Universal Invariant and Equivariant Graph Neural Networks

Nicolas Keriven
École Normale Supérieure
Paris, France
nicolas.keriven@ens.fr

Gabriel Peyré
CNRS and École Normale Supérieure
Paris, France
gabriel.peyre@ens.fr

Abstract

Graph Neural Networks (GNN) come in many flavors, but should always be either *invariant* (permutation of the nodes of the input graph does not affect the output) or *equivariant* (permutation of the input permutes the output). In this paper, we consider a specific class of invariant and equivariant networks, for which we prove new universality theorems. More precisely, we consider networks with a single hidden layer, obtained by summing channels formed by applying an equivariant linear operator, a pointwise non-linearity, and either an invariant or equivariant linear output layer. Recently, Maron et al. (2019b) showed that by allowing higher-order tensorization inside the network, universal *invariant* GNNs can be obtained. As a first contribution, we propose an alternative proof of this result, which relies on the Stone-Weierstrass theorem for algebra of real-valued functions. Our main contribution is then an extension of this result to the *equivariant* case, which appears in many practical applications but has been less studied from a theoretical point of view. The proof relies on a new generalized Stone-Weierstrass theorem for algebra of equivariant functions, which is of independent interest. Additionally, unlike many previous works that consider a fixed number of nodes, our results show that a GNN defined by a single set of parameters can approximate uniformly well a function defined on graphs of varying size.

1 Introduction

Designing Neural Networks (NN) to exhibit some *invariance* or *equivariance* to group operations is a central problem in machine learning (Shawe-Taylor, 1993). Among these, Graph Neural Networks (GNN) are primary examples that have gathered a lot of attention for a large range of applications. Indeed, since a graph is not changed by permutation of its nodes, GNNs must be either *invariant to permutation*, if they return a result that must not depend on the representation of the input, or *equivariant to permutation*, if the output must be permuted when the input is permuted, for instance when the network returns a *signal over the nodes* of the input graph. In this paper, we examine universal approximation theorems for invariant and equivariant GNNs.

From a theoretical point of view, invariant GNNs have been much more studied than their equivariant counterpart (see the following subsection). However, many practical applications deal with equivariance instead, such as community detection (Chen et al., 2019), recommender systems (Ying et al., 2018), interaction networks of physical systems (Battaglia et al., 2016), state prediction (Sanchez-Gonzalez et al., 2018), protein interface prediction (Fout et al., 2017), among many others. See (Zhou et al., 2018; Bronstein et al., 2017) for thorough reviews. It is therefore of great interest to increase our understanding of equivariant networks, in particular, by extending arguably one of the most classical result on neural networks, namely the universal approximation theorem for multi-layers perceptron (MLP) with a single hidden layer (Cybenko, 1989; Hornik et al., 1989; Pinkus, 1999).

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Maron et al. (2019b) recently proved that certain *invariant* GNNs were universal approximators of invariant continuous functions on graphs. The main goal of this paper is to extend this result to the *equivariant* case, for similar architectures.

Outline and contribution. The outline of our paper is as follows. After reviewing previous works and notations in the rest of the introduction, in Section 2 we provide an alternative proof of the result of (Maron et al., 2019b) for invariant GNNs (Theorem 1), which will serve as a basis for the equivariant case. It relies on a non-trivial application of the classical Stone-Weierstrass theorem for algebras of real-valued functions (recalled in Theorem 2). Then, as our main contribution, in Section 3 we prove this result for the equivariant case (Theorem 3), which to the best of our knowledge was not known before. The proof relies on a new version of Stone-Weierstrass theorem (Theorem 4). Unlike many works that consider a fixed number of nodes n, in both cases we will prove that a GNN described by a single set of parameters can approximate uniformly well a function that acts on graphs of varying size.

1.1 Previous works

The design of neural network architectures which are equivariant or invariant under group actions is an active area of research, see for instance (Ravanbakhsh et al., 2017; Gens and Domingos, 2014; Cohen and Welling, 2016) for finite groups and (Wood and Shawe-Taylor, 1996; Kondor and Trivedi, 2018) for infinite groups. We focus here our attention to discrete groups acting on the coordinates of the features, and more specifically to the action of the full set of permutations on tensors (order-1 tensors corresponding to sets, order-2 to graphs, order-3 to triangulations, etc).

Convolutional GNN. The most appealing construction of GNN architectures is through the use of local operators acting on vectors indexed by the vertices. Early definitions of these "message passing" architectures rely on fixed point iterations (Scarselli et al., 2009), while more recent constructions make use of non-linear functions of the adjacency matrix, for instance using spectral decompositions (Bruna et al., 2014) or polynomials (Defferrard et al., 2016). We refer to (Bronstein et al., 2017; Xu et al., 2019) for recent reviews. For regular-grid graphs, they match classical convolutional networks (LeCun et al., 1989) which by design can only approximate translation-invariant or equivariant functions (Yarotsky, 2018). It thus comes at no surprise that these convolutional GNN are not universal approximators (Xu et al., 2019) of permutation-invariant functions.

Fully-invariant GNN. Designing Graph (and their higher-dimensional generalizations) NN which are equivariant or invariant to the whole permutation group (as opposed to e.g. only translations) requires the use of a small sub-space of linear operators, which is identified in (Maron et al., 2019a). This generalizes several previous constructions, for instance for sets (Zaheer et al., 2017; Hartford et al., 2018) and points clouds (Qi et al., 2017). Universality results are known to hold in the special cases of sets, point clouds (Qi et al., 2017) and discrete measures (de Bie et al., 2019) networks.

In the *invariant* GNN case, the universality of architectures built using a single hidden layer of such equivariant operators followed by an invariant layer is proved in (Maron et al., 2019b) (see also (Kondor et al., 2018)). This is the closest work from our, and we will provide an alternative proof of this result in Section 2, as a basis for our main result in Section 3.

Universality in the equivariant case has been less studied. Most of the literature focuses on equivariance to *translation* and its relation to convolutions (Kondor et al., 2018; Cohen and Welling, 2016), which are ubiquitous in image processing. In this context, Yarotsky (2018) proved the universality of some translation-equivariant networks. Closer to our work, universality of NNs equivariant to permutations acting on point clouds has been recently proven in (Sannai et al., 2019), however their theorem does not allow for high-order inputs like graphs. It is the purpose of our paper to fill this missing piece and prove the universality of a class of equivariant GNNs for high-order inputs such as (hyper-)graphs.

1.2 Notations and definitions

Graphs. In this paper, (hyper-)graphs with n nodes are represented by tensors $G \in \mathbb{R}^{n^d}$ indexed by $1 \leqslant i_1, \dots, i_d \leqslant n$. For instance, "classical" graphs are represented by edge weight matrices ($d = 2$), and hyper-graphs by high-order tensors of "multi-edges" connecting more than two nodes.

Note that we do not impose G to be symmetric, or to contain only non-negative elements. In the rest of the paper, we fix some $d \geqslant 1$ for the order of the inputs, however we allow n to vary.

Permutations. Let $[n] \overset{\text{def.}}{=} \{1, \dots, n\}$. The set of permutations $\sigma : [n] \to [n]$ (bijections from $[n]$ to itself) is denoted by $\mathcal{O}_n$, or simply $\mathcal{O}$ when there is no ambiguity. Given a permutation σ and an order-k tensor $G \in \mathbb{R}^{n^k}$, a "permutation of nodes" on G is denoted by $\sigma \star G$ and defined as

$$(\sigma \star G)_{\sigma(i_1),\dots,\sigma(i_k)} = G_{i_1,\dots,i_k}.$$

We denote by $P_\sigma \in \{0,1\}^{n \times n}$ the permutation matrix corresponding to σ, or simply P when there is no ambiguity. For instance, for $G \in \mathbb{R}^{n^2}$ we have $\sigma \star G = PGP^\top$.

Two graphs G_1, G_2 are said isomorphic if there is a permutation σ such that $G_1 = \sigma \star G_2$. If $G = \sigma \star G$, we say that σ is a self-isomorphism of G. Finally, we denote by $\mathcal{O}(G) \overset{\text{def.}}{=} \{\sigma \star G \ ; \ \sigma \in \mathcal{O}\}$ the orbit of all the permuted versions of G.

Invariant and equivariant linear operators. A function $f : \mathbb{R}^{n^k} \to \mathbb{R}$ is said to be *invariant* if $f(\sigma \star G) = f(G)$ for every permutation σ. A function $f : \mathbb{R}^{n^k} \to \mathbb{R}^{n^\ell}$ is said to be *equivariant* if $f(\sigma \star G) = \sigma \star f(G)$. Our construction of GNNs alternates between *linear* operators that are invariant or equivariant to permutations, and non-linearities. Maron et al. (2019a) elegantly characterize all such linear functions, and prove that they live in vector spaces of dimension, respectively, exactly $b(k)$ and $b(k+\ell)$, where $b(i)$ is the i^{th} Bell number. An important corollary of this result is that the dimension of this space *does not depend on the number of nodes* n, but only on the order of the input and output tensors. Therefore one can parameterize linearly for all n such an operator by the same set of coefficients. For instance, a linear equivariant operator $F : \mathbb{R}^{n^2} \to \mathbb{R}^{n^2}$ from matrices to matrices is formed by a linear combination of $b(4) = 15$ basic operators such as "sum of rows replicated on the diagonal", "sum of columns replicated on the rows", and so on. The 15 coefficients used in this linear combination define the "same" linear operator for every n.

Invariant and equivariant Graph Neural Nets. As noted by Yarotsky (2018), it is in fact trivial to build invariant universal networks for finite groups of symmetry: just take a non-invariant universal architecture, and perform a group averaging. However, this holds little interest in practice, since the group of permutation is of size $n!$. Instead, researchers use architectures for which invariance is hard-coded into the construction of the network itself. The same remark holds for equivariance.

In this paper, we consider one-layer GNNs of the form:

$$f(G) = \sum_{s=1}^{S} H_s \Big[\rho(F_s[G] + B_s) \Big] + b, \tag{1}$$

where $F_s : \mathbb{R}^{n^d} \to \mathbb{R}^{n^{k_s}}$ are linear equivariant functions that yield k_s-tensors (i.e. they *potentially increase or decrease the order of the input tensor*), and H_s are invariant linear operators $H_s : \mathbb{R}^{n^{k_s}} \to \mathbb{R}$ (resp. equivariant linear operators $H_s : \mathbb{R}^{n^{k_s}} \to \mathbb{R}^n$), such that the GNN is globally invariant (resp. equivariant). The invariant case is studied in Section 2, and the equivariant in Section 3. The bias terms $B_s \in \mathbb{R}^{n^{k_s}}$ are equivariant, so that $B_s = \sigma \star B_s$ for all σ. They are also characterized by Maron et al. (2019a) and belong to a linear space of dimension $b(k_s)$. We illustrate this simple architecture in Fig. 1.

In light of the characterization by Maron et al. (2019a) of linear invariant and equivariant operators described in the previous paragraph, a GNN of the form (1) is described by $1+\sum_{s=1}^{S} b(k_s+d)+2b(k_s)$ parameters in the invariant case and $1 + \sum_{s=1}^{S} b(k_s + d) + b(k_s + 1) + b(k_s)$ in the equivariant. As mentioned earlier, this number of parameters does not depend on the number of nodes n, and a GNN described by a single set of parameters can be applied to graphs of any size. In particular, we are going to show that a GNN approximates uniformly well a continuous function for several n at once.

The function ρ is any locally Lipschitz pointwise non-linearity for which the Universal Approximation Theorem for MLP applies. We denote their set $\mathcal{F}_{\text{MLP}}$. This includes in particular any continuous function that is not a polynomial (Pinkus, 1999). Among these, we denote the sigmoid $\rho_{\text{sig}}(x) = e^x/(1+e^x)$.

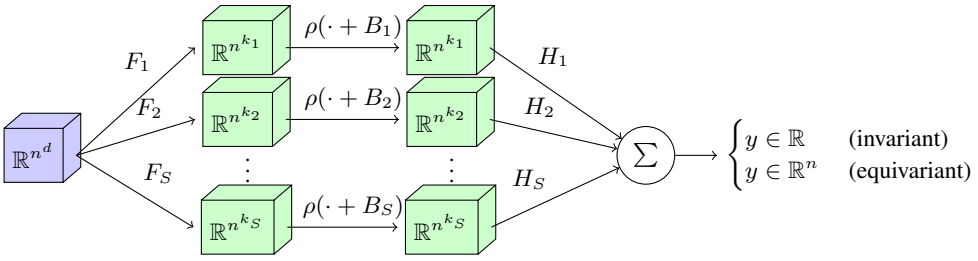

Figure 1: The model of GNNs studied in this paper. For each channel $s \leqslant S$, the input tensor is passed through an equivariant operator $F_s : \mathbb{R}^{n^d} \to \mathbb{R}^{n^{k_s}}$, a non-linearity with some added equivariant bias B_s, and a final operator H_s that is either invariant (Section 2) or equivariant (Section 3). These GNNs are universal approximators of invariant or equivariant continuous functions (Theorems 1 and 3).

We denote by $\mathcal{N}_{\text{inv.}}(\rho)$ (resp. $\mathcal{N}_{\text{eq.}}(\rho)$) the class of invariant (resp. equivariant) 1-layer networks of the form (1) (with S and k_s being arbitrarily large). Our contributions show that they are dense in the spaces of continuous invariant (resp. equivariant) functions.

2 The case of invariant functions

Maron et al. (2019b) recently proved that *invariant* GNNs similar to (1) are universal approximators of continuous invariant functions. As a warm-up, we propose an alternative proof of (a variant of) this result, that will serve as a basis for our main contribution, the equivariant case (Section 3).

Edit distance. For invariant functions, isomorphic graphs are undistinguishable, and therefore we work with a set of *equivalence classes* of graphs, where two graphs are equivalent if isomorphic. We define such a set for any number $n \leqslant n_{\max}$ of nodes and bounded G

$$\mathcal{G}_{\text{inv.}} \overset{\text{def.}}{=} \left\{ \mathcal{O}(G) \ ; \ G \in \mathbb{R}^{n^d} \text{ with } n \leqslant n_{\max}, \|G\| \leqslant R \right\},$$

where we recall that $\mathcal{O}(G) = \{\sigma \star G \ ; \ \sigma \in \mathcal{O}\}$ is the set of every permuted versions of G, here seen as an equivalence class.

We need to equip this set with a metric that takes into account graphs with different number of nodes. A distance often used in the literature is the *graph edit distance* (Sanfeliu and Fu, 1983). It relies on defining a set of elementary operations o and a cost $c(o)$ associated to each of them, here we consider node addition and edge weight modification. The distance is then defined as

$$d_{\text{edit}}(\mathcal{O}(G_1), \mathcal{O}(G_2)) \overset{\text{def.}}{=} \min_{(o_1,\dots,o_k) \in \mathcal{P}(G_1,G_2)} \sum_{i=1}^{k} c(o_i) \tag{2}$$

where $\mathcal{P}(G_1, G_2)$ contains every sequence of operation to transform G_1 into a graph isomorphic to G_2, or G_2 into G_1. Here we consider $c(node\_addition) = c$ for some constant $c > 0$, $c(edge\_weight\_change) = |w - w'|$ where the weight change is from w to w', and "edge" refers to any element of the tensor $G \in \mathbb{R}^{n^d}$. Note that, if we have $d_{\text{edit}}(\mathcal{O}(G_1), \mathcal{O}(G_2)) < c$, then G_1 and G_2 have the same number of nodes, and in that case $d_{\text{edit}}(\mathcal{O}(G_1), \mathcal{O}(G_2)) = \min_{\sigma \in \mathcal{O}_n} \|G_1 - \sigma \star G_2\|_1$, where $\|\cdot\|_1$ is the element-wise ℓ_1 norm, since each edge must be transformed into another.

We denote by $\mathcal{C}(\mathcal{G}_{\text{inv.}}, d_{\text{edit}})$ the space of real-valued functions on $\mathcal{G}_{\text{inv.}}$ that are continuous with respect to d_{edit}, equipped with the infinity norm of uniform convergence. We then have the following result.

Theorem 1. *For any $\rho \in \mathcal{F}_{\text{MLP}}$, $\mathcal{N}_{\text{inv.}}(\rho)$ is dense in $\mathcal{C}(\mathcal{G}_{\text{inv.}}, d_{edit})$.*

Comparison with (Maron et al., 2019b). A variant of Theorem 1 was proved in (Maron et al., 2019b). The two proofs are however different: their proof relies on the construction of a basis of invariant polynomials and on classical universality of MLPs, while our proof is a direct application of Stone-Weierstrass theorem for algebras of real-valued functions. See the next subsection for details.

One improvement of our result with respect to the one of (Maron et al., 2019b) is that it can handle graphs of varying sizes. As mentioned in the introduction, a single set of parameters defines a GNN that can be applied to graphs of any size. Theorem 1 shows that any continuous invariant function is *uniformly* well approximated by a GNN on the whole set $\mathcal{G}_{\text{inv.}}$, that is, for all numbers of nodes $n \leqslant n_{\max}$ simultaneously. On the contrary, Maron et al. (2019b) work with a fixed n, and it does not seem that their proof can extend easily to encompass several n at once. A weakness of our proof is that it does not provide an upper bound on the order of tensorization k_s. Indeed, through Noether's theorem on polynomials, the proof of Maron et al. (2019b) shows that $k_s \leqslant n^d(n^d-1)/2$ is sufficient for universality, which we cannot seem to deduce from our proof. Moreover, they provide a lower-bound $k_s \geqslant n^d$ below which universality cannot be achieved.

2.1 Sketch of proof of Theorem 1

The proof for the invariant case will serve as a basis for the equivariant case in the Section 3. It relies on Stone-Weierstrass theorem, which we recall below.

Theorem 2 (Stone-Weierstrass (Rudin (1991), Thm. 5.7))**.** *Suppose X is a compact Hausdorff space and A is a subalgebra of the space of continuous real-valued functions $\mathcal{C}(X)$ which contains a non-zero constant function. Then $\mathcal{A}$ is dense in $\mathcal{C}(X)$ if and only if it separates points, that is, for all $x \neq y$ in X there exists $f \in \mathcal{A}$ such that $f(x) \neq f(y)$.*

We will construct a class of GNNs that satisfy all these properties in $\mathcal{G}_{\text{inv.}}$. As we will see, unlike classical applications of this theorem to e.g. polynomials, the main difficulty here will be to prove the separation of points. We start by observing that $\mathcal{G}_{\text{inv.}}$ is indeed a compact set for d_{edit}.

Properties of $(\mathcal{G}_{\text{inv.}}, d_{\text{edit}})$. Let us first note that the metric space $(\mathcal{G}_{\text{inv.}}, d_{\text{edit}})$ is Hausdorff (i.e. separable, all metric spaces are). For each $\mathcal{O}(G_1), \mathcal{O}(G_2) \in \mathcal{G}_{\text{inv.}}$ we have: if $d_{\text{edit}}(\mathcal{O}(G_1), \mathcal{O}(G_2)) < c$, then the graphs have the same number of nodes, and in that case $d_{\text{edit}}(\mathcal{O}(G_1)\mathcal{O}(G_2)) \leqslant \|G_1 - G_2\|_1$. Therefore, the embedding $G \mapsto \mathcal{O}(G)$ is continuous (locally Lipschitz). As the continuous image of the compact $\bigcup_{n=1}^{n_{\max}} \left\{ G \in \mathbb{R}^{n^d} \;;\; \|G\| \leqslant R \right\}$, the set $\mathcal{G}_{\text{inv.}}$ is indeed compact.

Algebra of invariant GNNs. Unfortunately, $\mathcal{N}_{\text{inv.}}(\rho)$ is not a subalgebra. Following Hornik et al. (1989), we first need to extend it to be closed under multiplication. We do that by allowing Kronecker products inside the invariant functions:

$$f(G) = \sum_{s=1}^{S} H_s \Big[\rho\left(F_{s1}[G] + B_{s1}\right) \otimes \ldots \otimes \rho\left(F_{sT_s}[G] + B_{sT_s}\right) \Big] + b \tag{3}$$

where F_{st} yields k_{st}-tensors, $H_s : \mathbb{R}^{n^{\sum_t k_{st}}} \to \mathbb{R}$ are invariant, and B_{st} are equivariant bias. By $(\sigma \star G) \otimes (\sigma \star G') = \sigma \star (G \otimes G')$, they are indeed invariant. We denote by $\mathcal{N}^{\otimes}_{\text{inv.}}(\rho)$ the set of all GNNs of this form, with S, T_s, k_{st} arbitrarily large.

Lemma 1. *For any locally Lipschitz ρ, $\mathcal{N}^{\otimes}_{\text{inv.}}(\rho)$ is a subalgebra in $\mathcal{C}(\mathcal{G}_{\text{inv.}}, d_{edit})$.*

The proof, presented in Appendix A.1.1 follows from manipulations of Kronecker products.

Separability. The main difficulty in applying Stone-Weierstrass theorem is the separation of points, which we prove in the next Lemma.

Lemma 2. *$\mathcal{N}^{\otimes}_{\text{inv.}}(\rho_{sig})$ separates points.*

The proof, presented in Appendix A.1.2, proceeds by contradiction: we show that two graphs G, G' that coincides for every GNNs are necessarily permutation of each other. Applying Stone-Weierstrass theorem, we have thus proved that $\mathcal{N}^{\otimes}_{\text{inv.}}(\rho_{\text{sig}})$ is dense in $\mathcal{C}(\mathcal{G}_{\text{inv.}}, d_{\text{edit}})$.

Then, following Hornik et al. (1989), we go back to the original class $\mathcal{N}_{\text{inv.}}(\rho)$, by applying: (i) a Fourier approximation of ρ_{sig}, (ii) the fact that a product of $\cos$ is also a sum of $\cos$, and (iii) an approximation of $\cos$ by any other non-linearity. The following Lemma is proved in Appendix A.1.3, and concludes the proof of Thm 1.

Lemma 3. *We have the following: (i) $\mathcal{N}^{\otimes}_{\text{inv.}}(\cos)$ is dense in $\mathcal{N}^{\otimes}_{\text{inv.}}(\rho_{sig})$; (ii) $\mathcal{N}^{\otimes}_{\text{inv.}}(\cos) = \mathcal{N}_{\text{inv.}}(\cos)$; (iii) for any $\rho \in \mathcal{F}_{\text{MLP}}$, $\mathcal{N}_{\text{inv.}}(\rho)$ is dense in $\mathcal{N}_{\text{inv.}}(\cos)$.*

3 The case of equivariant functions

This section contains our main contribution. We examine the case of equivariant functions that return a vector $f(G) \in \mathbb{R}^n$ when G has n nodes, such that $f(\sigma \star G) = \sigma \star f(G)$. In that case, isomorphic graphs are not equivalent anymore. Hence we consider a compact set of graphs

$$\mathcal{G}_{\text{eq.}} \overset{\text{def.}}{=} \left\{ G \in \mathbb{R}^{n^d} \ ; \ n \leqslant n_{\max}, \|G\| \leqslant R \right\},$$

Like the invariant case, we consider several numbers of nodes $n \leqslant n_{\max}$ and will prove uniform approximation over them. We do not use the edit distance but a simpler metric:

$$d(G, G') = \begin{cases} \|G - G'\| & \text{if } G \text{ and } G' \text{ have the same number of nodes,} \\ \infty & \text{otherwise.} \end{cases}$$

for any norm $\|\cdot\|$ on $\mathbb{R}^{n^d}$.

The set of equivariant continuous functions is denoted by $\mathcal{C}_{\text{eq.}}(\mathcal{G}_{\text{eq.}}, d)$, equipped with the infinity norm $\|f\|_\infty = \sup_{G \in \mathcal{G}_{\text{eq.}}} \|f(G)\|_\infty$. We recall that $\mathcal{N}_{\text{eq.}}(\rho) \subset \mathcal{C}_{\text{eq.}}(\mathcal{G}_{\text{eq.}}, d)$ denotes one-layer GNNs of the form (1), with equivariant output operators H_s. Our main result is the following.

Theorem 3. *For any $\rho \in \mathcal{F}_{\text{MLP}}$, $\mathcal{N}_{\text{eq.}}(\rho)$ is dense in $\mathcal{C}_{\text{eq.}}(\mathcal{G}_{\text{eq.}}, d)$.*

The proof, detailed in the next section, follows closely the previous proof for invariant functions, but is significantly more involved. Indeed, the classical version of Stone-Weierstrass only provides density of a subalgebra of functions in the *whole space* of continuous functions, while in this case $\mathcal{C}_{\text{eq.}}(\mathcal{G}_{\text{eq.}}, d)$ is *already* a particular subset of continuous functions. On the other hand, it seems difficult to make use of fully general versions of Stone-Weierstrass theorem, for which some questions are still open (Glimm, 1960). Hence we prove a new, specialized Stone-Weierstrass theorem for equivariant functions (Theorem 4), obtained with a non-trivial adaptation of the constructive proof by Brosowski and Deutsch (1981).

Like the invariant case, our theorem proves uniform approximation for all numbers of nodes $n \leqslant n_{\max}$ at once by a single GNN. As is detailed in the next subsection, our proof of the generalized Stone-Weierstrass theorem relies on being able to *sort* the coordinates of the output space $\mathbb{R}^n$, and therefore our current proof technique does not extend to high-order *output* $\mathbb{R}^{n^\ell}$ (graph to graph mappings), which we leave for future work. For the same reason, while the previous invariant case could be easily extended to invariance to *subgroups* of $\mathcal{O}_n$, as is done by Maron et al. (2019b), for the equivariant case our theorem only applies when considering the full permuation group $\mathcal{O}_n$. Nevertheless, our generalized Stone-Weierstrass theorem may be applicable in other contexts where equivariance to permutation is a desirable property.

Comparison with (Sannai et al., 2019). Sannai et al. (2019) recently proved that equivariant NNs acting *on point clouds* are universal, that is, for $d = 1$ in our notations. Despite the apparent similarity with our result, there is a fundamental obstruction to extending their proof to high-order input tensors like graphs. Indeed, it strongly relies on Theorem 2 of (Zaheer et al., 2017) that characterizes invariant functions $\mathbb{R}^n \to \mathbb{R}$, which is no longer valid for high-order inputs.

3.1 Sketch of proof of Theorem 3: an equivariant version of Stone-Weierstrass theorem

We first need to introduce a few more notations. For a subset $I \subset [n]$, we define $\mathcal{O}_I \overset{\text{def.}}{=} \{\sigma \in \mathcal{O}_n \ ; \ \exists i \in I, j \in I^c, \sigma(i) = j \text{ or } \sigma(j) = i\}$ the set of permutations that exchange at least one index between I and I^c. Indexing of vectors (or multivariate functions) is denoted by brackets, e.g. $[x]_I$ or $[f]_I$, and inequalities $x \geqslant a$ are to be understood element-wise.

A new Stone-Weierstrass theorem. We define the "multiplication" of two multivariate functions using the Hadamard product $\odot$, i.e. the component-wise multiplication. Since $(\sigma \star x) \odot (\sigma \star x') = \sigma \star (x \odot x')$, it is easy to see that $\mathcal{C}_{\text{eq.}}(\mathcal{G}_{\text{eq.}}, d)$ is closed under multiplication, and is therefore a (strict) subalgebra of the set of all continuous functions that return a vector in $\mathbb{R}^n$ for an input graph with n nodes. As mentioned before, because of this last fact we cannot directly apply Stone-Weierstrass theorem. We therefore prove a new generalized version.

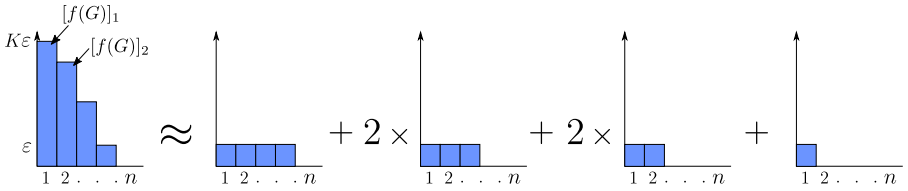

Figure 2: Illustration of strategy of proof for the equivariant Stone-Weierstrass theorem (Theorem 4). Considering a function f that we are trying to approximate and a graph G for which the coordinates of $f(G)$ are sorted by decreasing order, we approximate $f(G)$ by summing step-functions f_i, whose first coordinates are close to 1, and otherwise close to 0.

Theorem 4 (Stone-Weierstrass for equivariant functions)**.** *Let $\mathcal{A}$ be a subalgebra of $\mathcal{C}_{\text{eq.}}(\mathcal{G}_{\text{eq.}}, d)$, such that $\mathcal{A}$ contains the constant function $\mathbf{1}$ and:*

- (Separability) *for all $G, G' \in \mathcal{G}_{\text{eq.}}$ with number of nodes respectively n and n' such that $G \notin \mathcal{O}(G')$, for any $k \in [n]$, $k' \in [n']$, there exists $f \in \mathcal{A}$ such that $[f(G)]_k \neq [f(G')]_{k'}$;*
- ("Self"-separability) *for all number of nodes $n \leqslant n_{\max}$, $I \subset [n]$, $G \in \mathcal{G}_{\text{eq.}}$ with n nodes that has no self-isomorphism in $\mathcal{O}_I$, and $k \in I, \ell \in I^c$, there is $f \in \mathcal{A}$ such that $[f(G)]_k \neq [f(G)]_\ell$.*

Then $\mathcal{A}$ is dense in $\mathcal{C}_{\text{eq.}}(\mathcal{G}_{\text{eq.}}, d)$.

In addition to a "separability" hypothesis, which is similar to the classical one, Theorem 4 requires a "self"-separability condition, which guarantees that $f(G)$ can have different values on its coordinates under appropriate assumptions on G. We give below an overview of the proof of Theorem 4, the full details can be found in Appendix B.

Our proof is inspired by the one for the classical Stone-Weierstrass theorem (Thm. 2) of Brosowski and Deutsch (1981). Let us first give a bit of intuition on this earlier proof. It relies on the explicit construction of "step"-functions: given two disjoint closed sets A and B, they show that $\mathcal{A}$ contains functions that are approximately 0 on A and approximately 1 on B. Then, given a function $f : X \to \mathbb{R}$ (non-negative w.l.o.g.) that we are trying to approximate and $\varepsilon > 0$, they define $A_k = \{x \; ; \; f(x) \leqslant (k - 1/3)\varepsilon\}$ and $B_k = \{x \; ; \; f(x) \geqslant (k + 1/3)\varepsilon\}$ as the lower (resp. upper) *level sets* of f for a grid of values with precision ε. Then, taking the step-functions f_k between A_k and B_k, it is easy to prove that f is well-approximated by $g = \varepsilon \sum_k f_k$, since for each x only the right number of f_k is close to 1, the others are close to 0.

The situation is more complicated in our case. Given a function $f \in \mathcal{C}_{\text{eq.}}(\mathcal{G}_{\text{eq.}}, d)$ that we want to approximate, we work in the compact subset of $\mathcal{G}_{\text{eq.}}$ where the coordinates of f are *ordered*, since by permutation it covers every case: $\mathcal{G}_f \overset{\text{def.}}{=} \left\{ G \in \mathcal{G}_{\text{eq.}} \; ; \; \text{if } G \in \mathbb{R}^{n^d} \text{: } [f(G)]_1 \geqslant [f(G)]_2 \geqslant \ldots \geqslant [f(G)]_n \right\}$. Then, we will prove the existence of step-functions such that: when A and B satisfy some appropriate hypotheses, the step-function is close to 0 on A, and *only the first coordinates are close to* 1 on B, the others are close to 0. Indeed, by combining such functions, we can approximate a vector of ordered coordinates (Fig. 2). The construction of such step-functions is done in Lemma 7. Finally, we consider modified level-sets

$$A_k^{n,\ell} \overset{\text{def.}}{=} \left\{ G \in \mathcal{G}_f \cap \mathbb{R}^{n^d} \; ; \; [f(G)]_\ell - [f(G)]_{\ell+1} \leqslant (k - 1/3)\varepsilon \right\} \cup \bigcup_{n' \neq n} \left(\mathcal{G}_f \cap \mathbb{R}^{(n')^d} \right),$$

$$B_k^{n,\ell} \overset{\text{def.}}{=} \left\{ G \in \mathcal{G}_f \cap \mathbb{R}^{n^d} \; ; \; [f(G)]_\ell - [f(G)]_{\ell+1} \geqslant (k + 1/3)\varepsilon \right\}$$

that distinguish "jumps" between (ordered) coordinates. We define the associated step-functions $f_k^{n,\ell}$, and show that $g = \varepsilon \sum_{k,n,\ell} f_k^{n,\ell}$ is a valid approximation of f.

End of the proof. The rest of the proof of Theorem 3 is similar to the invariant case. We first build an algebra of GNNs, again by considering nets of the form (3), where we replace the H_s's by equivariant linear operators in this case. We denote this space by $\mathcal{N}_{\text{eq.}}^{\otimes}(\rho)$.

Lemma 4. *$\mathcal{N}_{\text{eq.}}^{\otimes}(\rho)$ is a subalgebra of $\mathcal{C}_{\text{eq.}}(\mathcal{G}_{\text{eq.}}, d)$.*

The proof, presented in Appendix A.2.1, is very similar to that of Lemma 1. Then we show the two separation conditions for equivariant GNNs.

Lemma 5. *$\mathcal{N}^{\otimes}_{\text{eq.}}(\rho_{sig.})$ satisfies both the separability and self-separability conditions.*

The proof is presented in Appendix A.2.2. The "normal" separability is in fact equivalent to the previous one (Lemma 2), since we can construct an equivariant network by simply stacking an invariant network on every coordinate. The self-separability condition is proved in a similar way. Finally we go back to $\mathcal{N}_{\text{eq.}}(\rho)$ in exactly the same way. The proof of Lemma 6 is exactly similar to that of Lemma 3 and is omitted.

Lemma 6. *We have the following: (i) $\mathcal{N}^{\otimes}_{\text{eq.}}(\cos)$ is dense in $\mathcal{N}^{\otimes}_{\text{eq.}}(\rho_{sig})$; (ii) $\mathcal{N}^{\otimes}_{\text{eq.}}(\cos) = \mathcal{N}_{\text{eq.}}(\cos)$; (iii) for any $\rho \in \mathcal{F}_{\text{MLP}}$, $\mathcal{N}_{\text{eq.}}(\rho)$ is dense in $\mathcal{N}_{\text{eq.}}(\cos)$.*

4 Numerical illustrations

This section provides numerical illustrations of our findings on simple synthetic examples. The goal is to examine the impact of the tensorization orders k_s and the width S. The code is available at `https://github.com/nkeriven/univgnn`. We emphasize that the contribution of the present paper is first and foremost theoretical, and that, like MLPs with a single hidden layer, we cannot expect the shallow GNNs (1) to be state-of-the-art and compete with deep models, despite their universality. A benchmarking of *deep* GNNs that use invariant and equivariant linear operators is done in (Maron et al., 2019a).

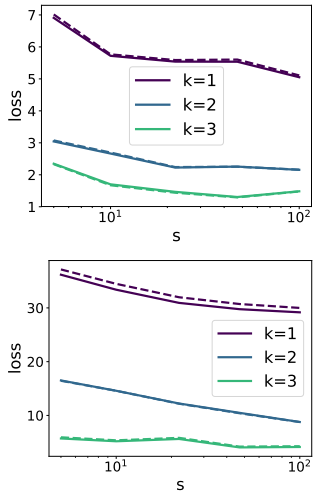

Figure 3: MSE results after 150 epochs, in the invariant (top) and equivariant (bottom) cases, averaged over 5 experiments. Dashed lines represent the testing error.

We consider graphs, represented using their adjacency matrices (i.e. 2-ways tensor, so that $d = 2$). The synthetic graphs are drawn uniformly among 5 graph topologies (complete graph, star, cycle, path or wheel) with edge weights drawn independently as the absolute value of a centered Gaussian variable. Since our approximation results are valid for several graph sizes simultaneously, both training and testing datasets contain $1.4 \cdot 10^4$ graphs, half with 5 nodes and half with 10 nodes. The training is performed by minimizing a square Euclidean loss (MSE) on the training dataset. The minimization is performed by stochastic gradient descent using the ADAM optimizer (Kingma and Ba, 2014). We consider two different regression tasks: (i) in the invariant case, the scalar to predict is the geodesic diameter of the graph, (ii) in the equivariant case, the vector to predict assigns to each node the length of the longest shortest-path emanating from it. While these functions can be computed using polynomial time all-pairs shortest paths algorithms, they are highly non-local, and are thus challenging to learn using neural network architectures. The GNNs (1) are implemented with a fixed tensorization order $k_s = k \in \{1, 2, 3\}$ and $\rho = \rho_{\text{sig.}}$.

Figure 3 shows that, on these two cases, when increasing the width S, the out-of-sample prediction error quickly stagnates (and sometime increasing too much S can slightly degrade performances by making the training harder). In sharp contrast, increasing the tensorization order k has a significant impact and lowers this optimal error value. This support the fact that universality relies on the use of higher tensorization order. This is a promising direction of research to integrate higher order tensors withing deeper architecture to better capture complex functions on graphs.

5 Conclusion

In this paper, we proved the universality of a class of one hidden layer equivariant networks. Handling this vector-valued setting required to extend the classical Stone-Weierstrass theorem. It remains an open problem to extend this technique of proof for more general equivariant networks whose outputs are graph-valued, which are useful for instance to model dynamic graphs using recurrent architectures (Battaglia et al., 2016). Another outstanding open question, formulated in (Maron et al., 2019b), is the characterization of the approximation power of networks whose tensorization orders k_s inside the layers are bounded, since they are much more likely to be implemented on large graphs in practice.

References

P. W. Battaglia, R. Pascanu, M. Lai, D. Rezende, and K. Kavukcuoglu. Interaction Networks for Learning about Objects, Relations and Physics. In *Advances in Neural Information and Processing Systems (NIPS)*, pages 4509–4517, 2016.

M. M. Bronstein, J. Bruna, Y. Lecun, A. Szlam, and P. Vandergheynst. Geometric Deep Learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

B. Brosowski and F. Deutsch. An elementary proof of the Stone-Weierstrass Theorem. *Proceedings of the American Mathematical Society*, 81(1):89–92, 1981.

J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun. Spectral Networks and Locally Connected Networks on Graphs. In *ICLR*, pages 1–14, 2014.

Z. Chen, X. Li, and J. Bruna. Supervised Community Detection with Line Graph Neural Networks. In *ICLR*, 2019.

T. Cohen and M. Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999, 2016.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mahematics of Control, Signals, and Systems*, 2:303–314, 1989.

G. de Bie, G. Peyré, and M. Cuturi. Stochastic deep networks. In *Proceedings of ICML 2019*, 2019.

M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In *Advances in Neural Information and Processing Systems (NIPS)*, 2016.

A. Fout, B. Shariat, J. Byrd, and A. Ben-Hur. Protein Interface Prediction using Graph Convolutional Networks. *Nips*, (Nips):6512–6521, 2017.

R. Gens and P. M. Domingos. Deep symmetry networks. In *Advances in neural information processing systems*, pages 2537–2545, 2014.

J. Glimm. A Stone-Weierstrass Theorem for C * -Algebras. *Annals of Mathematics*, 72(2):216–244, 1960.

J. Hartford, D. R. Graham, K. Leyton-Brown, and S. Ravanbakhsh. Deep models of interactions across sets. *arXiv preprint arXiv:1803.02879*, 2018.

K. Hornik, M. Stinchcombe, and H. White. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, 2:359–366, 1989.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2014.

R. Kondor and S. Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. *arXiv preprint arXiv:1802.03690*, 2018.

R. Kondor, H. T. Son, H. Pan, B. Anderson, and S. Trivedi. Covariant compositional networks for learning graphs. *arXiv preprint arXiv:1801.02144*, 2018.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

H. Maron, H. Ben-Hamu, N. Shamir, and Y. Lipman. Invariant and Equivariant Graph Networks. In *ICLR*, pages 1–13, 2019a.

H. Maron, E. Fetaya, N. Segol, and Y. Lipman. On the Universality of Invariant Networks. In *International Conference on Machine Learning (ICML)*, 2019b.

A. Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8(May): 143–195, 1999.

C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.

S. Ravanbakhsh, J. Schneider, and B. Poczos. Equivariance through parameter-sharing. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2892–2901. JMLR. org, 2017.

W. Rudin. *Functional Analysis*. 1991.

A. Sanchez-Gonzalez, N. Heess, J. T. Springenberg, J. Merel, M. Riedmiller, R. Hadsell, and P. Battaglia. Graph networks as learnable physics engines for inference and control. *arxiv:1806.01242*, 2018.

A. Sanfeliu and K.-S. Fu. A distance measure between attributed relational graphs for pattern recognition. *IEEE transactions on systems, man, and cybernetics*, (3):353–362, 1983.

A. Sannai, Y. Takai, and M. Cordonnier. Universal approximations of permutation invariant/equivariant functions by deep neural networks. *ArXiv: 1903.01939*, 2019.

F. Scarselli, M. Gori, A. C. Tsoi, G. Monfardini, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

J. Shawe-Taylor. Symmetries and discriminability in feedforward network architectures. *IEEE Transactions on Neural Networks*, 4(5):816–826, 1993.

J. Wood and J. Shawe-Taylor. Representation theory and invariant neural networks. *Discrete applied mathematics*, 69(1-2):33–60, 1996.

K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How Powerful are Graph Neural Networks? In *ICLR*, pages 1–15, 2019.

D. Yarotsky. Universal approximations of invariant maps by neural networks. *ArXiv: 1804.10306*, pages 1–64, 2018.

R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec. Graph Convolutional Neural Networks for Web-Scale Recommender Systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 974–983, 2018.

M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola. Deep sets. In *Advances in neural information processing systems*, pages 3391–3401, 2017.

J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun. Graph Neural Networks: A Review of Methods and Applications. *ArXiv: 1812.08434*, pages 1–20, 2018.

A Proofs

A.1 Invariant case

A.1.1 Proof of Lemma 1

We first prove that invariant GNNs are continuous with respect to d_{edit}. For two graphs G_1, G_2 such that $d_{\text{edit}}(\mathcal{O}(G_1), \mathcal{O}(G_2)) < c$, the graphs have the same number of nodes. Using the fact that ρ, H, F are (locally) Lipschitz in this case, we have $|f(G_1) - f(G_2)| \lesssim \|G_1 - G_2\|_1$, and by invariance by permutation:

$$|f(\mathcal{O}(G_1)) - f(\mathcal{O}(G_2))| \lesssim \min_{\sigma} \|G_1 - \sigma \star G_2\|_1 = d_{\text{edit}}(\mathcal{O}(G_1), \mathcal{O}(G_2))$$

and therefore we have indeed $\mathcal{N}^{\otimes}_{\text{inv.}}(\rho) \subset \mathcal{C}(\mathcal{G}_{\text{inv.}}, d_{\text{edit}})$.

Since $\mathcal{N}^{\otimes}_{\text{inv.}}(\rho)$ is obviously a vector space, we must now prove that it is closed by multiplication. For that, it is sufficient to prove that, for two invariant linear operators $H_1 : \mathbb{R}^{n^{k_1}} \to \mathbb{R}$ and $H_2 : \mathbb{R}^{n^{k_1}} \to \mathbb{R}$, there exists an invariant linear operator $H_3 : \mathbb{R}^{n^{k_1+k_2}} \to \mathbb{R}$ such that $H_1[G_1]H_2[G_2] = H_3[G_1 \otimes G_2]$. For this we recall that $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ and $\text{vec}(A) \otimes \text{vec}(B) = \text{vec}(A \otimes B)$, and thus that

$$\begin{aligned} H_1[G_1]H_2[G_2] &= \Big(\text{vec}(H_1)^\top \text{vec}(G_1)\Big)\Big(\text{vec}(H_2)^\top \text{vec}(G_2)\Big) \\ &= (\text{vec}(H_1)^\top \otimes \text{vec}(H_2)^\top)(\text{vec}(G_1) \otimes \text{vec}(G_2)) \\ &= \text{vec}(H_1 \otimes H_2)^\top \text{vec}(G_1 \otimes G_2) \end{aligned}$$

Hence we can define $H_3 = H_1 \otimes H_2$ and check that it is invariant by permutation. By Maron et al. (2019a) a necessary and sufficient condition is $P^{\otimes k_1+k_2}\text{vec}(H_3) = \text{vec}(H_3)$, which we can easily check:

$$P^{\otimes(k_1+k_2)}\text{vec}(H_3) = (P^{\otimes k_1}\text{vec}(H_1)) \otimes (P^{\otimes k_2}\text{vec}(H_2)) = \text{vec}(H_3)$$

since $P^{\otimes k_i}\text{vec}(H_i) = \text{vec}(H_i)$. □

A.1.2 Proof of Lemma 2

We proceed by contradiction, and show that if $f(\mathcal{O}(G)) = f(\mathcal{O}(G'))$ for any $f \in \mathcal{N}^{\otimes}_{\text{inv.}}(\rho_{\text{sig}})$, then $\mathcal{O}(G) = \mathcal{O}(G')$, i.e. G and G' are permutation of each other. Let G, G' be any two such graphs.

The first step if to show that G and G' have the same number of nodes $n = n'$. Consider $\tau = \min_{i_1,\dots,i_d}(\min(G_{i_1,\dots,i_d}, G'_{i_1,\dots,i_d})) - 1$ the minimal element of both G and G' minus 1, and the following family of networks:

$$f_\lambda(G) = H\left[\rho_{\text{sig}}\left(\lambda(G - \tau\mathbf{1}^{\otimes d})\right)\right] \quad \text{with} \quad H[z] = \sum_{i_1,\dots,i_d} z_{i_1,\dots,i_d}\,.$$

By letting $\lambda \to \infty$, the sigmoid produces 1 for every element in G that is above τ, that is, every element in G or G'. Hence we have $f_\lambda(G) \xrightarrow[\lambda\to\infty]{} n^d$ and $f_\lambda(G') \xrightarrow[\lambda\to\infty]{} (n')^d$, and therefore $n = n'$.

Then, we show similarly that the multiset (that is, set with multiplicity) of $\{G_{i_1,\dots,i_d}\}$ is the same as the multiset of $\{G'_{i_1,\dots,i_d}\}$. Consider them ordered: $G_{i_1^{(1)}\dots i_d^{(1)}} \leqslant \dots \leqslant G_{i_1^{(N)}\dots i_d^{(N)}}$, and $G'_{j_1^{(1)}\dots j_d^{(1)}} \leqslant \dots \leqslant G'_{i_1^{(N)}\dots i_d^{(N)}}$, where $N = n^d$. Then, by contradiction, if there is a q such that $G_{i_1^{(q)}\dots i_d^{(q)}} \neq G'_{j_1^{(q)}\dots j_d^{(q)}}$, say $G_{i_1^{(q)}\dots i_d^{(q)}} < G'_{j_1^{(q)}\dots j_d^{(q)}}$ w.l.o.g., set $\tau = (G_{i_1^{(q)}\dots i_d^{(q)}} + G'_{j_1^{(q)}\dots j_d^{(q)}})/2$. Then, for $\lambda > 0$, consider the same neural networks as above with this τ. Again, by letting $\lambda \to \infty$, the sigmoid produces 1 for every element in G that is above τ, and 0 otherwise. Hence $f_\lambda(G) \xrightarrow[\lambda\to\infty]{} n^d - q$, and $f_\lambda(G') \xrightarrow[\lambda\to\infty]{} n^d - q + 1$, which is a contradiction. Hence $G_{i_1^{(q)}\dots i_d^{(q)}} = G'_{j_1^{(q)}\dots j_d^{(q)}}$ for every q, and G and G' are formed by the same multiset of n^d real numbers.

Consider now the tensors $A = \rho_{\text{sig}}(G)$, $A' = \rho_{\text{sig}}(G')$ which have strictly positive elements. Since ρ_{sig} is a 1-to-1 mapping in $\mathbb{R}$, producing a permutation between A, A' yields a permutation for G, G'

and allow us to conclude. We consider the following class of neural nets in $\mathcal{N}^{\otimes}_{\text{inv.}}(\rho_{\text{sig}})$:

$$f(G) = H[A^{\otimes k}]$$

for every integer $k > 0$ and invariant H. Recall that $A^{\otimes k}$ is an dk-order tensor indexed such that

$$(A^{\otimes k})_{(i_{11},\dots,i_{1d}),\dots,(i_{k1},\dots,i_{kd})} = \prod_{\ell=1}^{k} a_{i_{\ell 1},\dots,i_{\ell d}}$$

for any $1 \leqslant i_{\ell q} \leqslant n$. Then, for any fixed set of such indices, it is not difficult to see that a valid invariant operator is the following:

$$H[A^{\otimes k}] = \sum_{\sigma \in \mathcal{O}_n} \prod_{\ell=1}^{k} a_{\sigma(i_{\ell 1}),\dots,\sigma(i_{\ell d})}$$

where $\mathcal{O}_n$ is the set of all permutations. Indeed, for all $\bar{\sigma} \in \mathcal{O}_n$:

$$\begin{aligned} H[(\bar{\sigma} \star A)^{\otimes k}] &= \sum_{\sigma \in \mathcal{O}_n} \prod_{\ell=1}^{k} a_{\bar{\sigma}^{-1}\sigma(i_{\ell 1}),\dots,\bar{\sigma}^{-1}\sigma(i_{\ell d})} \\ &= \sum_{\sigma \in \mathcal{O}_n} \prod_{\ell=1}^{k} a_{\sigma(i_{\ell 1}),\dots,\sigma(i_{\ell d})} = H[A^{\otimes k}] \end{aligned}$$

by a simple change of variable in the sum $\sum_{\sigma \in \mathcal{O}_n}$. In the same spirit, for any set of integers $k_{i_1,\dots,i_d} \geqslant 0$ where $1 \leqslant i_q \leqslant n$, the following is a valid invariant GNN in $\mathcal{N}^{\otimes}_{\text{inv.}}(\rho_{\text{sig}})$:

$$f(G) = H[A^{\otimes \sum_{i_1,\dots,i_d} k_{i_1,\dots,i_d}}] = \sum_{\sigma \in \mathcal{O}_n} \prod_{i_1,\dots,i_d=1}^{n} a^{k_{i_1,\dots,i_d}}_{\sigma(i_1),\dots,\sigma(i_d)}$$

Hence, we have that for any $k_{i_1,\dots,i_d}$:

$$\sum_{\sigma \in \mathcal{O}_n} \prod_{i_1,\dots,i_d=1}^{n} a^{k_{i_1,\dots,i_d}}_{\sigma(i_1),\dots,\sigma(i_d)} = \sum_{\sigma \in \mathcal{O}_n} \prod_{i_1,\dots,i_d=1}^{n} (a')^{k_{i_1,\dots,i_d}}_{\sigma(i_1),\dots,\sigma(i_d)}$$

Recalling that $\{a_{i_1,\dots,i_d}\}$ and $\{a'_{i_1,\dots,i_d}\}$ are the same multiset, we can apply Lemma 11 in Appendix C, which yields a permutation σ such that $a_{i_1,\dots,i_d} = a'_{\sigma(i_1),\dots,\sigma(i_d)}$ and concludes the proof. □

A.1.3 Proof of Lemma 3

(i) Consider any function in $\mathcal{N}^{\otimes}_{\text{inv.}}(\rho_{\text{sig}})$

$$f(G) = \sum_{s=1}^{S} H_s\Big[\rho_{\text{sig}}(F_{s1}[G] + B_{s1}) \otimes \dots \otimes \rho_{\text{sig}}(F_{sT_s}[G] + B_{sT_s})\Big] + b$$

and any $\varepsilon > 0$.

Given that we are on a bounded domain, there exists M such that $\sup_G \max_{s,t} \|F_{st}[G] + B_{st}\|_\infty \leqslant M$ for all s (where $\|\cdot\|_\infty$ is element-wise maximum). The Fourier development of ρ_{sig} on $[-M, M]$ yields that there exist a_i, b_i, c_i, $i \leqslant N$, such that for all $u \in [-M, M]$

$$\left|\rho_{\text{sig}}(u) - \sum_{i=1}^{N} a_i \cos(b_i u + c_i)\right| \leqslant \varepsilon$$

Defining

$$\begin{aligned} f_{st}(G) &= \rho_{\text{sig}}(F_{st}[G] + B_{st})\,, \\ h_{st}(G) &= \sum_{i=1}^{N} a_i \cos\left(b_i(F_{st}[G] + B_{st}) + c_i 1^{\otimes 2k_{st}}\right)\,, \end{aligned}$$

we have

$$\sup_G \max_{s,t} \|f_{st}(G) - h_{st}(G)\|_\infty \leqslant \varepsilon$$

Hence, for any s, is we define $e_t = \|f_{s1}(G) \otimes \ldots \otimes f_{st}(G) - h_{s1}(G) \otimes \ldots \otimes h_{st}(G)\|_\infty$, we have

$$\begin{aligned} e_{T_s} &\leqslant \|f_{s1}(G) \otimes \ldots \otimes f_{sT_s-1}(G) \otimes (f_{sT_s}(G) - h_{sT_s}(G))\|_\infty \\ &\quad + \|((f_{s1}(G) \otimes \ldots \otimes f_{sT_s-1}(G) - h_{s1}(G) \otimes \ldots \otimes h_{sT_s-1}(G)) \otimes h_{sT_s}(G)\|_\infty \\ &\leqslant \varepsilon + (1+\varepsilon)e_{T_s-1} \leqslant 3^{T_s} e_1 \leqslant 3^{T_s}\varepsilon \end{aligned}$$

Since the H_s are linear in finite dimension they are bounded operators and we call L_s such that $|H_s(W)| \leqslant L_s \|W\|_\infty$. Finally, if we define $g \in \mathcal{N}^{\otimes}_{\text{inv.}}(\cos)$ by

$$g(G) = \sum_{s=1}^{S} H_s\left[h_{s1}(G) \otimes \ldots \otimes h_{sT_s}(G)\right]$$

we have proved that we have $\sup_G |f(G) - g(G)| \leqslant (\sum_s L_s 3^{T_s})\varepsilon$, which concludes the proof.

(ii) The proof is based on the fact that $\cos(a)\cos(b) = \cos(a+b) + \cos(a-b)$. Hence:

$$\begin{aligned} &\cos(F_1[G] + B_1) \otimes \cos(F_2[G] + B_2) \\ &\quad = \Big(\cos(F_1[G] + B_1) \otimes 1_{n^{k_2}} 1_{n^{k_2}}^\top\Big) \odot \Big(1_{n^{k_1}} 1_{n^{k_1}}^\top \otimes \cos(F_2[G] + B_2)\Big) \\ &\quad = \cos(\bar{F}_1[G] + \bar{B}_1 + \bar{F}_2[G] + \bar{B}_2) \\ &\qquad\qquad + \cos(\bar{F}_1[G] + \bar{B}_1 - \bar{F}_2[G] - \bar{B}_2) \end{aligned}$$

where $\bar{F}_1[G] = F_1[G] \otimes 1_{n^{k_2}} 1_{n^{k_2}}^\top$ and $\bar{F}_2[G] = 1_{n^{k_1}} 1_{n^{k_1}}^\top \otimes F_2[G]$ and similarly for $\bar{B}_i$. Since $11^\top$ is invariant by permutation, it is easy to see that the $\bar{F}_i$ are equivariant linear functions outputting a $k_1 + k_2$-tensor, and $\bar{B}_i$ are equivariant biases, which proves the result.

(iii) Since $\rho \in \mathcal{F}_{\text{MLP}}$ and the universal approximation theorem applies, the cosine function on a compact of $\mathbb{R}$ can be uniformly approximated by a linear combination of ρ:

$$\sup_{x \in [-MM]} \left|\cos(x) - \sum_{i=1}^{N} a_i \rho(b_i x + c_i)\right| \leqslant \varepsilon$$

The rest of the proof is similar to (i).

□

A.2 Equivariant case

A.2.1 Proof of Lemma 4

Again we must prove that $\mathcal{N}^{\otimes}_{\text{eq.}}(\rho)$ is closed by "multiplication", that is, Hadamard product. For that, it is sufficient to show that for two equivariant linear operators $H_1 : \mathbb{R}^{n^k} \to \mathbb{R}^n$, $H_2 : \mathbb{R}^{n^\ell} \to \mathbb{R}^n$, there exists an equivariant linear operator $H_3 : \mathbb{R}^{n^{k+\ell}} \to \mathbb{R}^n$ such that

$$H_1[G_1] \odot H_2[G_2] = H_3[G_1 \otimes G_2]$$

For that, writing the matrices $H_1 \in \mathbb{R}^{n^k \times n}$ and $H_2 \in \mathbb{R}^{n^\ell \times n}$ by abuse of notation, we have

$$H_1[G_1] \odot H_2[G_2] = \operatorname{diag}\left(H_1 \operatorname{vec}(G_1) \operatorname{vec}(G_2)^\top H_2^\top\right)$$

Then, defining $\operatorname{mat}_{k,\ell}$ the operator that transforms a tensor $G \in \mathbb{R}^{n^{k+\ell}}$ to a $\mathbb{R}^{n^k \times n^\ell}$ matrix and the linear operator $H_3[G] = \operatorname{diag}\left(H_1 \operatorname{mat}_{k,\ell}(G) H_2^\top\right)$, we have indeed that $H_1[G_1] \odot H_2[G_2] = H_3[G_1 \otimes G_2]$. Then, for any permutation σ and corresponding matrix P, since $H_1 P^{\otimes k} = P H_1$, $H_2 P^{\otimes \ell} = P H_2$, and $\operatorname{mat}_{k,\ell}(\sigma \star G) = P^{\otimes k} \operatorname{mat}_{k,\ell}(G)(P^\top)^{\otimes \ell}$, we have

$$\begin{aligned} H_3[\sigma \star G] &= \operatorname{diag}\left(H_1 \operatorname{mat}_{k,\ell}(\sigma \star G) H_2^\top\right) \\ &= \operatorname{diag}\left(H_1 P^{\otimes k} \operatorname{mat}_{k,\ell}(G)(P^\top)^{\otimes \ell} H_2^\top\right) \\ &= \operatorname{diag}\left(P H_1 \operatorname{mat}_{k,\ell}(G) H_2^\top P^\top\right) = P H_3[G] \end{aligned}$$

and therefore H_3 is equivariant, which concludes the proof. □

A.2.2 Proof of Lemma 5

Separability. The separability condition is in fact exactly equivalent to the invariant case: indeed, we can construct linear equivariant operators H_s just by stacking linear invariant operators on every coordinate. Hence, for any invariant GNN $f \in \mathcal{N}_{\text{inv.}}^{\otimes}(\rho_{\text{sig.}})$, $h = [f, \dots, f] \in \mathcal{N}_{\text{eq.}}^{\otimes}(\rho_{\text{sig.}})$ is a valid equivariant operator. Hence, for any two graphs G, G' such that are not permutation of each other, by Lemma 2 there is $f \in \mathcal{N}_{\text{inv.}}^{\otimes}(\rho_{\text{sig.}})$ such that $f(G) \neq f(G')$, and by considering $h = [f, \dots, f]$ every coordinate of $h(G)$ is different from that of $h(G')$.

Self-separability. For the self-separability, consider any $G \in \mathcal{G}_{\text{eq.}}$ with n nodes, and any $I \subset [n]$. Once again we proceed by contradiction: we are going to show that if there exist $k \in I, \ell \in I^c$ such that for all $h \in \mathcal{N}_{\text{eq.}}^{\otimes}(\rho_{\text{sig.}})$ we have $[h(G)]_k = [h(G)]_\ell$, then $G \in \mathcal{G}_{\text{eq.}}(\mathcal{O}_I)$. Let G be such a graph, with the corresponding fixed k, ℓ.

Similar to the proof of the separability in the invariant case, we define $A = \rho_{\text{sig.}}(G)$, again keeping in mind that the sigmoid in a one-to-one mapping. Then, for any $k_{i_1,\dots,i_d}$, recall that the following is a valid *invariant* GNN:

$$H[A^{\otimes \sum_{i_1,\dots,i_d} k_{i_1,\dots,i_d}}] = \sum_{\sigma \in \mathcal{O}_n} \prod_{i_1,\dots,i_d=1}^{n} a_{\sigma(i_1),\dots,\sigma(i_d)}^{k_{i_1,\dots,i_d}}$$

Similarly, we are going to show that the following defines a valid equivariant GNN:

$$[f(G)]_q = \left[H[A^{\otimes \sum_{i_1,\dots,i_d} k_{i_1,\dots,i_d}}]\right]_q = \sum_{\sigma \in \mathcal{O}^{(q)}} \prod_{i_1,\dots,i_d=1}^{n} a_{\sigma(i_1),\dots,\sigma(i_d)}^{k_{i_1,\dots,i_d}}$$

where $\mathcal{O}^{(q)} \overset{\text{def.}}{=} \{\sigma \in \mathcal{O} \,;\, \sigma(k) = q\}$ (where we recall that k, ℓ are fixed and part of the hypothesis we have made on G). Indeed, for any permutation $\bar{\sigma}$, we have

$$\begin{aligned}
[f(\bar{\sigma} \star G)]_{\bar{\sigma}(q)} &= \sum_{\sigma \in \mathcal{O}^{(\sigma(q))}} \prod_{i_1,\dots,i_d=1}^{n} a_{\bar{\sigma}^{-1}(\sigma(i_1)),\dots,\bar{\sigma}^{-1}(\sigma(i_d))}^{k_{i_1,\dots,i_d}} \\
&= \sum_{\sigma \in \mathcal{O}, \bar{\sigma}^{-1}(\sigma(k)) = q} \prod_{i_1,\dots,i_d=1}^{n} a_{\bar{\sigma}^{-1}(\sigma(i_1)),\dots,\bar{\sigma}^{-1}(\sigma(i_d))}^{k_{i_1,\dots,i_d}} \\
&= \sum_{\sigma \in \mathcal{O}, \sigma(k) = q} \prod_{i_1,\dots,i_d=1}^{n} a_{\sigma(i_1),\dots,\sigma(i_d)}^{k_{i_1,\dots,i_d}} = [f(G)]_q
\end{aligned}$$

Hence, we have indeed $f(\bar{\sigma} \star G) = \bar{\sigma} \star f(G)$, and f is equivariant. Now, by hypothesis on G, it means that for all $k_{i_1,\dots,i_d}$, we have:

$$\sum_{\sigma \in \mathcal{O}^{(k)}} \prod_{i_1,\dots,i_d=1}^{n} a_{\sigma(i_1),\dots,\sigma(i_d)}^{k_{i_1,\dots,i_d}} = \sum_{\sigma \in \mathcal{O}^{(\ell)}} \prod_{i_1,\dots,i_d=1}^{n} a_{\sigma(i_1),\dots,\sigma(i_d)}^{k_{i_1,\dots,i_d}} .$$

Now, since $\mathcal{O}^{(k)}$ contains the identity and has the same cardinality as $\mathcal{O}^{(\ell)}$, by Lemma 11 it means that there is a permutation $\sigma \in \mathcal{O}^{(\ell)}$ such that $G = \sigma \star G$. Observing that $\mathcal{O}^{(\ell)} \subset \mathcal{O}_I$ concludes the proof. □

B Adapted Stone-Weierstrass theorem: proof of Theorem 4

Let us first introduce some more notations. For $\mathcal{O}' \subset \mathcal{O}$ and $\mathcal{G}$ a set of graphs, we define

$$\mathcal{O}'(\mathcal{G}) \overset{\text{def.}}{=} \{\sigma \star G \,;\, \sigma \in \mathcal{O}', G \in \mathcal{G}\}$$
$$\mathcal{G}(\mathcal{O}') \overset{\text{def.}}{=} \{G \in \mathcal{G} \,;\, \exists \sigma \in \mathcal{O}', G = \sigma \star G\}$$

that is, respectively, the set of permuted graphs in $\mathcal{G}$, and the set of graphs in $\mathcal{G}$ that have a self-isomorphism in $\mathcal{O}'$. Recall that we denote by $[f]_I$ and $[x]_I$ indexation of multivariate functions and vectors, and that inequalities $x \geqslant a$ are element-wise. A neighborhood of x is an open set V such that $x \in V$. Finally, for convenience we denote $\mathcal{G}_{\text{eq.}}^{(n)} = \mathcal{G}_{\text{eq.}} \cap \mathbb{R}^{n^d}$ the graphs in $\mathcal{G}_{\text{eq.}}$ that have n nodes.

As described in the paper, the key lemma is the construction of *step-functions*.

Lemma 7 (Existence of step-functions)**.** *Let $n \leqslant n_{\max}$, and $I \subset [n]$ be any subset of indices. Let $A \subset \mathcal{G}_{\text{eq.}}, B \subset \mathcal{G}_{\text{eq.}}^{(n)}$ be two closed sets such that $B \cap \mathcal{G}_{\text{eq.}}^{(n)}(\mathcal{O}_I) = \emptyset$ and $B \cap \mathcal{O}(A) = \emptyset$, that is, graphs in B have no self-isomorphism in $\mathcal{O}_I$ and no two graphs between A and B are isomorphic. Then, for all $\varepsilon > 0$, there exists $f \in \mathcal{A}$ such that:*

$$\begin{cases} \forall G, & 0 \leqslant f(G) \leqslant 1 \\ \forall G \in B, & [f(G)]_I \geqslant 1 - \varepsilon \quad \textit{and} \quad [f(G)]_{I^c} \leqslant \varepsilon \\ \forall G \in A, & f(G) \leqslant \varepsilon \end{cases}$$

We start the proof by a serie of three intermediate lemmas.

Lemma 8. *Let $n \leqslant n_{\max}$, and $I \subset [n]$ be any subset of indices. Let $G_0 \in \mathcal{G}_{\text{eq.}}^{(n)}$ such that $G \notin \mathcal{G}_{\text{eq.}}^{(n)}(\mathcal{O}_I)$, and T be a closed subset of $\mathcal{G}_{\text{eq.}}$ such that $T \cap \mathcal{O}(G_0) = \emptyset$. Then, there exists $V(G_0) \subset \mathbb{R}^{n^d}$ a neighborhood of G_0 such that the following holds: for all $\varepsilon > 0$, there exists $f \in \mathcal{A}$ such that:*

$$\begin{cases} \forall G, & f(G) \in [0,1] \\ \forall G \in V(G_0), & [f(G)]_I \geqslant 1 - \varepsilon \quad \textit{and} \quad [f(G)]_{I^c} \leqslant \varepsilon \\ \forall G \in T, & f(G) \leqslant \varepsilon \end{cases}$$

Proof. Our goal is to build a function $g \in \mathcal{A}$ along with a threshold $\delta > 0$ and $V(G_0)$ a neighborhood of G_0 such that:

$$\begin{cases} \forall G \in T, & g(G) \geqslant \delta \\ \forall G \in V(G_0), & [g(G)]_I \leqslant \delta/2 \text{ and } [g(G)]_{I^c} \geqslant \delta \end{cases}$$

Then we can conclude similarly to the end of the proof of Lemma 1 in (Brosowski and Deutsch, 1981).

Take any $k \in I, \ell \in I^c$ and $G \in T$. Note that G does not necessarily have n nodes, we denote n_G its number of nodes. Let $i \in [n_G]$ be any index. According to the two separability hypotheses, there exists $g_{G,k,i}, h_{k,\ell} \in \mathcal{A}$ such that $[g_{G,k,i}(G_0)]_k \neq [g_{G,k,i}(G)]_i$ and $[h_{k,\ell}(G_0)]_k \neq [h_{k,\ell}(G_0)]_\ell$. Then, consider

$$g_G = \prod_{k \in I} \left(\frac{1}{n_G} \sum_{i=1}^{n_G} \frac{(g_{G,k,i} - [g_{G,k,i}(G_0)]_k \mathbf{1})^2}{\|g_{G,k,i} - [g_{G,k,i}(G_0)]_k \mathbf{1}\|_\infty^2} \right) \in \mathcal{A}$$

$$h = \prod_{k \in I} \left(\frac{1}{|I^c|} \sum_{\ell \in I^c} \frac{(h_{k,\ell} - [h_{k,\ell}(G_0)]_k \mathbf{1})^2}{\|h_{k,\ell} - [h_{k,\ell}(G_0)]_k \mathbf{1}\|_\infty^2} \right) \in \mathcal{A}$$

where $\prod, (\cdot)^2$ are to be understood component-wise and $\|g\|_\infty = \sup_G \|g(G)\|_\infty$. These functions satisfy

$$\begin{cases} g_G, h \in [0,1]\,, \\ [g_G(G_0)]_I = [h(G_0)]_I = 0 \\ g_G(G) > 0, [h(G_0)]_{I^c} > 0 \end{cases}$$

By continuity, define $S(G) \subset \mathbb{R}^{n_G^d}$ a neighborhood of G such that $g_G > 0$ on $S(G)$. By compacity of T, there is a finite number of $G_1, \ldots, G_m$ such that $T \subset \bigcup_i S(G_i)$. Then, we define $g = \frac{1}{m+1}(\sum_i g_{G_i} + h) \in \mathcal{A}$, which satisfies:

$$\begin{cases} g \in [0,1] \\ g > 0 \text{ on } T \\ [g(G_0)]_I = 0 \text{ and } [g(G_0)]_{I^c} > 0\,. \end{cases}$$

Again, by compacity of T, there exists $\delta > 0$ such that $g \geqslant \delta$ on T and $[g(G_0)]_{I^c} \geqslant 2\delta$. Then, by continuity, we define $V(G_0)$ a neighborhood of G_0 such that $[g]_I \leqslant \delta/2$ and $[g]_{I^c} \geqslant \delta$ on $V(G_0)$.

We can now conclude. Assuming that $\delta < 1$ is small enough without lost of generality, let k be an integer such that $1/\delta < k < 2/\delta$, and define the following functions in $\mathcal{A}$:

$$q_p = (1 - g^p)^{k^p}$$

which are obviously such that $q_p \in [0,1]$.

Then, using the elementary Bernoulli inequality $(1+h)^p \geqslant 1+ph$ for all $h \geqslant -1$, we have for all $G \in V(G_0)$ and $i \in I$:

$$q_p(G)_i \geqslant 1 - (kg_i(G))^p \geqslant 1 - (k\delta/2)^p \xrightarrow[p\to\infty]{} 1$$

and similarly, for either $G \in T$ and any i, or $G \in V(G_0)$ and $i \in I^c$, we have

$$\begin{aligned} q_p(G)_i &\leqslant \frac{1+(kg_i(G))^p}{(kg_i(G))^p}(1-g_i(G)^p)^{k^p} \leqslant \frac{(1+g_i(G)^p)^{k^p}}{(kg_i(G))^p}(1-g_i(G)^p)^{k^p} \text{ by Bernoulli's inequality} \\ &= \frac{(1-g_i(G)^{2p})^{k^p}}{(kg_i(G))^p} \leqslant \frac{1}{(k\delta)^p} \xrightarrow[p\to\infty]{} 0 \end{aligned}$$

Hence, for all $\varepsilon > 0$, there exists an p such that $q_p \leqslant \varepsilon$ on T, $[q_p]_{I^c} \leqslant \varepsilon$ and $[q_p]_I \geqslant 1-\varepsilon$ on $V(G_0)$. Taking $f = 1 - q_p$ concludes the proof. □

A similar result without the interval I is the following.

Lemma 9. *Let G_0 be any graph and T be a closed subset of $\mathcal{G}_{\text{eq.}}$ such that $T \cap \mathcal{O}(G_0) = \emptyset$. Then, there exists $V(G_0)$ a neighborhood of G_0 such that the following holds: for all $\varepsilon > 0$, there exists $f \in \mathcal{A}$ such that:*

$$\begin{cases} \forall G, & f(G) \in [0,1] \\ \forall G \in V(G_0), & f(G) \geqslant 1-\varepsilon \\ \forall G \in T, & f(G) \leqslant \varepsilon \end{cases}$$

Proof. The proof is similar (but simpler) to that of Lemma 8, without introduction of the interval I and the function h. □

An easy consequence of the above Lemma is the following.

Lemma 10. *Let A, B be two closed sets such that $B \cap \mathcal{O}(A) = \emptyset$. Then, for all $\varepsilon > 0$, there exists $f \in \mathcal{A}$ such that:*

$$\begin{cases} \forall G, & f(G) \in [0,1] \\ \forall G \in B, & f(G) \geqslant 1-\varepsilon \\ \forall G \in A, & f(G) \leqslant \varepsilon \end{cases}$$

Proof. Let $G \in B$. By hypothesis, $A \cap \mathcal{O}(G) = \emptyset$, so by Lemma 9 there exists $V(G)$ a neighborhood of G such that for all $\varepsilon > 0$ there exists $f \in \mathcal{A}$ satisfying: $0 \leqslant f \leqslant 1$, $f \geqslant 1-\varepsilon$ on $V(G)$, and $f \leqslant \varepsilon$ on A. By compacity of B, there is a finite number of $G_1, \dots, G_m \in B$ such that $B \subset \cup_{i=1}^m V(G_i)$. Denote by f_i the associated functions produced by Lemma 9 for some $\varepsilon' > 0$, and denote $f = \prod_i (1-f_i)$. We have that $f \leqslant \varepsilon'$ on B and $f \geqslant (1-\varepsilon')^m$ on A. Hence by choosing appropriately ε' (note that ε' is authorized to depend on m), we obtain a function f such that $f \leqslant \varepsilon$ on B and $f \geqslant 1-\varepsilon$ on A, and taking $1-f$ concludes the proof. □

We can now show Lemma 7.

Proof of Lemma 7. Let $G \in B \subset \mathcal{G}_{\text{eq.}}^{(n)}$. By hypothesis, $G \notin \mathcal{G}_{\text{eq.}}^{((n))}(\mathcal{O}_I)$ and $A \cap \mathcal{O}(G) = \emptyset$, so by Lemma 8 there exists $V(G) \subset \mathbb{R}^{n^d}$ a neighborhood of G such that for all $\varepsilon > 0$ there exists $f \in \mathcal{A}$ satisfying:

$$\begin{aligned} &0 \leqslant f \leqslant 1 \\ &[f]_I \geqslant 1-\varepsilon \text{ and } [f]_{I^c} \leqslant \varepsilon \text{ on } V(G) \\ &f \leqslant \varepsilon \text{ on } A\,. \end{aligned}$$

By compacity of B, there is a finite number of $G_1, \dots, G_m \in B$ such that $B \subset \cup_{i=1}^m V(G_i)$. For some $\varepsilon > 0$ that we will choose later, denote the associated functions $f_1, \dots, f_m$ (note that the $V(G_i)$ do not depend on ε, but the f_i do).

We remark that we cannot just consider the function $\prod_i f_i$ and conclude: indeed, on each $V(G_i)$ only f_i will satisfy $[f_i]_I \geqslant 1-\varepsilon$, and the others f_j are not guaranteed to be lower bounded. For the same reason, we cannot consider $\frac{1}{m}\sum_i f_i$ either, due to the requirement that $[f]_{I^c} \leqslant \varepsilon$ on B. We need to introduce auxiliary functions $\tilde{f}_i$ such that we are guaranteed that for each $j \neq i$, $[\tilde{f}_j]_I \geqslant 1-\varepsilon$ on $V(G_i)$, and we can conclude with $\prod_i (f_i + \tilde{f}_i)$. We will construct such functions with Lemma 10. A final difficulty is that $V(G_i)$ are open sets, while Lemma 10 can only work with closed sets.

Hence, by continuity, consider the neighborhoods $V'(G_i) \subset \mathbb{R}^{n^d}$ such that

$$\begin{aligned}&\overline{V(G_i)} \subset V'(G_i)\\ &[f_i]_I \geqslant 1-2\varepsilon \text{ and } [f_i]_{I^c} \leqslant 2\varepsilon \text{ on } V'(G_i)\,.\end{aligned}$$

Note that the $V(G_i)$ do not depend on ε, but the $V'(G_i)$ do.

Then, for all $i \in [n]$ consider the closed sets $A_i = A \cup \overline{V(G_i)}$ and $B_i = B \backslash \mathcal{O}(V'(G_i))$. By construction of B_i and hypothesis on A, we have indeed that $\mathcal{O}(A_i) \cap B_i = \emptyset$, since $\overline{V(G_i)} \subset V'(G_i)$. Applying Lemma 10, we obtain a function $\tilde{f}_i \in \mathcal{A}$ such that $\tilde{f} \leqslant \varepsilon$ on A_i and $\tilde{f} \geqslant 1-\varepsilon$ on B_i.

Finally, consider the following function: $f = \frac{1}{2^m}\prod_i (f_i + \tilde{f}_i)$. Take any $G \in B$. Consider the index i such that $G \in V(G_i)$. We have $[f_i(G) + \tilde{f}_i(G)]_I \geqslant 1-\varepsilon$ by definition of f_i and $[f_i(G) + \tilde{f}_i(G)]_{I^c} \leqslant 2\varepsilon$ by definition of f_i and $\tilde{f}_i$ and the fact that $G \in V(G_i) \subset A_i$. For any $j \neq i$, we have the following: either $G \in \mathcal{O}(V'(G_j))$, in which case, by equivariance of f_j and the fact that $G \notin \mathcal{G}_{\text{eq.}}(\mathcal{O}_I)$, we have $[f_j(G)]_I \geqslant 1-2\varepsilon$; or $G \in B_j$, in which case $[\tilde{f}_j(G)]_I \geqslant 1-2\varepsilon$. Overall, we obtain that

$$\begin{aligned}&[f]_I \geqslant \frac{1}{2^m}(1-2\varepsilon)^m \text{ and } [f]_{I^c} \leqslant \frac{1}{2^m}2\varepsilon \text{ on } B\\ &f \leqslant \frac{1}{2^m}2\varepsilon \text{ on } A\,.\end{aligned}$$

We conclude by choosing ε such that $(1-2\varepsilon)^m > 2\varepsilon$ and proceeding similarly to the end of the proof of Lemma 8, resorting to Bernoulli's inequality. □

We are now ready to prove Theorem 4.

Proof of Theorem 4. Fix $f \in \mathcal{C}_{\text{eq.}}$ a continuous equivariant function and $\varepsilon > 0$. Our goal is to find a function $g \in \mathcal{A}$ such that for all $G \in \mathcal{G}_{\text{eq.}}$, $\|F(G) - f(G)\|_\infty \leqslant \varepsilon$. Since $\mathcal{G}_{\text{eq.}}$ is compact, f is bounded, and since we can add constants to g, without lost of generality we assume that $0 < f < f_{\max}$ on $\mathcal{G}_{\text{eq.}}$.

We first restrict the space to the compact set where the coordinates of F are ordered:

$$\mathcal{G}_f \overset{\text{def.}}{=} \bigcup_{n=1}^{n_{\max}} \mathcal{G}_f^{(n)} \quad \text{where} \quad \mathcal{G}_f^{(n)} \overset{\text{def.}}{=} \left\{ G \in \mathcal{G}_{\text{eq.}}^{(n)} \;;\; f_1(G) \geqslant f_2(G) \geqslant \ldots \geqslant f_n(G) \right\}$$

Indeed, by equivariance of f, every graph $G \in \mathcal{G}_{\text{eq.}}$ has a permuted representation in $\mathcal{G}_f$. Hence proving the uniform approximation of f on $\mathcal{G}_f$ is sufficient to prove it on the whole set $\mathcal{G}_{\text{eq.}}$.

Now, denote $K \in \mathbb{N}$ an integer such that $(K-1)\varepsilon \leqslant f_{\max} \leqslant K\varepsilon$. For $k = 1, \ldots, K$, $n = 1, \ldots, n_{\max}$ and $\ell = 1, \ldots, n$, define the following compact set:

$$\begin{aligned}A_k^{n,\ell} &= \left\{ G \in \mathcal{G}_f^{(n)} \;;\; f_\ell(G) - f_{\ell+1}(G) \leqslant (k-1/3)\varepsilon \right\} \cup \bigcup_{n' \neq n} \mathcal{G}_f^{(n')}\\ B_k^{n,\ell} &= \left\{ G \in \mathcal{G}_f^{(n)} \;;\; f_\ell(G) - f_{\ell+1}(G) \geqslant (k+1/3)\varepsilon \right\}\end{aligned}$$

where we use the convention that for $G \in \mathbb{R}^{n^d}$, $f_{n+1}(G) = 0$. Note that $A_k^{n,\ell} \subset A_{k+1}^{n,\ell}$, and $B_{k+1}^{n,\ell} \subset B_k^{n,\ell}$. For $\ell = 1, \ldots, n$ we denote the integer interval $I_\ell = [1, \ell]$.

Let us first show that $A_k^{n,\ell} \cap \mathcal{O}(B_k^{n,\ell}) = \emptyset$ and $B_k^\ell \cap \mathcal{G}_{\text{eq.}}^{(n)}(\mathcal{O}_{I_\ell}) = \emptyset$, so that we can apply Lemma 7. Consider $G \in B_k^{n,\ell}$, $G' \in A_k^{n,\ell}$, $y = F(G)$ and $y' = F(G')$. If $G' \in \mathcal{G}_{\text{eq.}}^{(n')}$ for $n' \neq n$, G and

G' are obviously not permutation of one another. If $G' \in \mathcal{G}_{\text{eq.}}^{(n)}$, the coordinates of both y and y' are sorted, and we have $y'_\ell - y'_{\ell+1} > y_\ell - y_{\ell+1}$ (again with the convention that $y_{n+1}, y'_{n+1} = 0$). It is therefore impossible for y and y' to be permutation of one another, and thus $A_k^{n,\ell} \cap \mathcal{O}(B_k^{n,\ell}) = \emptyset$. Furthermore, for any $\ell < n$, we have $y_i \geqslant y_\ell > y_{\ell+1} \geqslant y_j$ for any $i \leqslant \ell < j$, and therefore there is no self-permutation of y that exchange an index before ℓ and one after. In other words, we have $B_k^{n,\ell} \cap \mathcal{G}_{\text{eq.}}^{(n)}(\mathcal{O}_{I_\ell}) = \emptyset$.

Then, for all k and n, for $\ell < n$, by applying Lemma 7 for $\varepsilon > 0$ we obtain $f_k^{n,\ell}$ such that $0 \leqslant f_k^{n,\ell} \leqslant 1$, $[f_k^{n,\ell}]_{I_\ell} \geqslant 1 - \varepsilon$ and $[f_k^{n,\ell}]_{I_\ell^c} \leqslant \varepsilon$ on $B_k^{n,\ell}$, and $f_k^{n,\ell} \leqslant \varepsilon$ on $A_k^{n,\ell}$. Similarly, by applying Lemma 10, we obtain functions $f_k^{n,n} \in \mathcal{A}$ such that $0 \leqslant f_k^{n,n} \leqslant 1$, $f_k^{n,n} \geqslant 1 - \varepsilon$ on $B_k^{n,n}$ and $f_k^{n,n} \leqslant \varepsilon$ on $A_k^{n,n}$. Finally, we define $g = \sum_{k,n} \sum_{\ell=1}^n f_k^{n,\ell}$.

Now, take any $G \in \mathcal{G}_f$, denote by n its number of nodes. For every $\ell \leqslant n$, denote k_ℓ such that $k_\ell - \frac{2}{3} \leqslant \frac{f_\ell(G) - f_{\ell+1}(G)}{\varepsilon} \leqslant k_\ell + \frac{2}{3}$. By summing these equations, we obtain for all q:

$$\varepsilon \sum_{\ell=q}^n k_\ell - \frac{2n\varepsilon}{3} \leqslant \varepsilon \sum_{\ell=q}^n \left(k_\ell - \frac{2}{3}\right) \leqslant (f(G))_q \leqslant \varepsilon \sum_{\ell=q}^n \left(k_\ell + \frac{2}{3}\right) \leqslant \varepsilon \sum_{\ell=q}^n k_\ell + \frac{2n\varepsilon}{3} \tag{4}$$

We are going to show that g approximates these bounds. For all ℓ, we have $G \in A_{k_\ell+1}^{n,\ell} \cup B_{k_\ell-1}^{n,\ell}$. Moreover, it is obvious that $G \in A_k^{n',\ell}$ for all $n' \neq n$, and all k, ℓ. By construction of $f_k^{n,\ell}$, we have:

$$\begin{aligned}
&\forall n' \neq n, \forall k, \ell, f_k^{n',\ell}(G) \leqslant \varepsilon && \text{since } G \in A_k^{n',\ell} \\
&\forall \ell \leqslant n, \forall k \geqslant k_\ell + 1, \quad f_k^{n,\ell}(G) \leqslant \varepsilon && \text{since } G \in A_{k_\ell+1}^{n,\ell} \subset A_k^{n,\ell} \\
&\forall \ell \leqslant n, \forall k \leqslant k_\ell - 1, \begin{cases} [f_k^{n,\ell}(G)]_{[1,\ell]} & \geqslant 1 - \varepsilon \\ [f_k^{n,\ell}(G)]_{[\ell+1,n]} & \leqslant \varepsilon \end{cases} && \text{since } G \in B_{k_\ell-1}^{n,\ell} \subset B_k^{n,\ell}
\end{aligned}$$

Then, we decompose

$$[f(G)]_q = \varepsilon \left(\sum_{n' \neq n} \sum_{\ell=1}^{n'} \sum_k [f_k^{n',\ell}(G)]_q + \sum_{\ell=1}^{q-1} \sum_k [f_k^{n,\ell}(G)]_q + \sum_{\ell=q}^{n} \sum_k [f_k^{n,\ell}(G)]_q \right) \tag{5}$$

By what precedes the first term is bounded by

$$0 \leqslant \sum_{n' \neq n} \sum_{\ell=1}^{n'} \sum_k [f_k^{n',\ell}(G)]_q \leqslant n_{\max}^2 K \varepsilon < n_{\max}^2 f_{\max}$$

For the second term, we have

$$0 \leqslant \sum_{\ell=1}^{q-1} \sum_k [f_k^{n,\ell}(G)]_q \leqslant (q-1)(1 + (K-1)\varepsilon) < n_{\max}(1 + f_{\max})$$

since for all $\ell \leqslant q-1$ and any $k \neq k_q$, we have $[f_k^{n,\ell}(G)]_q \leqslant \varepsilon$. Finally, for the third term:

$$\begin{aligned}
&\sum_{\ell=q}^n \sum_k [f_k^{n,\ell}(G)]_q \geqslant (1-\varepsilon) \sum_{\ell=q}^n (k_\ell - 1) \geqslant \sum_{\ell=q}^n k_\ell - n_{\max}(f_{\max} + 1), \\
&\sum_{\ell=q}^n \sum_k [f_k^{n,\ell}(G)]_q \leqslant \sum_{\ell=q}^n (k_\ell + 1 + (K - k_\ell - 1)\varepsilon) \leqslant \sum_{\ell=q}^n k_\ell + n_{\max}(1 + f_{\max})
\end{aligned}$$

Hence, combining (4) and (5) with the bounds above we obtain

$$-\varepsilon \left(\frac{2n_{\max}}{3} + n_{\max}(f_{\max} + 1) \right) \leqslant [f(G)]_q - [F(G)]_q \leqslant \varepsilon \left(\frac{2n_{\max}}{3} + 2n_{\max}(1 + f_{\max}) + n_{\max}^2 f_{\max} \right)$$

Hence appropriately choosing ε concludes the proof of the theorem. □

C Additional technical lemma

The next technical lemma is used in proving separation of points.

Lemma 11. *Let $a_{i_1,\dots,i_d}, a'_{i_1,\dots,i_d} > 0$ for $1 \leqslant i_1, \dots, i_d \leqslant n$ be n^d positive numbers such that $\{a_{i_1,\dots,i_d}\}$ and $\{a'_{i_1,\dots,i_d}\}$ are the same multisets. Let $\mathcal{O}', \mathcal{O}'' \subset \mathcal{O}_n$ be sets of permutations such that $\mathrm{Id} \in \mathcal{O}'$ and $|\mathcal{O}'| = |\mathcal{O}''|$. If, for every set of integers $k_{i_1,\dots,i_d} > 0$, we have*

$$\sum_{\sigma \in \mathcal{O}'} \prod_{i_1,\dots,i_d} a^{k_{i_1,\dots,i_d}}_{\sigma(i_1),\dots,\sigma(i_d)} = \sum_{\sigma \in \mathcal{O}''} \prod_{i_1,\dots,i_d} (a'_{\sigma(i_1),\dots,\sigma(i_d)})^{k_{i_1,\dots,i_d}} , \tag{6}$$

then there exists a permutation $\sigma \in \mathcal{O}''$ such that $a_{i_1,\dots,i_d} = a'_{\sigma(i_1),\dots,\sigma(i_d)}$.

Proof. The proof is ultimately based on the simple fact that for two vectors $x, y \in \mathbb{R}^p$, the inner product $\langle x, y \rangle$ is maximum if the elements of x and y have the same ordering (the largest element x_i is multiplied to the largest element y_i, and so on[1]).

Let us begin by fixing any $k_{i_1,\dots,i_d}$ and showing that there is a bijection $\varphi : \mathcal{O}' \to \mathcal{O}''$ between the two considered sets of permutations such that for all $\sigma \in \mathcal{O}'$, $\prod_{i_1,\dots,i_d} a^{k_{i_1,\dots,i_d}}_{\sigma(i_1),\dots,\sigma(i_d)} = \prod_{i_1,\dots,i_d} (a'_{\varphi(\sigma)(i_1),\dots,\varphi(\sigma)(i_d)})^{k_{i_1,\dots,i_d}}$, that is, there is a bijection between each additive term of (6). Denoting $A_\sigma = \prod_{i_1,\dots,i_d} a^{k_{i_1,\dots,i_d}}_{\sigma(i_1),\dots,\sigma(i_d)}$ and similarly A'_σ for a', a consequence of (6) is that

$$\sum_{\sigma \in \mathcal{O}'} A^k_\sigma = \sum_{\sigma \in \mathcal{O}''} (A'_\sigma)^k \tag{7}$$

for all k. Hence, considering the maximum elements $\max_\sigma A_\sigma$ and $\max_{\sigma'} A'_{\sigma'}$ (with arbitrary choice in case of multiple maxima): if they are different, by dividing the equation by the largest of the two and letting $k \to \infty$, we have that one side goes to 0 while the other tends to a positive constant. Hence the maximal elements are the same, we can substract them from the equation and reiterate. Hence we have proven that there is indeed a bijection between the A_σ and $A'_{\sigma'}$.

Then, considering the multiset of n^d numbers $\{a_{i_1,\dots,i_d}\}$, pick $k_{i_1,\dots,i_d} > 0$ in the same order than these numbers: $a_{i_1,\dots,i_d} \leqslant a_{i'_1,\dots,i'_d}$ implies $k_{i_1,\dots,i_d} \leqslant k_{i'_1,\dots,i'_d}$. Using the previously proved property, consider the permutation $\sigma \in \mathcal{O}''$ (which depends on the $k_{i_1,\dots,i_d}$) such that

$$\prod_{i_1,\dots,i_d} a^{k_{i_1,\dots,i_d}}_{i_1,\dots,i_d} = \prod_{i_1,\dots,i_d} (a'_{\sigma(i_1),\dots,\sigma(i_d)})^{k_{i_1,\dots,i_d}}$$

(that is, we have isolated the term corresponding to $\mathrm{Id} \in \mathcal{O}'$ in the l.h.s. of (7) and located the component in the r.h.s. that is in bijection with it). Then, remembering that the $a_{i_1,\dots,i_d}$ and $a'_{i_1,\dots,i_d}$ are taken from the same pool of real numbers, we claim that having the $a'_{\sigma(i_1),\dots,\sigma(i_d)}$ ordered as the $k_{i_1,\dots,i_d}$ is the only way to reach the maximum value (reached by the $a_{i_1,\dots,i_d}$) among all orderings of the $\{a'_{\sigma(i_1),\dots,\sigma(i_d)}\}$: indeed, for that, take the logarithm of the equation above, and we use the fact that the scalar product between two vectors formed by a fixed set of elements is maximal when they are ordered in the same fashion. Hence we have proven that: $a_{i_1,\dots,i_d} \leqslant a_{i'_1,\dots,i'_d}$ implies $k_{i_1,\dots,i_d} \leqslant k_{i'_1,\dots,i'_d}$ which implies $a'_{\sigma(i_1),\dots,\sigma(i_d)} \leqslant a'_{\sigma(i'_1),\dots,\sigma(i'_d)}$. Since the a, a' are drawn from the same pool of numbers, we have proven that $a_{i_1,\dots,i_d} = a'_{\sigma(i_1),\dots,\sigma(i_d)}$, which concludes the proof. □

[1] If there are i, j such that $x_i < x_j$ and $y_j < y_i$, then we can form y' by swapping y_i and y_j, and we have $x^\top y' - x^\top y = x_i y_j + x_j y_i - x_i y_i - x_j y_j = (x_j - x_i)(y_i - y_j) > 0$, hence the swapping strictly increases the scalar product, which is maximal when x and y are ordered in the same fashion.