{K_{f^{(1)}}}{2} \left(\frac{\ln(nd_1/\rho)}{n\kappa} \right)^{1/d} \leqslant H^{(1)}(\rho)\,,$$

for $\rho \geqslant ne^{-n\kappa r_0^d 2^d}$, and, in particular, for $\rho \geqslant 2ne^{-n\kappa r_0^d 2^d}$. That concludes the case $L=1$.

Now suppose the result true for $L \geqslant 1$, fix a node $i \in \{1, \dots, n\}$, we have the following bound,

$$\begin{aligned}
&\left\| z_i^{(L+1)} - f^{(L+1)}(X_i) \right\|_\infty = \left\| \max_{j\neq i} W(X_i,X_j) m^{(L+1)}\left(z_j^{(L)}\right) - \sup_{y\in\mathcal{X}} W(X_i,y) m^{(L+1)}(f^{(L)}(y)) \right\|_\infty \\
&\leqslant \left\| \max_{j\neq i} W(X_i,X_j) m^{(L+1)}(z_j^{(L)}) - \max_{j\neq i} W(X_i,X_j) m^{(L+1)}(f^{(L)}(X_j)) \right\|_\infty \\
&\quad + \left\| \max_{j\neq i} W(X_i,X_j) m^{(L+1)}(f^{(L)}(X_j)) - \sup_{y\in\mathcal{X}} W(X_i,y) m^{(L+1)}(f^{(L)}(y)) \right\|_\infty \\
&\leqslant K_{m^{(L+1)}} \max_{j\neq i} \left\| z_j^{(L)} - f^{(L)}(X_j) \right\|_\infty \\
&\quad + \left\| \max_{j\neq i} W(X_i,X_j) m^{(L+1)}(f^{(L)}(X_j)) - \sup_{y\in\mathcal{X}} W(X_i,y) m^{(L+1)}(f^{(L)}(y)) \right\|_\infty\,,
\end{aligned} \tag{37}$$

where the last Inequality (37) uses Lemma 28, the fact that $|W| \leqslant 1$, and the Lipschitz continuity of $m^{(L+1)}$. Thus taking the maximum over i, we get

$$\begin{aligned}
&\max_i \left\| z_i^{(L+1)} - f^{(L+1)}(X_i) \right\|_\infty \\
&\leqslant K_{m^{(L+1)}} \max_i \left\| z_j^{(L)} - f^{(L)}(X_j) \right\|_\infty \\
&\quad + \max_i \left\| \max_{j\neq i} W(X_i,X_j) m^{(L+1)}(f^{(L)}(X_j)) - \sup_{y\in\mathcal{X}} W(X_i,y) m^{(L+1)}(f^{(L)}(y)) \right\|_\infty .
\end{aligned} \tag{38}$$

Now we bound Inequality (38) with high probability. We use the induction hypothesis for the first term. For the second term, we set $g^{(L+1)}(x,y) = W(x,y) m^{(L+1)}(f^{(L)}(y))$ and use Lemma 20 on g which is $K_{f^{(L+1)}} = K_{m^{(L+1)}} K_{f^{(L)}} + \|m^{(L+1)} \circ f^{(L)}\|_\infty K_W$ Lipschitz. The process is the same as in the case $L=1$. By conditioning over X_i followed by a union bound, we obtain that for $\rho \geqslant 2e^{-n\kappa r_0^d 2^d}$, with probability at least $1-\rho$, it holds,

$$\begin{aligned}
&\max_i \left\| z_i^{(L+1)} - f^{(L+1)}(X_i) \right\|_\infty \\
&\leqslant K_{m^{(L+1)}} H^{(L)}(\rho/2) + \frac{K_{f^{(L+1)}}}{2} \left(\frac{\ln(2nd_{L+1}/\rho)}{n\kappa} \right)^{1/d} \\
&= \sum_{l=1}^{L} K_{m^{(L+1)}} B^{(l,L)} \frac{K_{f^{(l)}}}{2} \left(\frac{\ln\left(2^{L+2-l} n d_l/\rho\right)}{n\kappa} \right)^{1/d} + \frac{K_{f^{(L+1)}}}{2} \left(\frac{\ln(2nd_{L+1}/\rho)}{n\kappa} \right)^{1/d} \\
&= H^{(L+1)}(\rho)\,.
\end{aligned}$$

where we have used the recursive expression

$$\begin{cases} B^{(l,l)} = 1 \\ B^{(l,L+1)} = K_{m^{(L+1)}} B^{(l,L)} \end{cases} \tag{39}$$

■

We finish with the proof of the invariant case, which follows from the previous result.
Proof [Proof of the invariant case, Inequality (32) of Theorem 22] The final readout is simply a maximum (respectively a supremum), over the nodes of the graph, hence a triangular inequality gives,

$$\begin{aligned}
&\left\|\overline{\Theta}_{G_n}\left(\iota_X f\right) - \overline{\Theta}_{W,P}(f)\right\|_\infty = \left\|\max_i z_i^{(L)} - \sup f^{(L)}\right\|_\infty \\
&\leqslant \left\|\max_i z_i^{(L)} - \max_i f^{(L)}(X_i)\right\|_\infty + \left\|\max_i f^{(L)}(X_i) - \sup f^{(L)}\right\|_\infty \\
&\leqslant \max_i \left\|z_i^{(L)} - f^{(L)}(X_i)\right\|_\infty + \left\|\max_i f^{(L)}(X_i) - \sup f^{(L)}\right\|_\infty .
\end{aligned}$$

Using the bound for the equivariant case and Lemma 20 on $f^{(L)}$, we obtain the result. ■

Appendix C. Examples

In this section, we put all computations related to the examples, we prove Propositions 9 and 18. For notational convenience, we drop any subscript or superscript (l) referring to layers. Recall that $m = m^{(l)}$ is supposed Lipschitz and bounded, we denote K_m its Lipschitz constant and $\|m\|_\infty = \sup_x \|m(x)\|_\infty$ (on the respective domain of x at each layer). We divide this section into subsections for each example.

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

We prove Propositions 9 and 18 on all the examples. We verify Assumptions 11, 12 and 13 on Examples 1 to 4, but not Example 5. For the latter, we simply verify that the bounded differences fail to be sharp enough.

Proposition 9 for Examples 1 and a. By independence and identical distribution of the random variables and linearity of the expected value, the convergence in Equation (14) is actually an equality for all integer n. For all x,

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

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

Assumption 12 for Examples 1 and a. The above calculation yields $s_n = 0$.

Assumption 11 for Examples 1 and a. We consider $\psi = \mathrm{id}$, such that $\widetilde{D}_n = 0$ and $K_\psi = 1$.

Let $x_1, \dots, x_n$ and $x'_2, \dots, x'_n$ be such that $x_i = x'_i$ except at $k = 2$.

$$
\begin{aligned}
&\left\| F\left(f(x_1), \{\!\!\{(f(x_i), W(x_1, x_i))\}\!\!\}_{2 \leqslant i \leqslant n}\right) - F\left(f(x_1), \{\!\!\{(f(x'_i), W(x_1, x'_i))\}\!\!\}_{2 \leqslant i \leqslant n}\right) \right\|_\infty \\
&= \frac{1}{n-1} \left\| W(x_1, x_2) m(f(x_2)) - W(x_1, x'_2) m(f(x'_2)) \right\|_\infty \\
&= 2\|m\|_\infty / (n-1) \\
&= O(1/n)\,,
\end{aligned}
$$

since m is bounded. Hence, $D_n = O(1/n)$.

Assumption 13 for Examples 1 and a. We have

$$
\begin{aligned}
&\left\| F(z_1, \{\!\!\{(z_i, w_i)\}\!\!\}_{i \geqslant 2}) - F(z'_1, \{\!\!\{z'_i, w_i\}\!\!\}_{i \geqslant 2}) \right\|_\infty \\
&\leqslant \left\| \frac{1}{n-1} \sum_{2 \leqslant i \leqslant n} w_i m(z_i) - \frac{1}{n-1} \sum_{2 \leqslant i \leqslant n} w_i m(z'_i) \right\|_\infty \\
&\leqslant K_m \max_i \left\| z_i - z'_i \right\|_\infty .
\end{aligned}
$$

Hence Assumption 13 is satisfied with $K_{F,n} = K_m$ and $\alpha_F = 1$.

C.2 Examples 2 and b: Degree normalized message-passing

Recall here that we assume that $W(\cdot, \cdot) \geqslant a > 0$.

Proposition 9 for Examples 2 and b. We have for all x,

$$
\int_{\mathcal{X}} \frac{W(x, y) m(f(y))}{\int_{\mathcal{X}} W(x, t) \mathrm{d}P(t)} \mathrm{d}P(y) = \frac{\mathbb{E}\left[W(x, X) m(f(X))\right]}{\mathbb{E}\left[W(x, X)\right]}\,,
$$

where $X \sim P$. Let $x \in \mathcal{X}$ be fixed. For simplicity, we compute s_{n+1} (*i.e.,* we consider n random variables and not $n-1$). Then,

$$\begin{aligned}
&\left\| \mathbb{E}_{X_i}\left[\frac{\frac{1}{n}\sum_i W(x,X_i)m(f(X_i))}{\frac{1}{n}\sum_i W(x,X_i)}\right] - \frac{\mathbb{E}\left[W(x,X)m(f(X))\right]}{\mathbb{E}\left[W(x,X)\right]}\right\|_\infty \\
&= \left\| \mathbb{E}_{X_i}\left[\frac{\frac{1}{n}\sum_i W(x,X_i)m(f(X_i))}{\frac{1}{n}\sum_i W(x,X_i)} - \frac{\mathbb{E}\left[W(x,X)m(f(X))\right]}{\mathbb{E}\left[W(x,X)\right]}\right]\right\|_\infty \\
&\leqslant \mathbb{E}_{X_i}\left[\left\| \frac{\frac{1}{n}\sum_i W(x,X_i)m(f(X_i))}{\frac{1}{n}\sum_i W(x,X_i)} - \frac{\mathbb{E}\left[W(x,X)m(f(X))\right]}{\mathbb{E}\left[W(x,X)\right]}\right\|_\infty\right] \\
&\leqslant \mathbb{E}\left[\left\|\frac{1}{n}\sum_i W(x,X_i)m(f(X_i))\right\|_\infty \left|\frac{1}{\frac{1}{n}\sum_i W(x,X_i)} - \frac{1}{\mathbb{E}\left[W(x,X)\right]}\right|\right. \\
&\qquad \left. + \left\|\frac{1}{n}\sum_i W(x,X_i)m(f(X_i)) - \mathbb{E}\left[W(x,X)m(f(X))\right]\right\|_\infty \left|\frac{1}{\mathbb{E}\left[W(x,X)\right]}\right|\right] \\
&\leqslant \frac{\|m\|_\infty}{a^2}\mathbb{E}\left[\left|\frac{1}{n}\sum_i W(x,X_i) - \mathbb{E}\left[W(x,X)\right]\right|\right] \\
&\qquad + \frac{1}{a}\mathbb{E}\left[\left\|\frac{1}{n}\sum_i W(x,X_i)m(f(X_i)) - \mathbb{E}\left[W(x,X)m(f(X))\right]\right\|_\infty\right], \qquad (40)
\end{aligned}$$

since $0 < a \leqslant W \leqslant 1$. Consequently, using the formula $\mathbb{E}(X) = \int_{t>0} \mathbb{P}(X > t)\mathrm{d}t$ for X nonnegative, we get that this last quantity (40) is equal to

$$\begin{aligned}
&\frac{\|m\|_\infty}{a^2}\int_{t>0}\mathbb{P}\left(\left|\mathbb{E}\left[W(x,X)\right] - \frac{1}{n}\sum_i W(x,X_i)\right| > t\right)\mathrm{d}t \\
&\qquad + \frac{1}{a}\int_{t>0}\mathbb{P}\left(\left\|\frac{1}{n}\sum_i W(x,X_i)m(f(X_i)) - \mathbb{E}\left[W(x,X)m(f(X))\right]\right\|_\infty > t\right)\mathrm{d}t\,. \qquad (41)
\end{aligned}$$

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 multivariate maps have bounded differences of the form $c_i = C/n$ for all i. Therefore, there are some positive constants C_1, C_2 *independent of* x such that Equation (41) is bounded by

$$C_1 \int_{t>0} e^{-nC_2 t^2}\mathrm{d}t = O\left(1/\sqrt{n}\right) \to 0.$$

Since this is true for all x, and that this bound is independent of x, we obtain convergence in L^∞_P norm. This remains true when replacing P by $\phi \cdot P$, f by $\phi \cdot f$ and W by $\phi \cdot W$ for any $\phi \in \mathrm{Aut}(\mathcal{X})$. Hence we have shown Proposition 9.

Assumption 12 for Examples 2 and b. The above calculation yields $s_n = O\left(1/\sqrt{n}\right)$.

Assumption 11 for Examples 2 and b. Let us first introduce some intermediate computations. Let $x_1, \dots, x_n$ and $x'_2, \dots, x'_n$ be such that $x_i = x'_i$ except at $i = 2$. Denote by $z_i = f(x_i)$, $z'_i = f(x_i)$, $w_i = W(x_1, x_i)$, and $w'_i = W(x'_1, x'_i)$ for short.

$$\begin{aligned}
&\left\| F\left(f(x_1), \{\!\!\{(f(x_i), W(x_1, x_i))\}\!\!\}_{2\leqslant i\leqslant n}\right) - F\left(f(x_1), \{\!\!\{\left(f(x'_i), W(x_1, x'_i)\right)\}\!\!\}_{2\leqslant i\leqslant n}\right) \right\|_\infty \\
&= \left\| F\left(z_1, \{\!\!\{z_i, w_i\}\!\!\}\right) - F\left(z_1, \{\!\!\{z'_i, w'_i\}\!\!\}\right) \right\|_\infty \\
&= \left\| \frac{\sum_{2\leqslant i\leqslant n} w_i m(z_i)}{\sum_i w_i} - \frac{\sum_{2\leqslant i\leqslant n} w'_i m(z'_i)}{\sum_i w'_i} \right\|_\infty \\
&\leqslant \frac{1}{\sum_i w_i} \left\| \sum_i w_i m(z_i) - \sum_i w'_i m(z'_i) \right\|_\infty + \left\| \sum_i w'_i m(z'_i) \right\|_\infty \left| \frac{1}{\sum_i w_i} - \frac{1}{\sum_i w'_i} \right| \\
&\leqslant \frac{1}{(n-1)a} \left\| \sum_i w_i m(z_i) - \sum_i w'_i m(z'_i) \right\|_\infty + \frac{\|m\|_\infty}{(n-1)a^2} \left| \sum_i w_i - \sum_i w'_i \right| . \qquad (42)
\end{aligned}$$

Then, again we consider $\psi = \mathrm{id}$, such that $\widetilde{D}_n = 0$ and $K_\psi = 1$. Applying Inequality (42) with $z_i = z'_i$ and $w_i = w'_i$ except for $i = 2$, we prove Assumption 11 with

$$D_n = \frac{1}{(n-1)a} + \frac{\|m\|_\infty}{(n-1)a^2} = O\left(1/n\right) .$$

Assumption 13 for Examples 2 and b. Again using Inequality (42) with $w_i = w'_i$ and using the Lipschitz property of m,

$$\begin{aligned}
&\left\| F\left(z_1, \{\!\!\{z_i, w_i\}\!\!\}\right) - F\left(z'_1, \{\!\!\{z'_i, w_i\}\!\!\}\right) \right\|_\infty \\
&\leqslant \frac{1}{(n-1)a} \left\| \sum_i w_i m(z_i) - \sum_i w_i m(z'_i) \right\|_\infty \\
&\leqslant \frac{K_m}{a} \max_i \left\| z_i - z'_i \right\|_\infty .
\end{aligned}$$

Hence we prove Assumption 13 with $K_F = \frac{K_m}{a}$ and $\alpha_F = 1$.

C.3 Examples 3 and c: Attentional message-passing

Call $V(x, y) = c\left(f(x), f(y), W(x, y)\right)$. We are basically brought to the previous example with V instead of W.

Proposition 9 for Examples 3 and c. Using V instead of W, we are brought to the previous example and therefore Proposition 9 is satisfied.

Assumption 12 for Examples 3 and c. By the same argument, Assumption 12, is fulfilled with $s_n = O\left(1/\sqrt{n}\right)$.

Assumption 11 for Examples 3 and c. From the previous Example, equation Inequality (42) remains valid with $w_i = V(x_1, x_i), w'_i = V(x'_1, x'_i)$ instead of W. Hence Assumption 11 is proven with $D_n = O\left(1/\sqrt{n}\right)$ and $\psi = \mathrm{id}$.

Assumption 13 for Examples 3 and c. Again $\psi = \mathrm{id}$. Using again Inequality (42) but with $v_i = c(z_1, z_i, w_i)$, where $c(x,y,w) \leqslant b$, and $|c(x,y,w) - c(x',y',w)| \leqslant K_c(\|x-x'\|_\infty + \|y-y'\|_\infty)$, we get:

$$
\begin{aligned}
&\left\| F\left(z_1, \{\!\{(z_i,w_i)\}\!\}_{2\leqslant k\leqslant n}\right) - F\left(z_1', \{\!\{(z_i',w_i)\}\!\}_{2\leqslant k\leqslant n}\right)\right\|_\infty \\
&\leqslant \frac{b}{(n-1)a}\left\|\sum_i v_i m(z_i) - \sum_i v_i' m(z_i')\right\|_\infty + \frac{b\|m\|_\infty}{(n-1)a^2}\left|\sum_i v_i - \sum_i v_i'\right| \\
&\leqslant \frac{b}{a}\max_i \left\|v_i m(z_i) - v_i' m(z_i')\right\|_\infty + \frac{b\|m\|_\infty}{a^2}\max_i |v_i - v_i'| \\
&\leqslant \frac{b}{a}\left(\|m\|_\infty K_c\left(\|z_1 - z_1'\|_\infty + \max_i \|z_i - z_i'\|_\infty\right) + bK_m \max_i \|z_i - z_i'\|_\infty\right) \\
&+ \frac{b\|m\|_\infty}{a^2} K_c\left(\|z_1 - z_1'\|_\infty + \max_i \|z_i - z_i'\|_\infty\right) \\
&\leqslant \left(\frac{2b\|n\|_\infty K_c + b^2 K_m}{a} + \frac{2bmK_c}{a^2}\right)\max_i \|z_i - z_i'\|_\infty \,.
\end{aligned}
$$

Hence we have shown Assumption 13 with $\alpha_F = 1$.

C.4 Examples 4 and d: Generalized mean

Proposition 9 for Examples 4 and d. We call $g(x,y) = h^{-1}(W(x,y)m(f(y)))$ and again consider n neighbors (instead of $n-1$). For any fixed $x \in \mathcal{X}$, we have

$$
\begin{aligned}
&\left\| \mathbb{E}_{X_i}\left[h\Big(\frac{1}{n}\sum_i g(x,X_i)\Big)\right] - h\big(\mathbb{E}_X\left[g(x,X)\right]\big)\right\|_\infty \\
&\leqslant \mathbb{E}_{X_i}\left[\left\|h\Big(\frac{1}{n}\sum_i g(x,X_i)\Big) - h\big(\mathbb{E}_X\left[g(x,X)\right]\big)\right\|_\infty\right] \\
&\leqslant K_h \mathbb{E}_{X_i}\left[\left\|\frac{1}{n}\sum_i g(x,X_i) - \mathbb{E}_X g(x,X)\right\|_\infty^{\alpha_h}\right] .
\end{aligned}
$$

Again, this is equal to

$$
K_h \int_t \mathbb{P}\left(\left\|\frac{1}{n}\sum_i g(x,X_i) - \mathbb{E}\left[g(x,X)\right]\right\|_\infty^{\alpha_h} > t\right)\mathrm{d}t\,,
$$

and since h^{-1}, therefore g, is bounded we can use Hoeffding inequality to show that this is bounded for some constants C_1, C_2 by

$$
C_1 \int_{t>0} e^{-nC_2 t^{2/\alpha_h}}\mathrm{d}t = O(n^{-\alpha_h/2}) \to 0\,.
$$

Since this is true for all x and this bound is independent of x, we obtain convergence in L_P^∞ norm. This remains true when replacing P by $\phi \cdot P$, f by $\phi \cdot f$ and W by $\phi \cdot W$ for any $\phi \in \mathrm{Aut}(\mathcal{X})$. Hence we have shown Proposition 9.

Assumption 12 for Examples 4 and d. The above calculation yields $s_n = O\left(n^{-\frac{\alpha_h}{2}}\right)$.

Assumption 11 for Examples 4 and d. Here we consider $\psi = h$. For Item (i) of Assumption 11, we have

$$\psi^{-1}\left(F(f(x_1), \{\!\{(f(x_i), W(x_1, x_i))\}\!\}_{i\geqslant 2})\right) = \frac{1}{n-1}\sum_{i\geqslant 2} h^{-1}(W(x_1, x_i)m(f(x_i))\,.$$

So we are brought back to Example 1 when h^{-1} is bounded, and therefore $D_n = O(1/n)$.

For Item (ii), calling again $g(x,y) = h^{-1}(W(x,y), m(f(y)))$, we have

$$\begin{aligned}
&\left\|\psi\left(\mathbb{E}\psi^{-1}\left(F\left(f(x_1), \{\!\{(f(X_i), W(x_1, X_i))\}\!\}_{i\geqslant 2}\right)\right)\right) - \mathbb{E}F(f(x_1), \{\!\{(f(X_i), W(x_1, X_i))\}\!\}_{i\geqslant 2})\right\|_\infty \\
&= \left\|h\left(\mathbb{E}\frac{1}{n-1}\sum_i g(x, X_i)\right) - \mathbb{E}\left[h\bigg(\frac{1}{n-1}\sum_i g(x, X_i)\bigg)\right]\right\|_\infty \\
&= \left\|h\left(\mathbb{E}_X\left[g(x,X)\right]\right) - \mathbb{E}_{X_i}\left[h\bigg(\frac{1}{n-1}\sum_i g(x, X_i)\bigg)\right]\right\|_\infty \\
&= O\left(n^{-\alpha_h/2}\right)\,,
\end{aligned}$$

as per the computation above for Proposition 9. Hence $\widetilde{D}_n = O\left(n^{-\alpha_h/2}\right)$.

Finally, $\psi = h$ is α_h-Holder so Item (iii) from Assumption 11 is satisfied with $\alpha_\psi = \alpha_h$.

Assumption 13 for Examples 4 and d. This is to be treated in a case-by-case manner. For the examples mentioned in Proposition 18.

Power mean. Consider $h = x \mapsto x^{1/p}$, and m nonnegative of dimension 1 (the reasoning can be done in each dimension). Define vectors S, S' of size $n-1$ such that $S_i = w_i m(f(x_i))$ and similarly for S'. Denoting by $\|S\|_p = \left(\sum_i S_i^p\right)^{1/p}$ the p-norm, we get

$$\begin{aligned}
&\left|F\left(f(x_1), \{\!\{(f(x_i), W(x_1, x_i))\}\!\}_{2\leqslant i\leqslant n}\right) - F\left(f(x_1), \{\!\{(f(x_i'), W(x_1, x_i'))\}\!\}_{2\leqslant i\leqslant n}\right)\right| \\
&= \frac{1}{(n-1)^{1/p}}\left|\|S\|_p - \|S'\|_p\right| \leqslant \frac{1}{(n-1)^{1/p}}\|S - S'\|_p \leqslant \|S - S'\|_\infty \\
&\leqslant K_m \max_i \|z_i - z_i'\|_\infty\,.
\end{aligned}$$

Hence $\alpha_F = 1$.

Lipschitz h^{-1}. We have

$$\begin{aligned}
&\left\|F\left(f(x_1), \{\!\{(f(x_i), W(x_1, x_i))\}\!\}_{2\leqslant i\leqslant n}\right) - F\left(f(x_1), \{\!\{(f(x_i'), W(x_1, x_i'))\}\!\}_{2\leqslant i\leqslant n}\right)\right\|_\infty \\
&\leqslant K_h\left(\frac{1}{n-1}\left\|\sum_i h^{-1}(w_i m(z_i)) - h^{-1}(w_i m(z_i'))\right\|_\infty\right)^{\alpha_h} \\
&\leqslant K_h(K_m \max_i \|z_i - z_i'\|_\infty)^{\alpha_h}\,,
\end{aligned}$$

since h^{-1} is Lipschitz.

Lipschitz h and h^{-1} It is the previous case with $\alpha_h = 1$.

C.5 Examples 5 and e: Convolutional Message-Passing with *max* aggregation

For this example, we need to prove Proposition 9 and that the Mcdiarmid's method fails, *i.e.,* that the bounded differences do not tend to zero.

Proposition 9 for Examples 5 and e. We call $g(x, y) = W(x, y)m(f(y))$. We start by the case when g is real-valued, since g is continuous and P is strictly positive, $\operatorname{ess\,sup}_P g(x, \cdot) = \sup g(x, \cdot)$ for all x by Lemma 27. Let $\varepsilon > 0$ and $x \in \mathcal{X}$. By definition of the supremum and by independence of the X_i, we have that

$$\begin{aligned} &\mathbb{P}(|\max_i g(x, X_i) - \sup g(x, \cdot)| \geqslant \varepsilon) \\ &= \mathbb{P}(\max_i g(x, X_i) \leqslant \sup g(x, \cdot) - \varepsilon) \\ &= \mathbb{P}(g(x, X_1) \leqslant \sup g(x, \cdot) - \varepsilon)^n \\ &= \mathbb{P}(|g(x, X_1) - \sup g(x, \cdot)| \geqslant \varepsilon)^n . \end{aligned} \tag{43}$$

By continuity and compactness, there is $x^* \in \mathcal{X}$ such that $\sup g(x, \cdot) = g(x, x^*)$, so Equation (43) is equal to

$$\begin{aligned} &\mathbb{P}(|g(x, X_1) - g(x, x^*)| \geqslant \varepsilon)^n \\ &= \left(1 - \mathbb{P}(|g(x, X_1) - g(x, x^*)| < \varepsilon)\right)^n . \end{aligned} \tag{44}$$

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 $|g(x, X_1) - g(x, x^*))| < \varepsilon$. Thus, Equation (44) is bounded from above by

$$(1 - \mathbb{P}(\|X_1 - x^*\| < \delta))^n = \left(1 - P(B(x^*, \delta) \cap \mathcal{X})\right)^n , \tag{45}$$

where $B(x^*, \delta)$ is the open ball of center x^* and radius δ in $\mathbb{R}^d$. To finish let us justify that the measure of the $B(x^*, \delta) \cap \mathcal{X}$ 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 arbitrarily small. By compactness, up to sub-sequence extraction, we can assume there is $(x_k) \in \mathcal{X}^{\mathbb{N}}$ such that $x_n \to x \in \mathcal{X}$ and $P(B(x_k, \delta) \cap \mathcal{X}) \leqslant 1/2^k$. Call $U = B(x, \delta/2) \cap \mathcal{X}$, there is rank k_0 such that $\forall k \geqslant k_0$, $x_k \in U$. Thus $U \subset B(x_k, \delta) \cap \mathcal{X}$ $\forall k \geqslant k_0$ yielding $P(U) \leqslant 1/2^k$ $\forall k \geqslant k_0$, that is, $P(U) = 0$. Impossible since U is a nonempty relative open set of $\mathcal{X}$.

So there is $\eta > 0$ independent of x such that $P(B(x^*, \delta) \cap \mathcal{X}) > \eta$ and, coming back to Equation (45):

$$\mathbb{P}(|\max_i g(x, X_i) - \sup g(x, \cdot)| \geqslant \varepsilon) \leqslant (1 - \eta)^n . \tag{46}$$

If g is vector-valued, say in $\mathbb{R}^q$, call $g_1, \ldots, g_q$ its components and η_k such that g_k satisfies Equation (46) with $\eta = \eta_k$. Then by a union bound we have

$$\mathbb{P}(\|\max_i g(x, X_i) - \sup g(x, \cdot)\|_\infty \geqslant \varepsilon) \leqslant \sum_{k=1}^{q} (1 - \eta_k)^n . \tag{47}$$

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

$$\begin{aligned} \|\mathbb{E}(\max_i g(x, X_i)) - \sup g(x, \cdot)\|_\infty &\leqslant \mathbb{E}(Z) \\ &= \mathbb{E}(Z\mathbb{1}_{Z\geqslant\varepsilon}) + \mathbb{E}(Z\mathbb{1}_{Z<\varepsilon}) \\ &\leqslant 2\|g\|_\infty \sum_{k=1}^{q} (1-\eta_k)^n + \varepsilon. \end{aligned} \tag{48}$$

Again, since the right-hand side does not depend on x, 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 nonempty open of $\mathcal{X}$,

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

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

Bounded differences are $\Omega(1)$. Here we check the bounded differences are $\Omega(1)$, *i.e.,* they do not tend to zero.

Call $g(x, y) = W(x, y)f^{(l-1)}(y)$, and $(g_1, \ldots, g_{d_l})$ 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 $x_1, \ldots, x_n$ such that $g(x, x_1), \ldots, g(x, x_n)$ are all strictly smaller that $g(x, x^*)$. Up to reordering them, we suppose $g(x, x_1) = \max_{2\leqslant i\leqslant n} g(x, x_i)$ and call $a = |g_1(x, x^*) - g_1(x, x_1)| > 0$.

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

Overall, for any n,

$$a < \sup \| \max\{g(x, x_1), \ldots, g(x, x_n)\} - \max\{g(x, x_1'), \ldots, \ldots, g(x, x_n')\}\|_\infty\,,$$

where the supremum is taken over $x, x_2, \ldots, x_n, x_2', \ldots, x_n' \in \mathcal{X}$ such that $(x_2, \ldots, x_n)$ and $(x_2', \ldots, x_n')$ differ from only one component. This proves that the bounded differences are $\Omega(1)$.

Appendix D. Useful results

In this section, we gather useful theorems, and other third-party lemmas.

Theorem 24 (McDiarmid inequality (Boucheron et al., 2013))**.** *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 $c_1, \ldots, c_n$. Then for any independent random variables $X_1, \ldots, X_n$ in $\mathcal{E}$, for any $\varepsilon > 0$:*

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

Notice that the X_i are not required to be identically distributed. By a union bound and reformulating Theorem 24 to a bound with high probability one can obtain the following result for function taking multidimensional values.

Corollary 25 (Multi-dimensional McDiarmid inequality). *Suppose that $f : \mathcal{E}^n \to \mathbb{R}^d$ satisfies a vectorial version on the bounded difference: $\|f(x) - f(x')\|_\infty \leqslant c_i$ whenever x and x' differ only from the ith component. Then for any independent random variables $X_1, \dots, X_n$ in $\mathcal{E}$, for any $\rho > 0$*

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

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

Corollary 26 (McDiarmid inequality with Holder transform). *Let $f : \mathcal{E}^n \to \mathcal{C} \subset \mathbb{R}^d$ and $\psi : \mathcal{C} \to \mathcal{D}$ an invertible function which is α-Hölder,*

$$\|\psi(y) - \psi(y')\|_\infty \leqslant K_\psi \|y - y'\|_\infty^\alpha\,,$$

and such that,

$$\|\psi^{-1}(f(x)) - \psi^{-1}(f(x'))\|_\infty \leqslant c_i$$

whenever x and x' differ only from the ith component, and

$$\|\psi(\mathbb{E}\psi^{-1}(f(X_1, \dots, X_n))) - \mathbb{E}(f(X_1, \dots, X_n))\|_\infty = u_n \to 0$$

for independent random variables $X_1, \dots, X_n$ in $\mathcal{E}$. Then, for any $\rho > 0$:

$$\|f(X_1, \dots, X_n) - \mathbb{E}(f(X_1, \dots, X_n))\|_\infty \leqslant K_\psi \left(\frac{1}{2}\sum_{i=1}^{n} c_i^2 \ln\left(\frac{2d}{\rho}\right)\right)^{\alpha/2} + u_n\,.$$

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

Proof We write $f(X_1, \dots, X_n) = f(X)$ and:

$$\begin{aligned}\|f(X) - \mathbb{E}(f(X))\|_\infty &\leqslant \|\psi(\psi^{-1}(f(X))) - \psi(\mathbb{E}\psi^{-1}(f(X)))\|_\infty \\ &\quad + \|\psi(\mathbb{E}\psi^{-1}(f(X))) - \psi(\psi^{-1}(\mathbb{E}f(X)))\|_\infty \\ &\leqslant K_\psi \|\psi^{-1}(f(X)) - \mathbb{E}\psi^{-1}(f(X))\|_\infty^\alpha + \|\psi\left(\mathbb{E}\psi^{-1}(f(X))\right) - \mathbb{E}f(X)\|_\infty\end{aligned}$$

We apply McDiarmid's inequality (Corollary 25) on the first term and bound the second by u_n to obtain the result. ■

Lemma 27. *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 nonempty. Then for any continuous map $f : \mathcal{X} \to \mathbb{R}$,*

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

Proof Clearly $\operatorname{ess\,sup}_P f \leqslant \sup f$ 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 a strictly positive measure, which yields a contradiction with the fact that $\operatorname{ess\,sup}_P f < M$. ■

Lemma 28. *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) $\|\max_i a_i\|_\infty \leqslant \max_i \|a_i\|_\infty$.

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

Where the maximums are the be understood component-wise.

Proof We start with Item (i). Suppose $m = 1$, trivially, $a_i \leqslant |a_i|$ which implies $\max_i a_i \leqslant \max_i |a_i|$, and therefore

$$|\max_i a_i| \leqslant |\max_i |a_i|| = \max_i |a_i| \,. \tag{49}$$

For $m \geqslant 1$, denote $a_i^{(k)}$ (resp. $b_i^{(k)}$) the kth coordinate of a_i (reap. b_i) for $1 \leqslant k \leqslant m$. Using Inequality (49), we easily verify that

$$\|\max_i a_i\|_\infty = \max_k |\max_i a_i^{(k)}| \leqslant \max_k \max_i |a_i^{(k)}| = \max_i \max_k |a_i^{(k)}| = \max_i \|a_i\|_\infty \,.$$

We turn to Item (ii). We start with the case $m = 1$. Let i_a (resp. i_b) be an index that realizes $\max_i a_i$ (resp. $\max_i b_i$), we have

$$\max_i a_i - \max_i b_i = a_{i_a} - b_{i_b} = a_{i_a} - b_{i_a} + b_{i_a} - b_{i_b} \,.$$

However, by definition of i_b, the last member $b_{i_a} - b_{i_b} \leqslant 0$, which yields

$$\max_i a_i - \max_i b_i \leqslant a_{i_a} - b_{i_a} \leqslant \max_i a_i - b_i \leqslant |\max_i a_i - b_i| \leqslant \max_i |a_i - b_i| \,,$$

where the last inequality follows from Inequality (49). Then, the same calculation will yield

$$\max_i b_i - \max_i a_i \leqslant \max_i |b_i - a_i| = \max_i |a_i - b_i| \,,$$

which proves the desired result. For $m \geqslant 1$, simply apply the previous result to each coordinate,

$$\begin{aligned}
\|\max_i a_i - \max_i b_i\|_\infty &= \max_k |\max_i a_i^{(k)} - \max_i b_i^{(k)}| \\
&\leqslant \max_k \max_i |a_i^{(k)} - b_i^{(k)}| \\
&\leqslant \max_i \max_k |a_i^{(k)} - b_i^{(k)}| \\
&= \max_i \|a_i - b_i\|_\infty \,.
\end{aligned}$$

■

Lemma 29. *Let $\mathcal{X}$ be compact and $g : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^m$ be K_g-Lipschitz continuous. Then, f defined by*

$$f : \left| \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathbb{R}^m \\ x & \longmapsto & \sup_{y \in \mathcal{X}} g(x, y) \end{array} \right. .$$

where the supremum is taken component-wise, is also K_g-Lipschitz continuous.

Proof Start with the case $m = 1$. Let $x, x' \in \mathcal{X}$, by continuity and compactness, the supremums defining $f(x)$ and $f(x')$ are reached, *i.e.,* there exist x^*, x'^* such that $f(x) = g(x, x^*)$ and $f(x') = g(x', x'^*)$. Then,

$$f(x) - f(x') = g(x, x^*) - g(x', x^*) + g(x', x^*) - g(x', x'^*) \,,$$

However, by definition of the supremum, $g(x', x^*) - g(x', x'^*) \leqslant 0$, which yields

$$f(x) - f(x') \leqslant g(x, x^*) - g(x', x^*) \leqslant K_g \|x - x'\|_\infty \,,$$

by Lipschitz continuity of g on $\mathcal{X}^2$. Permuting x and x' and repeating the same computation will yield to

$$f(x') - f(x) \leqslant K_g \|x - x'\|_\infty \,,$$

such that we obtain the desired Lipschitz condition on f.

For $m \geqslant 1$, denote g_i the components of g for $1 \leqslant i \leqslant m$. It is immediate that, since g is K_g-Lipschitz, each real-valued g_i is K_g-Lipschitz too. Therefore, according to the case $m = 1$, each $f_i \colon x \mapsto \sup_y g_i(x, y)$ is also K_g-Lipschitz,. We conclude that for any $x, x' \in \mathcal{X}$,

$$\|f(x) - f(x')\|_\infty = \max_i |f_i(x) - f_i(x')| \leqslant K_g \|x - x'\|_\infty \,.$$

■

Lemma 30. *Let $(X_n)_n$ be a sequence of random variables. If the series of general term $\mathbb{P}\left(|X_n| > \varepsilon\right)$ is summable for all $\varepsilon > 0$, then X_n tends to zero almost surely.*

Proof Let Ω be an abstract probability space which implicitly defines the random variables X_n. In the language of probabilities, the assertion "X_n tends to zero" is equivalent to,

$$\forall k \geqslant 1,\ \forall \omega \in \Omega,\ \exists N(\omega) \mid \forall n \geqslant N(\omega),\ |X_n(\omega)| \leqslant \frac{1}{k} \,.$$

This translates into the set-theoretic language of events by

$$(X_n \to 0) = \bigcap_{k \geqslant 1} \bigcup_{N \geqslant 0} \bigcap_{n \geqslant N} \left(|X_n| \leqslant \frac{1}{k} \right) \,.$$

We claim that this event is almost sure. Indeed,

$$\mathbb{P}\left(X_n \to 0\right) = 1 - \mathbb{P}\left(\bigcup_{k \geqslant 1} \bigcap_{N \geqslant 0} \bigcup_{n \geqslant N} \left(|X_n| > \frac{1}{k} \right) \right) = 1 - \mathbb{P}\left(\bigcup_{k \geqslant 1} \left(\overline{\lim}\, |X_n| > \frac{1}{k} \right) \right) \,.$$

However, by the Borel-Cantelli lemma, $\mathbb{P}\left(\overline{\lim}\, |X_n| > \epsilon\right) = 0$ for all ε, therefore, for $\varepsilon = 1/k$ in particular. Finally, since a countable union of null sets is null, we obtain the result. ■

References

Jan Böker, Ron Levie, Ningyuan Huang, Soledad Villar, and Christopher Morris. Fine-grained expressivity of graph neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: a nonasymptotic theory of independence*. Oxford university press, Oxford, 2013. ISBN 978-0-19-953525-5.

Peter S Bullen. *Handbook of means and their inequalities*, volume 560. Springer Science & Business Media, 2013.

Juan Cerviño, Luana Ruiz, and Alejandro Ribeiro. Learning by transference: Training graph neural networks on growing graphs. *IEEE Transactions on Signal Processing*, 71:233–247, 2023. doi: 10.1109/TSP.2023.3242374.

Frédéric Chazal, Brittany Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Subsampling methods for persistent homology. In *International Conference on Machine Learning*, pages 2143–2151. PMLR, 2015a.

Frédéric Chazal, Marc Glisse, Catherine Labruère, and Bertrand Michel. Convergence rates for persistence diagram estimation in topological data analysis. *Journal of Machine Learning Research*, 16(110):3603–3635, 2015b. URL `http://jmlr.org/papers/v16/chazal15a.html`.

Frédéric Chazal, Pascal Massart, and Bertrand Michel. Rates of convergence for robust geometric inference. *Electronic Journal of Statistics*, 10(2):2243 – 2286, 2016. doi: 10.1214/16-EJS1161. URL `https://doi.org/10.1214/16-EJS1161`.

Zhengdao Chen, Lisha Li, and Joan Bruna. Supervised community detection with line graph neural networks. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=H1g0Z3A9Fm`.

Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal neighbourhood aggregation for graph nets. *Advances in Neural Information Processing Systems*, 33:13260–13271, 2020.

H. Crane. *Probabilistic Foundations of Statistical Network Analysis*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. CRC Press, 2018. ISBN 9781351807326. URL `https://books.google.fr/books?id=LERnDwAAQBAJ`.

Antonio Cuevas. On pattern analysis in the non-convex case. *Kybernetes*, 19(6):26–33, 1990.

Antonio Cuevas and Ricardo Fraiman. A plug-in approach to support estimation. *The Annals of Statistics*, 25(6):2300–2312, 1997. ISSN 00905364. URL `http://www.jstor.org/stable/2959033`.

Antonio Cuevas and Alberto Rodríguez-Casal. On boundary estimation. *Advances in Applied Probability*, 36(2):340–354, 2004. doi: 10.1239/aap/1086957575.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989. doi: 10.1007/BF02551274.

Miguel de Carvalho. Mean, What do You Mean? *The American Statistician*, 70(3):270–274, July 2016. ISSN 0003-1305, 1537-2731. doi: 10.1080/00031305.2016.1148632. URL https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1148632.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/04df4d434d481c5bb723be1b6df1ee65-Paper.pdf.

Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR 2019 Workshop on Representation Learning on Graphs and Manifolds*, 2019. URL https://arxiv.org/abs/1903.02428.

Alex Fout, Jonathon Byrd, Basir Shariat, and Asa Ben-Hur. Protein interface prediction using graph convolutional networks. In *NIPS*, 2017.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/gilmer17a.html.

Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, Edoardo M Airoldi, et al. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

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.*, volume 2, pages 729–734 vol. 2, July 2005. doi: 10.1109/IJCNN.2005.1555942. ISSN: 2161-4407.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991. doi: 10.1016/0893-6080(91)90009-T. URL http://www.sciencedirect.com/science/article/pii/089360809190009T.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22118–22133.

Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/fb60d411a5c5b72b2e7d3527cfc84fd0-Paper.pdf.

Stefanie Jegelka. Theory of Graph Neural Networks: Representation and Learning, April 2022. URL http://arxiv.org/abs/2204.07697. arXiv:2204.07697 [cs, stat].

Nicolas Keriven. Not too little, not too much: a theoretical analysis of graph (over)smoothing. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 2268–2281. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/0f956ca6f667c62e0f71511773c86a59-Paper-Conference.pdf.

Nicolas Keriven and Gabriel Peyré. Universal invariant and equivariant graph neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/ea9268cb43f55d1d12380fb6ea5bf572-Paper.pdf.

Nicolas Keriven and Samuel Vaiter. What functions can graph neural networks compute on random graphs? the role of positional encoding. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 11823–11849. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/271ec4d1a9ff5e6b81a6e21d38b1ba96-Paper-Conference.pdf.

Nicolas Keriven, Alberto Bietti, and Samuel Vaiter. Convergence and Stability of Graph Convolutional Networks on Large Random Graphs. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21512–21523. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/f5a14d4963acf488e3a24780a84ac96c-Paper.pdf.

Nicolas Keriven, Alberto Bietti, and Samuel Vaiter. On the universality of graph neural networks on large random graphs. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6960–6971. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/38181d991caac98be8fb2ecb8bd0f166-Paper.pdf.

Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

A.N. Kolmogorov and G. Castelnuovo. *Sur la notion de la moyenne*. G. Bardi, tip. della R. Accad. dei Lincei, 1930. URL https://books.google.fr/books?id=iUqLnQEACAAJ.

Ryan Kortvelesy, Steven Morad, and Amanda Prorok. Generalised f-mean aggregation for graph neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=JMrIeKjTAe.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *Ann. Statist.*, 43(1), February 2015. ISSN 0090-5364. doi: 10.1214/14-AOS1274. URL `http://arxiv.org/abs/1312.2050`. arXiv: 1312.2050.

Ron Levie. A graphon-signal analysis of graph neural networks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 64482–64525. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/cb7943be26bb34f036c7e4068c490903-Paper-Conference.pdf`.

Ron Levie, Wei Huang, Lorenzo Bucci, Michael Bronstein, and Gitta Kutyniok. Transferability of spectral graph convolutional neural networks. *Journal of Machine Learning Research*, 22(272):1–59, 2021. URL `http://jmlr.org/papers/v22/20-213.html`.

László Lovász. *Large Networks and Graph Limits*, volume 60 of *Colloquium Publications*. American Mathematical Society, Providence, Rhode Island, December 2012. ISBN 978-0-8218-9085-1 978-1-4704-1583-9. doi: 10.1090/coll/060. URL `http://www.ams.org/coll/060`.

Stéphane Mallat. Group Invariant Scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012. ISSN 00103640. doi: 10.1002/cpa.21413.

Haggai Maron, Heli Ben-Hamu, Hadar Serviansky, and Yaron Lipman. Provably powerful graph networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/bb04af0f7ecaee4aae62035497da1387-Paper.pdf`.

Haggai Maron, Ethan Fetaya, Nimrod Segol, and Yaron Lipman. On the universality of invariant networks. *Proceedings of the 36th International Conference on Machine Learning*, 97, 2019b.

Sohir Maskey, Ron Levie, Yunseok Lee, and Gitta Kutyniok. Generalization analysis of message passing neural networks on large random graphs. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 4805–4817. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/1eeaae7c89d9484926db6974b6ece564-Paper-Conference.pdf`.

Sohir Maskey, Ron Levie, and Gitta Kutyniok. Transferability of graph neural networks: An extended graphon approach. *Applied and Computational Harmonic Analysis*, 63:48–83, 2023. ISSN 10635203. doi: 10.1016/j.acha.2022.11.008. URL `https://linkinghub.elsevier.com/retrieve/pii/S1063520322000987`.

Colin McDiarmid. *On the method of bounded differences*, page 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989. doi: 10.1017/CBO9781107359949.008.

Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4602–4609, 2019.

Christopher Morris, Fabrizio Frasca, Nadav Dym, Haggai Maron, İsmail İlkan Ceylan, Ron Levie, Derek Lim, Michael Bronstein, Martin Grohe, and Stefanie Jegelka. Future directions in the theory of graph machine learning. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=wBr5ozDEKp`.

Pál András Papp and Roger Wattenhofer. A theoretical comparison of graph neural network extensions. In *International Conference on Machine Learning*, pages 17323–17345. PMLR, 2022.

Luana Ruiz, Luiz Chamon, and Alejandro Ribeiro. Graphon neural networks and the transferability of graph neural networks. *Advances in Neural Information Processing Systems*, 33:1702–1712, 2020.

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

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2008.

Nicolas Tremblay, Paulo Gonçalves, and Pierre Borgnat. Design of graph filters and filterbanks. In Petar M. Djurić and Cédric Richard, editors, *Cooperative and Graph Signal Processing*, pages 299–324. Academic Press, June 2018. doi: 10.1016/B978-0-12-813677-5.00011-0. URL `https://inria.hal.science/hal-01675375`.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *6th International Conference on Learning Representations*, 2017.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Clément Vignac, Andreas Loukas, and Pascal Frossard. Building powerful and equivariant graph neural networks with structural message-passing. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14143–14155. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/a32d7eeaae19821fd9ce317f3ce952a7-Paper.pdf`.

Ulrike Von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *Annals of Statistics*, 36(2):555–586, 2008. ISSN 00905364. doi: 10.1214/009053607000000640.

Boris Weisfeiler and Andrei Leman. The reduction of a graph to canonical form and the algebra which appears therein. *nti, Series*, 2(9):12–16, 1968.

Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6861–6871. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/wu19e.html`.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Netw. Learning Syst.*, 32(1):4–24, January 2021. ISSN 2162-237X, 2162-2388. doi: 10.1109/TNNLS.2020.2978386. URL `http://arxiv.org/abs/1901.00596`. arXiv: 1901.00596.

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