Convergence of Invariant Graph Networks

Chen Cai[1] Yusu Wang[1]

Abstract

Although theoretical properties such as expressive power and over-smoothing of graph neural networks (GNN) have been extensively studied recently, its convergence property is a relatively new direction. In this paper, we investigate the convergence of one powerful GNN, Invariant Graph Network (IGN) over graphs sampled from graphons. We first prove the stability of linear layers for general k-IGN (of order k) based on a novel interpretation of linear equivariant layers. Building upon this result, we prove the convergence of k-IGN under the model of Ruiz et al. (2020), where we access the edge weight but the convergence error is measured for graphon inputs. Under the more natural (and more challenging) setting of Keriven et al. (2020) where one can only access 0-1 adjacency matrix sampled according to edge probability, we first show a negative result that the convergence of any IGN is not possible. We then obtain the convergence of a subset of IGNs, denoted as IGN-small, after the edge probability estimation. We show that IGN-small still contains function class rich enough that can approximate spectral GNNs arbitrarily well. Lastly, we perform experiments on various graphon models to verify our statements.

1. Introduction

Graph neural networks (GNNs) have recently become a key framework for the learning and analysis of graph type of data, leading to progress on link prediction, knowledge graph embedding, and property prediction to name a few (Wu et al., 2020; Zhou et al., 2020). Although theoretical properties such as expressive power (Maron et al., 2019b; Keriven & Peyré, 2019; Maron et al., 2019a; Garg et al., 2020; Azizian & Lelarge, 2020; Geerts, 2020; Bevilacqua

et al., 2021) and over-smoothing (Li et al., 2018; Oono & Suzuki, 2019; Cai & Wang, 2020; Zhou et al., 2021) of GNNs have received much attention, their convergence property is less understood. In this paper, we systematically investigate the convergence of one of the most powerful families of GNNs, the *Invariant Graph Network (IGN)* (Maron et al., 2018). Different from message passing neural network (MPNN) (Gilmer et al., 2017), it treats graphs and associated node/edge features as monolithic tensors and processes them in a permutation equivariant manner. 2-IGN can approximate the message passing neural network (MPNN) arbitrarily well on the compact domain. When allowing the use of high-order tensor as the intermediate representation, k-IGN is shown at least as powerful as k-WL test. As the tensor order k goes to $\mathcal{O}(n^4)$, it achieves the universality and can distinguish all graphs of size n (Maron et al., 2019b; Keriven & Peyré, 2019; Azizian & Lelarge, 2020).

The high level question we are interested in is the convergence and stability of GNNs. In particular, given a sequence of graphs sampled from some generative models, does a GNN performed on them also converge to a limiting object? This problem has been considered recently, however, so far, the studies (Ruiz et al., 2020; Keriven et al., 2020) focus on the convergence of *spectral GNNs*, which encompasses several models (Bruna et al., 2013; Defferrard et al., 2016) including GCNs with order-1 filters (Kipf & Welling, 2016). However, it is known that the expressive power of GCN is limited. Given that 2(k)-IGN is strictly more powerful than GCN (Xu et al., 2018) in terms of separating graphs[1] and its ability to achieve universality, it is of great interest to study the convergence of such powerful GNN. In fact, it is posted as an open question in Keriven et al. (2021) to study convergence for models more powerful than spectral GNNs and higher order GNNs. This is the question we aim to study in this paper.

Contributions. We present the first convergence study of the powerful k-IGNs (strictly more powerful than the Spectral GNN which previous work studied). We first analyze the building block of IGNs: linear equivariant layers, and develop a stability result for such layers. The case of 2-IGN is proved via case analysis while the general case of k-IGN

---

[1]University of California San Diego, San Diego, USA. Correspondence to: Chen Cai <c1cai@ucsd.edu>.

[1]In terms of separating graphs, k-IGN > 2-IGN = GIN > GCN for $k > 2$.

uses a novel interpretation of the linear equivariant layers which we believe is of independent interest.

There have been two existing models of convergence of spectral GNNs for graphs sampled from graphons developed in Ruiz et al. (2020) and Keriven et al. (2020), respectively. Using the model of Ruiz et al. (2020) (denoted by the *edge weight continuous model*) where we access the edge weight but the convergence error is measured between *graphon inputs* (see Section 5 for details), we obtain analogous convergence results for k-IGNs. The results cover both deterministic and random sampling for k-**IGN** while Ruiz et al. (2020) only covers deterministic sampling for the much weaker **Spectral GNN**s.

Under more natural (and more challenging) setting of Keriven et al. (2020) where one can only access 0-1 adjacency matrix sampled according to edge probability (called the *edge probability discrete model*), we first show a negative result that in general the convergence of all IGNs is not possible. Building upon our earlier stability result, we obtain the convergence of a subset of IGN, denoted as IGN-small, after a step of edge probability estimation. We show that IGN-small still contains rich function class that can approximate Spectral GNN arbitrarily well. Lastly, we perform experiments on various graphon models to verify our statements.

2. Related Work

One type of convergence in deep learning concerns the limiting behavior of neural networks when the width goes to infinity (Jacot et al., 2018; Du et al., 2018; Arora et al., 2019; Lee et al., 2019; Du et al., 2019). In that regime, the gradient flow on a normally initialized, fully connected neural network with a linear output layer in the infinite-width limit turns out to be equivalent to kernel regression with respect to the Neural Tangent Kernel (Jacot et al., 2018).

Another type of convergence concerns the limiting behavior of neural networks when the depth goes to infinity. In the continuous limit, models such as residual networks, recurrent neural network decoders, and normalizing flows can be seen as an Euler discretization of an ordinary differential equation (Weinan, 2017; Chen et al., 2018; Lu et al., 2018; Ruthotto & Haber, 2020).

The type of convergence we consider in this paper concerns when the input objects converge to a limit, does the output of some neural network over such sequence of objects also converge to a limit? In the context of GNNs, such convergence and related notion of stability and transferability have been studied in both graphon (Ruiz et al., 2020; Keriven et al., 2020; Gama et al., 2020; Ruiz et al., 2021) and manifold setting Kostrikov et al. (2018); Levie et al. (2021). In the manifold setting, the analysis is closely re-

lated to the literature on convergence of Laplacian operator (Xu, 2004; Wardetzky, 2008; Belkin et al., 2008; 2009; Dey et al., 2010).

3. Preliminaries

3.1. Notations

To talk about convergence/stability, we will consider graphs of different sizes sampled from a generative model. Similar to the earlier work in this direction, the specific general model we consider is a graphon model.

Graphons. A graphon is a bounded, symmetric and measurable function $W : [0, 1]^2 \rightarrow [0, 1]$. We denote the space of graphon as $\mathcal{W}$. It can be intuitively thought of as an undirected weighted graph with an uncountable number of nodes: roughly speaking, given $u_i, u_j \in [0, 1]$, we can consider there is an edge (i, j) with weight $W(u_i, u_j)$. Given a graphon W, we can sample **unweighted** graphs of any size from W, either in a deterministic or stochastic manner. We defer the definition of the sampling process until we introduce the edge weight continuous model in Section 5 and edge probability discrete model in Section 6.

Tensor. Let $[n]$ denote $\{1, ..., n\}$. A tensor X of order k, called a k-*tensor*, is a map from $[n]^{\otimes k}$ to $\mathbb{R}^d$. If we specifiy a name name_i for each axis, we then say X is indexed by $(\text{name}_1, ..., \text{name}_k)$. With slight abuse of notation, we also write that $X \in \mathbb{R}^{n^k \times d}$. We refer to d as the *feature dimensions* or the *channel dimensions*. If $d = 1$, then we have a k-tensor $\mathbb{R}^{n^k \times 1} = \mathbb{R}^{n^k}$. Although the name for each axis acts as an identifier and can be given arbitrarily, we will use *set* to name each axis in this paper. For example, given a 3-tensor X, we use $\{1\}$ to name the first axis, $\{2\}$ for the second axis, and so on. The benefits of doing so will be clear in Section 4.2.

Partition. A partition of $[k]$, denoted as γ, is defined to be a set of disjoint sets $\gamma := \{\gamma_1, ..., \gamma_s\}$ with $s \leqslant k$ such that the following condition satisfies, 1) for all $i \in [s], \gamma_i \subset [k]$, 2) $\gamma_i \cap \gamma_j = \emptyset, \forall\, i, j \in [s]$, and 3) $\cup_{i=1}^s \gamma_i = [k]$. We denote the space of all partitions of $[k]$ as Γ_k. Its cardinality is called the k-th *bell number* $\text{bell}(k) = |\Gamma_k|$.

Other conventions. By default, we use 2-norm (Frobenius norm) to refer ℓ_2 norm for all vectors/matrices and L_2 norm for functions on $[0, 1]$ and $[0, 1]^2$. $\| \cdot \|_2$ or $\| \cdot \|$ denotes the 2 norm for discrete objects while $\|W\|_{L_2} := \int \int W(u, v) du dv$ denotes the norm for continuous objects. Similarly, we use $\| \cdot \|_\infty$ and $\| \cdot \|_{L_\infty}$ to denotes the infinity norm. When necessary, we use $\| \cdot \|_{L_2([0,1])}$ to specify the support explicitly. We use $\| \cdot \|_{\text{spec}}$ to denote spectral norm. Φ_c and Φ_d refers to the continuous IGN and discrete IGN respectively. We sometimes call a function $f : [0, 1] \rightarrow \mathbb{R}^d$ a *graphon signal*. Given $A \in \mathbb{R}^{n^k \times d_1}, B \in \mathbb{R}^{n^k \times d_2}, [A, B]$

Table 1: Linear equivariant maps for $\mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$ and $\mathbb{R}^{[0,1]^2} \to \mathbb{R}^{[0,1]^2}$. $\mathbf{1}$ is a all-one vector of size $n \times 1$ and $\mathrm{I}_{u=v}$ is the indicator function.

Operations	Discrete	Continuous	Partitions
1-2: The identity and transpose operations	$T(A) = A$ $T(A) = A^T$	$T(W) = W$ $T(W) = W^T$	$\{\{1,3\}, \{2,4\}\}$ $\{\{1,4\}, \{2,3\}\}$
3: The diag operation	$T(A) = \text{Diag}(\text{Diag}^*(A))$	$T(W)(u,v) = W(u,v)\mathrm{I}_{u=v}$	$\{\{1,2,3,4\}\}$
4-6: Average of rows replicated on rows/ columns/ diagonal	$T(A) = \frac{1}{n} A \mathbf{1} \mathbf{1}^T$ $T(A) = \frac{1}{n} \mathbf{1} (A\mathbf{1})^T$ $T(A) = \frac{1}{n} \text{Diag}(A\mathbf{1})$	$T(W)(u,*) = \int W(u,v)dv$ $T(W)(*,u) = \int W(u,v)dv$ $T(W)(u,v) = \mathrm{I}_{u=v} \int W(u,v')dv'$	$\{\{1,4\}, \{2\}, \{3\}\}$ $\{\{1,3\}, \{2\}, \{4\}\}$ $\{\{1,3,4\}, \{2\}\}$
7-9: Average of columns replicated on rows/ columns/ diagonal	$T(A) = \frac{1}{n} A^T \mathbf{1} \mathbf{1}^T$ $T(A) = \frac{1}{n} \mathbf{1} (A^T \mathbf{1})^T$ $T(A) = \frac{1}{n} \text{Diag}(A^T \mathbf{1})$.	$T(W)(*,v) = \int W(u,v)du$ $T(W)(v,*) = \int W(u,v)du$ $T(W)(u,v) = \mathrm{I}_{u=v} \int W(u',v)du'$	$\{\{1\}, \{2,4\}, \{3\}\}$ $\{\{1\}, \{2,3\}, \{4\}\}$ $\{\{1\}, \{2,3,4\}\}$
10-11: Average of all elements replicated on all matrix/ diagonal	$T(A) = \frac{1}{n^2} (\mathbf{1}^T A \mathbf{1}) \cdot \mathbf{1}\mathbf{1}^T$ $T(A) = \frac{1}{n^2} (\mathbf{1}^T A \mathbf{1}) \cdot \text{Diag}(\mathbf{1})$.	$T(W)(*,*) = \int W(u,v)dudv$ $T(W)(u,v) = \mathrm{I}_{u=v} \int W(u',v')du'dv'$	$\{\{1\}, \{2\}, \{3\}, \{4\}\}$ $\{\{1\}, \{2\}, \{3,4\}\}$
12-13: Average of diagonal elements replicated on all matrix/diagonal	$T(A) = \frac{1}{n} (\mathbf{1}^T \text{Diag}^*(A)) \cdot \mathbf{1}\mathbf{1}^T$ $T(A) = \frac{1}{n} (\mathbf{1}^T \text{Diag}^*(A)) \cdot \text{Diag}(\mathbf{1})$	$T(W)(*,*) = \int \mathrm{I}_{u=v} W(u,v)dudv$ $T(W)(u,v) = \mathrm{I}_{u=v} \int W(u',u')du'$	$\{\{1,2\}, \{3\}, \{4\}\}$ $\{\{1,2\}, \{3,4\}\}$
14-15: Replicate diagonal elements on rows/columns	$T(A) = \text{Diag}^*(A)\mathbf{1}^T$ $T(A) = \mathbf{1} \text{Diag}^*(A)^T$	$T(W)(u,v) = W(u,u)$ $T(W)(u,v) = W(v,v)$	$\{\{1,2,4\}, \{3\}\}$ $\{\{1,2,3\}, \{4\}\}$

is defined to be the concatenation of A and B along feature dimensions, i.e., $[A, B] \in \mathbb{R}^{n^k \times (d_1 + d_2)}$. See Table 4 in Appendix for the full symbol list.

3.2. Invariant Graph Network

Definition 1. *An Invariant Graph Network (IGN) is a function* $\Phi : \mathbb{R}^{n^2 \times d_0} \to \mathbb{R}^d$ *of the following form:*

$$F = h \circ L^{(T)} \circ \sigma \circ \cdots \circ \sigma \circ L^{(1)}, \tag{1}$$

where each $L^{(t)}$ *is a linear equivariant (LE) layer (Maron et al., 2018) from* $\mathbb{R}^{n^{k_{t-1}} \times d_{t-1}}$ *to* $\mathbb{R}^{n^{k_t} \times d_t}$ *(i.e., mapping a* k_{t-1} *tensor with* d_{t-1} *channels to a* k_t *tensor with* d_t *channels),* σ *is nonlinear activation function,* h *is a linear invariant layer from* k_T*-tensor* $\mathbb{R}^{n^{k_T} \times d_T}$ *to vector in* $\mathbb{R}^d$. d_t *is the channel number, and* k_t *is tensor order in t-th layer.*

Let $\text{Diag}(\cdot)$ be the operator of constructing a diagonal matrix from vector and $\text{Diag}^*(\cdot)$ be the operation of extracting a diagonal from a matrix. Under the IGN framework, we view a graph with n nodes as a 2-tensor: In particular, given its adjacency matrix A_n of size $n \times n$ with node features $X_n \in \mathbb{R}^{n \times d_{\text{node}}}$ and edge features $E_{n \times n} \in \mathbb{R}^{n^2 \times d_{\text{edge}}}$, the input of IGN is the concatenation of $[A_n, \text{Diag}(X_n), E_{n \times n}] \in \mathbb{R}^{n^2 \times (1 + d_{\text{node}} + d_{\text{edge}})}$ along different channels. We drop the subscript when there is no confusion. We use 2-**IGN** to denote the IGN whose largest tensor order within any intermediate layer is 2, while k-**IGN** is one whose largest tensor order across all layers is k. We use IGN to refer to the general IGN for any order k.

Without loss of generality, we consider input and output tensor to have a single channel. The extension to multiple channels case is presented in Appendix G.2. Consider all lin-

ear equivariant maps from $\mathbb{R}^{n^\ell}$ to $\mathbb{R}^{n^m}$, denoted as $\text{LE}_{\ell+m}$. Maron et al. (2018) characterizes the basis of the space of $\text{LE}_{\ell,m}$. It turns out that the cardinality of the basis equals to the bell number $\text{bell}(\ell + m)$, thus depending only on the order of input/output tensor and independent from graph size n. As an example, we list a specific basis of the space of LE maps for 2-IGN (thus with tensor order at most 2) in Tables 1, 2 and 3 when input/output channel numbers are both 1. Extending the LE layers to multiple input/output channels is straightforward, and can be achieved by parametrizing the LE layers according to indices of input/output channel. See Remark 8 in Appendix. Note that one difference of the operators in Tables 1, 2 and 3 from those given in the original paper is that here we normalize all operators appropriately w.r.t. the graph size n. (This normalization is also in the official implementation of the IGN paper.) This is necessary when we consider the continuous limiting case.

To talk about convergence, one has to define the continuous analog of IGN for graphons. In Tables 1, 2 and 3 we extend all LE operators defined for graphs to graphons, resulting in the continuous analog of 2-IGN, denoted as 2-cIGN or Φ_c in the remaining text. Similar operation can be done in general for k-IGN as well, where the basis elements for k-IGNs will be described in Section 4.2.

Definition 2 (2-cIGN). *By extending all* LE *layers for* 2-*IGN to the graphon case as shown in Tables 1, 2 and 3, we can definite the corresponding 2-cIGN via Eq.* (1).

4. Stability of Linear Layers in IGN

In this section, we first show a stability result for a single linear layer of IGN. That is, given two graphon W_1, W_2, we show that if $\|W_1 - W_2\|_{\text{pn}}$ is small, then the distance

between the objects after applying a single LE layer remain close. Here $\|\cdot\|_{\mathrm{pn}}$ is a partition-norm that will be introduced in a moment. Similar statements also hold for the discrete case when the input is a graph. We first describe how to prove stability for 2-(c)IGN as a warm-up. We then prove it for k-(c)IGN, which is significantly more interesting and requires a new interpretation of the elements in a specific basis of the space of LE operators in Maron et al. (2018).

A the general LE layer $T : \mathbb{R}^{n^\ell} \to \mathbb{R}^{n^m}$ can be written as $T = \sum_\gamma c_\gamma T_\gamma$, where $T_\gamma \in \mathcal{B} := \{T_\gamma | \gamma \in \Gamma_{\ell+m}\}$ is the basis element of the space of $\mathrm{LE}_{\ell,m}$ and c_γ are denoted as filter coefficients. Hence proving the stability of T can be reduced to showing the stability for each element in $\mathcal{B}$, which we focus from now on.

4.1. Stability of Linear Layers of 2-IGN

A natural way to show stability is by showing that the spectral norm of each LE operator in a basis is bounded. However, even for 2-IGN, as we see some LE operator requires replicating "diagonal elements to all rows" (e.g., operator 14-15 in Table 1), and has unbounded spectral norm. To address this challenge, we need a more refined analysis. In particular, below we will introduce a "new" norm that treats the diagonal differently from non-diagonal elements for the 2-tensor case. We term it *partition-norm* as later when handling high order k-IGN, we will see that this norm arises naturally w.r.t. the partition of index set of tensors.

Definition 3 (Partition-norm). *The* partition-norm *of 2-tensor* $A \in \mathbb{R}^{n^2}$ *is defined as* $\|A\|_{\mathrm{pn}} := \left(\frac{\|Diag^*(A)\|_2}{\sqrt{n}}, \frac{\|A\|_2}{n} \right)$. *The continuous analog of the partition-norm for graphon* $W \in \mathcal{W}$ *is defined as* $\|W\|_{\mathrm{pn}} = \left(\sqrt{\int W^2(u,u)du}, \sqrt{\iint W^2(u,v)dudv} \right)$.

We refer to the first term as the normalized diagonal norm *and the second term as the* normalized matrix norm. *Furthermore, we define operations like addition/comparasion on the partition-norm simply as component-wise operations. For example,* $\|A\|_{pn} \le \|B\|_{pn}$ *if each of the two terms of A is at most the corresponding term of B.*

As each term in partition-norm is a norm on different parts of the input, the partition-norm is also a norm. By summing over the finite feature dimension both for finite and infinite cases, the definition of the partition-norm can be extended to multi-channel tensors $\mathbb{R}^{n^2 \times d}$ and its continuous version $\mathbb{R}^{[0,1]^2 \times d}$. See Appendix B.1 for details.

The following result shows that each basis operation for 2-IGN, shown in Tables 1, 2 and 3, is stable w.r.t. the partition-norm. Hence a LE layer consisting of a finite combination of these operations will remain stable. The proof is via a case-by-case analysis and can be found in Appendix B.2.

Proposition 1. *For all LE operators $T_i : \mathbb{R}^{n^2} \to \mathbb{R}^{n^2}$ of discrete 2-IGN listed in Table 1, $\|T_i(A)\|_{\mathrm{pn}} \le \|A\|_{pn}$ for any $A \in \mathbb{R}^{n^2}$. Similar statements hold for $T_i : \mathbb{R}^n \to \mathbb{R}^{n^2}$ and $T_i : \mathbb{R}^{n^2} \to \mathbb{R}^n$ in Tables 2 and 3 in Appendix A. In the case of continuous 2-cIGN, the stability also holds.*

Remark 1. *Note that this also implies that given $W_1, W_2 \in \mathcal{W}$, we have that $\|T_i(W_1) - T_i(W_2)\|_{\mathrm{pn}} \le \|W_1 - W_2\|_{pn}$. Similarly, given $A_1, A_2 \in \mathbb{R}^{n^2 \times 1} = \mathbb{R}^{n^2}$, we have $\|T_i(A_1) - T_i(A_2)\|_{\mathrm{pn}} \le \|A_1 - A_2\|_{pn}$.*

4.2. Stability of Linear Layers of k-IGN

We now consider the more general case of k-IGN. In principle, the proof of 2-IGN can still be extended to k-IGN, but going through all bell(k) number of elements of LE basis of k-IGN one by one can be quite cumbersome. In the next two subsections, we provide a new interpretation of elements of the basis of space of $\mathrm{LE}_{\ell,m}$ in a unified framework so that we can avoid a case-by-case analysis. Such an interpretation, detailed in Section 4.3, is potentially of independent interest. First, we need some notations.

Definition 4 (Equivalence pattern). *Given a k-tensor X, denote the space of its indices $\{(i_1, ..., i_k) \mid i_1 \in [n], ..., i_k \in [n]\}$ by $\mathcal{I}_k$. Given X, $\gamma = \{\gamma_1, ..., \gamma_d\} \in \Gamma_k$ and an element $\boldsymbol{a} = (a_1, ..., a_k) \in \mathcal{I}_k$, we say $\boldsymbol{a} \in \gamma$ if $i, j \in \gamma_l$ for some $l \in [d]$ always implies $a_i = a_j$. Alternatively, we also say $\boldsymbol{a}$ satisfies the equivalence pattern of γ if $\boldsymbol{a} \in \gamma$.*

As an example, suppose $\gamma = \{\{1,2\},\{3\}\}$. Then $(x,x,y) \in \gamma$ while $(x,y,z) \notin \gamma$. Equivalence patterns can induce "slices"/sub-tensors of a tensor.

Definition 5 (Slice/sub-tensor of $X \in \mathbb{R}^{n^k \times 1}$ for $\gamma \in \Gamma_k$). *Let $X \in \mathbb{R}^{n^k \times 1}$ be a k-tensor indexed by $(\{1\}, ..., \{k\})$. Consider a partition $\gamma = \{\gamma_1, ..., \gamma_{k'}\} \in \Gamma_k$ of cardinality $k' \le k$. The slice (sub-tensor) of X induced by γ is a k'-tensor X_γ, indexed by $(\gamma_1, ..., \gamma_{k'})$, and defined to be $X_\gamma(j_1, ..., j_{k'}) := X(\iota_\gamma(j_1, ..., j_{k'}))$ where $j. \in [n]$ and $\iota_\gamma(j_1, ..., j_{k'}) \in \gamma$. $\iota_\gamma : [n]^{k'} \to [n]^k$ is defined to be $\iota_\gamma(j_1, ..., j_{k'}) := (i_1, ..., i_k)$ such that $\{a, b\} \subseteq \gamma_c$ implies $i_a = i_b := j_c$. Here $a, b \in [k], c \in [k']$. As an example, we show five slices of a 3-tensor in Figure 1.*

Consider the LE operators from $\mathbb{R}^{n^\ell}$ to $\mathbb{R}^{n^m}$. Each such map T_γ can be represented by a matrix of size $n^\ell \times n^m$ which can further considered as a $(\ell + m)$-tensor $\boldsymbol{B}_\gamma$. Maron et al. (2018) showed that a specific basis for such operators can be characterized as follows: Each basis element will correspond to one of the bell$(\ell + m)$ partitions in $\Gamma_{\ell+m}$. In particular, given a partition $\gamma \in \Gamma_{\ell+m}$, we have a corresponding basis LE operator T_γ and its tensor representation $\boldsymbol{B}_\gamma$ defined as follows:

$$\text{for any } \boldsymbol{a} \in \mathcal{I}_{\ell+m}, \ \ \mathbf{B}_\gamma(\boldsymbol{a}) = \begin{cases} 1 & \boldsymbol{a} \in \gamma \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The collection $\mathcal{B} = \{T_\gamma \mid \gamma \in \Gamma_{\ell+m}\}$ form a basis for all $\mathrm{LE}_{\ell,m}$ maps. In Section 4.3, we will provide an intereptation of each element of $\mathcal{B}$, making it easy to reason its effect on an input tensor using a unified framework.

Before the main theorem, we also need to extend the partition-norm in Definition 3 from 2-tensor to high-order tensor. Intuitively, for $X \in \mathbb{R}^{n^k}$, $\|X\|_{\mathrm{pn}}$ has $\mathrm{bell}(k)$ components, where each component corresponds to the normalized norm of X_γ, the slice of X induced by $\gamma \in \Gamma_k$. See Figure 1 for examples of slices of a 3-tensor. The partition-norm of input and output of a $\mathrm{LE}_{\ell,m}$ will be of dimension $\mathrm{bell}(\ell)$ and $\mathrm{bell}(m)$ respectively. See Appendix B.1 for details.
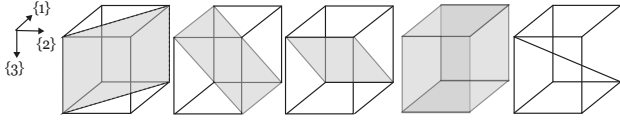

Figure 1: Five possible "slices" of a 3-tensor, corresponding to $\mathrm{bell}(3) = 5$ paritions of [3]. From left to right: a) $\{\{1,2\},\{3\}\}$ b) $\{\{1\},\{2,3\}\}$ c) $\{\{1,3\},\{2\}\}$ d) $\{\{1\},\{2\},\{3\}\}$ e) $\{\{1,2,3\}\}$.

The following theorem characterizes the effect of each operator in $\mathcal{B}$ in terms of partition-norm of input and output, generalizing Proposition 1 from matrix to high order tensor.

Theorem 1 (Stability of LE layers for k-IGN). *Let* $T_\gamma :$ $\mathbb{R}^{[0,1]^\ell} \to \mathbb{R}^{[0,1]^m}$ *be a basis element of the space of* $\mathrm{LE}_{\ell,m}$ *maps where* $\gamma \in \Gamma_{\ell+m}$. *If* $\|X\|_{pn} \leqslant \epsilon \mathbf{1}_{\mathrm{bell}(\ell)}$, *then the partition-norm of* $Y := T_\gamma(X)$ *satisfies* $\|Y\|_{pn} \leqslant \epsilon \mathbf{1}_{\mathrm{bell}(m)}$ *for all* $\gamma \in \Gamma_{\ell+m}$.

The proof relies on a new interpretation of elements of $\mathcal{B}$ in k-IGN. We give only an intuitive sketch using an example in the next subsection. See Appendix B.3 for the proof.

4.3. Interetation of Basis Elements

For better understanding, we color the input axis $\{1, ..., \ell\}$ as red and output axis $\{\ell+1, ..., \ell+m\}$ as blue. Each T_γ corresponds to one partition γ of $[\ell+m]$.

For any partition $\gamma \in \Gamma_{l+k}$, we can write this set as disjoint union $\gamma = S_1 \cup S_2 \cup S_3$ where S_1 is a set of set(s) of input axis, and S_3 is a set of set(s) of output axis. S_2 is a set of set(s) where each set contains both input and output axis. With slight abuse of notation, we omit the subscript γ for S_1, S_3, S_3 when its choice is fixed or clear, and denote $\{\ell+1, ..., \ell+m\}$ as $\ell+[m]$. As an example, one basis element of the space of $\mathrm{LE}_{3,3}$ maps is $\gamma = \{\{1,2\},\{3,6\},\{4\},\{5\}\}$

$$\underbrace{S_1 = \{\{1,2\}\}}_{\text{Only has input axis}} \cup \underbrace{S_2 = \{\{3,6\}\}}_{\substack{\text{has both} \\ \text{input and output axis}}} \cup \underbrace{S_3 = \{\{4\},\{5\}\}}_{\text{only has output axis}} \quad (3)$$

where $1, 2, 3$ specifies the axis of input tensor and $4, 5, 6$

specifies the axis of the output tensor. Recall that there
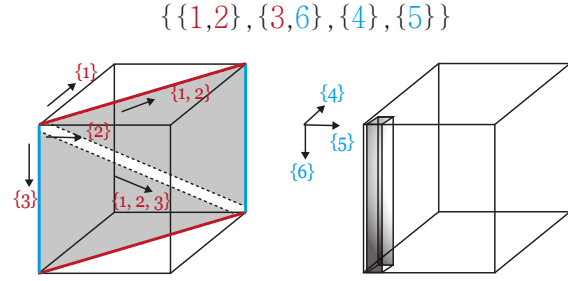
$$\{\{1,2\},\{3,6\},\{4\},\{5\}\}$$

Figure 2: An illustration of the one basis element of the space of $\mathrm{LE}_{3,3}$. The partition is $\{\{1,2\},\{3,6\},\{4\},\{5\}\}$. It selects area spanned by axis $\{1,2\}$ and $\{3\}$ (grey shaded), average over the (red) axis $\{1,2\}$, and then align the resulting 1D tensor with axis $\{6\}$ in the output tensor, and finally replicate the slices along axis $\{4\}$ and $\{5\}$ to fill in the whole cube on the right.

is a one-to-one correspondence between the partitions over $[\ell+m]$ and the base elements in $\mathcal{B}$ as in Eqn (4.2). The basis element T_γ corresponding to $\gamma = S_1 \cup S_2 \cup S_3$ operates on an input tensor $X \in \mathbb{R}^{n^\ell}$ and produce an output tensor $Y \in \mathbb{R}^{n^m}$ as follows:

> Given input X, (step 1) obtain its slice X_γ on Π_1 (selection axis), (step 2) average X_γ over Π_2 (reduction axis), resulting in $X_{\gamma,\mathrm{reduction}}$. (step 3) Align $X_{\gamma,\mathrm{reduction}}$ on Π_3 (alignment axis) with Y_γ and (step 4) replicate Y_γ along Π_4 (replication axis), resulting $Y_{\gamma,\mathrm{replication}}$, a slice of Y. Entries of Y outside $Y_{\gamma,\mathrm{replication}}$ will be set to be 0.

In general, Π_i can be read off from S_1-S_3. See Appendix B.3 for details. As a running example, Figure 2 illustrates the basis element corresponding to $\gamma = S_1 \cup S_2 \cup S_3$ where $S_1 = \{\{1,2\}\} \cup S_2 = \{\{3,6\}\} \cup S_3 = \{\{4\},\{5\}\}$. In the first step, given 3-tensor X, indexed by $\{\{1\},\{2\},\{3\}\}$ we select slices of interest X_γ on $\Pi_1 = \{\{1,2\},\{3\}\}$, colored in grey in the left cube of Figure 2. In the second step, we average X_γ over axis $\Pi_2 = \{\{1,2\}\}$ to reduce 2-tensor X_γ, indexed by $\{\{1,2\},\{3\}\}$ to a 1-tensor $X_{\gamma,\mathrm{reduction}}$, indexed by $\{\{3\}\}$. In the third step, the $X_{\gamma,\mathrm{reduction}}$ is aligned with $\Pi_3 = \{\{6\}\}$, resulting in the grey cuboid Y_γ indexed by $\{\{6\}\}$, shown in the right cube in Figure 2. Here the only difference between $X_{\gamma,\mathrm{reduction}}$ and Y_γ is the index name of two tensors. In the fourth step, we replicate the grey cuboid Y_γ over axis $\Pi_4 = \{\{4\},\{5\}\}$ to fill in the cube, resulting in $Y_{\gamma,\mathrm{replication}}$, indexed by $\{\{3\},\{4\},\{5\}\}$. Note in general $Y_{\gamma,\mathrm{replication}}$ is a slice of Y and does have to be the same as Y.

These steps are defined formally in the Appendix. For each of the four steps, we can control the partition-norm of output for each step (shown in Lemma 3 in Appendix), and

therefore control the partition-norm of the final output for every basis element. See Appendix B.3 for full proofs.

5. Convergence of IGN in the Edge Weight Continuous Model

Ruiz et al. (2020) consider the convergence of $\|\Phi_c(W) - \Phi_c(W_n)\|_{L_2}$ in the *graphon space*, where W is the orignal graphon and W_n is a piecewise constant graphon induced from graphs of size n sampled from W (to be defined soon). We call this model as the *edge weight continuous model*. The main result of Ruiz et al. (2021) is the convergence of *continuous* spectral GNN in the *deterministic* sampling case where graphs are sampled from W deterministically. Leveraging our earlier stability result of linear layers of continuous IGNs in Theorem 1, we can prove an analogous convergence result of cIGNs in the edge weight continuous model for both the deterministic and random sampling cases.

Setup of the edge weight continuous model. Given a graphon $W \in \mathcal{W}$ and a signal $X \in \mathbb{R}^{[0,1] \times d}$, the input of cIGN will be $[W, \mathrm{Diag}(X)] \in \mathbb{R}^{[0,1]^2 \times (1+d)}$. In the random sampling setting, we sample a graph of size n from W by setting the following edge weight matrix and discrete signal:

$$[\widetilde{A_n}]_{ij} := W(u_i, u_j) \quad \text{and} \quad [\widetilde{x_n}]_i := X(u_i) \quad (4)$$

where u_i is the i-th smallest point from n i.i.d points sampled from uniform distribution on $[0, 1]$. We further lift the discrete graph $(\widetilde{A_n}, \widetilde{x_n})$ to a piecewise-constant graphon $\widetilde{W_n}$ with signal $\widetilde{X_n}$. Specifically, partition $[0, 1]$ to be $I_1 \cup \ldots \cup I_n$ with $I_i = (u_i, u_{i+1}]$. We then define

$$\widetilde{W_n}(u, v) := [\widetilde{A_n}]_{ij} \times \mathrm{I}(u \in I_i)\mathrm{I}(v \in I_j) \quad \text{and}$$
$$\widetilde{X_n}(u) := [\widetilde{x_n}]_i \times \mathrm{I}(u \in I_i) \quad (5)$$

where I is the indicator function. Replacing the random sampling with fixed grid, i.e., let $u_i = \frac{i-1}{n}$, we can get the deterministic edge weight continuous model, where W_n and X_n can be defined similarly as the lifting of a discrete sampled graph to a piecewise constant graphon. Note that $\widetilde{W_n}$ is a piecewise constant graphon where each block is not of the same size, while all blocks W_n are of size $\frac{1}{n} \times \frac{1}{n}$. We use $\widetilde{\cdot}$ to emphasize that $\widetilde{W_n}/\widetilde{X_n}$ are random variables, in contrast to the deterministic W_n/X_n.

We also need a few assumptions on the input and IGN.

AS1. *The graphon W is A_1-Lipschitz, i.e. $|W(u_2, v_2) - W(u_1, v_1)| \leqslant A_1(|u_2 - u_1| + |v_2 - v_1|)$.*

AS2. *The filter coefficients c_γ are upper bounded by A_2.*

AS3. *The graphon signal X is A_3-Lipschitz.*

AS4. *The activation functions in IGNs are normalized Lipschitz, i.e. $|\rho(x) - \rho(y)| \leqslant |x - y|$, and $\rho(0) = 0$.*

Such four assumptions are quite natural and also adopted in Ruiz et al. (2020). With AS 1-4, we have the following key proposition. The proof leverages the stability of linear layers for k-IGN from Theorem 1; see Appendix C for details.

Proposition 2 (Stability of Φ_c). *If cIGN $\Phi_c : \mathbb{R}^{[0,1]^2 \times d_{\mathrm{in}}} \to \mathbb{R}^{d_{\mathrm{out}}}$ satisfy AS2, AS4 and $\|W_1 - W_2\|_{pn} \leqslant \epsilon \mathbf{1}_2$, then $\|\Phi_c(W_1) - \Phi_c(W_2)\|_{pn} = \|\Phi_c(W_1) - \Phi_c(W_2)\|_{L_2} \leqslant C(A_2)\epsilon$. The same statement still holds if we change the underlying norm of Partition-norm from L_2 to L_∞.*

Remark 2. *Statements in Proposition 2 holds for discrete IGN Φ_d as well.*

From AS3 we can also bound the difference between the original signal X and the induced signal (X_n and $\widetilde{X_n}$).

Lemma 1. *Let $X \in \mathbb{R}^{[0,1] \times d}$ be an A_3-Lipschitz graphon signal satisfying AS3, and let $\widetilde{X_n}$ and X_n be the induced graphon signal as in Eqs. (4) and (5). Then we have i) $\|X - X_n\|_{pn}$ converges to 0 and ii) $\|X - \widetilde{X_n}\|_{pn}$ converges to 0 in probability.*

We have the similar statements for W as well.

Lemma 2. *If W satisfies AS1, $\|W - W_n\|_{\mathrm{pn}}$ converges to 0. $\|W - \widetilde{W_n}\|_{\mathrm{pn}}$ converges to 0 in probability.*

The following main theorem (for k-cIGN of any order k) of this section can be shown by combining Proposition 2 with Lemmas 1 and 2; see Appendix C for details.

Theorem 2 (Convergence of cIGN in the edge weight continuous model). *Under the fixed sampling condition, IGN converges to cIGN, i.e., $\|\Phi_c([W, Diag(X)]) - \Phi_c([W_n, Diag(X_n)])\|_{L_2}$ converges to 0.*

An analogous statement hold for the random sampling setting, where $\|\Phi_c([W, Diag(X)]) - \Phi_c([\widetilde{W_n}, Diag(\widetilde{X_n})])\|_{L_2}$ converges to 0 in probability.

6. Convergence of IGN in the Edge Probability Discrete Model

In this section, we will consider the convergence setup of Keriven et al. (2020), which we call the *edge probability discrete model*. The major difference from the edge weight continuous model of Ruiz et al. (2020) is that (1) we only access 0-1 adjacency matrix instead of full edge weights and (2) the convergence error is measured in the graph space (instead of graphon space).

This model is more natural. However, we will first show a negative result that in general IGN does not converge in the edge probability discrete model in Section 6.2. This motivates us to consider a relaxed setting where we estimate the edge probability from data. With this extra assumption, we can prove the convergence of IGN-small, a subset of IGN, in the edge probability discrete model in Section 6.3.

Although this is not entirely satisfactory, we show that nevertheless, the family of functions that can be represented by IGN-small is still rich enough to for example approximate any spectral GNN arbitrarily well.

6.1. Setup: Edge Probability Continuous Model

We first state the setup and results of Keriven et al. (2020). We keep the notation close to the original paper for consistency. A random graph model (P, W, f) is represented as a probability distribution P uniform over latent space $\mathcal{U} = [0, 1]$, a symmetric kernel $W : \mathcal{U} \times \mathcal{U} \to [0, 1]$ and a bounded function (graph signal) $f : \mathcal{U} \to \mathbb{R}^{d_z}$. A random graph G_n with n nodes is then generated from (P, W, f) according to latent variables $U := \{u_1, ..., u_n\}$ as follows:

$$\forall j < i \leqslant n: \quad \text{graph node } u_i \overset{iid}{\sim} P, \quad z_i = f(u_i),$$
$$\text{graph edge } a_{ij} \sim \text{Ber}(\alpha_n W(u_i, u_j)) \quad (6)$$

where Ber is the Bernoulli distribution and α_n controls the sparsity of sampled graph. Note that in our case, we assume that the sparsification factor $\alpha_n = 1$ (which is the classical graphon model). We define a degree function by $d_{W,P}(\cdot) := \int W(\cdot, u) dP(u)$. We assume the following

$$\|W(\cdot, u)\|_{L_\infty} \leqslant c_{\max}, \quad d_{W,P}(u) \geqslant c_{\min},$$
$$W(\cdot, u) \text{ is } (c_{\text{Lip.}}, n_{\mathcal{U}}) \text{-piecewise Lipschitz.} \quad (7)$$

We introduce two normalized sampling operator S_U and S_n that sample a continuous function to a discrete one over n points. For a function $W' : \mathcal{U}^{\otimes k} \to \mathbb{R}^{d_{\text{out}}}$, $S_U W'(i_1, ..., i_k) := (\frac{1}{\sqrt{n}})^k (W'(u_{(i_1)}), ..., W'(u_{(i_k)})$ where $u_{(i)}$ is the i-th smallest number over n uniform random samples over $[0, 1]$ and $i_1, ..., i_k \in [n]$. Similarly, $S_n W'(i_1, ..., i_k) := (\frac{1}{\sqrt{n}})^k (W'(\frac{i_1}{n}), ..., W'(\frac{i_k}{n}))$ Note that the normalizing constant will depend on the dimension of the support of W'. We have $\|S_U W'\|_2 \leqslant \|W'\|_{L_\infty}$ and $\|S_n W'\|_2 \leqslant \|W'\|_{L_\infty}$.

To measure the convergence error, we consider root mean square error at the node level: for a signal $x \in \mathbb{R}^{n^2 \times d_{\text{out}}}$ and latent variables U, we define $\text{RMSE}_U(f, x) := \|S_U f - \frac{x}{n}\|_2 = (n^{-2} \sum_{i=1}^n \sum_{j=1}^n \|f(u_i, u_j) - x(i, j)\|^2)^{1/2}$. Again, there is a dependency on the input dimension – the normalization term n^{-2} will need to be adjusted when the input order is different from 2.

6.2. Negative Result

Theorem 3. *Given any graphon W with $c_{max} < 1$ and an IGN architecture, there exists a set of parameters θ such that convergence of IGN_θ to $cIGN_\theta$ is not possible, i.e., $\text{RMSE}_U(\Phi_c([W, Diag(X)]), \Phi_d([A_n, Diag(\widetilde{X_n})]))$ does not converge to 0 as $n \to \infty$, where A_n is 0-1 matrix generated according to Eq. (6), i.e., $A_n[i][j] = a_{i,j}$.*

The proof of Theorem 3 hinges on the fact that the input to IGN in discrete case is 0-1 matrix while the input to cIGN in the continuous case has edge weight upper bounded by $c_{\max} < 1$. The margin between 1 and $c_{\max}$ makes it easy to construct counterexamples. See Appendix D.1 for details.

Theorem 3 states that we cannot expect every IGN will converge to its continuous version cIGN. As the proof of this theorem crucially uses the fact that we can only access 0-1 adjacency matrix, a natural question is what if we can estimate the edge probability from the data? Interestingly, we can obtain the convergence of for a subset of IGNs (which is still rich enough), called IGN-small, in this case.

6.3. Convergence of IGN-small

Let $\widehat{W}_{n\times n}$ be the estimated $n \times n$ edge probability matrix from A_n. $\widehat{W}_n$ is the induced graphon defined in Eq. (5). To analysize the convergence error for general IGN after edge probability estimation, we first decompose the convergence error of the interest using triangle inequality. Assuming the output is 1-tensor, then

$$\text{RMSE}_U(\Phi_c(W), \Phi_d(\widehat{W}_{n\times n}))$$
$$= \|S_U \Phi_c(W) - \frac{1}{\sqrt{n}} \Phi_d(\widehat{W}_{n\times n})\|$$
$$\leqslant \underbrace{\|S_U \Phi_c(W) - S_U \Phi_c(\widetilde{W}_n)\|}_{\text{First term: discrization error}}$$
$$+ \underbrace{\|S_U \Phi_c(\widetilde{W}_n) - \Phi_d S_U(\widetilde{W}_n)\|}_{\text{Second term: sampling error}}$$
$$+ \underbrace{\|\Phi_d S_U(\widetilde{W}_n) - \frac{1}{\sqrt{n}} \Phi_d(\widehat{W}_{n\times n})\|}_{\text{Third term: estimation error}} \quad (8)$$

The three terms measure the different sources of error. First-term is concerned with the discretization error, which can be controlled via a property of S_U and Proposition 2. The Second term concerns the sampling error from the randomness of U. This term will vanish if we consider only S_n instead of S_U under the extra condition stated below. The third term concerns the edge probability estimation error, which can also be controlled by leveraging existing literature on the statistical guarantee of the *edge probability estimation* algorithm from Zhang et al. (2015). [2]

Controlling the second term is more involved. This is also the place where we have to add an extra assumption to constrain the IGN space in order to achieve convergence after edge smoothing.

Definition 6 (IGN-small). *Let $\widetilde{W_{n,E}}$ be a graphon with*

---

[2]For better readability, here we only use the W as input instead of $[W, \text{Diag}(X)]$. Adding $\text{Diag}(X)$ into the input is easy and is included in the full proof in Appendix D.2.
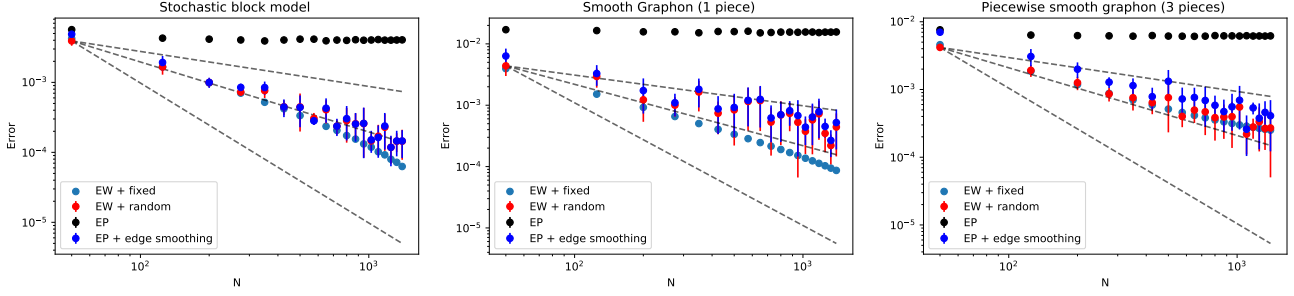
Figure 3: The convergence error for three generative models: (left) stochastic block model, (middle) smooth graphon, (right) piece-wise smooth graphon. EW and EP stands for edge weight continuous model (Eq. (4)) and edge probability discrete model (Eq. (6)). Three dashed line in each figure indicates the decay rate of $n^{-0.5}, n^{-1}$ and n^{-2}.

"chessboard pattern" [3], *i.e., it is a piecewise constant graphon where each block is of the same size. Similarly, define $\widetilde{X_{n,E}}$ as the 1D analog. IGN-small denotes a subset of IGN that satisfies $S_n \Phi_c([\widetilde{[W_{n,E}}, Diag(\widetilde{X_{n,E}})]]) = \Phi_d S_n([\widetilde{[W_{n,E}}, Diag(\widetilde{X_{n,E}})]]).$*

Theorem 4 (convergence of IGN-small in the edge probability discrete model). *Assume AS 1-4, and let $\widehat{W}_{n\times n}$ be the estimated edge probability that satisfies $\frac{1}{n}\|W_{n\times n} - \widehat{W}_{n\times n}\|_2$ converges to 0 in probability. Let Φ_c, Φ_d be continuous and discrete IGN-small. Then $\text{RMSE}_U\left(\Phi_c\left([W, Diag(X)]\right), \Phi_d\left([\widehat{W}_{n\times n}, Diag(\widetilde{x_n})]\right)\right)$ converges to 0 in probability.*

We leave the detailed proofs in Appendix D.2 with some discussion on the challenges for achieving full convergence results in the Remark 4. We note that Theorem 4 has a practical implication: It suggests that in practice, for a given unweighted graph (potentially sampled from some graphon), it may be beneficial to first perform edge probability estimation before feeding into the general IGN framework, to improve the architecture's stability and convergence.

Finally, although the convergence of IGN-small is not entirely satisfactory, it contains some interesting class of functions that can approximate any spectral GNN arbitrarily well. See Appendix E for proof details.

Theorem 5. *IGN-small can approximates spectral GNN (both discrete and continuous ones) arbitrarily well on the compact domain in the $\|\cdot\|_{L_\infty}$ sense.*

7. Experiments

We experiment 2-IGN on three graphon models of increasing complexity: Erdoes Renyi graph with $p = 0.1$, stochastic block model of 2 blocks of equal size and probability matrix $[[0.1, 0.25], [0.25, 0.4]]$, a Lipschitz graphon model with $W(u, v) = \frac{u+v+1}{4}$, and a piecewise Lipschitz graphon

---

[3]See full definition in Definition 10 in Appendix.

with $W(u, v) = \frac{u\%\frac{1}{3}+v\%\frac{1}{3}+1}{4}$ where $\%$ is modulo operation. Similar to (Keriven et al., 2020), we consider untrained IGN with random weights to assess how convergence depends on the choice of architecture rather than learning. We use a 5-layer IGN with hidden dimension 16. We take graphs of different sizes as input and plot the error in terms of the norm of the output difference. The results are plotted in Figure 3. See Appendix F for full details and results.

As suggested by the Theorem 2, for both deterministic and random sampling, the error decreases as we increase the size of the sampled graph. Interestingly, if we take the 0-1 adjacency matrix as the input, the error does not decrease, which aligns with the negative result in Theorem 3. We further implement the edge smoothing algorithm (Eldridge et al., 2016) and find that after the edge probability estimation, the error again decreases, as implied by Theorem 4. We remark that although Theorem 4 works only for IGN-small, our experiments for the general 2-IGN with randomized initialized weights still show encouraging convergence results. Understanding the convergence of general IGN after edge smoothing is an important direction that we will leave for further investigation.

8. Conclusion

In this paper, we investigate the convergence property of a powerful GNN, Invariant Graph Network. We first prove a general stability result of linear layers in IGNs. We then prove a convergence result under the model of Ruiz et al. (2020) for both 2-IGN and high order k-IGN. Under the model of Keriven et al. (2020) we first show a negative result that in general the convergence of every IGN is not possible. Nevertheless, we pinpoint the major roadblock and prove that if we preprocess input graphs by edge smoothing (Zhang et al., 2015), the convergence of a subfamily of IGNs, called IGN-small, can be obtained. As an attempt to quantify the size of IGN-small, we also show that IGN-small contains a rich class of functions that can approximate any spectral GNN.

In the future, we would like to (1) further explore the expressive power of IGN-small and (2) investigate the convergence for the general IGNs under the edge probability discrete model model, or design variants with convergence property but are equally powerful.

Acknowledgement

References

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

Azizian, W. and Lelarge, M. Expressive power of invariant and equivariant graph neural networks. *arXiv preprint arXiv:2006.15646*, 2020.

Belkin, M., Sun, J., and Wang, Y. Discrete laplace operator on meshed surfaces. In *Proceedings of the twenty-fourth annual symposium on Computational geometry*, pp. 278–287, 2008.

Belkin, M., Sun, J., and Wang, Y. Constructing laplace operator from point clouds in r^d. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1031–1040. SIAM, 2009.

Bevilacqua, B., Frasca, F., Lim, D., Srinivasan, B., Cai, C., Balamurugan, G., Bronstein, M. M., and Maron, H. Equivariant subgraph aggregation networks. *arXiv preprint arXiv:2110.02910*, 2021.

Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

Cai, C. and Wang, Y. A note on over-smoothing for graph neural networks. *arXiv preprint arXiv:2006.13318*, 2020.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. Neural ordinary differential equations. *arXiv preprint arXiv:1806.07366*, 2018.

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

Defferrard, M., Bresson, X., and Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29:3844–3852, 2016.

Dey, T. K., Ranjan, P., and Wang, Y. Convergence, stability, and discrete approximation of laplace spectra. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 650–663. SIAM, 2010.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pp. 1675–1685. PMLR, 2019.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

Eldridge, J., Belkin, M., and Wang, Y. Graphons, mergeons, and so on! In *Advances in Neural Information Processing Systems*, pp. 2307–2315, 2016.

Finzi, M., Welling, M., and Wilson, A. G. A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups. *arXiv preprint arXiv:2104.09459*, 2021.

Gama, F., Bruna, J., and Ribeiro, A. Stability properties of graph neural networks. *IEEE Transactions on Signal Processing*, 68:5680–5695, 2020.

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

Geerts, F. The expressive power of kth-order invariant graph networks. *arXiv preprint arXiv:2007.12035*, 2020.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

Holst, L. On the lengths of the pieces of a stick broken at random. *Journal of Applied Probability*, 17(3):623–634, 1980.

Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

Keriven, N. and Peyré, G. Universal invariant and equivariant graph neural networks. *Advances in Neural Information Processing Systems*, 32:7092–7101, 2019.

Keriven, N., Bietti, A., and Vaiter, S. Convergence and stability of graph convolutional networks on large random graphs. *arXiv preprint arXiv:2006.01868*, 2020.

Keriven, N., Bietti, A., and Vaiter, S. On the universality of graph neural networks on large random graphs. *arXiv preprint arXiv:2105.13099*, 2021.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Kostrikov, I., Jiang, Z., Panozzo, D., Zorin, D., and Bruna, J. Surface networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2540–2548, 2018.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32: 8572–8583, 2019.

Levie, R., Huang, W., Bucci, L., Bronstein, M., and Kutyniok, G. Transferability of spectral graph convolutional neural networks. *Journal of Machine Learning Research*, 22(272):1–59, 2021.

Li, Q., Han, Z., and Wu, X.-M. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI conference on artificial intelligence*, 2018.

Lu, Y., Zhong, A., Li, Q., and Dong, B. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In *International Conference on Machine Learning*, pp. 3276–3285. PMLR, 2018.

Maron, H., Ben-Hamu, H., Shamir, N., and Lipman, Y. Invariant and equivariant graph networks. *arXiv preprint arXiv:1812.09902*, 2018.

Maron, H., Ben-Hamu, H., Serviansky, H., and Lipman, Y. Provably powerful graph networks. *arXiv preprint arXiv:1905.11136*, 2019a.

Maron, H., Fetaya, E., Segol, N., and Lipman, Y. On the universality of invariant networks. In *International conference on machine learning*, pp. 4363–4371. PMLR, 2019b.

Oono, K. and Suzuki, T. Graph neural networks exponentially lose expressive power for node classification. *arXiv preprint arXiv:1905.10947*, 2019.

Pyke, R. Spacings. *Journal of the Royal Statistical Society: Series B (Methodological)*, 27(3):395–436, 1965.

Rényi, A. On the theory of order statistics. *Acta Mathematica Academiae Scientiarum Hungarica*, 4(3-4):191–231, 1953.

Ruiz, L., Chamon, L., and Ribeiro, A. Graphon neural networks and the transferability of graph neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

Ruiz, L., Gama, F., and Ribeiro, A. Graph neural networks: Architectures, stability, and transferability. *Proceedings of the IEEE*, 109(5):660–682, 2021.

Ruthotto, L. and Haber, E. Deep neural networks motivated by partial differential equations. *Journal of Mathematical Imaging and Vision*, 62(3):352–364, 2020.

Wardetzky, M. Convergence of the cotangent formula: An overview. *Discrete differential geometry*, pp. 275–286, 2008.

Weinan, E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Philip, S. Y. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

Xu, G. Discrete laplace–beltrami operators and their convergence. *Computer aided geometric design*, 21(8):767–784, 2004.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Zhang, Y., Levina, E., and Zhu, J. Estimating network edge probabilities by neighborhood smoothing. *arXiv preprint arXiv:1509.08588*, 2015.

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

Zhou, K., Huang, X., Zha, D., Chen, R., Li, L., Choi, S.-H., and Hu, X. Dirichlet energy constrained learning for deep graph neural networks. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Table 2: Linear equivariant maps for $\mathbb{R}^n \to \mathbb{R}^{n \times n}$ and $\mathbb{R}^{[0,1]} \to \mathbb{R}^{[0,1]^2}$.

Operations	Discrete	Continuous	Partitions
1-3: Replicate to diagonal/rows/columns	$T(A) = \text{Diag}(A)$ $T(A)_{i,j} = A_i$ $T(A)_{i,j} = A_j$	$T(W)(u,v) = I_{u=v} W(u)$ $T(W)(u,v) = W(u)$ $T(W)(u,v) = W(v)$	$\{\{1,2,3\}\}$ $\{\{1,3\},\{2\}\}$ $\{\{1,2\},\{3\}\}$
4-5: Replicate mean to diagonal/all matrix	$T(A)_{i,i} = \frac{1}{n} A\mathbf{1}$ $T(A)_{i,j} = \frac{1}{n} A\mathbf{1}$	$T(W)(u,v) = I_{u=v} \int W(u)du$ $T(W)(u,v) = \int W(u)du$	$\{\{1\},\{2,3\}\}$ $\{\{1\},\{2\},\{3\}\}$

Table 3: Linear equivariant maps for $\mathbb{R}^{n \times n} \to \mathbb{R}^n$ and $\mathbb{R}^{[0,1]^2} \to \mathbb{R}^{[0,1]}$.

Operations	Discrete	Continuous	Partitions
1-3: Replicate diagonal/row mean/ columns mean	$T(A) = \text{Diag}^*(A)$ $T(A) = \frac{1}{n} A\mathbf{1}$ $T(A) = \frac{1}{n} A^T \mathbf{1}$	$T(W)(u) = W(u,u)$ $T(W)(u) = \int W(u,v)dv$ $T(W)(u) = \int W(u,v)du$	$\{\{1,2,3\}\}$ $\{\{1,2\},\{3\}\}$ $\{\{1,3\},\{2\}\}$
4-5: Replicate mean of all elements/ mean of diagonal	$T(A)_i = \frac{1}{n^2} \mathbf{1}^T A\mathbf{1}$ $T(A)_i = \frac{1}{n} \mathbf{1}^T \text{Diag}(\text{Diag}^*(A))\mathbf{1}$	$T(W)(u) = \int W(u,v)dudv$ $T(W)(u) = \int I_{u,v} W(u,v)dudv$	$\{\{1\},\{2\},\{3\}\}$ $\{\{1,2\},\{3\}\}$

A. Tables

We list the all LE maps for $\mathbb{R}^n \to \mathbb{R}^{n \times n}$ and $\mathbb{R}^{n \times n} \to \mathbb{R}^n$ in Table 2 and Table 3 respectively.

We also summarize the notations used throughout the paper in Table 4.

B. Missing Proofs from Section 4

B.1. Extension of Partition-norm

There are three ways of extending Partition-norm 1) extend the definition of partition-norm to multiple channels 2) changing the underlying norm from L_2 norm to L_∞ norm, and 3) extend Partition-norm defined for 2-tensor to k-tensor.

First recall the definition partition-norm.

Definition 3 (Partition-norm). *The partition-norm of 2-tensor $A \in \mathbb{R}^{n^2}$ is defined as $\|A\|_{\text{pn}} := \left(\frac{\|Diag^*(A)\|_2}{\sqrt{n}}, \frac{\|A\|_2}{n}\right)$. The continuous analog of the partition-norm for graphon $W \in \mathcal{W}$ is defined as $\|W\|_{\text{pn}} = \left(\sqrt{\int W^2(u,u)du}, \sqrt{\iint W^2(u,v)dudv}\right)$.*

We refer to the first term as the normalized diagonal norm *and the second term as the* normalized matrix norm*. Furthermore, we define operations like addition/comparasion on the partition-norm simply as component-wise operations. For example, $\|A\|_{pn} \leq \|B\|_{pn}$ if each of the two terms of A is at most the corresponding term of B.*

To extend partition-norm to signal $A \in \mathbb{R}^{n^2 \times d}$ of multiple channels, we denote $A = [A_{\cdot,1} \in \mathbb{R}^{n^2 \times 1}, ..., A_{\cdot,d} \in \mathbb{R}^{n^2 \times 1}]$ where $[\cdot, \cdot]$ is the concatenation along channels. $\|A\|_{\text{pn}} := \sum_{i=1}^d \|A_{\cdot,i}\|_{\text{pn}}$. As $\|\cdot\|_{\text{pn}}$ is defined by both continous and discrete case, we can extending $\|\cdot\|_{\text{pn}}$ both for multi-channel signal both for graphs and graphons.

Another way of generalizing Partition-norm is to change the L_2 to L_∞ norm. We denote the resulting norm as $\|\cdot\|_{\text{pn}-\infty}$. For $W \in \mathcal{W}, \|W\|_{\text{pn}-\infty} := (\max_{u \in [0,1]} W(u,u), \max_{u \in [0,1], v \in [0,1]} W(u,v))$. The discrete case and high order tensor case can be defined similarily as the L_2 case.

The last way of extending Partition-norm to k-tensor $X \in \mathbb{R}^{n^k \times 1}$ is to define the norm for each slice of X, i.e., $\|X\|_{\text{pn}} := \left(\left(\frac{1}{\sqrt{n}}\right)^{|\gamma_1|} \|X_{\gamma_1}\|_2, ..., \frac{1}{\sqrt{n}}^{|\gamma_{\text{bell}(k)}|} \|X_{\gamma_{\text{bell}(k)}}\|_2\right)$ where $\gamma_\cdot \in \Gamma_k$. Note how we order $(\gamma_1, ..., \gamma_{\text{bell}(k)})$ can be arbitrary as long as the order is used consistent.

Table 4: Summary of important notations.

Symbol	Meaning
$\mathbf{1}_n$	all-one vector of size $n \times 1$
$\|\cdot\|_2/\|\cdot\|_{L_2}$	2-norm for matrix/ graphon
$\|\cdot\|_\infty/\|\cdot\|_{L_\infty}$	infnity-norm for matrix/graphon
$[\cdot,\cdot]$	Given $A \in \mathbb{R}^{n^k \times d_1}$, $B \in \mathbb{R}^{n^k \times d_2}$, $[A,B]$ is the concatenation of A and B along feature dimension. $[A,B] \in \mathbb{R}^{n^k \times (d_1+d_2)}$.
$W : [0,1]^2 \to [0,1]$	graphon
$X \in \mathbb{R}^{[0,1] \times d}$	1D signal
$\mathcal{W}$	space of graphons
$\|\cdot\|_{\mathrm{pn}}$	partition-norm. When the underlying norm is L_∞ norm, we also use $\|\cdot\|_{\mathrm{pn}-\infty}$.
I	indicator function
I	interval
SGNN	spectral graph neuarl networks, defined in Equation (12)
$\mathrm{LE}_{\ell,m}$	linear equivariant maps from ℓ-tensor to m-tensor
	Notations related to sampling
W_n	Induced piecewise constant graphon from fixed grid
$\widetilde{W_n}$	Induced piecewise constant graphon from random grid
$\widetilde{W_{n,E}}$	Induced piecewise constant graphon from random grid, but resize the all individual blocks to be of equal size (also called chessboard graphon in the paper). $\widetilde{W_{n,E}}(I_i \times I_j) := W(u_{(i)}, u_{(j)})$
$W_{n \times n}$	$n \times n$ matrix sampled from W; $W_{n \times n}(i,j) = W(u_i, u_j)$
$\widehat{W}_{n \times n} \in \mathbb{R}^{n \times n}$	the estimated edge probablity from graphs sampled according to edge probability discrete model from Zhang et al. (2015)
$\widetilde{x_n} \in \mathbb{R}^{n \times d}$	sampled singal $[\widetilde{x_n}]_i := X(u_i)$
X_n	induced 1D piecewise graphon signal from fixed grid
$\widetilde{X_n}$	induced 1D piecewise graphon signal from random grid
S_U	normalized sampling operator for random grid. $S_U f(i,j) = \frac{1}{n}(f(u_{(i)}), f(u_{(j)}))$
S_n	normalized sampling operator for fixed grid. $S_n f(i,j) = \frac{1}{n}(f(\frac{i}{n}), f(\frac{j}{n}))$
$\mathrm{RMSE}_U(x,f)$	$\left(n^{-1} \sum_{i=1}^n \|x_i - f(u_i)\|^2\right)^{1/2}$ for 1D signal; $\left(n^{-2} \sum_i \sum_j \|x_{i,j} - f(u_i, u_j)\|^2\right)^{1/2}$ for 2D case
α_n	a parameter that controls the sparsity of sample graphs. Set to be 1 in the paer.
	Notations related to IGN
$\mathrm{bell}(k)$	Bell number: number of partitions of $[k]$. $\mathrm{bell}(2) = 2, \mathrm{bell}(3) = 5, \mathrm{bell}(4) = 15, \mathrm{bell}(5) = 52...$
Γ_k	space of all partitions of $[k]$
$\mathcal{I}_k$	the space of indices. $\mathcal{I}_k := \{(i_1,...,i_k)|i_1 \in [n],...,i_k \in [n]\}$. Elements of $\mathcal{I}_k$ is denoted as $\boldsymbol{a}$
$\gamma \in [k]$	partition of $[k]$. For example $\{\{1,2\},\{3\}\}$ is a partion of $[3]$. The totoal number of partitions of $[k]$ is $\mathrm{bell}(k)$.
$\boldsymbol{a} \in \gamma$	$\boldsymbol{a}$ satisfies the equivalence pattern of γ. For example, $(x,x,y) \in \{\{1,2\},\{3\}\}$ where $x,y,z \in [n]$.
$\gamma < \beta$	given two partitions $\gamma, \beta \in \Gamma_k$, $\gamma < \beta$ if γ is finer than β. For example, $\{1,2,3\} < \{\{1,2\},\{3\}\}$.
$\boldsymbol{B}_\gamma$	$l+m$ tensor; tensor representation of $\mathrm{LE}_{l,m}$ maps. we differentiate T_γ (operators) from $\boldsymbol{B}_\gamma$ (tensor representation of operators)
$\mathcal{B}$	a basis of the space of linear equivariant operations from ℓ-tensor to m-tensor. $\mathcal{B} = \{T_\gamma | \gamma \in \Gamma_{l+k}\}$
T_c/T_d	linear equivariant layers for graphon (continous) and graphs (discrete)
Φ_c/Φ_d	IGN for graphon (continous) and graphs (discrete)
$L^{(i)}$	i-th linear equivariant layer of IGN
L	normalized graph Laplacian
T_i	basis element of the space of linear equivariant maps; sometimes also written as T_γ.

B.2. Proof of stability of linear layer for 2-IGN

Proposition 1. *For all* LE *operators* $T_i : \mathbb{R}^{n^2} \to \mathbb{R}^{n^2}$ *of discrete* 2-IGN *listed in Table 1,* $\|T_i(A)\|_{\mathrm{pn}} \leqslant \|A\|_{pn}$ *for any* $A \in \mathbb{R}^{n^2}$. *Similar statements hold for* $T_i : \mathbb{R}^n \to \mathbb{R}^{n^2}$ *and* $T_i : \mathbb{R}^{n^2} \to \mathbb{R}^n$ *in Tables 2 and 3 in Appendix A. In the case of continuous* 2-cIGN, *the stability also holds.*

Proof. The statements hold in both discrete and continuous cases. Without loss of generality, we only prove the continuous case by going over all linear equivariant maps $\mathbb{R}^{[0,1]^2} \to \mathbb{R}^{[0,1]^2}$ in Table 1.

- 1-3: It is easy to see that the partition-norm does not increase for all three cases.

- 4-6: It is enough to prove case 4 only. For diagonal norm $\|\mathrm{Diag}(T(W))\|_{L_2}^2 = \int (\int W(u,v)dv)^2 du \leqslant \iint W^2(u,v)dudv = \epsilon$. For matrix norm: $\|T(W)\|_{L_2}^2 = \|\mathrm{Diag}(T(W))\|_{L_2} \leqslant \epsilon$. Therefore the statement holds for this linear equivariant operation.

- 7-9: same as case 4-6.

- 10-11: It is enough to prove the first case: average of all elements replicated on the whole matrix. The diagonal norm is the same as the matrix norm. Both norms are decreasing so we are done.

- 12-13: It is enough to prove only case 12. Since diagonal norm is equal to matrix norm, and diagonal norm is decreasing by Jensen's inequality we are done.

- 14-15: Since matrix norm is the same as diagonal norm, which stays the same so we are done.

As shown in all cases for any $W \in \mathcal{W}$ with $\|W\|_{\mathrm{pn}} < (\epsilon, \epsilon)$, $\|T_i(W)\|_{\mathrm{pn}} < (\epsilon, \epsilon)$. Therefore we finish the proof for $\mathbb{R}^{[0,1]^2} \to \mathbb{R}^{[0,1]^2}$. We next go over all linear equivariant maps $\mathbb{R}^{[0,1]} \to \mathbb{R}^{[0,1]^2}$ in Table 2 and prove it case by case.

- 1-3: It is enough to prove the second case. It is easy to see diagonal norm is preserved and $\|T(W)\|_2 = \|W\|_2 \leqslant \epsilon$. Therefore $\|T(W)\|_{\mathrm{pn}} \leqslant (\epsilon, \epsilon)$.

- 4-5: It is enough to prove the second case. Norm on diagonal is no larger than $\|W\|$ by Jensen's inequality. The matrix norm is the same as the diagonal norm therefore also no large than ϵ. Therefore $\|T(W)\|_{\mathrm{pn}} \leqslant (\epsilon, \epsilon)$.

Last, we prove the cases for $\mathbb{R}^{[0,1]^2} \to \mathbb{R}^{[0,1]}$.

For cases 1-3, it is enough to prove case 2. Since the norm of the output is no large than the matrix norm of input by Jensen's inequality, we are done. Similar reasoning applies to cases 4-5 as well. □

B.3. Proof of Theorem 1

We need a few definitions and lemmas first.

Definition 7 (axis of a tensor). *Given a k-tensor* $X \in \mathbb{R}^{n^k \times 1}$ *indexed by* $(\mathrm{name}_1, ..., \mathrm{name}_k)$. *The axis of* X, *denoted as* $ax(X)$, *is defined to be* $ax(X) := (\mathrm{name}_1, ..., \mathrm{name}_k)$.

As an example, the aixs of the first grey sub-tensor in Figure 5a is $\{\{1, 2\}, \{3\}\}$.

Definition 8 (partial order of partitions). *Given two paritions of* $[k]$, *denoted as* $\gamma = \{\gamma_1, ..., \gamma_{d_1}\}$ *and* $\beta = \{\beta_1, ..., \beta_{d_2}\}$, *we say* γ *is finer than* β, *denoted as* $\gamma < \beta$, *if and only if 1)* $\gamma \neq \beta$ *and 2) for any* $\beta_j \in \beta$, *there exists* $\gamma_i \in \gamma$ *such that* $\beta_j \subseteq \gamma_i$.

For example, $\{\{1, 2, 3\}\}$ is finer than $\{\{1, 2\}, \{3\}\}$ but $\{\{1, 2\}, \{3\}\}$ is not comparable with $\{\{1, 3\}, \{2\}\}$. Note that space of partitions forms a hasse diagram under the partial order defined above (each set of elements has a least upper bound and a greatest lower bound, so that it forms a lattice). See Figure 4 for an example.
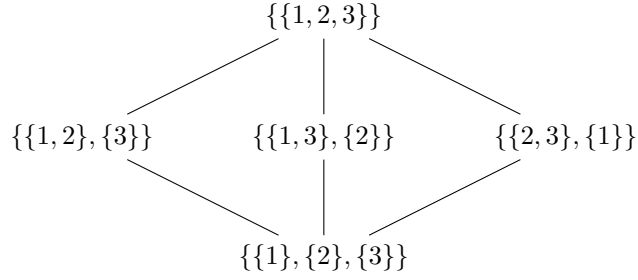
Figure 4: Space of partitions forms a hasse diagram under the partial order defined in Definition 8.

Definition 9 (average a k-tensor X over Π). *Let $X \in \mathbb{R}^{n^k \times 1}$ be a k-tensor indexed by $\{\{1\}, ..., \{k\}\}$. Without loss of generality, let $\Pi = \{\{1\}, ..., \{d\}\}$. Denote the resulting $(k-d)$-tensor X', indexed by $\{\{d+1\}, ..., \{k\}\}$. By averaging X over Π, we mean*

$$X'(\cdot) := \frac{1}{n^d} \sum_{t \in \mathcal{I}_d} X(t, \cdot).$$

The definition can be extended to $\mathbb{R}^{[0,1]^k}$ by replacing average with integral.

Lemma 3 (properties of partition-norm). *We list some properties of partition-norm. Although all lemmas are stated in the discrete case, the continous version also holds. The statements also holds for $\| \cdot \|_{pn-\infty}$ as well.*

1. *Let $X \in \mathbb{R}^{n^k \times 1}$ be a k-tensor and denote one of its slices $X' \in \mathbb{R}^{n^{k'} \times 1}$ with $k' \leqslant k$. If $\|X\|_{pn} \leqslant \epsilon \mathbf{1}_{\text{bell}(k)}$, then $\|X'\|_{pn} \leqslant \epsilon \mathbf{1}_{\text{bell}(k')}$.*

2. *Let $k' < k$. Let $X \in \mathbb{R}^{n^k \times 1}$ be a k-tensor and $X' \in \mathbb{R}^{n^{k'} \times 1}$ be the resulting k'-tensor after averaging over $k - k'$ axis of X. If $\|X\|_{pn} \leqslant \epsilon \mathbf{1}_{\text{bell}(k)}$, then $\|X'\|_{pn} \leqslant \epsilon \mathbf{1}_{\text{bell}(k')}$.*

3. *Let $k' > k$. Let $X \in \mathbb{R}^{n^k \times 1}$ be a k-tensor and X' be the resulting k'-tensor after replicating X over $k' - k$ axis of X'. If $\|X\|_{pn} \leqslant \epsilon \mathbf{1}_{\text{bell}(k)}$, then $\|X'\|_{pn} \leqslant \epsilon \mathbf{1}_{\text{bell}(k')}$.*

4. *Let $k' < k$ and $X \in \mathbb{R}^{n^k \times 1}$ be a k-tensor and $\gamma \in \Gamma_k$ be one parition of $[k]$. Let X has only one non-zero slice X_γ of order k', i.e., if $\boldsymbol{a} \in \mathcal{I}_k, X(\boldsymbol{a}) \neq 0$ implies $\boldsymbol{a} \in \gamma$. If $\|X_\gamma\|_{pn} \leqslant \epsilon \mathbf{1}_{\text{bell}(k')}$, then $\|X\|_{pn} \leqslant \epsilon \mathbf{1}_{\text{bell}(k)}$.*

Proof. We prove statements one by one. Note that although the proof is done for L_2 norm, we do not make use of any specific property of L_2 norm and the same proof can be applied to L_∞ as well. Therefore all statements in the lemma apply to $\| \cdot \|_{pn-\infty}$ as well.

1. By the definition of partition-norm and slice in Definition 5, we know that any slice of X' is also a slice of X, therefore any component of $\|X'\|_{pn}$ will be upper bounded by ϵ, which concludes the proof.

2. Without loss of generality, we can assume that $k' = k - 1$ and the axis of X that is averaged over is axis $\{1\}$. To bound $\|X'\|_{pn}$, we need to bound the normalized norm of any slice of X'. Let $X'_{\gamma'}$ be arbitrary slice of X'. Since X' is obtained by averaging over axis 1 of X, we know that $X'_{\gamma'}$ is the obtained by averaging over axis of 1 of X_γ, a slice of X, where $\gamma := \gamma' \cup \{\{1\}\}$. Since $\|X\|_{pn} \leqslant \epsilon \mathbf{1}_{\text{bell}(k)}$, we know that $(\frac{1}{\sqrt{n}})^{|\gamma|} \|X_\gamma\| \leqslant \epsilon$. By Jensen's inequality, we have $(\frac{1}{\sqrt{n}})^{|\gamma'|} \|X'_{\gamma'}\| \leqslant (\frac{1}{\sqrt{n}})^{|\gamma|} \|X_\gamma\|$, and therefore $(\frac{1}{\sqrt{n}})^{|\gamma'|} \|X'_{\gamma'}\| \leqslant \epsilon$. Since $(\frac{1}{\sqrt{n}})^{|\gamma'|} \|X'_{\gamma'}\| \leqslant \epsilon$ holds for arbitrary slice of X', we conclude that $\|X'\|_{pn} \leqslant \epsilon \mathbf{1}_{\text{bell}(k')}$.

   The proof above only handles the case of $k' = k - 1$. The general case where $k - k' > 1$ can be handled by evoking the proof above multiple times for different reduction axis.

3. We assume X is indexed by $(\{1\}, ..., \{k\})$ and X' is indexed by $(\{1\}, ..., \{k+1\})$. Just as the last case, without loss of generality we assume that X' is obtained by replicating X over 1 new axis, denoted as $\{k+1\}$. In other words, $\text{ax}(X') = \text{ax}(X) \cup \{\{k+1\}\}$.

To control $\|X'\|_{\text{pn}}$, we need to bound $(\frac{1}{\sqrt{n}})^{|\gamma|}\|X'_\gamma\|$ where $\gamma \in \Gamma_{k+1}$. Since X' is obtained from X by replicating it over $\{k+1\}$, $(\frac{1}{\sqrt{n}})^{|\gamma|}\|X'_\gamma\| = (\frac{1}{\sqrt{n}})^{|\beta|}\|X_\beta\|$ where $\beta = \gamma|_{[k]}$. As $\|X\|_{\text{pn}} \leqslant \epsilon\mathbf{1}_{\text{bell}(k)}$, it implies that $(\frac{1}{\sqrt{n}})^{|\gamma|}\|X'_\gamma\| \leqslant \epsilon$ holds for any $\gamma \in \Gamma_{k'}$. Therefore we conclude that $\|X'\|_{\text{pn}} \leqslant \epsilon\mathbf{1}_{\text{bell}(k')}$.

4. To bound $\|X\|_{\text{pn}}$, we need to bound the normalized norm of any slice of X. Let X_β be arbitrarily slice of X where $\beta \in \Gamma_k$. Since γ and β are partitions of $[k]$, there exist partitions that are finer than both β and γ, where the notion of finer between two partitions is defined in Definition 8. Among all partitions that satisfy such conditions, denote the most coarse one as $\alpha \in \Gamma_k$. This can be done because the Γ_k is finite. Note that $|\alpha| < |\beta|$ and $|\alpha| < |\gamma|$.

Since X_α is a slice of X_γ and $\|X_\gamma\|_{\text{pn}} \leqslant \epsilon\mathbf{1}_{\text{bell}(k')}$, $(\frac{1}{\sqrt{n}})^{|\alpha|}\|X_\alpha\| \leqslant \epsilon$ according to Lemma 3.1. As X_α is the slice of X_β (implies $\|X_\alpha \leqslant X_\beta\|$) and α is the most coarse partition that is finer than β and γ (implies $\|X_\alpha\| \geqslant \|X_\beta\|$), we have $\|X_\beta\| = \|X_\alpha\|$. This implies $(\frac{1}{\sqrt{n}})^{|\beta|}\|X_\beta\| \leqslant (\frac{1}{\sqrt{n}})^{|\alpha|}\|X_\alpha\| \leqslant \epsilon$. As $(\frac{1}{\sqrt{n}})^{k'}\|X_\beta\| \leqslant \epsilon$ holds for arbitrary slice β of X, we conclue that $\|X\|_{\text{pn}} \leqslant \epsilon\mathbf{1}_{\text{bell}(k)}$.

$\square$

Now we are ready to prove the main theorem.

Theorem 1 (Stability of LE layers for k-IGN). *Let* $T_\gamma : \mathbb{R}^{[0,1]^\ell} \to \mathbb{R}^{[0,1]^m}$ *be a basis element of the space of* $\text{LE}_{\ell,m}$ *maps where* $\gamma \in \Gamma_{\ell+m}$. *If* $\|X\|_{pn} \leqslant \epsilon\mathbf{1}_{\text{bell}(\ell)}$, *then the partition-norm of* $Y := T_\gamma(X)$ *satisfies* $\|Y\|_{pn} \leqslant \epsilon\mathbf{1}_{\text{bell}(m)}$ *for all* $\gamma \in \Gamma_{\ell+m}$.

Proof. Without loss of generality, we first consider discrete cases of mapping from $X \in \mathbb{R}^{n^\ell}$ to $Y \in \mathbb{R}^{n^m}$. In general, each element T_γ of linear permutation equivariant basis can be identified with the following operation on input/output tensors.

Given input X, (step 1) obtain its subtensor $X_\gamma \subset X$ on a certain Π_1 (selection axis), (step 2) average X_γ over Π_2 (reduction axis), resulting in $X_{\gamma,\text{reduction}}$. (step 3) Align $X_{\gamma,\text{reduction}}$ on Π_3 (alignment axis) with Y_γ and (step 4) replicate Y_γ along Π_4 (replication axis), resulting $Y_{\gamma,\text{replication}}$, a slice of Y. Entries of Y outside $Y_{\gamma,\text{replication}}$ will be set to be 0. In general, Π_i can be read off from S_1-S_3.

Π_1-Π_4 corresponds to different axis of input/output tensor and can be read off from different parts of $S_\gamma = S_1 \cup S_2 \cup S_3$. Note such operation can be naturally extended to the continuous case, as done in Tables 1, 2 and 3 for 2-IGN. We next give detailed explnations of each step.
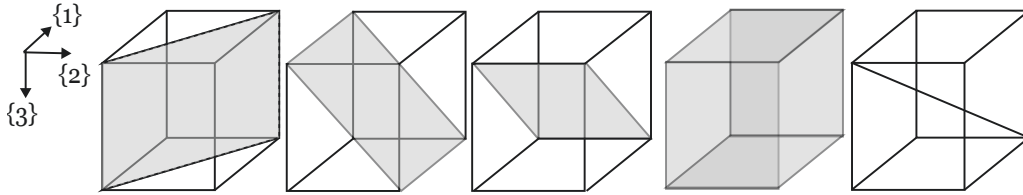

Figure 5: Five "slices" of a 3-tensor, corresponding to bell(3) = 5 paritions of [3]. From left to right: a) $\{\{1,2\},\{3\}\}$ b) $\{\{1\},\{2,3\}\}$ c) $\{\{1,3\},\{2\}\}$ d) $\{\{1\},\{2\},\{3\}\}$ e) $\{\{1,2,3\}\}$.

First step ($X \to X_\gamma$)**: select X_γ from X via Π_1.** Π_1 corresponds to

$$S|_{[\ell]} := \{s \cap [l] \mid s \in S \text{ and } s \cap [l] \neq \emptyset\}.$$

It specifies the what parts (such as diagonal part or off-diagonal part for 2-tensor) of the input ℓ-tensor is under consideration. We denote the resulting subtensor as X_γ. See Definition 5 for formal definition.

As an example in Equation (3), Π_1 corresponds to $\{\{1,2\},\{3\}\}$, meaing we select a 2-tensor with axises $\{1,2\}$ and $\{3\}$. Note that the cardinality $|S|_{[\ell]}| = |(S_1 \cup S_2)|_{[\ell]}| \leqslant l$ encodes the order of X_γ.

Second step ($X_\gamma \to X_{\gamma,\text{reduction}}$): **average of** X_γ **over** Π_2. Π_2 corresponds to $S_1 \subset S|_{[\ell]}$, which tells us along what axis to average over X_γ. It will reduce the tensor X_γ of order $|S_1| + |S_2|$, indexed by $S|_{[\ell]}$, to a tensor of order $|S|_{[\ell]}| - |S_1| = |S_2|$, indexed by $S_2|_{[l]}$.

In the example of Figure 6, this corresponds to averaging over axis $\{\{1,2\}\}$, reducing 2-tensor (indexed by axis $\{1,2\}$ and $\{3\}$) to 1-tensor (indexed by axis $\{3\}$). The normalization factor in the discrete case is $n^{|S_1|}$. We denote the tensor after reduction as $X_{\gamma,\text{reduction}}$.

As the second step performs tensor order reduction, we end up with a tensor $X_{\gamma,\text{reduction}}$ of order $|S_2|$. Last two steps specify how to fill in the output tensor Y with $X_{\gamma,\text{reduction}}$. To fill in Y, we will first align $X_{\gamma,\text{reduction}}$ with Y_γ, a subtensor of Y, on Π_3 and then replicate Y_γ on Π_4, resulting in $Y_{\gamma,\text{replication}}$, a sub-tensor of Y.

Third step ($X_{\gamma,\text{reduction}} \to Y_\gamma$): **align** $X_{\gamma,\text{reduction}}$ **with** Y_γ. To fill in Y_γ, we need to specify how the resuting $|S_2|$-tensor $X_{\gamma,\text{reduction}}$ is *aligned* with the $|S_2|$-subtensor of Y_γ. After all, there are many ways of selecting a $|S_2|$-tensor from Y, which is indexed by $\{\{l\}, \{l+1\}, ..., \{\ell+m\}\}$. This is specified by the third step. Let Y_γ be the $|S_2|$-tensor. We next define the precise relationship between $X_{\gamma,\text{reduction}}$ and Y_γ. $X_{\gamma,\text{reduction}}$ is indexed by $S_2|_{[l]}$ while Y_γ is indexed by $S_2|_{l+[m]}$ and defined to be $Y_\gamma(\cdot) = X_{\gamma,\text{reduction}}(\cdot)$.

In the example of Figure 6, $X_{\gamma,\text{reduction}}$ is a 1D tensor indexed by $\{3\}$ and Y_γ (the grey cuboid on the right cube of Figure 6) is indexed by $\{6\}$.

Fourth step ($Y_\gamma \to Y_{\gamma,\text{replication}}$): **replicating** Y_γ **over** Π_4. Π_4 denotes the S_3 and specifies to what axis we will replicate the $|S_2|$-tensor Y_γ over. Recall that Y_γ is indexed by $S_2|_{l+[m]}$. Let the resulting tensor be $Y_{\gamma,\text{replication}}$, which is a subtensor of $Y \in \mathbb{R}^{n^l}$. $Y_{\gamma,\text{replication}}$ is defined to be indexed by $(S_2 \cup S_3)|_{l+[m]}$. Without loss of generality, let the first $|S_2|$ component are indexed by $S_2|_{l+[m]}$ and the rest components are indexed by $S_3|_{l+[m]}$. The mathematical definition of the fourth step is then $Y_{\gamma,\text{replication}}(\cdot, t) := Y_\gamma(\cdot)$ for all $t \in [n]^{|S_3|}$. Note that the order of $Y_{\gamma,\text{replication}}$ can be smaller than order of Y.

The example in Equation (3) has $S_3 = \{\{4\}, \{5\}\}$, which means that we will replicate the 1-tensor along axis $\{4\}$ and $\{5\}$. Note that in general, we do not have to fill in the whole m-tensor (think about copy row average to diagonal in Table 1).
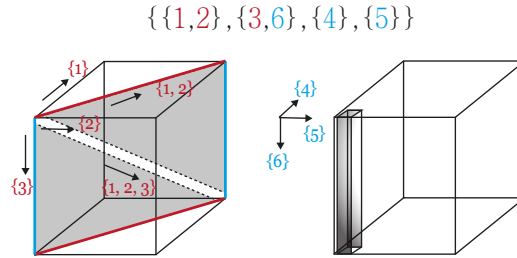
$$\{\{1,2\},\{3,6\},\{4\},\{5\}\}$$

Figure 6: An illustration of the one linear equivariant basis from $\mathbb{R}^{n^3} \to \mathbb{R}^{n^3}$. The partition is $\{\{1,2\},\{3,6\},\{4\},\{5\}\}$. It selects area spanned by axis $\{1,2\}$ and $\{3\}$ (grey shaded), average over the (red) axis $\{1,2\}$, and then align the resulting 1D slice with axis $\{6\}$ in the output tensor, and finally replicate the slices along axis $\{4\}$ and $\{5\}$ to fill in the whole cube on the right.

After the interpretation of general linear equivariant maps in k-IGN, We now show that if $\|X\|_{\text{pn}} \leqslant \epsilon \mathbf{1}_{\text{bell}(\ell)}$, then $T_\gamma(X) \leqslant \epsilon \mathbf{1}_{\text{bell}(m)}$ holds for all γ. This can be done easily with the use of Lemma 3.

For any partition of $[\ell + m]$ γ, according to the first step we are mainly concerned about the $\|X_\gamma\|_{\text{pn}}$ instead of $\|X\|_{\text{pn}}$. Since X_γ is a slice of X, then if $\|X\|_{\text{pn}} \leqslant \epsilon \mathbf{1}_{\text{bell}(\text{ord}(X))}$, by Lemma 3.1, then $\|X_\gamma\|_{\text{pn}} \leqslant \epsilon \mathbf{1}_{\text{bell}(|S_1|+|S_2|)}$.

According to the second step and Lemma 3.2, we can also conclude that $\|X_{\gamma,\text{reduction}}\|_{\text{pn}} \leqslant \epsilon \mathbf{1}_{\text{bell}(|S_2|)}$.

For the third step of align $X_{\gamma,\text{reduction}}$ with Y_γ, it is quite obvious that $\|Y_\gamma\|_{\text{pn}} = \|X_{\gamma,\text{reduction}}\|_{\text{pn}} \leqslant \epsilon \mathbf{1}_{\text{bell}(|S_2|)}$.

For the fourth step of replicating Y_γ over Π_4 to get $Y_{\gamma,\text{replication}}$, by Lemma 3.3, we have $\|Y_{\gamma,\text{replication}}\|_{\text{pn}} \leqslant \epsilon \mathbf{1}_{\text{bell}(|S_2|+|S_3|)}$.

Lastly, we envoke Lemma 3.4 to get $\|Y\|_{\text{pn}} \leqslant \epsilon \mathbf{1}_{\text{bell}(m)}$, which concludes our proof.

$\square$

C. Missing Proofs from Section 5 (Edge Weight Continuous Model)

First we need a lemma on the distrbution of gaps between n uniform sampled points on $[0, 1]$.

Lemma 4. *Let $u_{(i)}$ be n points unformly sampled on $[0, 1]$, sorted from small to large with $u_{(0)} = 0$ and $u_{(n+1)} = 1$. Let $D_i = u_{(i)} - u_{(i-1)}$. All D_is have same distribution, which is* $\text{Beta}(1, n - 1)$. *In particular, $E(D_i) = \frac{1}{n}$, $E(D_i^2) = \frac{2}{n(n+1)}$, $E(D_i^2) = \frac{6}{n(n+1)(n+2)}$.*

Proof. By a symmetry argument, it is easy to see that all intervals follow the same distribution. For the first interval, the probability all the $n - 1$ points are above x is $(1 - x)^{n-1}$ so the density of the length of the first (and so each) interval is $(n - 1)(1 - x)^{1-x}$. This is a Beta distribution with parameters $\alpha = 1$ and $\beta = n - 1$. The expectation of higher moments follows easily. Note that although the intervals are identically distributed, they are not independently distributed, since their sum is 1. □

Lemma 1. *Let $X \in \mathbb{R}^{[0,1] \times d}$ be an A_3-Lipschitz graphon signal satisfying AS3, and let $\widetilde{X_n}$ and X_n be the induced graphon signal as in Eqs.* (4) *and* (5)*. Then we have i) $\|X - X_n\|_{pn}$ converges to 0 and ii) $\|X - \widetilde{X_n}\|_{pn}$ converges to 0 in probability.*

Proof. We first bound the $\|X - X_n\|_{L_2[0,1]}$ and $\|X - \widetilde{X_n}\|_{L_2[0,1]}$. For the first case, partitioning the unit interval as $I_i = [(i - 1)/n, i/n]$ for $1 \leqslant i \leqslant n$ (the same partition used to obtain x_n, and thus X_n, from X), we can use the Lipschitz property of X to derive

$$\|X - X_n\|_{L_2(I_i)}^2 \leqslant A_3^2 \int_0^{1/n} u^2 du = \frac{A_3^2}{3n^3}$$

We can then write $\|X - X_n\|_{L_2([0,1])}^2 = \sum_i \|X - X_n\|_{L_2(I_i)}^2 \leqslant \frac{A_3}{3n^2}$, which implies that $\|X - X_n\|_{L_2([0,1])} \leqslant \sqrt{\frac{A_3}{3n^2}}$.

For the second case, since $\|X - \widetilde{X_n}\|_{L_2([0,1])}^2 = \sum_i \|X - \widetilde{X_n}\|_{L_2(I_i)}^2$, we will bound the $\left\|X - \widetilde{X_n}\right\|_{L_2(I_i)}^2$. As

$$\left\|X - \widetilde{X_n}\right\|_{L_2(I_i)}^2 \leqslant A_3^2 \int_0^{D_i} u^2 du = A_3 D_i^3/3$$

therefore

$$\left\|X - \widetilde{X_n}\right\|_{L_2(I)}^2 = \sum_i \left\|X - \widetilde{X_n}\right\|_{L_2(I_i)}^2 \leqslant A_3/3 \sum_i D_i^3$$

where D_i stands for the length of I_i, which is a random variable due to the random sampling.

According to Lemma 4, all D_i are identically distributed and follows the Beta distribution $B(1, n - 1)$. The expectation $E(D_i^3) = \frac{6}{n(n+1)(n+2)}$. Since by Jensen's inequality $E(\sqrt{Y}) \leqslant \sqrt{E(Y)}$ holds for any positive random variable Y, $E(\sqrt{\frac{A_3}{3} \sum_i D_i^3}) \leqslant \sqrt{E(\frac{A_3}{3} \sum_i D_i^3)} = \sqrt{\frac{A_3}{3} \frac{1}{n(n+2)}} = \Theta(\frac{1}{n})$. Using Markov inequality, we can then upper bound the

$$P(\|X - \widetilde{X_n}\|_{L_2(I)} \geq \epsilon) \leqslant P(\sqrt{\frac{A_3}{3} \sum_i D_i^3} \geq \epsilon) \leq \frac{E(\sqrt{\frac{A_3}{3} \sum_i D_i^3})}{\epsilon} = \Theta(\frac{1}{n\epsilon}) \tag{9}$$

Since the $P(\|X - \widetilde{X_n}\|_{L_2(I)} \geq \epsilon)$ goes to 0 as n increases, we conclude that $\|X - \widetilde{X_n}\|_{\text{pn}}$ converges to 0 in probability. □

Lemma 2. *If W satisfies AS1, $\|W - W_n\|_{\text{pn}}$ converges to 0. $\|W - \widetilde{W_n}\|_{\text{pn}}$ converges to 0 in probability.*

Proof. For the first case, partitioning the unit interval as $I_i = [(i - 1)/n, i/n]$ for $1 \leqslant i \leqslant n$, we can use the graphon's Lipschitz property to derive

$$\|W - W_n\|_{L_1(I_i \times I_j)} \leqslant A_1 \int_0^{1/n} \int_0^{1/n} |u| du dv + A_1 \int_0^{1/n} \int_0^{1/n} |v| dv du = \frac{A_1}{2n^3} + \frac{A_1}{2n^3} = \frac{A_1}{n^3}.$$

We can then write $\|W - W_n\|_{L_1([0,1]^2)} = \sum_{i,j} \|W - W_n\|_{L_1(I_i \times I_j)} \leqslant n^2 \frac{A_1}{n^3} = \frac{A_1}{n}$ which, since $W - W_n : [0,1]^2 \to$ $[-1,1]$, implies $\|W - W_n\|_{L_2([0,1]^2)} \leqslant \sqrt{\|W - W_n\|_{L_1([0,1]^2)}} \leqslant \sqrt{\frac{A_1}{n}}$. The second last inequality holds because all entries of $W - W_n$ lies in $[-1,1]$.

Similarly, $\|\text{Diag}(W - W_n)\|_{L_2[0,1]} \leqslant \sqrt{\|\text{Diag}(W - W_n)\|_{L_1[0,1]}} \leqslant \sqrt{2nA_1 \int_0^{1/n} u\,du} = \sqrt{\frac{A_1}{n}}$. Therefore we conclude the first part of the proof.

For the second case, diagonal norm is similar to the proof of Lemma 1 so we only focus on the $\|W - W_n\|_{L_2([0,1]^2)}$. Since $W - \widetilde{W_n} : [0,1]^2 \to [-1,1]$ implies

$$\|W - \widetilde{W_n}\|_{L_2([0,1]^2)} \leqslant \sqrt{\|W - \widetilde{W_n}\|_{L_1([0,1]^2)}} = \sqrt{\sum_{i,j} \|W - \widetilde{W_n}\|_{L_1(I_i \times I_j)}}$$

where

$$\|W - \widetilde{W_n}\|_{L_1(I_i \times I_j)} \leqslant A_1 \int_{I_v} \int_{I_u} |u|\,du\,dv + A_1 \int_{I_u} \int_{I_v} |v|\,dv\,du = \frac{A_1}{2}(D_i D_j^2 + D_j D_i^2)$$

Therefore

$$\|W - \widetilde{W_n}\|_{L_2([0,1]^2)} \leqslant \sqrt{\|W - \widetilde{W_n}\|_{L_1([0,1]^2)}} = \sqrt{\sum_{i,j} \frac{A_1}{2}(D_j D_i^2 + D_i D_j^2)} = \sqrt{A_1 \sum_i D_i^2} \tag{10}$$

where we use the $\sum_i D_i = 1$ for the last equality. Since by Jensen's inequality $E(\sqrt{Y}) \leqslant \sqrt{E(Y)}$ for any positive random variable Y, $E(\sqrt{\sum_i D_i^2}) \leqslant \sqrt{E(\sum_i D_i^2)} = \Theta(\frac{1}{\sqrt{n}})$ since $E(D_i^2) = \Theta(\frac{1}{n^2})$ by Lemma 4. By Markov inequality, we then bound

$$P(\|W - \widetilde{W_n}\|_{L_2([0,1]^2)} > \epsilon) \leq P(\sqrt{\|W - \widetilde{W_n}\|_{L_1([0,1]^2)}} > \epsilon) \leqslant \frac{E(\sqrt{\sum_i D_i^2})}{\epsilon} \leq \Theta(\frac{1}{\sqrt{n}\epsilon})$$

$\square$

Therefore, we conclude that both $\|W - W_n\|_{\text{pn}}$ and $\|W - \widetilde{W_n}\|_{\text{pn}}$ converges to 0.

Proposition 2 (Stability of Φ_c). *If cIGN $\Phi_c : \mathbb{R}^{[0,1]^2 \times d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$ satisfy AS2, AS4 and $\|W_1 - W_2\|_{\text{pn}} \leqslant \epsilon \mathbf{1}_2$, then $\|\Phi_c(W_1) - \Phi_c(W_2)\|_{\text{pn}} = \|\Phi_c(W_1) - \Phi_c(W_2)\|_{L_2} \leqslant C(A_2)\epsilon$. The same statement still holds if we change the underlying norm of Partition-norm from L_2 to L_∞.*

Proof. Without loss of generality, it suffices to prove for 2-IGN as k-IGN follows the same proof with the constant being slightly different. Since we have proved stability of every linear layers of IGN in Theorem 1, the general linear layer T is just a linear combinations of individual linear basis, i.e. $T = \sum_\gamma c_\gamma T_\gamma$ where $c_i \leqslant A_2$ for all i according to AS2. Without loss of generality, We can assume $T(X)$ is of order 2 and have

$$\|T(W_1) - T(W_2)\|_{\text{pn}} = \|\sum_i c_\gamma T_\gamma (W_1 - W_2)\|_{\text{pn}}$$

$$\leqslant \sum_i \|c_\gamma T_\gamma (W_1 - W_2)\|_{\text{pn}}$$

$$\leqslant (\sum |c_\gamma|\epsilon, \sum |c_\gamma|\epsilon) = (15 A_2 \epsilon, 15 A_2 \epsilon)$$

To extend the result to nonlinear layer, note that AS4 ensures the 2-norm shrinks after passing through nonlinear layers. Therefore $\|\sigma \circ T(X) - \sigma \circ T(Y)\|_{\text{pn}} \leqslant \|T(X) - T(Y)\|_{\text{pn}} = \|T(X - Y)\|_{\text{pn}} \leqslant 15 A_2 \|X - Y\|_{\text{pn}}$. Repeating such process across layers, we finish the proof of the L_2 case.

The extension to L_∞ is similar to the case of L_2 norm. The main modification is to change the definition of the partition-norm from L_2 norm on different slices (corresponding to different partitions of $[\ell]$ where ℓ is the order of input) to L_∞ norm. The extension to the case where input and output tensor is of order ℓ and m is also straightforward according to Theorem 1.

$\square$

Theorem 2 (Convergence of cIGN in the edge weight continuous model). *Under the fixed sampling condition, IGN converges to cIGN, i.e., $\|\Phi_c([W, Diag(X)]) - \Phi_c([W_n, Diag(X_n)])\|_{L_2}$ converges to 0.*

An analogous statement hold for the random sampling setting, where $\|\Phi_c([W, Diag(X)]) - \Phi_c([\widetilde{W_n}, Diag(\widetilde{X_n})])\|_{L_2}$ converges to 0 in probability.

Proof. By Proposition 2, it suffices to prove that $\|[W, Diag(X)]) - [W_n, Diag(X_n)]\|_{pn}$ and $\|[W, Diag(X)]) - [\widetilde{W_n}, Diag(\widetilde{X_n})]\|_{pn}$ goes to 0.

$\|[W, Diag(X)]) - [W_n, Diag(X_n)]\|_{pn}$ is upper bounded by $(\Theta(\frac{1}{n^{1.5}}), \Theta(\frac{1}{n^{1.5}}))$ according to Lemmas 1 and 2, which decrease to 0 as n increases. Therefore we finish the proof of convergence for the deterministic case.

For the random sampling case, by Lemmas 1 and 2, we know that both $\|W - \widetilde{W_n}\|_{L_2([0,1]^2)}$ and $\|X - \widetilde{X_n}\|_{L_2(I)}$ goes to 0 as n increases in probability at the rate of $\Theta(\frac{1}{n^{1.5}})$. Therefore we can also conclude that the convergence of IGN in probability according to Proposition 2. $\qquad\square$

D. Missing Proof from Section 6 (Edge Probability Continuous Model)

D.1. Missing Proof for Section 6.2

Theorem 3. *Given any graphon W with $c_{max} < 1$ and an IGN architecture, there exists a set of parameters θ such that convergence of IGN_θ to $cIGN_\theta$ is not possible, i.e., $\mathrm{RMSE}_U(\Phi_c([W, Diag(X)]), \Phi_d([A_n, Diag(\widetilde{X_n})]))$ does not converge to 0 as $n \to \infty$, where A_n is 0-1 matrix generated according to Eq. (6), i.e., $A_n[i][j] = a_{i,j}$.*

Proof. Without loss of generality, let the network be $IGN = L^{(2)} \circ \sigma \circ L^{(1)}$, and let the input to IGN be A in the discrete case and W in the continuous case. For simplicity, we assume that graphon W is constant p on $[0, 1]^2$. As A consists of only 0 and 1 and all entries of W is below c_{max}, it is easy to construct a linear layer $L^{(1)}$ from $\mathbb{R}$ to $\mathbb{R}$ via the right choice of bias such that $L^{(1)}$ map any number no large than c_{max} to negative and maps 1 to positive. Therefore $L^{(1)}(W) = 0$ and $L^{(1)}(A)$ is a postive number $c \in \mathbb{R}^+$ on entries (i, j) where $A(i, j) = 1$. Let σ be ReLU and $L^{(2)}$ be average of all entries. We can see that $cIGN(W) = 0$ for all n while $IGN(A)$ converges to $\sigma(c)p$ as n increases.

As the construction above only relies on the fact that there is a separation between c_{max} and 1 (but not on size n), it can be extended to any IGN for any n, which means the gap between $cIGN(W)$ and $IGN(A)$ will not decrease as n increases. In the general case of W not being constant, the only difference is that $IGN(A)$ will converge to be $\sigma(c)p^*$ where p^* is a different constant that depends on W. Therefore we conclude the proof. $\qquad\square$

Remark 3. *The reason that the same argument does not work for spectral GNN is that spectral GNN always maintains Ax in the intermediate layer. In contrast, IGN keeps both A and $Diag(x)$ in separate channels, which makes it easy to isolate them to construct counterexamples.*

D.2. Missing Proofs from Section 6.3

Notation. For any $P, Q \in \mathbb{R}^{n \times n}$, define $d_{2,\infty}$, the normalized $2, \infty$ matrix norm, by $d_{2,\infty}(P, Q) = n^{-1/2}\|P - Q\|_{2,\infty} := \max_i n^{-1/2}\|P_{i,\cdot} - Q_{i,\cdot}\|_2$ where $P_{i,\cdot}, P_{i,\cdot}$ are i-th row of P and Q. Note that $d_{2,\infty}(P, Q) \geq \frac{1}{n}\|P - Q\|_2$.

Let S_U be the sampling operator for W, i.e., $S_U(W) = \frac{1}{n}[W(U_i, U_j)]_{n \times n}$. Note that as U is randomly sampled, S_U is a random operator. Denote S_n as sampling on a fixed equally spaced grid of size $n \times n$, i.e. $S_n W = \frac{1}{n}[W(\frac{i}{n}, \frac{j}{n})]_{n \times n}$. S_n is a fixed operator when n is fixed.

Let $\widehat{W}_{n \times n}$ be the estimated edge probability from graphs A sampled from W. Let $\widetilde{W_n}$ be the piece-wise constant graphon induced from sample U as Eq. (5). Similarly, denote $W_{n \times n}$ be the $n \times n$ matrix realized on sample U, i.e., $W_{n \times n}[i, j] = W(u_i, u_j)$. It is easy to see that $S_U(W) = \frac{1}{n}W_{n \times n}$. Let $\widetilde{W_{n,E}}$ be the graphon induced by $W_{n \times n}$ with $n \times n$ blocks of the same size. In particular, $\widetilde{W_{n,E}}(I_i \times I_j) := W(u_{(i)}, u_{(j)})$ where $I_i = [\frac{i-1}{n}, \frac{i}{n}]$. E in the subscript is the shorthand for the "blocks of equal size". Similarly we can also define the 1D analog of $\widetilde{W_n}$ and $\widetilde{W_{n,E}}$, $\widetilde{X_n}$ and $\widetilde{X_{n,E}}$.

Proof strategy. We first state five lemmas that will be used in the proof of Theorem 4. Lemma 5 concerns the property of normalized sampling operator S_U and S_n. Lemmas 6 and 7 concerns the norm convergence of $\|\widetilde{W_n} - W\|_{L_\infty}$ and

$\|\widetilde{W_{n,E}} - W\|_{L_\infty}$. Lemma 8 characterize the effects of linear equivariant layers T and IGN Φ on L_∞ norm of the input and output. Lemma 9 bounds the L_∞ norm of the difference of stochastic sampling operator S_U and the deterministic sampling operator S_n. Theorem 4 is built on the results from five lemmeas and the existing result on the theoretical guarantee of edge probability estimation from Zhang et al. (2015).

The convergence of some lemmas is almost surely convergence. Convergence almost surely implies convergence in probability, and in this paper, all theorems concern convergence in probability. Note that proofs of Lemmas 5 to 7 and 9 for the W and X are almost the same. Therefore without loss of generality, we mainly prove the case of W.

Definition 10 (Chessboard pattern). *Let $u_i = \frac{i-1}{n}$ for all $i \in [n]$. A graphon W is defined to have chessboard pattern if and only if there exists a n such that W is a piecewise constant on $[u_i, u_{i+1}] \times [u_j, u_{j+1}]$ for all $i, j \in [n]$. Similarily, $f : [0,1] \to \mathbb{R}$ has 1D chessboard pattern if there exists n such that f is a piecewise constant on $[u_i, u_{i+1}]$ for all $i \in [n]$.*
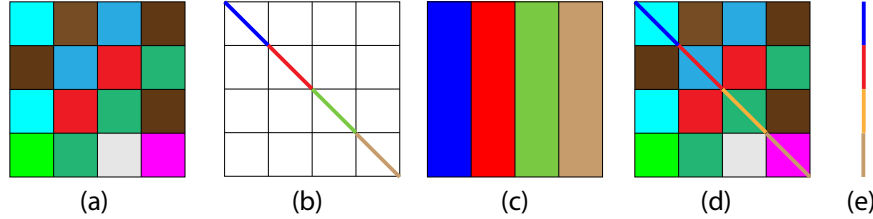
See Figure 7 for examples and counterexamples.

Figure 7: (a) and (c) has chessboard pattern. (e) has 1D chessboard pattern. (d) does not has the chessboard pattern. (b) is of form $\text{Diag}(\widetilde{f_{n,E}})$ and also does not have chessboard pattern, but in the case of IGN approximating Spectral GNN, (b) is reprerepresented in the form of c) via a linear equivariant layers of 2-IGN.

Lemma 5 (Property of S_n and S_U). *We list the properties of sampling operator S_U and S_n*

1. $S_U \circ \sigma = \sigma \circ S_U$. *Similar result holds for S_n as well.*

2. $\|S_U f_{1d}\| \leqslant \|f\|_{L_\infty}$ *where $f_{1d} : [0,1] \to \mathbb{R}$. Similar result holds for $f_{2d} : [0,1]^2 \to \mathbb{R}$ and S_n as well.*

Lemma 6. *Let W be $[0,1]^2 \to \mathbb{R}$ and X be $[0,1] \to \mathbb{R}$. If W is Lipschitz, $\|\widetilde{W_n} - W\|_{L_\infty}$ converges to 0 in probability. If X is Lipschitz, $\|\widetilde{X_n} - X\|_{L_\infty}$ converges to 0 in probability.*

Proof. Without loss of generality, we only prove the case for W. By the Lipschitz condition of W, if suffices to bound the $Z_n := \max_{i=1}^n D_i$ where D_i is the length of i-th interval $|u_{(i)} - u_{(i-1)}|$. Characterizing the distribution of the length of largest interval is a well studied problem (Rényi, 1953; Pyke, 1965; Holst, 1980). It can be shown that Z_n follows $P(Z_n \leqslant x) = \sum_{j=0}^{n+1} \binom{n+1}{j} (-1)^j (1-jx)_+^n$ with the expectation $E(Z_k) = \frac{1}{n+1} \sum_{i=1}^{n+1} \frac{1}{i} = \Theta(\frac{\log n}{n})$. By Markov inequality, we conclude that $\|\widetilde{W_n} - W\|_{L_\infty}$ converges to 0 in probability.

$\square$

Lemma 7. *Let W be $[0,1]^2 \to \mathbb{R}$ and X be $[0,1] \to \mathbb{R}$. If W is Lipschitz, $\|\widetilde{W_{n,E}} - W\|_{L_\infty}$ converges to 0 almost surely. If X is Lipschitz, $\|\widetilde{X_{n,E}} - X\|_{L_\infty}$ converges to 0 almost surely.*

Proof. As $\widetilde{W_{n,E}}$ is a piecewise constant graphon and W is Lipschitz according to AS1, we only need to examine $\max_{i,j} \|(W - \widetilde{W_{n,E}})(\frac{i}{n}, \frac{j}{n})\|$.

It is easy to see that $(W - \widetilde{W_{n,E}})(\frac{i}{n}, \frac{j}{n}) = W(\frac{i}{n}, \frac{j}{n}) - W(u_{(i)}, u_{(j)})$ where $u_{(i)}$ stands for the i-th smallest random variable from uniform i.i.d. samples from $[0,1]$. By the Lipschitz condition of W, if suffices to bound $\|\frac{i}{n} - u_{(i)}\| + \|\frac{j}{n} - u_{(j)}\|$. Glivenko-Cantelli theorem tells us that the L_∞ of empirical distribution F_n and cumulative distribution function F converges to 0 almost surely, i.e., $\sup_{u \in [0,1]} |F(u) - F_n(u)| \to 0$ almost surely. Since $\max_i \|u_{(i)} - \frac{i}{n}\| = \sup_{u \in \{u_{(1)}, \dots, u_{(n)}\}} |F(u) - $

$F_n(u)| \leqslant \sup_{u \in [0,1]} |F(u) - F_n(u)|$ when $F(u) = u$ (cdf of uniform distribution), we conclude that $\|\widetilde{W_{n,E}} - W\|_{L_\infty}$ converges to 0 almost surely.

$\square$

We also need a lemma on the property of the linear equivariant layers T.

Lemma 8 (Property of T_c and σ). *Let σ be nonlinear layer. Let T_c be a linear combination of elements of basis of the space of linear equivariant layers of cIGN, with coefficients upper bounded. We have the following property about T_c and σ*

1. *If W is Lipschitz, $T_c(W)$ is piecewise Lipschitz on diagonal and off-diagonal. Same statement holds for $\Phi_c(W)$.*

2. $S_n \circ \sigma(\widetilde{W_{n,E}}) = \sigma \circ S_n(\widetilde{W_{n,E}}).$

Proof. We prove two statements one by one.

1. We examine the linear equivariant operators from $\mathbb{R}^{[0,1]^2}$ to $\mathbb{R}^{[0,1]^2}$ in Table 1. There are some operations such as "average of rows replicated on diagonal" will destroy the Lipschitz conditon of $T_c(W)$ but $T_c(W)$ will still be piecewise Lipschitz on diagonal and off-diagonal. Since σ will preserve the Lipschitzness, $\Phi_c(W)$ is piecewise Lipschitz on diagonal and off-diagonal.

2. This is easy to see as σ acts on input pointwise.

$\square$

Lemma 9. *Let W be $[0,1]^2 \to \mathbb{R}$*

1. *If W is Lipschitz, $\|S_U W - S_n W\|$ converges to 0 almost surely. Similarily, if X is Lipschitz, $\|S_U Diag(X) - S_n Diag(X)\|$ converges to 0 almost surely.*

2. *If W is piecewise Lipschitz on S_1 and S_2 where S_1 is the diagonal and S_2 is off-diagonal, then $\|S_U W - S_n W\|$ converges to 0 almost surely.*

Proof. Since the case of X is essentially the same with that of W, we only prove the case of W.

1. As $n\|S_U W - S_n W\|_\infty \geq \|S_U W - S_n W\|$, it suffices to prove that $n\|S_U W - S_n W\|_\infty = \max_{i,j} |W(u_{(i)}, u_{(j)}) - W(\frac{i}{n}, \frac{j}{n})|$ converges to 0 almost surely. Similar to Lemma 7, using Lipschitz condition of W and Glivenko-Cantelli theorem concludes the proof.

2. This statement is stronger than the one above. The proof of the last item can be adapted here. As W is A_1 Lipschitz on off-diagonal region and A_2 Lipschitz on diagonal,

$$n\|S_U W - S_n W\|_\infty = \max_{i,j} \left| W(u_{(i)}, u_{(j)}) - W(\frac{i}{n}, \frac{j}{n}) \right|$$

$$= \max\left(\max_{i \neq j} \left| W(u_{(i)}, u_{(j)}) - W(\frac{i}{n}, \frac{j}{n}) \right|, \max_{i=j} \left| W(u_{(i)}, u_{(j)}) - W(\frac{i}{n}, \frac{j}{n}) \right| \right).$$

Using Lipschitz condition on diagonal and off-diagonal part of W and Glivenko-Cantelli theorem concludes the proof.

$\square$

With all lemmas stated, we are ready to prove the main theorem.

Theorem 4 (convergence of IGN-small in the edge probability discrete model). *Assume AS 1-4, and let $\widehat{W}_{n \times n}$ be the estimated edge probability that satisfies $\frac{1}{n}\|W_{n \times n} - \widehat{W}_{n \times n}\|_2$ converges to 0 in probability. Let Φ_c, Φ_d be continuous and discrete IGN-small. Then $\mathrm{RMSE}_U \left(\Phi_c \left([W, Diag(X)] \right), \Phi_d \left([\widehat{W}_{n \times n}, Diag(\widetilde{x_n})] \right) \right)$ converges to 0 in probability.*

Proof. Using the triangle inequality

$$\mathrm{RMSE}_U\left(\Phi_c\left([W,\mathrm{Diag}(X)]\right),\Phi_d\left([\widehat{W}_{n\times n},\mathrm{Diag}(\widetilde{x_n})]\right)\right)$$

$$=\left\|S_U\Phi_c\left([W,\mathrm{Diag}(X)]\right)-\frac{1}{\sqrt{n}}\Phi_d\left([\widehat{W}_{n\times n},\mathrm{Diag}(\widetilde{x_n})]\right)\right\|$$

$$=\|S_U\Phi_c\left([W,\mathrm{Diag}(X)]\right)-S_U\Phi_c\left([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})]\right)+S_U\Phi_c\left([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})]\right)-\Phi_dS_U([\widetilde{W_n},\mathrm{Diag}(\widetilde{x_n})])$$

$$+\Phi_dS_U([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})])-\frac{1}{\sqrt{n}}\Phi_d([\widehat{W}_{n\times n},\mathrm{Diag}(\widetilde{X_n})])\|$$

$$\leqslant\underbrace{\left\|S_U\Phi_c\left([W,\mathrm{Diag}(X)]\right)-S_U\Phi_c\left([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})]\right)\right\|}_{\text{First term: discrization error}}+\underbrace{\left\|S_U\Phi_c\left([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})]\right)-\Phi_dS_U([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})])\right\|}_{\text{Second term: sampling error}}$$

$$+\underbrace{\left\|\Phi_dS_U([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})])-\frac{1}{\sqrt{n}}\Phi_d\left([\widehat{W}_{n\times n},\mathrm{Diag}(\widetilde{x_n})]\right)\right\|}_{\text{Third term: estimation error}} \tag{11}$$

The three terms measure the different sources of error. The first term is concerned with the discretization error. The second term concerns the sampling error from the randomness of U. This term will vanish if we consider only S_n instead of S_U for IGN-small. The third term concerns the edge probability estimation error.

For the first term, it is similar to the sketch in Section 6.3. $\|S_U\Phi_c([W,\mathrm{Diag}(X)]) - S_U\Phi_c([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})])\| = \|S_U(\Phi_c([W,\mathrm{Diag}(X)]) - \Phi_c([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})]))\|$, if suffices to upper bound $\|\Phi_c([W,\mathrm{Diag}(X)]) - \Phi_c([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})])\|_{L_\infty}$ according to property of S_U in Lemma 5. Since $\|\Phi_c([W,\mathrm{Diag}(X)]) - \Phi_c([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})])\|_{L_\infty} \leqslant C(\|W - \widetilde{W_n}\|_{L_\infty} + \|\mathrm{Diag}(X) - \mathrm{Diag}(\widetilde{X_n})\|_{L_\infty})$ by Proposition 2, and $\|W - \widetilde{W_n}\|_{L_\infty}$ converges to 0 in probability according to Lemma 6, we conclude that the first term will converges to 0 in probability.

For the third term $\|\Phi_dS_U([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})]) - \frac{1}{\sqrt{n}}\Phi_d([\widehat{W}_{n\times n},\mathrm{Diag}(\widetilde{x_n})])\| = \|\frac{1}{\sqrt{n}}(\Phi_d([W_{n\times n},\mathrm{Diag}(\widetilde{x_n})]) - \Phi_d([\widehat{W}_{n\times n},\mathrm{Diag}(\widetilde{x_n})]))\| = \|\Phi_d([W_{n\times n},\mathrm{Diag}(\widetilde{x_n})]) - \Phi_d([\widehat{W}_{n\times n},\mathrm{Diag}(\widetilde{x_n})])\|_{\mathrm{pn}}$, it suffices to control the $\|[W_{n\times n},\mathrm{Diag}(\widetilde{x_n})] - [\widehat{W}_{n\times n},\mathrm{Diag}(\widetilde{x_n})]\|_{\mathrm{pn}} = \frac{1}{n}\|W_{n\times n} - \widehat{W}_{n\times n}\|_2 \leqslant \|W_{n\times n} - \widehat{W}_{n\times n}\|_{2,\infty}$, which will also goes to 0 in probability as n increases according to the statistical guarantee of edge probability estimation of neighborhood smoothing algorithm (Zhang et al., 2015), stated in Theorem 8 in Appendix G. Therefore by Proposition 2, the third term also goes to 0 in probability.

Therefore the rest work is to control the second term $\|S_U\Phi_c\left([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})]\right) - \Phi_dS_U\left([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})]\right)\|$. Again, we use the triangle inequality

Second term
$$=\left\|S_U\Phi_c\left([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})]\right)-\Phi_dS_U\left([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})]\right)\right\|$$

$$\leqslant\left\|S_U\Phi_c\left([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})]\right)-S_n\Phi_c\left([\widetilde{W_{n,E}},\mathrm{Diag}(\widetilde{X_{n,E}})]\right)\right\|+\left\|S_n\Phi_c\left([\widetilde{W_{n,E}},\mathrm{Diag}(\widetilde{X_{n,E}})]\right)-\Phi_dS_U\left([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})]\right)\right\|$$

$$=\left\|S_U\Phi_c\left([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})]\right)-S_n\Phi_c\left([\widetilde{W_{n,E}},\mathrm{Diag}(\widetilde{X_{n,E}})]\right)\right\|+\left\|S_n\Phi_c\left([\widetilde{W_{n,E}},\mathrm{Diag}(\widetilde{X_{n,E}})]\right)-\Phi_dS_n([\widetilde{W_{n,E}},\mathrm{Diag}(\widetilde{X_{n,E}})])\right\|$$

$$=\left\|S_U\Phi_c\left([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})]\right)-S_n\Phi_c\left([\widetilde{W_{n,E}},\mathrm{Diag}(\widetilde{X_{n,E}})]\right)\right\|$$

$$\leqslant\left\|S_U\Phi_c\left([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})]\right)-S_U\Phi_c\left([\widetilde{W_{n,E}},\mathrm{Diag}(\widetilde{X_{n,E}})]\right)\right\|+\left\|S_U\Phi_c\left([\widetilde{W_{n,E}},\mathrm{Diag}(\widetilde{X_{n,E}})]\right)-S_n\Phi_c\left([\widetilde{W_{n,E}},\mathrm{Diag}(\widetilde{X_{n,E}})]\right)\right\|$$

$$=\underbrace{\left\|S_U\left(\Phi_c([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})])-\Phi_c\left([\widetilde{W_{n,E}},\mathrm{Diag}(\widetilde{X_{n,E}})]\right)\right)\right\|}_{\text{term }a}+\underbrace{\left\|(S_U-S_n)\Phi_c\left([\widetilde{W_{n,E}},\mathrm{Diag}(\widetilde{X_{n,E}})]\right)\right\|}_{\text{term }b}$$

The second equality holds because $S_U([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})]) = S_n([\widetilde{W_{n,E}},\widetilde{X_{n,E}}])$ by definition of $\widetilde{W_{n,E}}$ and IGN-small (See Remark 4 for more discussion). The third equality holds by the definition of IGN-small. We will bound the term a) $\|S_U(\Phi_c([\widetilde{W_n},\mathrm{Diag}(\widetilde{X_n})]) - \Phi_c([\widetilde{W_{n,E}},\widetilde{X_{n,E}}]))\|$ and b) $\|(S_U - S_n)\Phi_c([\widetilde{W_{n,E}},\mathrm{Diag}(\widetilde{X_{n,E}})])\|$ next.

For term a) $\|S_U(\Phi_c([\widetilde{W_n}, \mathrm{Diag}(\widetilde{X_n})]) - \Phi_c([\widetilde{W_{n,E}}, \widetilde{X_{n,E}}]))\|$, if suffices to prove that $\|\Phi_c([\widetilde{W_n}, \mathrm{Diag}(\widetilde{X_n})]) - \Phi_c([\widetilde{W_{n,E}}, \widetilde{X_{n,E}}]))\|_{L_\infty}$ converges to 0 in probability. According to Proposition 2, it suffices to bound the $\|[\widetilde{W_n}, \widetilde{X_n}] - [\widetilde{W_{n,E}}, \widetilde{X_{n,E}}]\|_{L_\infty}$. Because $[\widetilde{W_n}, \widetilde{X_n}] - [\widetilde{W_{n,E}}, \widetilde{X_{n,E}}]\|_{L_\infty} = \|\widetilde{W_n} - \widetilde{W_{n,E}}\|_{L_\infty} + \|\mathrm{Diag}(\widetilde{X_n}) - \mathrm{Diag}(\widetilde{X_{n,E}})\|_{L_\infty}) \leqslant \|\widetilde{W_n} - W\|_{L_\infty} + \|\widetilde{W_{n,E}} - W\|_{L_\infty} + \|\mathrm{Diag}(\widetilde{X_n}) - \mathrm{Diag}(X)\|_{L_\infty} + \|\mathrm{Diag}(\widetilde{X_{n,E}}) - \mathrm{Diag}(X)\|_{L_\infty}$, we only need to upper bound $\|\widetilde{W_n} - W\|_{L_\infty}$, $\|\widetilde{W_{n,E}} - W\|_{L_\infty}$, $\|\mathrm{Diag}(\widetilde{X_n}) - \mathrm{Diag}(X)\|_{L_\infty})$ and $\|\mathrm{Diag}(\widetilde{X_{n,E}}) - \mathrm{Diag}(X)\|_{L_\infty})$, which are proved by Lemma 6 and Lemma 7 respectively.

For term b) $\|(S_U - S_n)\Phi_c\left([\widetilde{W_{n,E}}, \mathrm{Diag}(\widetilde{X_{n,E}})]\right)\|$

$$
\begin{aligned}
&\left\|(S_U - S_n)\Phi_c\left([\widetilde{W_{n,E}}, \mathrm{Diag}(\widetilde{X_{n,E}})]\right)\right\| \\
&= \left\|(S_U\Phi_c\left([\widetilde{W_{n,E}}, \mathrm{Diag}(\widetilde{X_{n,E}})]\right) - S_n\Phi_c\left([\widetilde{W_{n,E}}, \mathrm{Diag}(\widetilde{X_{n,E}})]\right)\right\| \\
&\leqslant \left\|(S_U\Phi_c\left([\widetilde{W_{n,E}}, \mathrm{Diag}(\widetilde{X_{n,E}})]\right) - S_U\Phi_c([W, \mathrm{Diag}(X)])\right\| + \|S_U\Phi_c([W, \mathrm{Diag}(X)]) - S_n\Phi_c([W, \mathrm{Diag}(X)])\| \\
&\quad + \left\|S_n\Phi_c([W, \mathrm{Diag}(X)]) - S_n\Phi_c\left([\widetilde{W_{n,E}}, \mathrm{Diag}(\widetilde{X_{n,E}})]\right)\right\| \\
&= \left\|(S_U(\Phi_c\left([\widetilde{W_{n,E}}, \mathrm{Diag}(\widetilde{X_{n,E}})]\right) - \Phi_c([W, \mathrm{Diag}(X)]))\right\| + \|S_U\Phi_c([W, \mathrm{Diag}(X)]) - S_n\Phi_c([W, \mathrm{Diag}(X)])\| \\
&\quad + \left\|S_n(\Phi_c\left([\widetilde{W_{n,E}}, \mathrm{Diag}(\widetilde{X_{n,E}})]\right) - \Phi_c([W, \mathrm{Diag}(X)]))\right\|
\end{aligned}
$$

For the first and last term, by the property of S_U, S_n and Φ_c, it suffices to bound $\|W - \widetilde{W_{n,E}}\|_{L_\infty}$ and $\|\mathrm{Diag}(X) - \mathrm{Diag}(\widetilde{X_{n,E}})\|_{L_\infty}$. Without loss of generality, We only prove the case for W. As $\|W - \widetilde{W_{n,E}}\|_{L_\infty}$ converges to 0 almost surely by Lemma 7, we conclude that the first and last term converges to 0 almost surely (therefore in probability). For the second term $\|S_U\Phi_c([W, \mathrm{Diag}(X)]) - S_n\Phi_c([W, \mathrm{Diag}(X)])\|$, $\Phi_c([W, \mathrm{Diag}(X)])$ is piecewise Lipschitz on diagonal and off-diagonal according to Lemma 8, and it converges to 0 almost surely according to the second part of Lemma 9.

As all terms converge to 0 in the probability or almost surely, we conclude that $\|S_U\Phi_c([W, \mathrm{Diag}(X)]) - \Phi_d([\widehat{W_{n\times n}}, \mathrm{Diag}(\widetilde{X_n})])\|$ converges to 0 in probability. □

Remark 4. *Note that we can not prove $S_n \cdot \Phi_c(\widetilde{W_{n,E}}) = \Phi_d \cdot S_n(\widetilde{W_{n,E}})$ in general. The difficulty is that starting with $\widetilde{W_{n,E}}$ of chessboard pattern, after the first layer, pattern like Figure 7(e) may appear in $\sigma \circ T_1(\widetilde{W_n})$. If T_2 is just a average/integral to map $\mathbb{R}^{n^2 \times 1}$ to $\mathbb{R}$, then $S_n \circ T_2 \circ \sigma \circ T_1(\widetilde{W_n}) = T_2 \circ \sigma \circ T_1(\widetilde{W_n})$ will not be equal to $T_2 \circ \sigma \circ T_1(S_n\widetilde{W_n})$. The reason is that both $\sigma \circ T_1(\widetilde{W_n})$ and $\sigma \circ T_1(S_n\widetilde{W_n})$ will no longer be of chessboard pattern (Figure 7(e) may occur). The diagonal in the $\sigma \circ T_1(\widetilde{W_n})$ has no effect after taking integral in T_2 as it is of measure 0. On the other hand, the diagonal in the matrix $\sigma \circ T_1(S_n\widetilde{W_n})$ will affect the average. Therefore in general, $S_n\Phi_c(\widetilde{W_{n,E}}) = \Phi_d S_n(\widetilde{W_{n,E}})$ does not hold.*

E. IGN-small Can Approximate Spectral GNN

Definition of Spectral GNN. The spectral GNN (SGNN) here stands for GNN with multiple layers of the following form $\forall j = 1, \ldots d_{\ell+1}$,

$$
z_j^{(\ell+1)} = \rho\left(\sum_{i=1}^{d_\ell} h_{ij}^{(\ell)}(L)z_i^{(\ell)} + b_j^{(\ell)}1_n\right) \in \mathbb{R}^n \tag{12}
$$

where $L = D(A)^{-\frac{1}{2}}AD(A)^{-\frac{1}{2}}$ stands for normalized adjacency,[4] $z_j^\ell, b_j^\ell \in \mathbb{R}$ denotes the embedding and bias at layer ℓ. $h : \mathbb{R} \to \mathbb{R}, h(\lambda) = \sum_{k\geq 0}\beta_k\lambda^k, h(L) = \sum_k \beta_k L^k$, i.e., we apply h to the eigenvalues of L when it is diagonalizable. Extending h to multiple input output channels which are indexed in i and j, we have $h_{ij}^{(\ell)}(\lambda) = \sum_k \beta_{ijk}^{(\ell)}\lambda^k$. By defining all components of spectral GNN for graphon, the continuous version of spectral GNN can also be defined. See Keriven et al. (2020) for details.

---

[4] We follow the same notation as Keriven et al. (2020), which is different from the conventional notation.

We first prove IGN can approximate spectral GNN arbitrarily well, both for discrete SGNN and continuous SGNN as well. Next, we show that such IGN belongs to IGN-small. We need the following simple assumption to ensure the input lies in a compact domain.

AS5. *There exists an upper bound on $\|x\|_{L_\infty}$ for the discrete case and $\|X\|_{L_\infty}$ in the continuous case.*

AS6. *$min(D(A)_{\mathrm{mean}}) \geq c_{\min}$ where $D(A)_{\mathrm{mean}}$ is defined to be $\frac{1}{n}Diag(A\mathbf{1})$. The same lower bound holds for graphon case.*

Lemma 10. *Given diagonal matrix D, matrix M and vector x, 2-IGN can approximate matrix-vector multiplication $\frac{1}{n}Mx$ and DMD arbitrarily well in L_∞ sense on a compact domain.*

Proof. Given diagonal matrix D and matrix M, to implement DMD with linear equivariant layers of 2-IGN, we first use operation 14-15 in Table 1 to copy diagonal elements in D to rows and columns of two matrix D_{row} and D_{col}. Then calculating DMD becomes entry-wise multiplication of three matrix $D_{\mathrm{row}}, M, D_{\mathrm{col}}$. Assuming all entries of D and M lies in a compact domain, we can use MLP (which is part of IGN according to Remark 8) to approximate multiplication arbitrarily well (Cybenko, 1989; Hornik et al., 1989). Note that in general matrix multiplication can not be achieved by 2-IGN but in our case, we are exploiting that $D(A)_{\mathrm{mean}}$ is a diagonal matrix. See Remark 5 for illustration.

To implement $\frac{1}{n}Mx$ with linear equivariant layers of 2-IGN, first map x into a diagonal matrix $\mathrm{Diag}(x)$ and concatenate it with M as the input $[\mathrm{Diag}(x), M] \in \mathbb{R}^{n \times n \times 2}$ to 2-IGN. Apply "copy diagonal to all columns" to the first channel and use MLP to uniformly approximates up to arbitrary precision ϵ the multiplication of first channel with the second channel. Then use operation "copy row mean" to map $\mathbb{R}^{n \times n} \to \mathbb{R}^n$ to get the $\frac{1}{n}Mx$ within ϵ precision. See Figure 8. $\square$

Remark 5. *Linear layers in 2-IGN can not implement matrix-matrix multiplication in general. When we introduce the matrix multiplication component, the expressive power of GNN in terms of WL test provably increases from 2-WL to 3-WL (Maron et al., 2019a)).*

Theorem 6. *Given n, ϵ, and $\mathrm{SGNN}_{\theta_1}(n)$, there exists a 2-IGN $\mathrm{IGN}_{\theta_2}(n)$ such that it approximates $\mathrm{SGNN}_{\theta_1}(n)$ on a compact set (input feature x_n) arbitrarily well in L_∞ sense on a compact domain.*
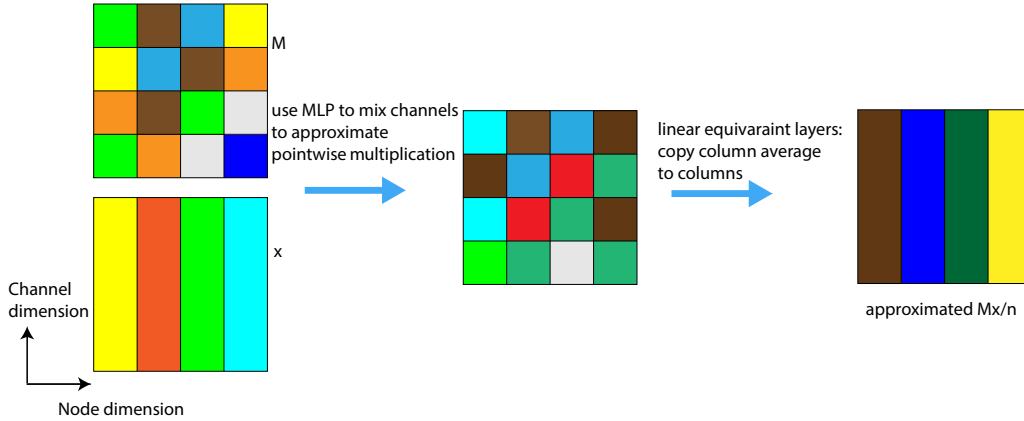

Figure 8: An illusttration of how we approximate the major building blocks of SGNN: $\frac{1}{n}Ax$.

Proof. Since IGN and SGNN has the same nonlinearity. To show that IGN can approximate SGNN, it suffices to show that IGN can approximate linear layer of SGNN, which further boils down to prove that IGN can approximate Lx.

Here we assume the input of 2-IGN is $A \in \mathbb{R}^{n \times n}$ and $x \in \mathbb{R}^{n \times d}$. We need to first show how $L = D(A)^{\frac{1}{2}}AD(A)^{\frac{1}{2}}$ can be implemented by linear layers of IGN. This is achieved by noting that $L = \frac{1}{n}D(A)_{\mathrm{mean}}^{-\frac{1}{2}}AD(A)_{\mathrm{mean}}^{-\frac{1}{2}}$ where $D(A)_{\mathrm{mean}}$ is normalized degree matrix $\frac{1}{n}\mathrm{Diag}(A\mathbf{1})$. Representing L as $\frac{1}{n}D(A)_{\mathrm{mean}}^{-\frac{1}{2}}AD(A)_{\mathrm{mean}}^{-\frac{1}{2}}$ ensures that all entries in A and $D(A)_{\mathrm{mean}}$ lies in a compact domain, which is crucial when we extending the approximation proof to the graphon case.

Now we show how $Lx = \frac{1}{n}D(A)_{\mathrm{mean}}^{-\frac{1}{2}}AD(A)_{\mathrm{mean}}^{-\frac{1}{2}}x$ is implemented. First, it is easy to see that 2-IGN can calculate exactly $D(A)_{\mathrm{mean}}$ using equivariant layers. Second, as approximating a) $f(a, b) = ab$ and b) $f(a) = \frac{1}{\sqrt{a}}$ can achieved by MLP on

compact domain, approximating $D(A)_{\text{mean}}^{-\frac{1}{2}}AD(A)_{\text{mean}}^{-\frac{1}{2}}$ can also achieved by 2-IGN layers according to Lemma 10. Third, we need to show $\frac{1}{n}D(A)_{\text{mean}}^{-\frac{1}{2}}AD(A)_{\text{mean}}^{-\frac{1}{2}}x$ can also be implemented. This is proved in Lemma 10.

There are two main functions we need to approximate with MLP: a) $f(x) = 1/\sqrt{a}$ and b) $f(a, b) = ab$.

For a) the input is then $D(A)_{\text{mean}}$ whose all entries lie in $[0, 1]$. By classical universal approximation theorem (Cybenko, 1989; Hornik et al., 1989), we know MLP can approximate a) arbitrarily well.

For b) the input is $(D(A)_{\text{mean}}^{-1/2}, A)$ for normalized adjacency matrix calculation, and (L, x) for graph signal convolution. To ensure the uniform approximation, we need to ensure all of them lie in a compact domain. This is indeed the case as all entries in $D(A)_{\text{mean}}, A, x$ are all upper bounded. We list the support of all components explicitly below

1. every entry in A is either 0 or 1 therefore lies in a compact domain.

2. similarly, all entries $D(A)_{\text{mean}}$ lies in $[c_{\min}, 1]$ by AS6, and therefore $D(A)_{\text{mean}}^{-\frac{1}{2}}$ also lies in a compact domain. As $L(A)$ is the multiplication of $D(A)_{\text{mean}}^{-1/2}, A, D(A)_{\text{mean}}^{-1/2}$, every entry of $L(A)$ also lies in compact domain.

3. input signal x has bounded l_∞-norm by assumption AS5.

4. all coefficient for operators is upper bounded and independent from n by AS2.

Since we showed the $L(A)x$ can be approximated arbitrarily well by IGN, repeating such processes and leveraging the fact that L has bounded spectral norm, we can then approximate $L^k(A)x$ up to ϵ precision. The errors ϵ depend on the approximation error of the MLP to the relevant function, the previous errors, and uniform bounds as well as uniform continuity of the approximated functions. □

Theorem 7. *Given ϵ, and a spectral GNN $c\text{SGNN}_{\theta_1}$, there exists a continous 2-IGN $c\text{IGN}_{\theta_2}$ such that it approximates $c\text{SGNN}_{\theta_1}$ on a compact set (input feature X) arbitrarily well.*

Proof. we show that all items listed in proof of Theorem 6 still holds in the continuous case

- we consider the W instead in the continuous case, where all entries still lies in a compact domain $[0, 1]$.

- similarly all entries of the continuous analog of $D(A)_{\text{mean}}, D(A)_{\text{mean}}^{-\frac{1}{2}}$, and $T(W)$ also lies in a compact domain according to AS6.

- the statements about input signal X and the coefficient for linear equivariant operators also holds in the continuous setting.

Therefore we conclude the proof. Now we are ready to prove that those IGN that can approximate SGNN well is a subset of IGN-small. □

Lemma 11. *Let $\widetilde{W_{n,E}}$ be graphon of chessboard pattern. Let $\widetilde{X_{n,E}}$ be a graphon signal with 1D chessboard pattern. $S_n \circ \widetilde{W_{n,E}X_{n,E}} = (S_n\widetilde{W_{n,E}})(S_n\widetilde{X_{n,E}})$*

Proof. Since $S_n \circ \widetilde{W_{n,E}X_{n,E}}(i) = S_n \circ \int_{j\in[0,1]} \widetilde{W_{n,E}}(i,j)\widetilde{X_{n,E}}(j)dj = \left(..., \frac{1}{\sqrt{n}}\int_{j\in[0,1]}\widetilde{W_{n,E}}(\frac{i}{n},j)\widetilde{X_{n,E}}(j), ...\right)$, it suffices to analysize $\frac{1}{\sqrt{n}}\int_{j\in[0,1]}\widetilde{W_{n,E}}(\frac{i}{n},j)\widetilde{X_{n,E}}(j)$.

Since $\widetilde{W_{n,E}}, \widetilde{X_{n,E}}$ are of chessboard pattern, we can replace integral with summation.

$$
\begin{aligned}
S_n \circ \widetilde{W_{n,E}}\widetilde{X_{n,E}}(i) &= \frac{1}{\sqrt{n}} \int_{j \in [0,1]} \widetilde{W_{n,E}}(\frac{i}{n}, j)\widetilde{X_{n,E}}(j) \\
&= \frac{1}{\sqrt{n}}\frac{1}{n} \sum_{j \in [n]} \widetilde{W_{n,E}}(\frac{i}{n}, \frac{j}{n})\widetilde{X_{n,E}}(\frac{j}{n}) \\
&= \sum_{j \in [n]} \frac{1}{n}\widetilde{W_{n,E}}(\frac{i}{n}, \frac{j}{n})(S_n\widetilde{X_{n,E}})(j) \\
&= \sum (S_n\widetilde{W_{n,E}})(i,j)(S_n\widetilde{X_{n,E}})(j) \\
&= \left((S_n\widetilde{W_{n,E}})(S_n\widetilde{X_{n,E}})\right)(i)
\end{aligned}
$$

Which concludes the proof. Note that our proof does make use of the property of multiplication between two numbers. $\square$

Remark 6. *The whole proof only relies on that $\widetilde{W_{n,E}}$ and $\widetilde{X_{n,E}}$ have checkboard patterns. Therefore replacing the multiplication with other operations (such as a MLP) will still hold.*

Theorem 5. *IGN-small can approximates spectral GNN (both discrete and continuous ones) arbitrarily well on the compact domain in the $\|\cdot\|_{L_\infty}$ sense.*

Proof. To prove this, we only need to show that $S_n\Phi_{c,\text{approx}}([\widetilde{W_{n,E}}, \widetilde{f_{n,E}}]) = \Phi_{d,\text{approx}}S_n([\widetilde{W_{n,E}}, \widetilde{f_{n,E}}])$. Here $\Phi_{c,\text{approx}}$ and $\Phi_{d,\text{approx}}$ denotes those specific IGN in Theorems 6 and 7 constructed to approximate SGNN.

To build up some intuition, let Φ_{SGNN} denotes the spectral GNN that Φ_{approx} approximates. it is easy to see that $S_n\Phi_{c,\text{SGNN}}([\widetilde{W_{n,E}}, \widetilde{f_{n,E}}]) = \Phi_{d,\text{SGNN}}S_n([\widetilde{W_{n,E}}, \widetilde{f_{n,E}}])$ due to Lemma 11 and Lemma 8.2. To show the same holds for Φ_{approx}, note that the only difference between $\widetilde{W_{n,E}}\widetilde{f_{n,E}}$ implemented by SGNN and approximated by Φ_{approx} is that Φ_{approx} use MLP to simulate multiplication between numbers. According to Remark 6, the approximated version of $\widetilde{W_{n,E}}\widetilde{f_{n,E}}$ still commutes with S_n.

Since nonlinear layer σ in Φ_{approx} also commutes with S_n according to Lemma 8.2, we can combine the result above and conclude that Φ_{approx} commutes with S_n. Therefore Φ_{approx} belongs to IGN-small, which finishs the proof. $\square$

F. More experiments

We next show full results to verify Theorems 2 and 4. The main procedure is described in Section 7.

As the ground truth is defined in the continuous regime, we use outputs of IGN on large graphs as the approximation of the unknown true limit. We experiment with two methods: a) we take the output of IGN from the deterministic edge weight continuous model as ground truth and b) we take graphs sampled from the stochastic edge weight continuous model as input to IGN and average the outputs over 10 random seeds. The case a) is shown In the main text. Here we include results for both a) and b).

G. Third-party results

G.1. Edge Probability Estimation from Zhang et al. (2015)

We next restate the setting and theorem regarding the theoretical guarantee of the edge probability estimation algorithm.

Definition 11. *For any $\delta, A_1 > 0$, let $\mathcal{F}_{\delta;L}$ de note a family of piecewise Lipschitz graphon functions $f : [0,1]^2 \to [0,1]$ such that (i) there exists an integer $K \geq 1$ and a sequence $0 = x_0 < \cdots < x_K = 1$ satisfying $\min_{0 \leqslant s \leqslant K-1} (x_{s+1} - x_s) \geq \delta$, and (ii) both $|f(u_1, v) - f(u_2, v)| \leqslant A_1|u_1 - u_2|$ and $|f(u, v_1) - f(u, v_2)| \leqslant A_1 | v_1 - v_2 |$ hold for all $u, u_1, u_2 \in [x_s, x_{s+1}], v, v_1, v_2 \in [x_t, x_{t+1}]$ and $0 \leqslant s, t \leqslant K - 1$*

Assume that $\alpha_n = 1$. It is easy to see that the setup considered in Zhang et al. (2015) is slightly more general than the setup in Keriven et al. (2020). The statistical guarantee of the edge smoothing algorithm is stated below.
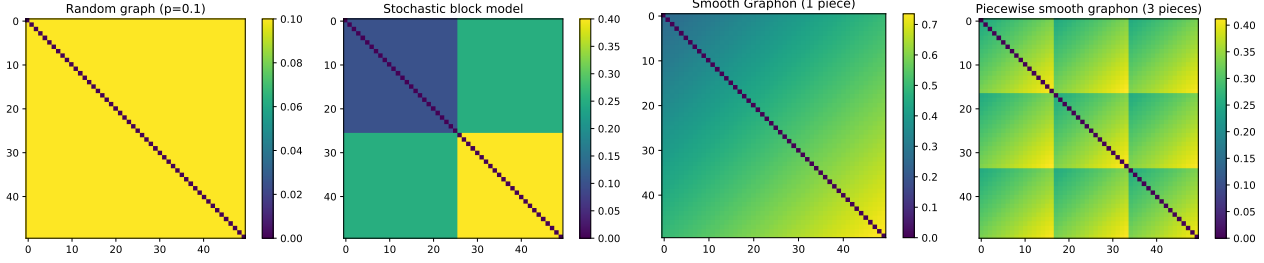
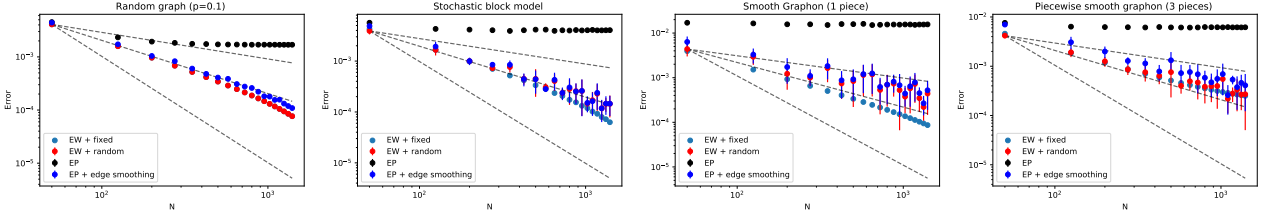Figure 9: Four graphons of increasing complexity.
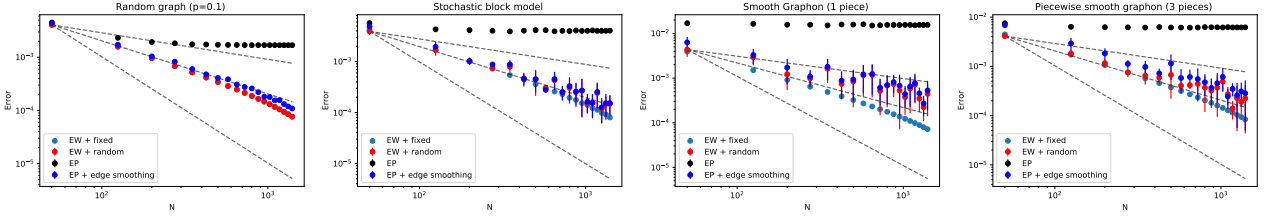

Figure 10: ground truth: random sample.

Figure 11: ground truth: grid sample.

Figure 12: The convergence error for four generative models under two ways of approximating ground truth. Three dashed line in each figure indicates the decay rate of $n^{-0.5}, n^{-1}$ and n^{-2}. EW stands for edge weight continuous model and EP stands for edge probability discrete model. As implied by Theorem 2, EW + fixed and EW + random both converges when n increases. On the other hand, EP does not converge, which is consistent with Theorem 3. After edge probability estimation, EP + edge smoothing again converges, which is consitent with Theorem 4.

Theorem 8 (Zhang et al. (2015)). *Assume that A_1 is a global constant and $\delta = \delta(n)$ depends on n, satisfying $\lim_{n\to\infty} \delta/(n^{-1}\log n)^{1/2} \to \infty$. Then the estimator $\tilde{P}$ with neighborhood $\mathcal{N}_i$ defined in Zhang et al. (2015) and $h = C(n^{-1}\log n)^{1/2}$ for any global constant $C \in (0,1]$, satisfies $\max_{f\in\mathcal{F}_{\delta;A_1}} \text{pr}\{d_{2,\infty}(\tilde{P}, P)^2 \geq C_1(\frac{\log n}{n})^{1/2}\} \leqslant n^{-C_2}$ where C_1 and C_2 are positive global constants. $d_{2,\infty}(P,Q) := n^{-1/2}\|P - Q\|_{2,\infty} = \max_i n^{-1/2}\|P_i - Q_i\|_2$.*

G.2. IGN Details

Remark 7 (independence from n). *Although for large n, the result in Maron et al. (2018) is correct. But as noted by Finzi et al. (2021), this does not hold when n is small, which is not an issue as we consider cases when n goes to infinity in this paper.*

Remark 8 (multi-channel IGN contains MLP). *For simplicity, in the main text, we focus on the case when the input and output tensor channel number is 1. The general case of multiple input and output channels is presented in Equation 9 of Maron et al. (2018). The main takeaway is that permutation equivariance does not constrain the mixing over feature channels, i.e., the space of linear equivariant maps from $\mathbb{R}^{n^\ell \times d_1} \to \mathbb{R}^{n^m \times d_2}$ if of dimension $d_1 d_2 \text{bell}(l + m)$. Therefore IGN contains MLP.*