A graphon-signal analysis of graph neural networks
Supplementary material

Note to reviewers on modified constants: when finalizing the writing of the proofs in the supplementary material, we realized that we can improve the constant in the regularity lemma from $9/4$ to 2. Hence, there is a difference in this constant between the appendix and the main paper. We also corrected the constant in the sampling lemmas. We will make the minor modification of changing the constants in the main paper in the revised paper.

A

B Basic definitions and properties of graphon-signals

In this appendix, we give basic properties of graphon-signals, cut norm, and cut distance.

B.1 Lebesgue spaces and signal spaces

For $1 \leq p < \infty$, the space $\mathcal{L}^p[0,1]$ is the space of (equivalence classes up to null-set) of measurable functions $f : [0,1] \to \mathbb{R}$, with finite L_1 norm

$$\|f\|_p = \left(\int_0^1 |f(x)|^p dx \right)^{1/p} < \infty.$$

The space $\mathcal{L}^\infty[0,1]$ is the space of (equivalence classes) of measurable functions with finite L_∞ norm

$$\|f\|_\infty = \operatorname{ess} \sup_{x\in[0,1]} |f(x)| = \inf\{a \geq 0 \mid |f(x)| \leq a \text{ for almost every } x \in [0,1]\}.$$

B.2 Properties of cut norm

Every $f \in \mathcal{L}_r^\infty[0,1]$ can be written as $f = f_+ - f_-$, where

$$f_+(x) = \begin{cases} f(x) & f(x) > 0 \\ 0 & f(x) \leq 0. \end{cases}$$

and f_- is defined similarly. It is easy to see that the supremum in (3) is attained for S which is either the support of f_+ or f_-, and

$$\|f\|_\square = \max\{\|f_+\|_1, \|f_-\|_1\}.$$

As a result, the signal cut norm is equivalent to the L_1 norm

$$\frac{1}{2}\|f\|_1 \leq \|f\|_\square \leq \|f\|_1. \tag{11}$$

Moreover, for every $r > 0$ and measurable function $W : [0,1]^2 \to [-r,r]$,

$$0 \leq \|W\|_\square \leq \|W\|_1 \leq \|W\|_2 \leq \|W\|_\infty \leq r.$$

The following lemma is from [23, Lemma 8.10].

Lemma B.1. *For every measurable $W : [0,1] \to \mathbb{R}$, the supremum*

$$\sup_{S,T\subset[0,1]} \left| \int_S \int_T W(x,y) dx dy \right|$$

is attained for some S, T.

B.3 Properties of cut distance and measure preserving bijections

Recall that we denote the standard Lebesgue measure of $[0,1]$ by μ. Let $S_{[0,1]}$ be the space of measurable bijections $[0,1] \to [0,1]$ with measurable inverse, that are measure preserving, namely, for every measurable $A \subset [0,1]$, $\mu(A) = \mu(\phi(A))$. Recall that $S'_{[0,1]}$ is the space of measurable bijections between co-null sets of $[0,1]$.

For $\phi \in S_{[0,1]}$ or $\phi \in S'_{[0,1]}$, we define $W^{\phi}(x,y) := W(\phi(x),\phi(y))$. In case $\phi \in S'_{[0,1]}$, W^{ϕ} is only define up to a null-set, and we arbitrarily set W to 0 in this null-set. This does not affect our analysis, as the cut norm is not affected by changes to the values of functions on a null sets. The *cut-metric* between graphons is then defined to be

$$\begin{aligned}\delta_{\square}(W, W^{\phi}) &= \inf_{\phi \in S_{[0,1]}} \|W - W^{\phi}\|_{\square} \\ &= \inf_{\phi \in S_{[0,1]}} \sup_{S,T \subseteq [0,1]} \left| \int_{S \times T} \big(W(x,y) - W(\phi(x),\phi(y))\big) dxdy \right|.\end{aligned}$$

Remark B.2. *Note that $\delta_{\square}$ can be defined equivalently with respect to $\phi \in S'_{[0,1]}$. Indeed, By [23, Equation (8.17) and Theorem 8.13], $\delta_{\square}$ can be defined equivalently with respect to the measure preserving maps that are not necessarily invertible. These include the extensions of mappings from $S'_{[0,1]}$ by defining $\phi(x) = 0$ for every x in the co-null set underlying ϕ.*

Similarly to the graphon case, the graphon-signal distance $\delta_{\square}$ is a pseudo-metric. By introducing an equivalence relation $(W,f) \sim (V,g)$ if $\delta_{\square}((W,f),(V,g)) = 0$, and the quotient space $\widetilde{\mathcal{WL}_r} := \mathcal{WL}_r/\sim$, $\widetilde{\mathcal{WL}_r}$ is a metric space with a metric $\delta_{\square}$ defined by $\delta_{\square}([(W,f)],[V,g)]) = d_{\square}(W,V)$ where $[(W,f)],[(V,g)]$, are the equivalence classes of (W,f) and (V,g) respectively. By abuse of terminology, we call elements of $\widetilde{\mathcal{WL}_r}$ also graphon-signals.

Remark B.3. *We note that $\widetilde{\mathcal{WL}_r} \neq \widetilde{\mathcal{W}_0} \times \widetilde{\mathcal{L}_r^{\infty}[0,1]}$ (for the natural definition of $\widetilde{\mathcal{L}_r^{\infty}[0,1]}$), since in $\widetilde{\mathcal{WL}_r}$ we require that the measure preserving bijection is shared between the graphon W and the signal f. Sharing the measure preserving bijetion between W and f is an important modelling requirement, as ϕ is seen as a "re-indexing" of the node set $[0,1]$. When re-indexing a node x, both the neighborhood $W(x,\cdot)$ of x and the signal value $f(x)$ at x should change together, otherwise, the graphon and the signal would fall out of alignment.*

We identify graphs with their induced graphons and signal with their induced signals

C Graphon-signal regularity lemmas

In this appendix, we prove a number of versions of the graphon-signal regularity lemma, where Theorem 3.4 is one version.

C.1 Properties of partitions and step functions

Given a partition $\mathcal{P}_k$ and $d \in \mathbb{N}$, the next lemma shows that there is an equiparition $\mathcal{E}_n$ such that the space $\mathcal{S}^d_{\mathcal{E}_n}$ uniformly approximates the space $\mathcal{S}^d_{\mathcal{P}_k}$ in $\mathcal{L}^1[0,1]^d$ norm (see Definition 3.3).

Lemma C.1 (Equitizing partitions)**.** *Let $\mathcal{P}_k$ be a partition of $[0,1]$ into k sets (generally not of the same measure). Then, for any $n > k$ there exists an equipartition $\mathcal{E}_n$ of $[0,1]$ into n sets such that any function $F \in \mathcal{S}^d_{\mathcal{P}_k}$ can be approximated in $L_1[0,1]^d$ by a function from $F \in \mathcal{S}^d_{\mathcal{E}_n}$ up to small error. Namely, for every $F \in \mathcal{S}^d_{\mathcal{P}_k}$ there exists $F' \in \mathcal{S}^d_{\mathcal{E}_n}$ such that*

$$\|F - F'\|_1 \leq d\|F\|_{\infty}\frac{k}{n}.$$

Proof. Let $\mathcal{P}_k = \{P_1, \ldots, P_k\}$ be a partition of $[0,1]$. For each i, we divide P_i into subsets $\mathbf{P}_i = \{P_{i,1}, \ldots, P_{i,m_i}\}$ of measure $1/n$ (up to the last set) with a residual, as follows. If $\mu(P_i) < 1/n$, we choose $\mathbf{P}_i = \{P_{i,1} = P_i\}$. Otherwise, we take $P_{i,1}, \ldots, P_{i,m_i-1}$ of measure $1/n$, and $\mu(P_{i,m_i}) \leq 1/n$. We call P_{i,m_i} the remainder.

We now define the sequence of sets of measure $1/n$

$$\mathcal{Q} := \{P_{1,1}, \ldots, P_{1,m_1-1}, P_{2,1}, \ldots, P_{2,m_2-1}, \ldots, P_{k,1}, \ldots, P_{k,m_k-1}\}, \tag{12}$$

where, by abuse of notation, for any i such that $m_i = 1$, we set $\{P_{i,1}, \ldots, P_{i,m_i-1}\} = \emptyset$ in the above formula. Note that in general $\cup\mathcal{Q} \neq [0,1]$. We moreover define the union of residuals $\Pi := P_{1,m_1} \cup P_{2,m_2} \cup \cdots \cup P_{k,m_k}$. Note that $\mu(\Pi) = 1 - \mu(\cup\mathcal{Q}) = 1 - k\frac{1}{n} = h/n$, where k is the number of elements in $\mathcal{Q}$, and $h = n - k$. Hence, we can partition Π into h parts $\{\Pi_1, \ldots \Pi_h\}$ of measure $1/n$ with no residual. Thus we have obtain the equipartition of $[0,1]$ to n sets of measure $1/n$

$$\mathcal{E}_n := \{P_{1,1}, \ldots, P_{1,m_1-1}, P_{2,1}, \ldots, P_{2,m_2-1}, \ldots, S_{k,1}, \ldots, S_{k,m_k-1}, \Pi_1, \Pi_2, \ldots, \Pi_h\}. \tag{13}$$

For convenience, we also denote $\mathcal{E}_n = \{Z_1, \ldots, Z_n\}$.

Let

$$F(x) = \sum_{j=(j_1,\ldots,j_d)\in[k]^d} c_j \prod_{l=1}^{d} \mathbb{1}_{P_{j_l}}(x_l) \in \mathcal{S}^d_{\mathcal{P}_k}.$$

We can write F with respect to the equipartition $\mathcal{E}_n$ as

$$F(x) = \sum_{j=(j_1,\ldots,j_d)\in[n]^d;\ \forall l=1,\ldots,d,\ Z_{j_l}\not\subset\Pi} \tilde{c}_j \prod_{l=1}^{d} \mathbb{1}_{Z_{j_l}}(x_l) \ + \ E(x),$$

for some $\{\tilde{c}_j\}$ with the same values as the values of $\{c_j\}$. Here, E is supported in the set $\Pi^{(d)} \subset [0,1]^d$, defied by

$$\Pi^{(d)} = \big(\Pi \times [0,1]^{d-1}\big) \cup \big([0,1] \times \Pi \times [0,1]^{d-2}\big) \cup \ldots \cup \big([0,1]^{d-1} \times \Pi\big).$$

Consider the step function

$$F'(x) = \sum_{j=(j_1,\ldots,j_d)\in[n]^d;\ \forall l=1,\ldots,d,\ Z_{j_l}\not\subset\Pi} \tilde{c}_j \prod_{l=1}^{d} \mathbb{1}_{Z_{j_l}}(x_l) \in \mathcal{S}^d_{\mathcal{E}_n}.$$

Since $\mu(\Pi) = k/n$, we have $\mu(\Pi^{(d)}) = dk/n$, and so

$$\|F - F'\|_1 \leq d\|F\|_\infty \frac{k}{n}.$$

■

Lemma C.2. *Let $\{Q_1, Q_2, \ldots, Q_m\}$ partition of $[0,1]$. Let $\{I_1, I_2, \ldots, I_m\}$ be a partition of $[0,1]$ into intervals, such that for every $j \in [m]$, $\mu(Q_j) = \mu(I_j)$. Then, there exists a measure preserving bijection $\phi : [0,1] \to [0,1] \in S'_{[0,1]}$ such that*[4]

$$\phi(Q_j) = I_j$$

Proof. By the definition of a standard probability space, the measure space induced by $[0,1]$ on a non-null subset $Q_j \subseteq [0,1]$ is a standard probability space. Moreover, each Q_j is atomless, since $[0,1]$ is atomless. Since there is a measure-preserving bijection (up to null-set) between any two atomless standard probability spaces, we obtain the result. ■

Lemma C.3. *Let $\mathcal{S} = \{S_j \subset [0,1]\}_{j=0}^{m-1}$ be a collection of measurable sets (that are not disjoint in general), and $d \in \mathbb{N}$. Let $\mathcal{C}^d_{\mathcal{S}}$ be the space of functions $F : [0,1]^d \to \mathbb{R}$ of the form*

$$F(x) = \sum_{j=(j_1,\ldots,j_d)\in[m]^d}^{m} c_j \prod_{l=1}^{d} \mathbb{1}_{S_{j_l}}(x_l),$$

for some choice of $\{c_j \in \mathbb{R}\}_{j\in[m]^d}$. Then, there exists a partition $\mathcal{P}_k = \{P_1, \ldots, P_k\}$ into $k = 2^m$ sets, that depends only on $\mathcal{S}$, such that

$$\mathcal{C}^d_{\mathcal{S}} \subset \mathcal{S}^d_{\mathcal{P}_k}.$$

[4]Namely, there is a measure preserving bijection ϕ between two co-null sets C_1 and C_2 of $[0,1]$, such that $\phi(Q_j \cap C_1) = I_j \cap C_2$.

Proof. The partition $\mathcal{P}_k = \{P_1, \dots, P_k\}$ is defined as follows. Let

$$\tilde{\mathcal{P}} = \big\{P \subset [0,1] \mid \exists\, x \in [0,1],\ P = \cap\{S_j \in \mathcal{S} | x \in S_j\}\big\}.$$

We must have $|\tilde{\mathcal{P}}| \leq 2^m$. Indeed, there are at most 2^m different subsets of $\mathcal{S}$ for the intersections. We endow an arbitrarily order to $\tilde{\mathcal{P}}$ and turn it into a sequence. If the size of $\tilde{\mathcal{P}}$ is strictly smaller than 2^m, we add enough copies of $\{\emptyset\}$ to $\tilde{\mathcal{P}}$ to make the size of the sequence 2^m, that we denote by $\mathcal{P}_k$, where $k = 2^m$. ■

The following simple lemma is proved similarly to Lemma C.3. We give it without proof.

Lemma C.4. *Let $\mathcal{P}_k = \{P_1, \dots, P_k\}, \mathcal{Q}_m = \{Q_1, \dots, Q_k\}$ be two partitions. Then, there exists a partition $\mathcal{Z}_{km}$ into km sets such that for every d,*

$$\mathcal{S}^d_{\mathcal{P}_k} \subset \mathcal{S}^d_{\mathcal{Z}_{mk}}, \quad \text{and} \quad \mathcal{S}^d_{\mathcal{Q}_m} \subset \mathcal{S}^d_{\mathcal{Z}_{mk}}.$$

C.2 List of graphon-signal regularity lemmas

The following lemma from [24, Lemma 4.1] is a tool in the proof of the weak regularity lemma.

Lemma C.5. *Let $\mathcal{K}_1, \mathcal{K}_2, \dots$ be arbitrary nonempty subsets (not necessarily subspaces) of a Hilbert space $\mathcal{H}$. Then, for every $\epsilon > 0$ and $v \in \mathcal{H}$ there is $m \leq \lceil 1/\epsilon^2 \rceil$ and $v_i \in \mathcal{K}_i$ and $\gamma_i \in \mathbb{R}$, $i \in [m]$, such that for every $w \in \mathcal{K}_{m+1}$*

$$\left| \left\langle w, v - (\sum_{i=1}^{m} \gamma_i v_i) \right\rangle \right| \leq \epsilon \, \|w\| \|v\|. \tag{14}$$

The following theorem is an extension of the graphon regularity lemma from [24] to the case of graphon-signals. Much of the proof follows the steps of [24].

Theorem C.6 (Weak regularity lemma for graphon-signals)**.** *Let $\epsilon, \rho > 0$. For every $(W, f) \in \mathcal{WL}_r$ there exists a partition $\mathcal{P}_k$ of $[0,1]$ into $k = \lceil r/\rho \rceil \left(2^{2^{\lceil 1/\epsilon^2 \rceil}}\right)$ sets, a step function graphon $W_k \in \mathcal{S}^2_{\mathcal{P}_k} \cap \mathcal{W}_0$ and a step function signal $f_k \in \mathcal{S}^1_{\mathcal{P}_k} \cap \mathcal{L}^\infty_r[0,1]$, such that*

$$\|W - W_k\|_\square \leq \epsilon \quad \text{and} \quad \|f - f_k\|_\square \leq \rho. \tag{15}$$

Proof. We first analyze the graphon part. In Lemma C.5, set $\mathcal{H} = \mathcal{L}^2([0,1]^2)$ and for all $i \in \mathbb{N}$, set

$$\mathcal{K}_i = \mathcal{K} = \big\{\mathbb{1}_{S \times T} \bigm| S, T \subset [0,1] \text{ measurable}\big\}.$$

Then, by Lemma C.5, there exists $m \leq \lceil 1/\epsilon^2 \rceil$ two sequences of sets $\mathcal{S}_m = \{S_i\}_{i=1}^m$, $\mathcal{T}_m = \{T_i\}_{i=1}^m$, a sequence of coefficients $\{\gamma_i \in \mathbb{R}\}_{i=1}^m$, and

$$W_\epsilon = \sum_{i=1}^{m} \gamma_i \mathbb{1}_{S_i \times T_i},$$

such that for any $V \in \mathcal{K}$, given by $V(x,y) = \mathbb{1}_S(x)\mathbb{1}_T(y)$, we have

$$\left| \int V(x,y)\big(W(x,y) - W_\epsilon(x,y)\big)dxdy \right| = \left| \int_S \int_T \big(W(x,y) - W_\epsilon(x,y)\big)dxdy \right| \tag{16}$$

$$\leq \epsilon \|\mathbb{1}_{S \times T}\| \|W\| \leq \epsilon. \tag{17}$$

We may choose exactly $m = \lceil 1/\epsilon^2 \rceil$ by adding copies of the empty set to $\mathcal{S}_m$ and $\mathcal{T}_m$, if the constant m guaranteed by Lemma C.5 is strictly less than $\lceil 1/\epsilon^2 \rceil$. Consider the concatenation of the two sequences $\mathcal{T}_m, \mathcal{S}_m$ given by $\mathcal{Y}_{2m} = \mathcal{T}_m \cup \mathcal{S}_m$. Note that in the notation of Lemma C.3, $W_\epsilon \in \mathcal{C}^2_{\mathcal{Y}_{2m}}$. Hence, by Lemma C.3, there exists a partition $\mathcal{Q}_n$ into $n = 2^{2m} = 2^{2\lceil \frac{1}{\epsilon^2} \rceil}$ sets, such that W_ϵ is a step graphon with respect to $\mathcal{Q}_n$.

To analyze the signal part, we partition the range of the signal $[-r, r]$ into $j = \lceil r/\rho \rceil$ intervals $\{J_i\}_{i=1}^j$ of length less or equal to 2ρ, where the left edge point of each J_i is $-r + (i-1)\frac{\rho}{r}$. Consider the partition of $[0,1]$ based on the preimages $\mathcal{Y}_j = \{Y_i = f^{-1}(J_i)\}_{i=1}^{j}$. It is easy to see that for the step signal

$$f_\rho(x) = \sum_{i=1}^{j} a_i \mathbb{1}_{Y_i}(x),$$

where a_i the midpoint of the interval Y_i, we have

$$\|f - f_\rho\|_\square \leq \|f - f_\rho\|_1 \leq \rho.$$

Lastly, by Lemma C.4, there is a partition $\mathcal{P}_k$ of $[0,1]$ into $k = \lceil r/\rho \rceil \left(2^{2^{\lceil 1/\epsilon^2 \rceil}}\right)$ sets such that $W_\epsilon \in \mathcal{S}^2_{\mathcal{P}_k}$ and $f_\rho \in \mathcal{S}^1_{\mathcal{P}_k}$.

■

Corollary C.7 (Weak regularity lemma for graphon-signals – version 2)**.** *Let $r > 0$ and $c > 1$. For every sufficiently small $\epsilon > 0$ (namely, ϵ that satisfies (19)), and for every $(W, f) \in \mathcal{WL}_r$ there exists a partition $\mathcal{P}_k$ of $[0,1]$ into $k = \left(2^{\lceil 2c/\epsilon^2 \rceil}\right)$ sets, a step graphon $W_k \in \mathcal{S}^2_{\mathcal{P}_k} \cap \mathcal{W}_0$ and a step signal $f_k \in \mathcal{S}^1_{\mathcal{P}_k} \cap \mathcal{L}^\infty_r[0,1]$, such that*

$$d_\square\big((W,f),(W_k,f_k)\big) \leq \epsilon.$$

Proof. First, evoke Theorem C.6, with errors $\|W - W_k\|_\square \leq \nu$ and $\|f - f_k\|_\square \leq \rho = \epsilon - \nu$. We now show that there is some $\epsilon_0 > 0$ such that for every $\epsilon < \epsilon_0$, there is a choice of ν such that the number of sets in the partition, guaranteed by Theorem C.6, satisfies

$$k(\nu) := \lceil r/(\epsilon - \nu) \rceil \left(2^{2^{\lceil 1/\nu^2 \rceil}}\right) \leq 2^{\lceil 2c/\epsilon^2 \rceil}.$$

Denote $c = 1 + t$. In case

$$\nu \geq \sqrt{\frac{2}{2(1+0.5t)/\epsilon^2 - 1}}, \tag{18}$$

we have

$$2^{2\lceil 1/\nu^2 \rceil} \leq 2^{2(1+0.5t)/\epsilon^2}.$$

On the other hand, for

$$\nu \leq \epsilon - \frac{r}{2^{t/\epsilon^2} - 1},$$

we have

$$\lceil r/(\epsilon - \nu) \rceil \leq 2^{2(0.5t)/\epsilon^2}.$$

The reconcile these two conditions, we restrict to ϵ such that

$$\epsilon - \frac{r}{2^{t/\epsilon^2} - 1} \geq \sqrt{\frac{2}{2(1+0.5t)/\epsilon^2 - 1}}. \tag{19}$$

There exists ϵ_0 that depends on c and r (and hence also on t) such that for every $\epsilon < \epsilon_0$ (19) is satisfied. Indeed, for small enough ϵ,

$$\frac{1}{2^{t/\epsilon^2} - 1} = \frac{2^{-t/\epsilon^2}}{1 - 2^{-t/\epsilon^2}} < 2^{-t/\epsilon^2} < \frac{\epsilon}{r}\left(1 - \frac{1}{1+0.1t}\right),$$

so

$$\epsilon - \frac{r}{2^{t/\epsilon^2} - 1} > \epsilon(1 + 0.1t).$$

Moreover, for small enough ϵ,

$$\sqrt{\frac{2}{2(1+0.5t)/\epsilon^2 - 1}} = \epsilon\sqrt{\frac{1}{(1+0.5t) - \epsilon^2}} < \epsilon/(1+0.4t).$$

Hence, for every $\epsilon < \epsilon_0$, there is a choice of ν such that

$$k(\nu) = \lceil r/(\epsilon - \nu) \rceil \left(2^{2\lceil 1/\nu^2 \rceil}\right) \leq 2^{2(0.5t)/\epsilon^2} 2^{2(1+0.5t)/\epsilon^2} \leq 2^{\lceil 2c/\epsilon^2 \rceil}.$$

Lastly, we add as many copies of $\emptyset$ to $\mathcal{P}_{k(\nu)}$ as needed so that we get a sequence of $k = 2^{\lceil 2c/\epsilon^2 \rceil}$ sets.

■

Theorem C.8 (Regularity lemma for graphon-signals – equipartition version)**.** *Let $c > 1$ and $r > 0$. For any sufficiently small $\epsilon > 0$, and every $(W, f) \in \mathcal{WL}_r$ there exists $\phi \in S'_{[0,1]}$, a step function graphon $[W^\phi]_n \in \mathcal{S}^2_{\mathcal{I}_n} \cap \mathcal{W}_0$ and a step signal $[f^\phi]_n \in \mathcal{S}^1_{\mathcal{I}_n} \cap \mathcal{L}^\infty_r[0,1]$, such that*

$$d_\square\Big(\left(W^\phi, f^\phi\right) , \left([W^\phi]_n, [f^\phi]_n\right) \Big) \leq \epsilon, \tag{20}$$

where $\mathcal{I}_n$ is the equipartition of $[0,1]$ into $n = 2^{\lceil 2c/\epsilon^2 \rceil}$ intervals.

Proof. Let $c = 1 + t > 1$, $\epsilon > 0$ and $0 < \alpha, \beta < 1$. In Corollary C.7, consider the approximation error

$$d_\square\big((W, f), (W_k, f_k)\big) \leq \alpha\epsilon.$$

with a partition $\mathcal{P}_k$ into $k = 2^{\lceil \frac{2(1+t/2)}{(\epsilon\alpha)^2} \rceil}$ sets. We next equatize the partition $\mathcal{P}_k$ up to error $\epsilon\beta$. More accurately, in Lemma C.1, we choose

$$n = \lceil 2^{\frac{2(1+0.5t)}{(\epsilon\alpha)^2}+1} / (\epsilon\beta) \rceil,$$

and note that

$$n \geq 2^{\lceil \frac{2(1+0.5t)}{(\epsilon\alpha)^2} \rceil} \lceil 1/\epsilon\beta \rceil = k \lceil 1/\epsilon\beta \rceil.$$

By Lemma C.1 and by the fact that the cut norm is bounded by L_1 norm, there exists an equipartition $\mathcal{E}_n$ into n sets, and step functions W_n and f_n with respect to $\mathcal{E}_n$ such that

$$\|W_k - W_n\|_\square \leq 2\epsilon\beta \quad \text{and} \quad \|f_k - f_n\|_1 \leq r\epsilon\beta.$$

Hence, by the triangle inequality,

$$d_\square\big((W, f), (W_n, f_n)\big) \leq d_\square\big((W, f), (W_k, f_k)\big) + d_\square\big((W_k, f_k), (W_n, f_n)\big) \leq \epsilon(\alpha + (2 + r)\beta).$$

In the following, we restrict to choices of α and β which satisfy $\alpha + (2 + r)\beta = 1$. Consider the function $n : (0, 1) \to \mathbb{N}$ defined by

$$n(\alpha) := \lceil 2^{\frac{4(1+0.5t)}{(\epsilon\alpha)^2}+1} / (\epsilon\beta) \rceil = \lceil (2 + r) \cdot 2^{\frac{9(1+0.5t)}{4(\epsilon\alpha)^2}+1} / (\epsilon(1 - \alpha)) \rceil.$$

Using a similar technique as in the proof of Corollary C.7, there is $\epsilon_0 > 0$ that depends on c and r (and hence also on t) such that for every $\epsilon < \epsilon_0$, we may choose α_0 (that depends on ϵ) which satisfies

$$n(\alpha_0) = \lceil (2 + r) \cdot 2^{\frac{2(1+0.5t)}{(\epsilon\alpha_0)^2}+1} / (\epsilon(1 - \alpha_0)) \rceil < 2^{\lceil \frac{2c}{\epsilon^2} \rceil}. \tag{21}$$

Moreover, there is a choice α_1 which satisfies

$$n(\alpha_1) = \lceil (2 + r) \cdot 2^{\frac{2(1+0.5t)}{(\epsilon\alpha_1)^2}+1} / (\epsilon(1 - \alpha_1)) \rceil > 2^{\lceil \frac{2c}{\epsilon^2} \rceil}. \tag{22}$$

We note that the function $n : (0, 1) \to \mathbb{N}$ satisfies the following intermediate value property. For every $0 < \alpha_1 < \alpha_2 < 1$ and every $m \in \mathbb{N}$ between $n(\alpha_1)$ and $n(\alpha_2)$, there is a point $\alpha \in [\alpha_1, \alpha_2]$ such that $n(\alpha) = m$. This follows the fact that $\alpha \mapsto (2 + r) \cdot 2^{\frac{2(1+0.5t)}{(\epsilon\alpha)^2}+1} / (\epsilon(1 - \alpha))$ is a continuous function. Hence, by (21) and (22), there is a point α (and β such that $\alpha + (2 + r)\beta = 1$) such that

$$n(\alpha) = n = \lceil 2^{\frac{2(1+0.5t)}{(\epsilon\alpha)^2}+1} / (\epsilon\beta) \rceil = 2^{\lceil 2c/\epsilon^2 \rceil}.$$

■

By a slight modification of the above proof, we can replace n with the constant $n = \lceil 2^{\frac{2c}{\epsilon^2}} \rceil$. As a result, we can easily prove that for any $n' \geq 2^{\lceil \frac{2c}{\epsilon^2} \rceil}$ we have the approximation property (20) with n' instead of n. This is done by choosing an appropriate $c' > c$ and using Theorem C.8 on c', giving a constant $n' = \lceil 2^{\frac{2c'}{\epsilon^2}} \rceil \geq 2^{\lceil \frac{2c}{\epsilon^2} \rceil} = n$. This leads to the following corollary.

Corollary C.9 (Regularity lemma for graphon-signals – equipartition version 2)**.** *Let $c > 1$ and $r > 0$. For any sufficiently small $\epsilon > 0$, for every $n \geq 2^{\lceil \frac{2c}{\epsilon^2} \rceil}$ and every $(W, f) \in \mathcal{WL}_r$, there exists $\phi \in S'_{[0,1]}$, a step function graphon $[W^\phi]_n \in \mathcal{S}^2_{\mathcal{I}_n} \cap \mathcal{W}_0$ and a step function signal $[f^\phi]_n \in \mathcal{S}^1_{\mathcal{I}_n} \cap \mathcal{L}^\infty_r[0,1]$, such that*

$$d_\square\Big(\left(W^\phi, f^\phi\right) , \left([W^\phi]_n, [f^\phi]_n\right) \Big) \leq \epsilon,$$

where $\mathcal{I}_n$ is the equipartition of $[0,1]$ into n intervals.

Next, we prove that we can use the average of the graphon and the signal in each part for the approximating graphon-signal. For that we define the projection of a graphon signal upon a partition.

Definition C.10. *Let $\mathcal{P}_n = \{P_1, \ldots, P_n\}$ be a partition of $[0,1]$, and $(W, f) \in \mathcal{WL}_r$. We define the* projection *of (W, f) upon $(\mathcal{S}^2_{\mathcal{P}} \times \mathcal{S}^1_{\mathcal{P}}) \cap \mathcal{WL}_r$ to be the step graphon-signal $(W, f)_{\mathcal{P}_n} = (W_{\mathcal{P}_n}, f_{\mathcal{P}_n})$ that attains the value*

$$W_{\mathcal{P}_n}(x,y) = \int_{P_i \times P_j} W(x,y)dxdy \ , \quad f_{\mathcal{P}_n}(x) = \int_{P_i} f(x)dx$$

for every $(x,y) \in P_i \times P_j$.

At the cost of replacing the error ϵ by 2ϵ, we can replace W' with its projection. This was shown in [1]. Since this paper does not use the exact same setting as us, for completeness, we write a proof of the claim below.

Corollary C.11 (Regularity lemma for graphon-signals – projection version)**.** *For any $c > 1$, and any sufficiently small $\epsilon > 0$, for every $n \geq 2^{\lceil \frac{8c}{\epsilon^2} \rceil}$ and every $(W, f) \in \mathcal{WL}_r$, there exists $\phi \in S'_{[0,1]}$, such that such that*

$$d_{\square}\Big(\ (W^{\phi}, f^{\phi}) \ , \ ([W^{\phi}]_{\mathcal{I}_n}, [f^{\phi}]_{\mathcal{I}_n}) \ \Big) \leq \epsilon.$$

where $\mathcal{I}_n$ is the equipartition of $[0,1]$ into n intervals.

We first prove a simple lemma.

Lemma C.12. *Let $\mathcal{P}_n = \{P_1, \ldots, P_n\}$ be a partition of $[0,1]$, and Let $V, R \in \mathcal{S}^2_{\mathcal{P}_n} \cap \mathcal{W}_0$. Then, the supremum of*

$$\sup_{S,T \subset [0,1]} \left| \int_S \int_T \big(V(x,y) - R(x,y)\big)dxdy \right| \tag{23}$$

is attained for S, T of the form

$$S = \bigcup_{i \in s} P_i \ , \quad T = \bigcup_{j \in t} P_j,$$

where $t, s \subset [n]$. Similarly for any two signals $f, g \in \mathcal{S}^1_{\mathcal{P}_n} \cap \mathcal{L}^{\infty}_r[0,1]$, the supremum of

$$\sup_{S \subset [0,1]} \left| \int_S \big(f(x) - g(x)\big)dx \right| \tag{24}$$

is attained for S of the form

$$S = \bigcup_{i \in s} P_i,$$

where $s \subset [n]$.

Proof. First, by Lemma B.1, the supremum of (23) is attained for some $S, T \subset [0,1]$. Given the maximizers S, T, without loss of generality, suppose that

$$\int_S \int_T \big(V(x,y) - R(x,y)\big)dxdy > 0.$$

we can improve T as follows. Consider the set $t \subset [n]$ such that for every $j \in t$

$$\int_S \int_{T \cap P_j} \big(V(x,y) - R(x,y)\big)dxdy > 0.$$

By increasing the set $T \cap P_j$ to P_j, we can only increase the size of the above integral. Indeed,

$$\begin{aligned}
\int_S \int_{P_j} \big(V(x,y) - R(x,y)\big)dxdy &= \frac{\mu(P_j)}{\mu(T \cap P_j)} \int_S \int_{T \cap P_j} \big(V(x,y) - R(x,y)\big)dxdy \\
&\geq \int_S \int_{T \cap P_j} \big(V(x,y) - R(x,y)\big)dxdy.
\end{aligned}$$

Hence, by increasing T to

$$T' = \bigcup_{\{j | T \cap P_j \neq \emptyset\}} P_j,$$

we get

$$\int_S \int_{T'} \big(V(x,y) - R(x,y)\big)dxdy \geq \int_S \int_T \big(V(x,y) - R(x,y)\big)dxdy.$$

We similarly replace each $T \cap P_j$ such that

$$\int_S \int_{T \cap P_j} \big(V(x,y) - R(x,y)\big)dxdy \leq 0$$

by the empty set. We now repeat this process for S, which concludes the proof for the graphon part.

For the signal case, let $f = f_+ - f_-$, and suppose without loss of generality that $\|f\|_\square = \|f\|_1$. It is easy to see that the supremum of (24) is attained for the support of f_+, which has the required form. ■

Proof. Proof of Corollary C.11 Let $W_n \in \mathcal{S}_{\mathcal{P}_n} \cap \mathcal{W}_0$ be the step graphon guaranteed by Corollary C.9, with error $\epsilon/2$ and measure preserving bijection $\phi \in S'_{[0,1]}$. Without loss of generality, we suppose that $W^\phi = W$. Otherwise, we just denote $W' = W^\phi$ and replace the notation W with W' in the following. By Lemma C.12, the infimum underlying $\|W_{\mathcal{P}_n} - W_n\|_\square$ is attained for for some

$$S = \bigcup_{i \in s} P_i\ , \quad T = \bigcup_{j \in t} P_j.$$

We now have, by definition of the projected graphon,

$$\begin{aligned}
\|W_n - W_{\mathcal{P}_n}\|_\square &= \left| \sum_{i \in s, j \in t} \int_{P_i} \int_{P_j} (W_{\mathcal{P}_n}(x,y) - W_n(x,y))dxdy \right| \\
&= \left| \sum_{i \in s, j \in t} \int_{P_i} \int_{P_j} (W(x,y) - W_n(x,y))dxdy \right| \\
&= \left| \int_S \int_T (W(x,y) - W_n(x,y))dxdy \right| = \|W_n - W\|_\square.
\end{aligned}$$

Hence, by the triangle inequality,

$$\|W - W_{\mathcal{P}_n}\|_\square \leq \|W - W_n\|_\square + \|W_n - W_{\mathcal{P}_n}\|_\square < 2\|W_n - W\|_\square.$$

A similar argument shows

$$\|f - f_{\mathcal{P}_n}\|_\square < 2\|f_n - f\|_\square.$$

Hence,

$$d_\square\Big(\big(W^\phi, f^\phi\big)\ ,\ \big([W^\phi]_{\mathcal{I}_n}, [f^\phi]_{\mathcal{I}_n}\big) \Big) \leq 2 d_\square\Big(\big(W^\phi, f^\phi\big)\ ,\ \big([W^\phi]_n, [f^\phi]_n\big) \Big) \leq \epsilon.$$

■

D Compactness and covering number of the graphon-signal space

In this appendix we prove Theorem 3.5.

Given a partition $\mathcal{P}_k$, recall that

$$[\mathcal{WL}_r]_{\mathcal{P}_k} := (\mathcal{W}_0 \cap \mathcal{S}^2_{\mathcal{P}_k}) \times (\mathcal{L}^\infty_r[0,1] \cap \mathcal{S}^1_{\mathcal{P}_k})$$

is called the space of SBMs or step graphon-signals with respect to $\mathcal{P}_k$. Recall that $\widetilde{\mathcal{WL}_r}$ is the space of equivalence classes of graphon-signals with zero $\delta_\square$ distance, with the $\delta_\square$ metric (defined on arbitrary representatives). By abuse of terminology, we call elements of $\widetilde{\mathcal{WL}_r}$ also graphon-signals.

Theorem D.1. *The metric space $(\widetilde{\mathcal{WL}_r}, \delta_\square)$ is compact.*

The proof is a simple extension of [24, Lemma 8] from the case of graphon to the case of graphon-signal. The proof relies on the notion of martingale. A martingale is a sequence of random variables for which, for each element in the sequence, the conditional expectation of the next value in the sequence is equal to the present value, regardless of all prior values. The Martingale convergence theorem states that for any bounded martingale $\{M_n\}_n$ over the probability pace X, the sequence $\{M_n(x)\}_n$ converges for almost every $x \in X$, and the limit function is bounded (see [11, 33]).

Proof. [Proof of Theorem D.1] Consider a sequence $\{[(W_n, f_n)]\}_{n\in\mathbb{N}} \subset \widetilde{\mathcal{WL}_r}$, with $(W_n, f_n) \in \mathcal{WL}_r$. For each k, consider the equipartition into m_k intervals $\mathcal{I}_{m_k}$, where $m_k = 2^{30\lceil(r^2+1)\rceil k^2}$. By Corollary C.11, there is a measure preserving bijection $\phi_{n,k}$ (up to nullset) such that

$$\|(W_n, f_n)^{\phi_{n,k}} - (W_n, f_n)^{\phi_{n,k}}_{\mathcal{I}_{m_k}}\|_{\square;r} < 1/k,$$

where $(W_n, f_n)^{\phi_{n,k}}_{\mathcal{I}_{m_k}}$ is the projection of $(W_n, f_n)^{\phi_{n,k}}$ upon $\mathcal{I}_{m_k}$ (Definition C.10). For every fixed k, each pair of functions $(W_n, f_n)^{\phi_{n,k}}_{\mathcal{I}_{m_k}}$ is defined via $m_k^2+m_k$ values in $[0,1]$. Hence, since $[0,1]^{m_k^2+m_k}$ is compact, there is a subsequence $\{n_j^k\}_{j\in\mathbb{N}}$, such that all of these values converge. Namely, for each k, the sequence

$$\{(W_{n_j^k}, f_{n_j^k})^{\phi_{n_j^k,k}}_{\mathcal{I}_{m_k}}\}_{j=1}^{\infty}$$

converges pointwise to some step graphon-signal (U_k, g_k) in $[\mathcal{WL}_r]_{\mathcal{P}_k}$ as $j \to \infty$. Note that $\mathcal{I}_{m_l}$ is a refinement of $\mathcal{I}_{m_k}$ for every $l > k$. As as a result, by the definition of projection of graphon-signals to partitions, for every $l > k$, the value of $(W_n^{\phi_{n,k}})_{\mathcal{I}_{m_k}}$ at each partition set $I^i_{m_k} \times I^j_{m_k}$ can be obtained by averaging the values of $(W_n^{\phi_{n,l}})_{\mathcal{I}_{m_l}}$ at all partition sets $I^{i'}_{m_l} \times I^{j'}_{m_l}$ that are subsets of $I^i_{m_k} \times I^j_{m_k}$. A similar property applies also to the signal. Moreover, by taking limits, it can be shown that the same property holds also for (U_k, g_k) and (U_l, g_l). We now see $\{(U_k, g_k)\}_{k=1}^{\infty}$ as a sequence of random variables over the standard probability space $[0,1]^2$. The above discussion shows that $\{(U_k, g_k)\}_{k=1}^{\infty}$ is a bounded martingale. By the martingale convergence theorem, the sequence $\{(U_k, g_k)\}_{k=1}^{\infty}$ converges almost everywhere pointwise to a limit (U, g), which must be in $\mathcal{WL}_r$.

Lastly, we show that there exist increasing sequences $\{k_z \in \mathbb{N}\}_{z=1}^{\infty}$ and $\{t_z = n_{j_z}^{k_z}\}_{z\in\mathbb{N}}$ such that $(W_{t_z}, f_{t_z})^{\phi_{t_z,k_z}}$ converges to (U, g) in cut distance. By the dominant convergence theorem, for each $z \in \mathbb{N}$ there exists a k_z such that

$$\|(U,g) - (U_{k_z}, g_{k_z})\|_1 < \frac{1}{3z}.$$

We choose such an increasing sequence $\{k_z\}_{z\in\mathbb{N}}$ with $k_z > 3z$. Similarly, for ever $z \in \mathbb{N}$, there is a j_z such that, with the notation $t_z = n_{j_z}^{k_z}$,

$$\|(U_{k_z}, g_{k_z}) - (W_{t_z}, f_{t_z})^{\phi_{t_z,k_z}}_{\mathcal{I}_{m_{k_z}}}\|_1 < \frac{1}{3z},$$

and we may choose the sequence $\{t_z\}_{z\in\mathbb{N}}$ increasing. Therefore, by the triangle inequality and by the fact that the L_1 norm bounds the cut norm,

$$\begin{aligned}
\delta_\square\big((U,g), (W_{t_z}, f_{t_z})\big) &\leq \|(U,g) - (W_{t_z}, f_{t_z})^{\phi_{t_z,k_z}}\|_\square \\
&\leq \|(U,g) - (U_{k_z}, g_{k_z})\|_1 + \|(U_{k_z}, g_{k_z}) - (W_{t_z}, f_{t_z})^{\phi_{t_z,k_z}}_{\mathcal{I}_{m_{k_z}}}\|_1 \\
&\quad + \|(W_{t_z}, f_{t_z})^{\phi_{t_z,k_z}}_{\mathcal{I}_{m_{k_z}}} - (W_{t_z}, f_{t_z})^{\phi_{t_z,k_z}}\|_\square \\
&\leq \frac{1}{3z} + \frac{1}{3z} + \frac{1}{3z} \leq \frac{1}{z}.
\end{aligned}$$

∎

The next theorem bounds the covering number of $\widetilde{\mathcal{WL}_r}$.

Theorem D.2. *Let $r > 0$ and $c > 1$. For every sufficiently small $\epsilon > 0$, the space $\widetilde{\mathcal{WL}_r}$ can be covered by*

$$\kappa(\epsilon) = 2^{k^2} \tag{25}$$

balls of radius ϵ in cut distance, where $k = \lceil 2^{2c/\epsilon^2} \rceil$.

Proof. Let $1 < c < c'$ and $0 < \alpha < 1$. Given an error tolerance $\alpha\epsilon > 0$, using Theorem C.8, we take the equipartition $\mathcal{I}_n$ into $n = 2^{\lceil \frac{2c}{\alpha^2\epsilon^2} \rceil}$ intervals, for which any graphon-signal $(W, f) \in \widetilde{\mathcal{WL}_r}$ can be approximated by some $(W, f)_n$ in $[\widetilde{\mathcal{WL}_r}]_{\mathcal{I}_n}$, up to error $\alpha\epsilon$. Consider the rectangle $\mathcal{R}_{n,r} = [0,1]^{n^2} \times [-r, r]^n$. We identify each element of $[\widetilde{\mathcal{WL}_r}]_{\mathcal{I}_n}$ with an element of $\mathcal{R}_{n,r}$ using the coefficients of (5). More accurately, the coefficients $c_{i,j}$ of the step graphon are identifies with the first n^2 entries of a point in $\mathcal{R}_{n,r}$, and the the coefficients b_i of the step signals are identifies with the last n entries of a point in $\mathcal{R}_{n,r}$. Now, consider the qunatized rectangle $\tilde{\mathcal{R}}_{n,r}$, defined as

$$\tilde{\mathcal{R}}_{n,r} = \big((1-\alpha)\epsilon\mathbb{Z}\big)^{n^2+2rn} \cap \mathcal{R}_{n,r}.$$

Note that $\tilde{\mathcal{R}}_n$ consists of

$$M \leq \lceil \frac{1}{(1-\alpha)\epsilon} \rceil^{n^2+2rn} \leq 2^{\left(-\log\left((1-\alpha)\epsilon\right)+1\right)(n^2+2rn)}$$

points. Now, every point $x \in \mathcal{R}_{n,r}$ can be approximated by a quantized version $x_Q \in \tilde{\mathcal{R}}_{n,r}$ up to error in normalized ℓ_1 norm

$$\|x - x_Q\|_1 := \frac{1}{M}\sum_{j=1}^{M}\left|x^j - x_Q^j\right| \leq (1-\alpha)\epsilon,$$

where we re-index the entries of x and x_Q in a 1D sequence. Let us denote by $(W, f)_Q$ the quantized version of (W_n, f_n), given by the above equivalence mapping between $(W, f)_n$ and $\mathcal{R}_{n,r}$. We hence have

$$\|(W, f) - (W, f)_Q\|_\square \leq \|(W, f) - (W_n, f_n)\|_\square + \|(W_n, f_n) - (W, f)_Q\|_\square \leq \epsilon.$$

We now choose the parameter α. Note that for any $c' > c$, there exists $\epsilon_0 > 0$ that depends on $c' - c$, such that for any $\epsilon < \epsilon_0$ there is a choice of α (close to 1) such that

$$M \leq \lceil \frac{1}{(1-\alpha)\epsilon} \rceil^{n^2+2rn} \leq 2^{\left(-\log\left((1-\alpha)\epsilon\right)+1\right)(n^2+2rn)} \leq 2^{k^2}$$

where $k = \lceil 2^{2c'/\epsilon^2} \rceil$. This is shown similarly to the proof of Corollary C.7 and Theorem C.8. We now replace the notation $c' \to c$, which concludes the proof.

■

E Graphon-signal sampling lemmas

In this appendix, we prove Theorem 3.6. We denote by $\mathcal{W}_1$ the space of measurable functions $U : [0, 1] \to [-1, 1]$, and call each $U \in \mathcal{W}_1$ a kernel.

E.1 Formal construction of sampled graph-signals

Let $W \in \mathcal{W}_0$ be a graphon, and $\Lambda' = (\lambda'_1, \ldots \lambda'_k) \in [0, 1]^k$. We denote by $W(\Lambda')$ the adjacency matrix

$$W(\Lambda') = \{W(\lambda'_i, \lambda'_j)\}_{i,j\in[k]}.$$

By abuse of notation, we also treat $W(\Lambda')$ as a weighted graph with k nodes and the adjacency matrix $W(\Lambda')$. We denote by $\Lambda = (\lambda_1, \ldots, \lambda_k) : (\lambda'_1, \ldots \lambda'_k) \mapsto (\lambda'_1, \ldots \lambda'_k)$ the identity random variable in $[0, 1]^k$. We hence call $(\lambda_1, \ldots, \lambda_k)$ random independent samples from $[0, 1]$. We call the random variable $W(\Lambda)$ a *random sampled weighted graph*.

Given $f \in \mathcal{L}^\infty_r[0, 1]$ and $\Lambda' = (\Lambda'_1, \ldots, \Lambda'_k) \in [0, 1]^k$, we denote by $f(\Lambda')$ the discrete signal with k nodes, and value $f(\lambda'_i)$ for each node $i = 1, \ldots, k$. We define the *sampled signal* as the random variable $f(\Lambda)$.

We then define the random sampled simple graph as follows. First, for a deterministic $\Lambda' \in [0, 1]^k$, we define a 2D array of Bernoulli random variables $\{e_{i,j}(\Lambda')\}_{i,j\in[k]}$ where $e_{i,j}(\Lambda') = 1$ in probability

$W(\lambda'_i, \lambda'_j)$, and zero otherwise, for $i, j \in [k]$. We define the the probability space $\{0,1\}^{k\times k}$ with normalized counting measure, defined for any $S \subset \{0,1\}^{k\times k}$ by

$$P_{\Lambda'}(S) = \sum_{\mathbf{z}\in S} \prod_{i,j\in[k]} P_{\Lambda';i,j}(z_{i,j}),$$

where

$$P_{\Lambda';i,j}(z_{i,j}) = \begin{cases} W(\lambda'_i, \lambda'_j) & \text{if } z_{i,j} = 1 \\ 1 - W(\lambda'_i, \lambda'_j) & \text{if } z_{i,j} = 0. \end{cases}$$

We denote the identity random variable by $\mathbb{G}(W, \Lambda') : \mathbf{z} \mapsto \mathbf{z}$, and call it a *random simple graph sampled from* $W(\Lambda')$.

Next we also allow to "plug" the random variable Λ into Λ'. For that, we define the joint probability space $\Omega = [0,1]^k \times \{0,1\}^{k\times k}$ with the product σ-algebra of the Lebesgue sets in $[0,1]^k$ with the power set σ-algebra of $\{0,1\}^{k\times k}$, with measure, for any measurable $S \subset \Omega$,

$$\mu(S) = \int_{[0,1]^k} P_{\Lambda'}\big(S(\Lambda')\big) d\Lambda',$$

where

$$S(\Lambda') \subset \{0,1\}^{k\times k} := \{\mathbf{z} = \{z_{i,j}\}_{i,j\in[k]} \in \{0,1\}^{k\times k} \mid (\Lambda', \mathbf{z}) \in S\},$$

We call the random variable $\mathbb{G}(W, \Lambda) : \Lambda' \times \mathbf{z} \mapsto \mathbf{z}$ the *random simple graph generated by* W. We extend the domains of the random variables $W(\Lambda)$, $f(\Lambda)$ and $\mathbb{G}(W, \Lambda')$ to Ω trivially (e.g., $f(\Lambda)(\Lambda', \mathbf{z}) = f(\Lambda)(\Lambda')$ and $\mathbb{G}(W, \Lambda')(\Lambda', \mathbf{z}) = \mathbb{G}(W, \Lambda')(\mathbf{z})$), so that all random variables are defined over the same space Ω. Note that the random sampled graphs and the random signal share the same sample points.

Given a kernel $U \in \mathcal{W}_1$, we define the random sampled kernel $U(\Lambda)$ similarly.

Similarly to the above construction, given a weighted graph H with k nodes and edge weights $h_{i,j}$, we define the *simple graph sampled from* H as the random variable simple graph $\mathbb{G}(H)$ with k nodes and independent Bernoulli variables $e_{i,j} \in \{0,1\}$, with $\mathbb{P}(e_{i,j} = 1) = h_{i,j}$, as the edge weights. The following lemma is taken from [23, Equation (10.9)].

Lemma E.1. *Let H be a weighted graph of k nodes. Then*

$$\mathbb{E}\big(d_\square(\mathbb{G}(H), H)\big) \le \frac{11}{\sqrt{k}}.$$

The following is a simple corollary of Lemma E.1, using the law of total probability.

Corollary E.2. *Let $W \in \mathcal{W}_0$ and $k \in \mathbb{N}$. Then*

$$\mathbb{E}\big(d_\square(\mathbb{G}(W, \Lambda), W(\Lambda))\big) \le \frac{11}{\sqrt{k}}.$$

E.2 Sampling lemmas of graphon-signals

The following lemma, from [23, Lemma 10.6], shows that the cut norm of a kernel is approximated by the cut norm of its sample.

Lemma E.3 (First Sampling Lemma for kernels)**.** *Let $U \in \mathcal{W}_1$, and $\Lambda \in [0,1]^k$ be uniform independent samples from $[0,1]$. Then, with probability at least $1 - 4e^{-\sqrt{k}/10}$,*

$$-\frac{3}{k} \le \|U[\Lambda]\|_\square - \|U\|_\square \le \frac{8}{k^{1/4}}.$$

We derive a version of Lemma E.3 with expected value using the following lemma.

Lemma E.4. *Let $z : \Omega \to [0,1]$ be a random variable over the probability space Ω. Suppose that in an event $\mathcal{E} \subset \Omega$ of probability $1 - \epsilon$ we have $z < \alpha$. Then*

$$\mathbb{E}(z) \le (1 - \epsilon)\alpha + \epsilon.$$

Proof.

$$\mathbb{E}(z) = \int_{\Omega} z(x)dx = \int_{\mathcal{E}} z(x)dx + \int_{\Omega\setminus\mathcal{E}} z(x)dx \leq (1-\epsilon)\alpha + \epsilon.$$

■

As a result of this lemma, we have a simple corollary of Lemma E.3.

Corollary E.5 (First sampling lemma - expected value version)**.** *Let $U \in \mathcal{W}_1$ and $\Lambda \in [0,1]^k$ be chosen uniformly at random, where $k \geq 1$. Then*

$$\mathbb{E}\left|\|U[\Lambda]\|_{\square} - \|U\|_{\square}\right| \leq \frac{14}{k^{1/4}}.$$

Proof. By Lemma E.4, and since $6/k^{1/4} > 4e^{-\sqrt{k}/10}$,

$$\mathbb{E}\big|\|U[\Lambda]\|_{\square} - \|U\|_{\square}\big| \leq \left(1 - 4e^{-\sqrt{k}/10}\right)\frac{8}{k^{1/4}} + 4e^{-\sqrt{k}/10} < \frac{14}{k^{1/4}}.$$

■

We note that a version of the first sampling lemma, Lemma E.3, for signals instead of kernels, is just a classical Monte Carlo approximation, when working with the $L_1[0,1]$ norm, which is equivalent to the signal cut norm.

Lemma E.6 (First sampling lemma for signals)**.** *Let $f \in \mathcal{L}_r^{\infty}[0,1]$. Then*

$$\mathbb{E}\left|\|f(\Lambda)\|_1 - \|f\|_1\right| \leq \frac{r}{k^{1/2}}.$$

Proof. By standard Monte Carlo theory, since r^2 bounds the variance of $f(\lambda)$, where λ is a random uniform sample from $[0,1]$, we have

$$\mathbb{V}(\|f(\Lambda)\|_1) = \mathbb{E}\big(\left|\|f(\Lambda)\|_1 - \|f\|_1\right|^2\big) \leq \frac{r^2}{k}.$$

Here, $\mathbb{V}$ denotes variance, and we note that $\mathbb{E}\|f(\Lambda)\|_1 = \frac{1}{k}\sum_{j=1}^{k}|f(\lambda_j)| = \|f\|_1$. Hence, by Cauchy Schwarz inequality,

$$\mathbb{E}\left|\|f(\Lambda)\|_1 - \|f\|_1\right| \leq \sqrt{\mathbb{E}\big(\left|\|f(\Lambda)\|_1 - \|f\|_1\right|^2\big)} \leq \frac{r}{k^{1/2}}.$$

■

We now extend [23, Lemma 10.16], which bounds the cut distance between a graphon and its sampled graph, to the case of a sampled graphon-signal.

Theorem E.7 (Second sampling lemma for graphon signals)**.** *Let $r > 1$. Let $k \geq K_0$, where K_0 is a constant that depends on r, and let $(W,f) \in \mathcal{WL}_r$. Then,*

$$\mathbb{E}\Big(\delta_{\square}\big((W,f),(W(\Lambda),f(\Lambda))\big)\Big) < \frac{15}{\sqrt{\log(k)}},$$

and

$$\mathbb{E}\Big(\delta_{\square}\big((W,f),(\mathbb{G}(W,\Lambda),f(\Lambda))\big)\Big) < \frac{15}{\sqrt{\log(k)}}.$$

The proof follows the steps of [23, Lemma 10.16] and [4]. We note that the main difference in our proof is that we explicitly write the measure preserving bijection that optimizes the cut distance. While this is not necessary in the classical case, where only a graphon is sampled, in our case we need to show that there is a measure preserving bijection that is shared by the graphon and the signal. We hence write the proof for completion.

Proof.

Denote a generic error bound, given by the regularity lemma Theorem C.8 by ϵ. If we take n intervals in the Theorem C.8 , then the error in the regularity lemma will be, for c such that $2c = 3$,

$$\lceil 3/\epsilon^2 \rceil = \log(n)$$

so

$$3/\epsilon^2 + 1 \geq \log(n).$$

For small enough ϵ, we increase the error bound in the regularity lemma to satisfy

$$4/\epsilon^2 > 3/\epsilon^2 + 1 \geq \log(n).$$

More accurately, for the equipartition to intervals $\mathcal{I}_n$, there is $\phi' \in S'_{[0,1]}$ and a piecewise constant graphon signal $([W^\phi]_n, [f^\phi]_n)$ such that

$$\|W^{\phi'} - [W^{\phi'}]_n\|_\square \leq \alpha \frac{2}{\sqrt{\log(n)}}$$

and

$$\|f^{\phi'} - [f^{\phi'}]_n\|_\square \leq (1-\alpha)\frac{2}{\sqrt{\log(n)}},$$

for some $0 \leq \alpha \leq 1$. If we choose n such that

$$n = \lceil \frac{\sqrt{k}}{r\log(k)} \rceil,$$

then an error bound in the regularity lemma is

$$\|W^{\phi'} - [W^{\phi'}]_n\|_\square \leq \alpha \frac{2}{\sqrt{\frac{1}{2}\log(k) - \log\big(\log(k)\big) - \log(r)}}$$

and

$$\|f^{\phi'} - [f^{\phi'}]_n\|_\square \leq (1-\alpha)\frac{2}{\sqrt{\frac{1}{2}\log(k) - \log\big(\log(k)\big) - \log(r)}},$$

for some $0 \leq \alpha \leq 1$. Without loss of generality, we suppose that ϕ' is the identity. This only means that we work with a different representative of $[(W,f)] \in \widetilde{\mathcal{WL}_r}$ throughout the proof. We hence have

$$d_\square(W, W_n) \leq \alpha \frac{2\sqrt{2}}{\sqrt{\log(k) - 2\log\big(\log(k)\big) - 2\log(r)}}$$

and

$$\|f - f_n\|_1 \leq (1-\alpha)\frac{4\sqrt{2}}{\sqrt{\log(k) - 2\log\big(\log(k)\big) - 2\log(r)}},$$

for some step graphon-signal $(W_n, f_n) \in [\mathcal{WL}_r]_{\mathcal{I}_n}$.

Now, by the first sampling lemma (Corollary E.5),

$$\mathbb{E}\big|d_\square\big(W(\Lambda), W_n(\Lambda)\big) - d_\square(W, W_n)\big| \leq \frac{14}{k^{1/4}}.$$

Moreover, by the fact that $f - f_n \in \mathcal{L}^\infty_{2r}[0,1]$, Lemma E.6 implies that

$$\mathbb{E}\big|\|f(\Lambda) - f_n(\Lambda)\|_1 - \|f - f_n\|_1\big| \leq \frac{2r}{k^{1/2}}.$$

Therefore,

$$\begin{aligned}\mathbb{E}\Big(d_\square\big(W(\Lambda), W_n(\Lambda)\big)\Big) &\leq \mathbb{E}\big|d_\square\big(W(\Lambda), W_n(\Lambda)\big) - d_\square(W, W_n)\big| + d_\square(W, W_n) \\ &\leq \frac{14}{k^{1/4}} + \alpha\frac{2\sqrt{2}}{\sqrt{\log(k) - 2\log\big(\log(k)\big) - 2\log(r)}}.\end{aligned}$$

Similarly, we have

$$\begin{aligned}\mathbb{E}\|f(\Lambda) - f_n(\Lambda)\|_1 &\leq \mathbb{E}\big|\|f(\Lambda) - f_n(\Lambda)\|_1 - \|f - f_n\|_1\big| + \|f - f_n\|_1 \\ &\leq \frac{2r}{k^{1/2}} + (1-\alpha)\frac{4\sqrt{2}}{\sqrt{\log(k) - 2\log\big(\log(k)\big) - 2\log(r)}}.\end{aligned}$$

Now, let π_Λ be a sorting permutation in $[k]$, such that

$$\pi_\Lambda(\Lambda) := \{\Lambda_{\pi_\Lambda^{-1}(i)}\}_{i=1}^k = (\lambda'_1, \ldots, \lambda'_k)$$

is a sequence in a non-decreasing order. Let $\{I_k^i = [i-1, i)/k\}_{i=1}^k$ be the intervals of the equipartition $\mathcal{I}_k$. The sorting permutation π_Λ induces a measure preserving bijection ϕ that sorts the intervals I_k^i. Namely, we define, for every $x \in [0,1]$,

$$\text{if } x \in I_k^i, \quad \phi(x) = J_{i,\pi_\Lambda(i)}(x), \tag{26}$$

where $J_{i,j} : I_k^i \to I_k^j$ are defined as $x \mapsto x - i/k + j/k$, for all $x \in I_k^i$.

By abuse of notation, we denote by $W_n(\Lambda)$ and $f_n(\Lambda)$ the induced graphon and signal from $W_n(\Lambda)$ and $f_n(\Lambda)$ respectively. Hence, $W_n(\Lambda)^\phi$ and $f_n(\Lambda)^\phi$ are well defined. Note that the graphons W_n and $W_n(\Lambda)^\phi$ are stepfunctions, where the set of values of $W_n(\Lambda)^\phi$ is a subset of the set of values of W_n. Intuitively, since $k \gg m$, we expect the partition $\{[\lambda'_i, \lambda'_{i+1})\}_{i=1}^k$ to be "close to a refinement" of $\mathcal{I}_n$ in high probability. Also, we expect the two sets of values of $W_n(\Lambda)^\phi$ and W_n to be identical in high probability. Moreover, since Λ' is sorted, when inducing a graphon from the graph $W_n(\Lambda)$ and "sorting" it to $W_n(\Lambda)^\phi$, we get a graphon that is roughly "aligned" with W_n. The same philosophy also applied to f_n and $f_n(\Lambda)^\phi$. We next formalize these observations.

For each $i \in [n]$, let λ'_{j_i} be the smaller point of Λ' that is in I_n^i, set $j_i = j_{i+1}$ if $\Lambda' \cap I_n^i = \emptyset$, and set $j_{n+1} = k+1$. For every $i = 1, \ldots, n$, we call

$$J_i := [j_i - 1, j_{i+1} - 1)/k$$

the i-th step of $W_n(\Lambda)^\phi$ (which can be the empty set). Let $a_i = \frac{j_i - 1}{k}$ be the left edge point of J_i. Note that $a_i = |\Lambda \cap [0, i/n)|/k$ is distributed binomially (up to the normalization k) with k trials and success in probability i/n.

$$\begin{aligned}
\|W_n - W_n(\Lambda)^\phi\|_\square \leq \|W_n - W_n(\Lambda)^\phi\|_1 & \\
= \sum_i \sum_k \int_{I_n^i \cap J_i} \int_{I_n^k \cap J_k} & \left|W_n(x,y) - W_n(\Lambda)^\phi(x,y)\right| dxdy \\
+ \sum_i \sum_{j \neq i} \sum_k \sum_{l \neq k} & \int_{I_n^i \cap J_j} \int_{I_n^k \cap J_l} \left|W_n(x,y) - W_n(\Lambda)^\phi(x,y)\right| dxdy \\
= \sum_i \sum_{j \neq i} \sum_k \sum_{l \neq k} & \int_{I_n^i \cap J_j} \int_{I_n^k \cap J_l} \left|W_n(x,y) - W_n(\Lambda)^\phi(x,y)\right| dxdy \\
= \sum_i \sum_k \int_{I_n^i \setminus J_i} \int_{I_n^k \setminus J_k} & \left|W_n(x,y) - W_n(\Lambda)^\phi(x,y)\right| dxdy \\
\leq \sum_i \sum_k \int_{I_n^i \setminus J_i} \int_{I_n^k \setminus J_k} & 1 dxdy \leq 2 \sum_i \int_{I_n^i \setminus J_i} 1 dxdy \\
\leq 2 \sum_i (|i/n - a_i| & + |(i+1)/n - a_{i+1}|).
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}\|W_n - W_n(\Lambda)^\phi\|_\square &\leq 2 \sum_i (\mathbb{E}\,|i/n - a_i| + \mathbb{E}\,|(i+1)/n - a_{i+1}|) \\
&\leq 2 \sum_i \left(\sqrt{\mathbb{E}(i/n - a_i)^2} + \sqrt{\mathbb{E}\big((i+1)/n - a_{i+1}\big)^2} \right)
\end{aligned}$$

By properties of the binomial distribution, we have $\mathbb{E}(ka_i) = ik/n$, so

$$\mathbb{E}(ik/n - ka_i)^2 = \mathbb{V}(ka_i) = k(i/n)(1 - i/n).$$

As a result

$$\begin{aligned}
\mathbb{E}\|W_n - W_n(\Lambda)^\phi\|_\square &\leq 5 \sum_{i=1}^n \sqrt{\frac{(i/n)(1 - i/n)}{k}} \\
&\leq 2 \int_1^n \sqrt{\frac{(i/n)(1 - i/n)}{k}} di,
\end{aligned}$$

and for $n > 10$,

$$\leq 5\frac{n}{\sqrt{k}}\int_0^{1.1}\sqrt{z-z^2}dz \leq 5\frac{n}{\sqrt{k}}\int_0^{1.1}\sqrt{z}dz \leq 10/3(1.1)^{3/2}\frac{n}{\sqrt{k}} < 4\frac{n}{\sqrt{k}}.$$

Now, by $n = \lceil \frac{\sqrt{k}}{r\log(k)} \rceil \leq \frac{\sqrt{k}}{r\log(k)} + 1$, for large enough k,

$$\mathbb{E}\|W_n - W_n(\Lambda)^\phi\|_\square \leq 4\frac{1}{r\log(k)} + 4\frac{1}{\sqrt{k}} \leq \frac{5}{r\log(k)}.$$

Similarly,

$$\mathbb{E}\|f_n - f_n(\Lambda)^\phi\|_1 \leq \frac{5}{\log(k)}.$$

Note that in the proof of Corollary C.7, in (18), α is chosen close to 1, and especially, for small enough ϵ, $\alpha > 1/2$. Hence, for large enough k,

$$\begin{aligned}
\mathbb{E}(d_\square(W, W(\Lambda)^\phi)) &\leq d_\square(W, W_n) + \mathbb{E}\big(d_\square(W_n, W_n(\Lambda)^\phi)\big) + \mathbb{E}(d_\square(W_n(\Lambda), W(\Lambda))) \\
&\leq \alpha\frac{2\sqrt{2}}{\sqrt{\log(k) - 2\log\big(\log(k)\big) - 2\log(r)}} + \frac{5}{r\log(k)} + \frac{14}{k^{1/4}} \\
&+ \alpha\frac{2\sqrt{2}}{\sqrt{\log(k) - 2\log\big(\log(k)\big) - 2\log(r)}} \\
&\leq \alpha\frac{6}{\sqrt{\log(k)}},
\end{aligned}$$

Similarly, for each k, if $1-\alpha < \frac{1}{\sqrt{\log(k)}}$, then

$$\begin{aligned}
\mathbb{E}(d_\square(f, f(\Lambda)^\phi)) \leq (1-\alpha)&\frac{2\sqrt{2}}{\sqrt{\log(k) - 2\log\big(\log(k)\big) - 2\log(r)}} + \frac{5}{\log(k)} \\
&+ \frac{2r}{k^{1/2}} + (1-\alpha)\frac{4\sqrt{2}}{\sqrt{\log(k) - 2\log\big(\log(k)\big) - 2\log(r)}} \leq \frac{14}{\log(k)}.
\end{aligned}$$

Moreover, for each k such that $1-\alpha > \frac{1}{\sqrt{\log(k)}}$, if k is large enough (where the lower bound of k depends on r), we have

$$\frac{5}{\log(k)} + \frac{2r}{k^{1/2}} < \frac{5.5}{\log(k)} < \frac{1}{\sqrt{\log(k)}}\frac{6}{\sqrt{\log(k)}} < (1-\alpha)\frac{6}{\sqrt{\log(k)}}$$

so, by $6\sqrt{2} < 9$,

$$\begin{aligned}
\mathbb{E}(d_\square(f, f(\Lambda)^\phi)) &\leq (1-\alpha)\frac{2\sqrt{2}}{\sqrt{\log(k) - 2\log\big(\log(k)\big) - 2\log(r)}} + \frac{2}{\log(k)} \\
&+ \frac{2r}{k^{1/2}} + (1-\alpha)\frac{4\sqrt{2}}{\sqrt{\log(k) - 2\log\big(\log(k)\big) - 2\log(r)}} \\
&\leq (1-\alpha)\frac{15}{\sqrt{\log(k)}}.
\end{aligned}$$

Lastly, by Corollary E.2,

$$\begin{aligned}
\mathbb{E}\Big(d_\square\big(W, \mathbb{G}(W,\Lambda)^\phi\big)\Big) &\leq \mathbb{E}\Big(d_\square\big(W, W(\Lambda)^\phi\big)\Big) + \mathbb{E}\Big(d_\square\big(W(\Lambda)^\phi, \mathbb{G}(W,\Lambda)^\phi\big)\Big) \\
&\leq \alpha\frac{6}{\sqrt{\log(k)}} + \frac{11}{\sqrt{k}} \leq \alpha\frac{7}{\sqrt{\log(k)}},
\end{aligned}$$

As a result, for large enough k,

$$\mathbb{E}\Big(\delta_\square\big((W,f),(W(\Lambda),f(\Lambda))\big)\Big) < \frac{15}{\sqrt{\log(k)}},$$

and

$$\mathbb{E}\Big(\delta_\square\big((W,f),(\mathbb{G}(W,\Lambda),f(\Lambda))\big)\Big) < \frac{15}{\sqrt{\log(k)}}.$$

■

F Graphon-signal MPNNs

In this appendix we give properties and examples of MPNNs.

F.1 Properties of graphon-signal MPNNs

Consider the construction of MPNN from Section 4.1. We first explain how a MPNN on a grpah is equivalent to a MPNN on the induced graphon.

Let G be a graph of n nodes, with adjacency matrix $A = \{a_{i,j}\}_{i,j\in[n]}$ and signal $\mathbf{f} \in \mathbb{R}^{n\times d}$. Consider a MPL θ, with receiver and transmitter message functions $\xi_{\mathrm{r}}^k, \xi_{\mathrm{t}}^k : \mathbb{R}^d \to \mathbb{R}^p$, for $k \in [K]$, where $K \in \mathbb{N}$, and update function $\mu : \mathbb{R}^{d+p} \to \mathbb{R}^s$. The application of the MPL on $(G,\mathbf{f})$ is defined as follows. We first define the message kernel $\Phi_{\mathbf{f}} : [n]^2 \to \mathbb{R}^p$, with entries

$$\Phi_{\mathbf{f}}(i,j) = \Phi(\mathbf{f}_i,\mathbf{f}_j) = \sum_{k=1}^{K} \xi_{\mathrm{r}}^k(\mathbf{f}_i)\xi_{\mathrm{t}}^k(\mathbf{f}_j).$$

We then aggregate the message kernel with normalized sum aggregation

$$\big(\mathrm{Agg}(G,\Phi_{\mathbf{f}})\big)_i = \frac{1}{n}\sum_{j\in[n]} a_{i,j}\Phi_{\mathbf{f}}(i,j).$$

Lastly, we apply the update function, to obtain the output $\theta(G,\mathbf{f})$ of the MPL with value at each node i

$$\theta(G,\mathbf{f})_i = \eta\Big(\mathbf{f}_i, \big(\mathrm{Agg}(G,\Phi_{\mathbf{f}})\big)_i\Big) \in \mathbb{R}^s.$$

Lemma F.1. *Consider a MPL θ as in the above setting. Then, for every graph signal $(G,A,\mathbf{f})$,*

$$\theta\Big((W,f)_{(G,\mathbf{f})}\Big) = (W,f)_{\theta(G,\mathbf{f})}.$$

Proof. Let $\{I_i,\ldots,I_n\}$ be the equipartition to intervals. For each $j\in[n]$, let $y_j \in I_j$ be an arbitrary point. Let $i\in[n]$ and $x\in I_i$. We have

$$\begin{aligned}\mathrm{Agg}(G,\Phi_{\mathbf{f}})_i &= \frac{1}{n}\sum_{j\in[n]} a_{i,j}\Phi_{\mathbf{f}}(i,j) = \frac{1}{n}\sum_{j\in[n]} W_G(x,y_j)\Phi_{f_{\mathbf{f}}}(x,y_j)\\ &= \int_0^1 W_G(x,y)\Phi_{f_{\mathbf{f}}}(x,y)dy = \mathrm{Agg}(W_G,\Phi_{f_{\mathbf{f}}})(x).\end{aligned}$$

Therefore, for every $i\in[n]$ and every $x\in I_i$,

$$\begin{aligned}f_{\theta(G,\mathbf{f})}(x) &= f_{\eta\left(\mathbf{f},\mathrm{Agg}(G,\Phi_{\mathbf{f}})\right)}(x) = \eta\big(\mathbf{f}_i, \mathrm{Agg}(G,\Phi_{\mathbf{f}})_i\big)\\ &= \eta\big(f_{\mathbf{f}}(x), \mathrm{Agg}(W_G,\Phi_{f_{\mathbf{f}}})(x)\big) = \theta(W_G,f_{\mathbf{f}})(x).\end{aligned}$$

■

F.2 Examples of MPNNs

The GIN convolutional layer [34] is defined as follows. First, the message function is

$$\Phi(a, b) = b$$

and the update function is

$$\eta(x, y) = M\big((1 + \epsilon)x + y\big).$$

where M is a multi-layer perceptron (MLP) and ϵ a constant. Each layer may have a different MLP and different constant ϵ. The standard GIN is defined with sum aggregation, but we use normalized sum aggregation.

Given a graph-signal $(G, \mathbf{f})$, with $\mathbf{f} \in \mathbb{R}^{n\times d}$ with adjacency matrix $A \in \mathbb{R}^{n\times n}$, a spectral convolutional layer based on a polynomial filter $p(\lambda) = \sum_{j=0}^{J} \lambda^j C_j$, where $C_j \in \mathbb{R}^{d\times p}$, is defined to be

$$p(A)\mathbf{f} = \sum_{j=0}^{J} A^j \mathbf{f} C_j,$$

followed by a pointwise non-linearity like ReLU. Such a convolutional layer can be seen as $J + 1$ MPLs. We first apply J MPLs, where each MPL is of the form

$$\theta(\mathbf{f}) = \big(\mathbf{f}, A\mathbf{f}\big).$$

We then apply an update layer

$$U(\mathbf{f}) = \mathbf{f}C$$

for some $C \in \mathbb{R}^{(J+1)d\times p}$, followed by the pointwise non-linearity. The message part of θ can be written in our formulation with $\Phi(a, b) = b$, and the update part of θ with $\eta(c, d) = (c, d)$. The last update layer U is linear followed by the pointwise non-linearity.

G Lipschitz continuity of MPNNs

In this appendix we prove Theorem 4.1. For $v \in \mathbb{R}^d$, we often denote by $|v| = \|v\|_\infty$. We define the L_1 norm of a measurable function $h : [0, 1] \to \mathbb{R}^d$ by

$$\|h\|_1 := \int_0^1 |h(x)|\, dx = \int_0^1 \|h(x)\|_\infty dx.$$

Similarly,

$$\|h\|_\infty := \sup_{x\in\mathbb{R}^d} |h(x)| = \sup_{x\in\mathbb{R}^d} \|h(x)\|_\infty.$$

We define Lipschitz continuity with respect to the infinity norm. Namely, $Z : \mathbb{R}^d \to \mathbb{R}^c$ is called Lipschitz continuous with Lipschitz constant L if

$$|Z(x) - Z(y)| = \|Z(x) - Z(y)\|_\infty \leq L\|x - z\|_\infty = L\,|x - z|\,.$$

We denote the minimal Lipschitz bound of the function Z by L_Z.

We extend $\mathcal{L}_r^\infty[0, 1]$ to the space of functions $f : [0, 1] \to \mathbb{R}^d$ with the above L_1 norm.

Define the space $\mathcal{K}_q$ of *kernels* bounded by $q > 0$ to be the space of measurable functions

$$K : [0, 1]^2 \to [-q, q].$$

The cut norm, cut metric, and cut distance are defined as usual for kernels in $\mathcal{K}_q$.

G.1 Lipschitz continuity of message passing and update layers

In this subsection we prove that message passing layers and update layers are Lipschitz continuous with respect to he graphon-signal cut metric.

Lemma G.1 (Product rule for message kernels)**.** *Let Φ_f, Φ_g be the message kernels corresponding to the signals f, g. Then*

$$\|\Phi_f - \Phi_g\|_{L^1[0,1]^2} \leq \sum_{k=1}^{K} \Big(L_{\xi_r^k}\|\xi_t^k\|_\infty + \|\xi_r^k\|_\infty L_{\xi_t^k}\Big)\|f - g\|_1.$$

Proof. Suppose $p = 1$ For every $x, y \in [0,1]^2$

$$|\Phi_f(x,y) - \Phi_g(x,y)| = \left|\sum_{k=1}^{K} \xi_r^k(f(x))\xi_t^k(f(y)) - \sum_{k=1}^{K} \xi_r^k(g(x))\xi_t^k(g(y))\right|$$
$$\leq \sum_{k=1}^{K} \left|\xi_r^k(f(x))\xi_t^k(f(y)) - \xi_r^k(g(x))\xi_t^k(g(y))\right|$$
$$\leq \sum_{k=1}^{K} \Big(\left|\xi_r^k(f(x))\xi_t^k(f(y)) - \xi_r^k(g(x))\xi_t^k(f(y))\right| + \left|\xi_r^k(g(x))\xi_t^k(f(y)) - \xi_r^k(g(x))\xi_t^k(g(y))\right| \Big)$$
$$\leq \sum_{k=1}^{K} \Big(L_{\xi_r^k} \left|f(x) - g(x)\right| \left|\xi_t^k(f(y))\right| + \left|\xi_r^k(g(x))\right| L_{\xi_t^k} \left|f(y) - g(y)\right| \Big).$$

Hence,

$$\|\Phi_f - \Phi_g\|_{L^1[0,1]^2}$$
$$\leq \sum_{k=1}^{K} \int_0^1 \int_0^1 \Big(L_{\xi_r^k} \left|f(x) - g(x)\right| \left|\xi_t^k(f(y))\right| + \left|\xi_r^k(g(x))\right| L_{\xi_t^k} \left|f(y) - g(y)\right| \Big) dxdy$$
$$\leq \sum_{k=1}^{K} \Big(L_{\xi_r^k} \|f - g\|_1 \|\xi_t^k\|_\infty + \|\xi_r^k\|_\infty L_{\xi_t^k} \|f - g\|_1 \Big)$$
$$= \sum_{k=1}^{K} \Big(L_{\xi_r^k} \|\xi_t^k\|_\infty + \|\xi_r^k\|_\infty L_{\xi_t^k} \Big) \|f - g\|_1.$$

■

Lemma G.2. *Let Q, V be two message kernels, and $W \in \mathcal{W}_0$. Then*

$$\|\mathrm{Agg}(W,Q) - \mathrm{Agg}(W,V)\|_1 \leq \|Q - V\|_1.$$

Proof.

$$\mathrm{Agg}(W,Q)(x) - \mathrm{Agg}(W,V)(x) = \int_0^1 W(x,y)(Q(x,y) - V(x,y))dy$$

So

$$\|\mathrm{Agg}(W,Q) - \mathrm{Agg}(W,V)\|_1 = \int_0^1 \left| \int_0^1 W(x,y)(Q(x,y) - V(x,y))dy \right| dx$$
$$\leq \int_0^1 \int_0^1 \left|W(x,y)(Q(x,y) - V(x,y))\right| dydx$$
$$\leq \int_0^1 \int_0^1 \left|(Q(x,y) - V(x,y))\right| dydx = \|Q - V\|_1.$$

■

As a result of Lemma G.2 and the product rule Lemma G.1, we have the following corollary, that computes the error in aggregating two message kernels with the same graphon.

Corollary G.3.

$$\|\mathrm{Agg}(W,\Phi_f) - \mathrm{Agg}(W,\Phi_g)\|_1 \leq \sum_{k=1}^{K} \Big(L_{\xi_r^k} \|\xi_t^k\|_\infty + \|\xi_r^k\|_\infty L_{\xi_t^k} \Big) \|f - g\|_1.$$

Next we fix the message kernel, and bound the difference between the aggregation of the message kernal with respect to two different graphons. Let $L^+[0,1]$ be the space of measurable function $f : [0,1] \to [0,1]$. The folliwing lemma is a trivial extension of [23, Lemma 8.10] from $\mathcal{K}_1$ to $\mathcal{K}_r$.

Lemma G.4. *For any kernel $Q \in \mathcal{K}_r$*

$$\|Q\|_{\square} = \sup_{f,g \in L^{+}[0,1]} \left| \int_{[0,1]^2} f(x)Q(x,y)g(y)dxdy \right|,$$

where the supremum is attained for some $f, g \in L^{+}[0,1]$.

The following Lemma is proven as part of the proof of [23, Lemma 8.11].

Lemma G.5. *For any kernel $Q \in \mathcal{K}_r$*

$$\sup_{f,g \in L^{\infty}_1[0,1]} \left| \int_{[0,1]^2} f(x)Q(x,y)g(y)dxdy \right| \leq 4\|Q\|_{\square}.$$

For completeness, we give here a self-contained proof.

Proof. Any function $f \in L^{\infty}_1[0,1]$ can be written as $f = f_+ - f_-$, where $f_+, f_- \in L^{+}[0,1]$. Hence, by Lemma G.4,

$$\begin{aligned}
&\sup_{f,g \in L^{\infty}_1[0,1]} \left| \int_{[0,1]^2} f(x)Q(x,y)g(y)dxdy \right| \\
&= \sup_{f_+,f_-,g_+,g_- \in L^{+}[0,1]} \left| \int_{[0,1]^2} (f_+(x) - f_-(x))Q(x,y)(g_+(y) - g_-(y))dxdy \right| \\
&\leq \sum_{s \in \{+,-\}} \sup_{f_s, g_s \in L^{+}[0,1]} \left| \int_{[0,1]^2} f_s(x)Q(x,y)g_s(y)dxdy \right| = 4\|Q\|_{\square}.
\end{aligned}$$

■

Next we state a simple lemma.

Lemma G.6. *Let $f = f_+ - f_-$ be a signal, where $f_+, f_- : [0,1] \to (0, \infty)$ are measurable. Then the supremum in the cut norm $\|f\|_{\square} = \sup_{S \subset [0,1]} \left| \int_S f(x)dx \right|$ is attained as the support of either f_+ or f_-.*

Lemma G.7. *Let $f \in \mathcal{L}^{\infty}_2[0,1]$, $W, V \in \mathcal{W}_0$, and suppose that $\left|\xi^k_{\mathrm{r}}(f(x))\right|, \left|\xi^k_{\mathrm{t}}(f(x))\right| \leq \rho$ for every $x \in [0,1]$ and $k = 1, \ldots, K$. Then*

$$\|\mathrm{Agg}(W, \Phi_f) - \mathrm{Agg}(V, \Phi_f)\|_{\square} \leq 4K\rho^2 \|W - V\|_{\square}.$$

Moreover, if ξ^k_{r} and ξ^k_{t} are non-negatively valued for every $k = 1, \ldots, K$, then

$$\|\mathrm{Agg}(W, \Phi_f) - \mathrm{Agg}(V, \Phi_f)\|_{\square} \leq K\rho^2 \|W - V\|_{\square}.$$

Proof. Let $T = W - V$. Let S be the minimizer of the infimum underlying the cut norm of $\mathrm{Agg}(T, \Phi_f)$. Suppose without loss of generality that $\int_S \mathrm{Agg}(T, \Phi_f)(x)dx > 0$. Denote $q^k_{\mathrm{r}}(x) = \xi^k_{\mathrm{r}}(f(x))$ and $q^k_{\mathrm{t}}(x) = \xi^k_{\mathrm{t}}(f(x))$. We have

$$\begin{aligned}
\int_S \big(\mathrm{Agg}(W, \Phi_f)(x) - \mathrm{Agg}(W, \Phi_f)(x)\big)dx &= \int_S \mathrm{Agg}(T, \Phi_f)(x)dx \\
&= \sum_{k=1}^{K} \int_S \int_0^1 q^k_{\mathrm{r}}(x)T(x,y)q^k_{\mathrm{t}}(y)dydx.
\end{aligned}$$

Let

$$v^k_{\mathrm{r}}(x) = \begin{cases} q^k_{\mathrm{r}}(x)/\rho & x \in S \\ 0 & x \notin S. \end{cases} \tag{27}$$

Moreover, define $v^k_{\mathrm{t}} = q^k_{\mathrm{t}}/\rho$, and note that $v^k_{\mathrm{r}}, v^k_{\mathrm{t}} \in L^{\infty}_1[0,1]$. We hence have, by Lemma G.5,

$$\begin{aligned}
\int_S \mathrm{Agg}(T, \Phi_f)(x)dx &= \sum_{k=1}^{K} \rho^2 \int_0^1 \int_0^1 v^k_{\mathrm{r}}(x)T(x,y)v^k_{\mathrm{t}}(y)dydx \\
&\leq \sum_{k=1}^{K} \rho^2 \left| \int_0^1 \int_0^1 v^k_{\mathrm{r}}(x)T(x,y)v^k_{\mathrm{t}}(y)dydx \right| \\
&\leq 4K\rho^2 \|T\|_{\square}.
\end{aligned}$$

Hence,

$$\|\mathrm{Agg}(W,\Phi_f)-\mathrm{Agg}(V,\Phi_f)\|_\square \leq 4K\rho^2\|T\|_\square$$

Lastly, in case $\xi_{\mathrm{r}}^k, \xi_{\mathrm{t}}^k$ are nonnegatively valued, so are $q_{\mathrm{r}}^k, q_{\mathrm{t}}^k$, and hence by Lemma G.4,

$$\int_S \mathrm{Agg}(T,\Phi_f)(x)dx \leq K\rho^2\|T\|_\square.$$

■

Theorem G.8. *Let $(W,f),(V,g)\in\mathcal{WL}_r$, and suppose that $\left|\xi_{\mathrm{r}}^k(f(x))\right|, \left|\xi_{\mathrm{t}}^k(f(x))\right| \leq \rho$ and $L_{\xi_{\mathrm{t}}^k}, L_{\xi_{\mathrm{t}}^k} < L$ for every $x\in[0,1]$ and $k=1,\ldots,K$. Then,*

$$\|\mathrm{Agg}(W,\Phi_f)-\mathrm{Agg}(V,\Phi_g)\|_\square \leq 4KL\rho\|f-g\|_\square + 4K\rho^2\|W-V\|_\square.$$

Proof. By Lemma G.1, Lemma G.2 and Lemma G.7,

$$\begin{aligned}
&\|\mathrm{Agg}(W,\Phi_f)-\mathrm{Agg}(V,\Phi_g)\|_\square \\
&\leq \|\mathrm{Agg}(W,\Phi_f)-\mathrm{Agg}(W,\Phi_g)\|_\square + \|\mathrm{Agg}(W,\Phi_g)-\mathrm{Agg}(V,\Phi_g)\|_\square \\
&\leq \sum_{k=1}^{K}\left(L_{\xi_r^k}\|\xi_t^k\|_\infty + \|\xi_r^k\|_\infty L_{\xi_t^k}\right)\|f-g\|_1 + 4K\rho^2\|W-V\|_\square \\
&\leq 4KL\rho\|f-g\|_\square + 4K\rho^2\|W-V\|_\square.
\end{aligned}$$

■

Lastly, we show that update layers are Lipschitz continuous. Since the update function takes two functions $f:[0,1]\to\mathbb{R}^{d_i}$ (for generally two different output dimensions d_1, d_2), we "concatenate" these two inputs and treat it as one input $f:[0,1]\to\mathbb{R}^{d_1+d_2}$.

Lemma G.9. *Let $\eta:\mathbb{R}^{d+p}\to\mathbb{R}^s$ be Lipschitz with Lipschitz constant L_η, and let $f,g\in\mathcal{L}_r^\infty[0,1]$ with values in $\mathbb{R}^{d+p}$ for some $d,p\in\mathbb{N}$.*

Then

$$\|\eta(f)-\eta(g)\|_1 \leq L_\eta\|f-g\|_1.$$

Proof.

$$\begin{aligned}
\|\eta(f)-\eta(g)\|_1 &= \int_0^1 \left|\eta\big(f(x)\big)-\eta\big(g(x)\big)\right| dx \\
&\leq \int_0^1 L_\eta\left|f(x)-g(x)\right| dx = L_\eta\|f-g\|_1.
\end{aligned}$$

■

G.2 Bounds of signals and MPLs with Lipschitz message and update functions

We will consider three settings for the MPNN Lipschitz bounds. In all setting, the transmitter, receiver, and update functions are Lipschitz. In the first setting all message and update functions are assumed to be bounded. In the second setting, there is no additional assumption over Lipschtzness of the transmitter, receiver, and update functions. In the third setting, we assume that the message function Φ is also Lipschitz with Lipschitz bound L_Φ, and that all receiver and transmitter functions are non-negatively bounded (e.g., via an application of ReLU or sigmoid in their implementation). Note that in case $K=1$ and all functions are differentiable, by the product rule, Φ can be Lipschitz only in two cases: if both ξ_{r} and ξ_{t} are bounded and Lipschitz, or if either ξ_{r} or ξ_{t} is constant, and the other function is Lipschitz. When $K>1$, we can have combinations of these cases.

We next derive bounds for the different settings. A bound for setting 1 is given in Theorem G.8. Moreover, When the receiver and transmitter message functions and the update functions are bounded, so is the signal at each layer.

Bounds for setting 2.

Next we show boundedness when the reciever and transmitter message and update functions are only assumed to be Lipschitz.

Define the *formal bias* B_ξ of a function $\xi : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ to be $\xi(0)$ [25]. We note that the formal bias of an affine-linear operator is its classical bias.

Lemma G.10. *Let $(W, f) \in \mathcal{WL}_r$, and suppose that for every $\mathrm{y} \in \{\mathrm{r}, \mathrm{t}\}$ and $k = 1, \ldots, K$*

$$\left|\xi_{\mathrm{y}}^k(0)\right| \leq B, \quad L_{\xi_{\mathrm{y}}^k} < L.$$

Then,

$$\|\xi_{\mathrm{y}}^k \circ f\|_\infty \leq Lr + B$$

and

$$\|\mathrm{Agg}(W, \Phi_f)\|_\infty \leq K(Lr + B)^2.$$

Proof. Let $\mathrm{y} \in \{\mathrm{r}, \mathrm{t}\}$. We have

$$\left|\xi_{\mathrm{y}}^k(f(x))\right| \leq \left|\xi_{\mathrm{y}}^k(f(x)) - \xi_{\mathrm{y}}^k(0)\right| + B \leq L_{\xi_{\mathrm{y}}^k} |f(x)| + B \leq Lr + B,$$

so,

$$\begin{aligned} |\mathrm{Agg}(W, \Phi_f)(x)| &= \left|\sum_{k=1}^K \int_0^1 \xi_{\mathrm{r}}^k(f(x)) W(x, y) \xi_{\mathrm{t}}^k(f(y)) dy\right| \\ &\leq K(Lr + B)^2. \end{aligned}$$

■

Next, we have a direct result of Theorem G.8.

Corollary G.11. *Suppose that for every $\mathrm{y} \in \{\mathrm{r}, \mathrm{t}\}$ and $k = 1, \ldots, K$*

$$\left|\xi_{\mathrm{y}}^k(0)\right| \leq B, \quad L_{\xi_{\mathrm{y}}^k} < L.$$

Then, for every $(W, f), (V, g) \in \mathcal{WL}_r$,

$$\|\mathrm{Agg}(W, \Phi_f) - \mathrm{Agg}(V, \Phi_g)\|_\square \leq 4K(L^2 r + LB)\|f - g\|_\square + 4K(Lr + B)^2 \|W - V\|_\square.$$

Bound for setting 3.

Lemma G.12. *Let $(W, f) \in \mathcal{WL}_r$, and suppose that*

$$|\Phi(0, 0)| < B, \quad L_\Phi < L.$$

Then,

$$\|\Phi_f\|_\infty \leq Lr + B$$

and

$$\|\mathrm{Agg}(W, \Phi_f)\|_\infty \leq Lr + B.$$

Proof. We have

$$|\Phi(f(x), f(y))| \leq |\Phi(f(x), f(y)) - \Phi(0, 0)| + B \leq L_\Phi \left|(f(x), f(y))\right| + B \leq Lr + B,$$

so,

$$\begin{aligned} |\mathrm{Agg}(W, \Phi_f)(x)| &= \left|\int_0^1 W(x, y) \Phi(f(x), f(y)) dy\right| \\ &\leq Lr + B. \end{aligned}$$

■

Additional bounds.

Lemma G.13. *Let f be a signal, $W, V \in \mathcal{W}_0$, and suppose that $\|\Phi_f\|_\infty \leq \rho$ for every $k = 1, \ldots, K$, and that ξ_{r}^k and ξ_{t}^k are non-negatively valued. Then*

$$\|\mathrm{Agg}(W, \Phi_f) - \mathrm{Agg}(V, \Phi_f)\|_\square \leq K\rho\|W - V\|_\square.$$

Proof. The proof follows the steps of Lemma G.7 until (27), from where we proceed differently. Since all of the functions q_{r}^k and q_{t}^k, $k \in [K]$, and since $\|\Phi_f\|_\infty \leq \rho$, the product of each $q_{\mathrm{r}}^k(x)q_{\mathrm{t}}^k(y)$ must be also bounded by ρ for every $x \in [0, 1]$ and $k \in [K]$. Hence, we may replace the normalization in (27) with

$$v_{\mathrm{r}}^k(x) = \begin{cases} q_{\mathrm{r}}^k(x)/\rho_{\mathrm{r}}^k & x \in S \\ 0 & x \notin S \end{cases}, \quad v_{\mathrm{t}}^k(y) = \begin{cases} q_{\mathrm{t}}^k(y)/\rho_{\mathrm{t}}^k & y \in S \\ 0 & y \notin S, \end{cases}$$

where for every $k \in [K]$, $\rho_{\mathrm{r}}^k \rho_{\mathrm{t}}^k = \rho$. This guarantees that $v_{\mathrm{r}}^k, v_{\mathrm{t}}^k \in L_1^\infty[0, 1]$. Hence,

$$\int_S \mathrm{Agg}(T, \Phi_f)(x)dx = \sum_{k=1}^K \int_0^1 \int_0^1 \rho_{\mathrm{r}}^k v_{\mathrm{r}}^k(x) T(x, y) \rho_{\mathrm{t}}^k v_{\mathrm{t}}^k(y) dy dx$$

$$\leq \sum_{k=1}^K \rho \left| \int_0^1 \int_0^1 v_{\mathrm{r}}^k(x) T(x, y) v_{\mathrm{t}}^k(y) dy dx \right| \leq K\rho\|T\|_\square.$$

■

Theorem G.14. *Let $(W, f), (V, g) \in \mathcal{WL}_r$, and suppose that $\|\Phi\|_\infty, \|\xi_{\mathrm{r}}^k\|_\infty, \|\xi_{\mathrm{t}}^k\|_\infty \leq \rho$, all message fucntions ξ are non-neagative valued, and $L_{\xi_{\mathrm{t}}^k}, L_{\xi_{\mathrm{t}}^k} < L$, for every $k = 1, \ldots, K$. Then,*

$$\|\mathrm{Agg}(W, \Phi_f) - \mathrm{Agg}(V, \Phi_g)\|_\square \leq 4KL\rho\|f - g\|_\square + K\rho\|W - V\|_\square.$$

The proof follows the steps of Theorem G.8.

Corollary G.15. *Suppose that for every $\mathrm{y} \in \{\mathrm{r}, \mathrm{t}\}$ and $k = 1, \ldots, K$*

$$|\Phi(0, 0)|, \left|\xi_{\mathrm{y}}^k(0)\right| \leq B, \quad L_\phi, L_{\xi_{\mathrm{y}}^k} < L,$$

and ξ, Φ are all non-negatively valued. Then, for every $(W, f), (V, g) \in \mathcal{WL}_r$,

$$\|\mathrm{Agg}(W, \Phi_f) - \mathrm{Agg}(V, \Phi_g)\|_\square \leq 4K(L^2 r + LB)\|f - g\|_\square + K(Lr + B)\|W - V\|_\square.$$

The proof follows the steps of Corollary G.11.

G.3 Lipschitz continuity theorems for MPNNs

The following recurrence sequence will govern the propagation of the Lipschitz constant of the MPNN and the bound of signal along the layers.

Lemma G.16. *Let $\mathbf{a} = (a_1, a_2, \ldots)$ and $\mathbf{b} = (b_1, b_2, \ldots)$. The solution to $e_{t+1} = a_t e_t + b_t$, with initialization e_0, is*

$$e_t = Z_t(\mathbf{a}, \mathbf{b}, e_0) := \prod_{j=0}^{t-1} a_j e_0 + \sum_{j=1}^{t-1} \prod_{i=1}^{j-1} a_{t-i} b_{t-j}, \tag{28}$$

where, by convention,

$$\prod_{i=1}^{0} a_{t-i} := 1.$$

In case there exist $a, b \in \mathbb{R}$ such that $a_i = a$ and $b_i = b$ for every i,

$$e_t = a^t e_0 + \sum_{j=0}^{t-1} a^j b.$$

Setting 1.

Theorem G.17. *Let Θ be a MPNN with T layers. Suppose that for every layer and every y and k,*

$$\|{}^t\xi^k_{\mathrm{y}}\|_\infty,\ \|\eta^t\|_\infty \leq \rho, \quad L_{\eta^t}, L_{{}^t\xi^k_{\mathrm{y}}} < L.$$

Let $(W,f),(V,g) \in \mathcal{WL}_r$. Then, for MPNN with no update function

$$\|\Theta_t(W,f) - \Theta_t(V,g)\|_\square \leq (4KL\rho)^t\|f-g\|_\square + \sum_{j=0}^{t-1}(4KL\rho)^j 4K\rho^2\|W-V\|_\square,$$

and for MPNN with update function

$$\|\Theta_t(W,f) - \Theta_t(V,g)\|_\square \leq (4KL^2\rho)^t\|f-g\|_\square + \sum_{j=0}^{t-1}(4KL^2\rho)^j 4K\rho^2 L\|W-V\|_\square.$$

Proof. We prove for MPNNs with update function, where the proof without update function is similar. We can write a recurrence sequence for a bound $\|\Theta_t(W,f) - \Theta_t(V,g)\|_\square \leq e_t$, by Theorem G.8 and Lemma G.9, as

$$e_{t+1} = 4KL^2\rho e_t + 4K\rho^2 L\|W-V\|_\square.$$

The proof now follows by applying Lemma G.16 with $a = 4KL^2\rho$ and $b = 4K\rho^2 L$. ∎

Setting 2.

Lemma G.18. *Let Θ be a MPNN with T layers. Suppose that for every layer t and every $\mathrm{y} \in \{\mathrm{r},\mathrm{t}\}$ and $k \in [K]$,*

$$\left|\eta^t(0)\right|,\ \left|{}^t\xi^k_{\mathrm{y}}(0)\right| \leq B, \quad L_{\eta^t},\ L_{{}^t\xi^k_{\mathrm{y}}} < L$$

with $L, B > 1$. Let $(W,f) \in \mathcal{WL}_r$. Then, for MPNN without update function, for every layer t,

$$\|\Theta_t(W,f)\|_\infty \leq (2KL^2B^2)^{2^t}\|f\|_\infty^{2^t},$$

and for MPNN with update function, for every layer t,

$$\|\Theta_t(W,f)\|_\infty \leq (2KL^3B^2)^{2^t}\|f\|_\infty^{2^t},$$

Proof. We first prove for MPNNs without update functions. Denote by C_t a bound on $\|{}^tf\|_\infty$, and let C_0 be a bound on $\|f\|_\infty$. By Lemma G.10, we may choose bounds such that

$$C_{t+1} \leq K(LC_t + B)^2 = KL^2C_t^2 + 2KLBC_t + KB^2.$$

We can always choose $C_t, K, L > 1$, and therefore,

$$C_{t+1} \leq KL^2C_t^2 + 2KLBC_t + KB^2 \leq 2KL^2B^2C_t^2.$$

Denote $a = 2KL^2B^2$. We have

$$\begin{aligned} C_{t+1} &= a(C_t)^2 = a(aC_{t-1}^2)^2 = a^{1+2}C_{t-1}^4 = a^{1+2}(a(C_{t-2})^2)^4 \\ &= a^{1+2+4}(C_{t-2})^8 = a^{1+2+4+8}(C_{t-3})^{16} \leq a^{2^t}C_0^{2^t}. \end{aligned}$$

Now, for MPNNs with update function, we have

$$\begin{aligned} C_{t+1} &\leq LK(LC_t + B)^2 + B \\ &= KL^3C_t^2 + 2KL^2BC_t + KB^2L + B \\ &\leq 2KL^3B^2C_t^2, \end{aligned}$$

and we proceed similarly.

∎

Theorem G.19. *Let Θ be a MPNN with T layers. Suppose that for every layer t and every $\mathrm{y} \in \{\mathrm{r}, \mathrm{t}\}$ and $k \in [K]$,*

$$\left|\eta^t(0)\right|, \ \left|{}^t\xi^k_{\mathrm{y}}(0)\right| \leq B, \quad L_{\eta^t}, \ L_{{}^t\xi^k_{\mathrm{y}}} < L,$$

with $L, B > 1$. Let $(W, g), (V, g) \in \mathcal{WL}_r$. Then, for MPNNs without update functions

$$\|\Theta_t(W, f) - \Theta_t(V, g)\|_\square \leq \prod_{j=0}^{t-1} 4K(L^2 r_j + LB)\|f - g\|_\square + \sum_{j=1}^{t-1}\prod_{i=1}^{j-1} 4K(L^2 r_{t-i} + LB)4K(Lr_{t-j} + B)^2\|W - V\|_\square,$$

where

$$r_i = (2KL^2B^2)^{2^i}\|f\|_\infty^{2^i},$$

and for MPNNs with update functions

$$\|\Theta_t(W, f) - \Theta_t(V, g)\|_\square \leq \prod_{j=0}^{t-1} 4K(L^3 r_j + L^2 B)\|f - g\|_\square + \sum_{j=1}^{t-1}\prod_{i=1}^{j-1} 4K(L^3 r_{t-i} + L^2 B)4KL(Lr_{t-j} + B)^2\|W - V\|_\square,$$

where

$$r_i = (2KL^3B^2)^{2^i}\|f\|_\infty^{2^i}.$$

Proof. We prove for MPNNs without update functions. The proof for the other case is similar. By Corollary G.11, since the signals at layer t are bounded by

$$r_t = (2KL^2B^2)^{2^t}\|f\|_\infty^{2^t},$$

we have

$$\|\Theta_{t+1}(W, f) - \Theta_{t+1}(V, g)\|_\square \leq 4K(L^2 r_t + LB)\|\Theta_t(W, f) - \Theta_t(V, g)\|_\square + 4K(Lr_t + B)^2\|W - V\|_\square.$$

We hence derive a recurrence sequence for a bound $\|\Theta_t(W, f) - \Theta_t(V, g)\|_\square \leq e_t$, as

$$e_{t+1} = 4K(L^2 r_t + LB)e_t + 4K(Lr_t + B)^2\|W - V\|_\square.$$

We now apply Lemma G.16. ∎

Setting 3.

Lemma G.20. *Suppose that for every layer t and every $\mathrm{y} \in \{\mathrm{r}, \mathrm{t}\}$ and $k = 1, \ldots, K$,*

$$\left|\eta^t(0)\right|, \ \left|\Phi^t(0,0)\right|, \ \left|{}^t\xi^k_{\mathrm{y}}(0)\right| \leq B, \quad L_{\eta^t}, \ L_{\Phi^t}, \ L_{{}^t\xi^k_{\mathrm{y}}} < L,$$

and ξ, Φ are all non-negatively valued. Then, for MPNNs without update function

$$\|\Theta^t(W, f)\|_\infty \leq L^t\|f\|_\infty + \sum_{j=1}^{t-1} L^j B,$$

and for MPNNs with update function

$$\|\Theta^t(W, f)\|_\infty \leq L^{2t}\|f\|_\infty + \sum_{j=1}^{t-1} L^{2j}(LB + B),$$

Proof. We first prove for MPNNs without update functions. By Lemma G.10, there is a bound e_t of $\|\Theta^t(W, f)\|_\infty$ that satisfies

$$e_t = Le_{t-1} + B.$$

Solving this recurent sequence via Lemma G.16 concludes the proof.

Lastly, for MPNN with update functions, we have a bound that satisfies

$$e_t = L^2 e_{t-1} + LB + B,$$

and we proceed as before. ∎

Lemma G.21. *Suppose that for every* $\mathrm{y} \in \{\mathrm{r}, \mathrm{t}\}$ *and* $k = 1, \ldots, K$

$$\left|\eta^t(0)\right|, \ \left|\Phi(0,0)\right|, \left|\xi^k_{\mathrm{y}}(0)\right| \leq B, \quad L_\Phi, L_{\xi^k_{\mathrm{y}}} < L,$$

and ξ, Φ *are all non-negatively valued. Let* $(W, g), (V, g) \in \mathcal{WL}_r$*. Then, for MPNNs without update functions*

$$\|\Theta^t(W, \Phi_f) - \Theta^t(V, \Phi_g)\|_\square = O(K^t L^{2t+t^2} r^t B^t)\Big(\|W - V\|_\square + \|f - g\|_\square\Big),$$

and for MPNNs with update functions

$$\|\Theta^t(W, \Phi_f) - \Theta^t(V, \Phi_g)\|_\square = O(K^t L^{3t+2t^2} r^t B^t)\Big(\|W - V\|_\square + \|f - g\|_\square\Big)$$

Proof. We start with MPNNs without update functions. By Corollary G.15 and Lemma G.20, there is a bound e_t on the error $\|\Theta^t(W, \Phi_f) - \Theta^t(V, \Phi_g)\|_\square$ at step t that satisfies

$$\begin{aligned} e_t &= 4K(L^2 r_{t-1} + LB)e_{t-1} + K(Lr + B)\|W - V\|_\square \\ &= 4K\Big(L^2\big(L^t\|f\|_\infty + \sum_{j=1}^{t-1} L^j B\big) + LB\Big)e_{t-1} + K\Big(L\big(L^t\|f\|_\infty + \sum_{j=1}^{t-1} L^j B\big) + B\Big)\|W - V\|_\square. \end{aligned}$$

Hence, by Lemma G.16, and Z defined by (28),

$$e_t = Z_t(\mathbf{a}, \mathbf{b}, \|f - g\|_\square) = O(K^t L^{2t+t^2} r^t B^t)\big(\|f - g\|_\square + \|W - V\|_\square\big),$$

where in the notations of Lemma G.16,

$$a_t = 4K\Big(L^2(L^t\|f\|_\infty + \sum_{j=1}^{t-1} L^j B) + LB\Big)$$

and

$$b_t = K\Big(L(L^t\|f\|_\infty + \sum_{j=1}^{t-1} L^j B) + B\Big)\|W - V\|_\square.$$

Next, for MPNNs with update functions, there is a bound that satisfies

$$\begin{aligned} e_t &= 4K(L^3 r_{t-1} + L^2 B)e_{t-1} + K(L^2 r + LB)\|W - V\|_\square \\ &= 4K\Big(L^3\big(L^{2t}\|f\|_\infty + \sum_{j=1}^{t-1} L^{2j}(LB + B)\big) + L^2 B\Big)e_{t-1} \\ &\quad + K\Big(L^2\big(L^{2t}\|f\|_\infty + \sum_{j=1}^{t-1} L^{2j}(LB + B)\big) + LB\Big)\|W - V\|_\square. \end{aligned}$$

Hence, by Lemma G.16, and Z defined by (28),

$$e_t = O(K^t L^{3t+2t^2} r^t B^t)\big(\|f - g\|_\square + \|W - V\|_\square\big).$$

■

H Generalization bound for MPNNs

In this appendix we prove Theorem 4.2.

H.1 Statistical learning and generalization analysis

In the statistical setting of learning, we suppose that the dataset comprises independent random samples from a probability space that describes all possible data $\mathcal{P}$. We suppose that for each $x \in \mathcal{P}$ there is a ground truth value $y_x \in \mathcal{Y}$, e.g., the ground truth class or value of x, where $\mathcal{Y}$ is, in general, some measure space. The *loss* is a measurable function $\mathcal{L} : \mathcal{Y}^2 \to \mathbb{R}_+$ that defines similarity in $\mathcal{Y}$. Given a measurable function $\Theta : \mathcal{P} \to \mathcal{Y}$, that we call the *model* or *network*, its accuracy on all potential inputs is defined as the *statistical risk* $R_{\rm stat}(\Theta) = \mathbb{E}_{x\sim\mathcal{P}}\Big(\mathcal{L}(\Theta(x), y_x)\Big)$. The goal in learning is to find a network Θ, from some *hypothesis space* $\mathcal{T}$, that has a low statistical risk. In practice, the statistical risk cannot be computed analytically. Instead, we suppose that a dataset $\mathcal{X} = \{x_m\}_{m=1}^M \subset \mathcal{P}$ of $M \in \mathbb{N}$ random independent samples with corresponding values $\{y_m\}_{m=1}^M \subset \mathcal{Y}$ is given. We estimate the statistical risk via a "Monte Carlo approximation," called the *empirical risk* $R_{\rm emp}(\Theta) = \frac{1}{M}\sum_{m=1}^M \mathcal{L}(\Theta(x_m), y_m)$. The network Θ is chosen in practice by optimizing the empirical risk. The goal in generalization analysis is to show that if a learned Θ attains a low empirical risk, then it is also guaranteed to have a low statistical risk.

One technique for bounding the statistical risk in terms of the empirical risk is to use the bound $R_{\rm stat}(\Theta) \le R_{\rm emp}(\Theta) + E$, where E is the *generalization error* $E = \sup_{\Theta\in\mathcal{T}} |R_{\rm stat}(\Theta) - R_{\rm emp}(\Theta)|$, and to find a bound for E. Since the trained network $\Theta = \Theta_{\mathcal{X}}$ depends on the data $\mathcal{X}$, the network is not a constant when varying the dataset, and hence the empirical risk is not really a Monte Carlo approximation of the statistical risk in the learning setting. If the network Θ was fixed, then Monte Carlo theory would have given us a bound of E^2 of order $O\big(\kappa(p)/M\big)$ in an event of probability $1-p$, where, for example, in Hoeffding's inequality Theorem H.2, $\kappa(p) = \log(2/p)$. Let us call such an event a *good sampling event*. Since the good sampling event depends on Θ, computing a naive bound to the generalization error would require intersecting all good sampling events for all $\Theta \in \mathcal{T}$. Uniform convergence bounds are approaches for intersecting adequate sampling events that allow bounding the generalization error more efficiently. This intersection of events leads to a term in the generalization bound, called the *complexity/capacity*, that describes the richness of the hypothesis space $\mathcal{T}$. This is the philosophy behind approaches such as VC-dimension, Rademacher dimension, fat-shattering dimension, pseudo-dimension, and uniform covering number (see, e.g., [32]).

H.2 Classification setting

We define a ground truth classifier into C classes as follows. Let $\mathcal{C} : \widetilde{\mathcal{WL}_r} \to \mathbb{R}^C$ be a measurable piecewise constant function of the following form. There is a partition of $\mathcal{WL}_r$ into disjoint measurable sets $B_1, \ldots, B_C \subset \widetilde{\mathcal{WL}_r}$ such that $\bigcup_{i=1}^C B_i = \widetilde{\mathcal{WL}_r}$, and for every $i \in [C]$ and every $x \in B_i$,

$$\mathcal{C}(x) = e_i,$$

where $e_i \in \mathbb{R}^C$ is the standard basis element with entries $(e_i)_j = \delta_{i,j}$, where $\delta_{i,j}$ is the Kronecker delta.

We define an arbitrary data distribution as follows. Let $\mathcal{B}$ be the Borel σ-algebra of $\widetilde{\mathcal{WL}_r}$, and ν be any probability measure on the measurable space $(\widetilde{\mathcal{WL}_r}, \mathcal{B})$. We may assume that we complete $\mathcal{B}$ with respect to ν, obtaining the σ-algebra Σ. If we do not complete the measure, we just denote $\Sigma = \mathcal{B}$. Defining $(\widetilde{\mathcal{WL}_r}, \Sigma, \nu)$ as a complete measure space or not will not affect our construction.

Let $\mathcal{S}$ be a metric space. Let $\mathrm{Lip}(\mathcal{S}, L)$ be the space of Lipschitz cintinuous mappings $\Upsilon : \mathcal{S} \to \mathbb{R}^C$ with Lipschitz constant L. Note that by Theorem 4.1, for every $i \in [C]$, the space of MPNN with Lipschitz continuous input and output message functions and Lipschitz update functions, restricted to B_i, is a subset of $\mathrm{Lip}(B_i, L_1)$ which is the restriction of $\mathrm{Lip}(\widetilde{\mathcal{WL}_r}, L_1)$ to $B_i \subset \widetilde{\mathcal{WL}_r}$, for some $L_1 > 0$. Moreover, B_i has finite covering $\kappa(\epsilon)$ given in (25). Let $\mathcal{E}$ be a Lipschitz continuous loss function with Lipschitz constant L_2. Therefore, since $\mathcal{C}|_{B_i}$ is in $\mathrm{Lip}(B_i, 0)$, for any $\Upsilon \in \mathrm{Lip}(\widetilde{\mathcal{WL}_r}, L_1)$, the function $\mathcal{E}(\Upsilon|_{B_i}, \mathcal{C}|_{B_i})$ is in $\mathrm{Lip}(B_i, L)$ with $L = L_1 L_2$.

H.3 Uniform Monte Carlo approximation of Lipschitz continuous functions

The proof of Theorem 4.2 is based on the following Theorem H.3, which studies uniform Monte Carlo approximations of Lipschitz continuous functions over metric spaces with finite covering.

Definition H.1. *A metric space $\mathcal{M}$ is said to have* covering number $\kappa : (0,\infty) \to \mathbb{N}$*, if for every $\epsilon > 0$, the space $\mathcal{M}$ can be covered by $\kappa(\epsilon)$ ball of radius ϵ.*

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

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

The following theorem is an extended version of [25, Lemma B.3], where the difference is that we use a general covering number $\kappa(\epsilon)$, where in [25, Lemma B.3] the covering number is exponential in ϵ. For completion, we repeat here the proof, with the required modification.

Theorem H.3 (Uniform Monte Carlo approximation for Lipschitz continuous functions)**.** *Let $\mathcal{X}$ be a probability metric space[5], with probability measure μ, and covering number $\kappa(\epsilon)$. Let $X_1, \ldots, X_N$ be drawn i.i.d. from $\mathcal{X}$. Then, for every $p > 0$, there exists an event $\mathcal{E}^p_{\text{Lip}} \subset \mathcal{X}^N$ (regarding the choice of $(X_1, \ldots, X_N)$), with probability*

$$\mu^N(\mathcal{E}^p_{\text{Lip}}) \geq 1 - p,$$

such that for every $(X_1, \ldots, X_N) \in \mathcal{E}^p_{\text{Lip}}$, for every bounded Lipschitz continuous function $F : \mathcal{X} \to \mathbb{R}^d$ with Lipschitz constant L_F, we have

$$\left\|\int F(x)d\mu(x) - \frac{1}{N}\sum_{i=1}^{N}F(X_i)\right\|_\infty \leq 2\xi^{-1}(N)L_f + \frac{1}{\sqrt{2}}\xi^{-1}(N)\|F\|_\infty(1 + \sqrt{\log(2/p)}), \quad (29)$$

where $\xi(r) = \frac{\kappa(r)^2\log(\kappa(r))}{r^2}$ and ξ^{-1} is the inverse function of ξ.

Proof. Let $r > 0$. There exists a covering of $\mathcal{X}$ by a set of balls $\{B_j\}_{j\in[J]}$ of radius r, where $J = \kappa(r)$. For $j = 2, \ldots, J$, we define $I_j := B_j \setminus \cup_{i<j}B_i$, and define $I_1 = B_1$. Hence, $\{I_j\}_{j\in[J]}$ is a family of measurable sets such that $I_j \cap I_i = \emptyset$ for all $i \neq j \in [J]$, $\bigcup_{j\in[J]} I_j = \chi$, and $\text{diam}(I_j) \leq 2r$ for all $j \in [J]$, where by convention $\text{diam}(\emptyset) = 0$. For each $j \in [J]$, let z_j be the center of the ball B_j.

Next, we compute a concentration of error bound on the difference between the measure of I_j and its Monte Carlo approximation, which is uniform in $j \in [J]$. Let $j \in [J]$ and $q \in (0, 1)$. By Hoeffding's inequality Theorem H.2, there is an event $\mathcal{E}^q_j$ with probability $\mu(\mathcal{E}^q_j) \geq 1 - q$, in which

$$\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{I_j}(X_i) - \mu(I_k)\right\|_\infty \leq \frac{1}{\sqrt{2}}\frac{\sqrt{\log(2/q)}}{\sqrt{N}}. \quad (30)$$

Consider the event

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

with probability $\mu^N(\mathcal{E}^{Jq}_{\text{Lip}}) \geq 1 - Jq$. In this event, (30) holds for all $j \in \mathcal{J}$. We change the failure probability variable $p = Jq$, and denote $\mathcal{E}^p_{\text{Lip}} = \mathcal{E}^{Jq}_{\text{Lip}}$.

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

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

[5] A metric space with a probability Borel measure, where we either take the completion of the measure space with respect to μ (adding all subsets of null-sets to the σ-algebra) or not.

Then,

$$
\begin{aligned}
\left\|\frac{1}{N}\sum_{i=1}^{N}F(X_i)-\int_{\chi}F(y)d\mu(y)\right\|_\infty \leq & \left\|\frac{1}{N}\sum_{i=1}^{N}F(X_i)-\frac{1}{N}\sum_{i=1}^{N}F^r(X_i)\right\|_\infty \\
& + \left\|\frac{1}{N}\sum_{i=1}^{N}F^r(X_i)-\int_{\chi}F^r(y)d\mu(y)\right\|_\infty \\
& + \left\|\int_{\chi}F^r(y)d\mu(y)-\int_{\chi}F(y)d\mu(y)\right\|_\infty \\
& =: (1)+(2)+(3).
\end{aligned} \tag{31}
$$

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

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

We proceed by bounding (2). In the event of $\mathcal{E}^p_{\mathrm{Lip}}$, which holds with probability at least $1-p$, equation (30) holds for all $j \in \mathcal{J}$. In this event, we get

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

Recall that $J=\kappa(r)$. Then, with probability at least $1-p$

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

To bound (3), we calculate

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

By plugging the bounds of (1), (2) and (3) into (31), we get

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

Lastly, choosing $r=\xi^{-1}(N)$ for $\xi(r)=\frac{\kappa(r)^2\log(\kappa(r))}{r^2}$, gives $\frac{\kappa(r)\sqrt{\log(\kappa(r))}}{\sqrt{N}}=r$, so

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

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

H.4 A generalization theorem for MPNNs

The following generalization theorem of MPNN is now a direct result of Theorem H.3.

Let $\text{Lip}(\widetilde{\mathcal{WL}_r},L_1)$ denote the space of Lipschitz continuous functions $\Theta:\mathcal{WL}_r\to\mathbb{R}^C$ with Lipschitz bound bounded by L_1 and $\|\Theta\|_{\infty}\leq L_1$. We note that the theorems of Appendix G.2 prove that MPNN with Lipschitz continuous message and update functions, and bounded formal biases, are in $\text{Lip}(\widetilde{\mathcal{WL}_r},L_1)$.

Theorem H.4 (MPNN generalization theorem)**.** *Consider the classification setting of Appendix H.2. Let $X_1,\ldots,X_N$ be independent random samples from the data distribution $(\widetilde{\mathcal{WL}_r},\Sigma,\nu)$. Then, for every $p>0$, there exists an event $\mathcal{E}^p\subset\widetilde{\mathcal{WL}_r}^N$ regarding the choice of $(X_1,\ldots,X_N)$, with probability*

$$\nu^N(\mathcal{E}^p)\geq 1-Cp-2\frac{C^2}{N},$$

in which for every function Υ in the hypothesis class $\text{Lip}(\widetilde{\mathcal{WL}_r},L_1)$, with we have

$$\left|\mathcal{R}(\Upsilon_{\mathbf{X}})-\hat{\mathcal{R}}(\Upsilon_{\mathbf{X}},\mathbf{X})\right|\leq\xi^{-1}(N/2C)\Big(2L+\frac{1}{\sqrt{2}}\big(L+\mathcal{E}(0,0)\big)\big(1+\sqrt{\log(2/p)}\big)\Big),\tag{32}$$

where $\xi(r)=\frac{\kappa(r)^2\log(\kappa(r))}{r^2}$, κ is the covering number of $\widetilde{\mathcal{WL}_r}$ given in (25), and ξ^{-1} is the inverse function of ξ.

Proof. For each $i\in[C]$, let S_i be the number of samples of $\mathbf{X}$ that falls within B_i. The random variable $(S_1,\ldots,S_C)$ is multinomial, with expected value $(N/C,\ldots,N/C)$ and variance $(\frac{N(C-1)}{C^2},\ldots,\frac{N(C-1)}{C^2})\leq(\frac{N}{C},\ldots,\frac{N}{C})$. We now use Chebyshev's inequality, which states that for any $a>0$,

$$P\Big(|S_i-N/C|>a\sqrt{\frac{N}{C}}\Big)<a^{-2}.$$

We choose $a\sqrt{\frac{N}{C}}=\frac{N}{2C}$, so $a=\frac{N^{1/2}}{2C^{1/2}}$, and

$$P(|S_i-N/C|>\frac{N}{2C})<\frac{2C}{N}.$$

Therefore,

$$P(S_i>\frac{N}{2C})>1-\frac{2C}{N}.$$

We intersect these events of $i \in [C]$, and get an event $\mathcal{E}_{\text{mult}}$ of probability more than $1 - 2\frac{C^2}{N}$ in which $S_i > \frac{N}{2C}$ for every $i \in [C]$. In the following, given a set B_i we consider a realization $M = S_i$, and then use the law of total probability.

From Theorem H.3 we get the following. For every $p > 0$, there exists an event $\mathcal{E}_i^p \subset B_i^M$ regarding the choice of $(X_1, \ldots, X_M) \subset B_i$, with probability

$$\nu^M(\mathcal{E}_{\text{Lip}}^p) \geq 1 - p,$$

such that for every function Υ' in the hypothesis class $\text{Lip}(\widetilde{\mathcal{WL}}_r, L_1)$, we have

$$\left| \int \mathcal{E}\big(\Upsilon'(x), \mathcal{C}(x)\big) d\nu(x) - \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}\big(\Upsilon'(X_i), \mathcal{C}(X_i)\big) \right| \tag{33}$$

$$\leq 2\xi^{-1}(M)L + \frac{1}{\sqrt{2}}\xi^{-1}(M)\|\mathcal{E}\big(\Upsilon'(\cdot), \mathcal{C}(\cdot)\big)\|_\infty(1 + \sqrt{\log(2/p)}) \tag{34}$$

$$\leq 2\xi^{-1}(N/2C)L + \frac{1}{\sqrt{2}}\xi^{-1}(N/2C)(L + \mathcal{E}(0,0))(1 + \sqrt{\log(2/p)}), \tag{35}$$

where $\xi(r) = \frac{\kappa(r)^2 \log(\kappa(r))}{r^2}$, κ is the covering number of $\widetilde{\mathcal{WL}}_r$ given in (25), and ξ^{-1} is the inverse function of ξ. In the last inequality, we use the bound, for every $x \in \tilde{\mathcal{WL}}_r$,

$$\left|\mathcal{E}\big(\Upsilon'(x), \mathcal{C}(x)\big)\right| \leq \left|\mathcal{E}\big(\Upsilon'(x), \mathcal{C}(x)\big) - \mathcal{E}(0,0)\right| + |\mathcal{E}(0,0)| \leq L_2 \, |L_1 - 0| + |\mathcal{E}(0,0)| \, .$$

Since (33) is true for any $\Upsilon' \in \text{Lip}(\widetilde{\mathcal{WL}}_r, L_1)$, it is also true for $\Upsilon_{\mathbf{X}}$ for any realization of $\mathbf{X}$, so we also have

$$\left|\mathcal{R}(\Upsilon_{\mathbf{X}}) - \hat{\mathcal{R}}(\Upsilon_{\mathbf{X}}, \mathbf{X})\right| \leq 2\xi^{-1}(N/2C)L + \frac{1}{\sqrt{2}}\xi^{-1}(N/2C)(L + \mathcal{E}(0,0))(1 + \sqrt{\log(2/p)}).$$

Lastly, we denote

$$\mathcal{E}^p = \mathcal{E}_{\text{mult}} \cap \Big(\bigcup_{i=1}^{C} \mathcal{E}_i^p\Big).$$

■

I Stability of MPNNs to graph subsampling

Lastly, we prove Theorem 4.3.

Theorem I.1. *Consider the setting of Theorem 4.2, and let Θ be a MPNN with Lipschitz constant L. Denote*

$$\Sigma = \big(W, \Theta(W, f)\big), \quad \textit{and} \quad \Sigma(\Lambda) = \Big(\mathbb{G}(W, \Lambda), \Theta\big(\mathbb{G}(W, \Lambda), f(\Lambda)\big)\Big).$$

Then

$$\mathbb{E}\Big(\delta_\square\big(\Sigma, \Sigma(\Lambda)\big)\Big) < \frac{15}{\sqrt{\log(k)}}L.$$

Proof. By Lipschitz continuity of Θ,

$$\delta_\square\big(\Sigma, \Sigma(\Lambda)\big) \leq L\delta_\square\Big(\big(W, f\big), \big(\mathbb{G}(W, \Lambda), f(\Lambda)\big)\Big).$$

Hence,

$$\mathbb{E}\Big(\delta_\square\big(\Sigma, \Sigma(\Lambda)\big)\Big) \leq L\mathbb{E}\Big(\delta_\square\Big(\big(W, f\big), \big(\mathbb{G}(W, \Lambda), f(\Lambda)\big)\Big)\Big),$$

and the claim of the theorem follows from Theorem 3.6. ■

As explained in Section 3.5, the above theorem of stability of MPNNs to graphon-signal sampling also applies to subsampling graph-signals.

References

[1] N. Alon, W. de la Vega, R. Kannan, and M. Karpinski. Random sampling and approximation of max-csps. *Journal of Computer and System Sciences*, 67(2):212–243, 2003. ISSN 0022-0000. doi: https://doi.org/10.1016/S0022-0000(03)00008-4. Special Issue on STOC 2002.

[2] K. Atz, F. Grisoni, and G. Schneider. Geometric deep learning on molecular representations. *Nature Machine Intelligence*, 3:1023–1032, 2021.

[3] W. Azizian and M. Lelarge. Expressive power of invariant and equivariant graph neural networks. In *ICLR*, 2021.

[4] C. Borgs, J. T. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi. Convergent sequences of dense graphs i: Subgraph frequencies, metric properties and testing. *Advances in Mathematics*, 219(6):1801–1851, 2008.

[5] J. Chen, T. Ma, and C. Xiao. FastGCN: Fast learning with graph convolutional networks via importance sampling. In *International Conference on Learning Representations*, 2018.

[6] Z. Chen, S. Villar, L. Chen, and J. Bruna. On the equivalence between graph isomorphism testing and function approximation with gnns. In *NeurIPS*. Curran Associates, Inc., 2019.

[7] W.-L. Chiang, X. Liu, S. Si, Y. Li, S. Bengio, and C.-J. Hsieh. Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '19, page 257–266, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330925.

[8] M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *NeurIPS*. Curran Associates Inc., 2016. ISBN 9781510838819.

[9] J. M. S. et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688 – 702.e13, 2020. ISSN 0092-8674.

[10] M. Fey and J. E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR 2019 Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[11] G. B. Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

[12] A. M. Frieze and R. Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19:175–220, 1999.

[13] V. Garg, S. Jegelka, and T. Jaakkola. Generalization and representational limits of graph neural networks. In H. D. III and A. Singh, editors, *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 3419–3430. PMLR, 13–18 Jul 2020.

[14] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pages 1263–1272, 2017.

[15] W. L. Hamilton, R. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, page 1025–1035. Curran Associates Inc., 2017. ISBN 9781510860964.

[16] J. M. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. A. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021.

[17] N. Keriven, A. Bietti, and S. Vaiter. Convergence and stability of graph convolutional networks on large random graphs. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020.

[18] N. Keriven, A. Bietti, and S. Vaiter. On the universality of graph neural networks on large random graphs. In *NeurIPS*. Curran Associates, Inc., 2021.

[19] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[20] R. Levie, F. Monti, X. Bresson, and M. M. Bronstein. Cayleynets: Graph convolutional neural networks with complex rational spectral filters. *IEEE Transactions on Signal Processing*, 67(1): 97–109, 2019. doi: 10.1109/TSP.2018.2879624.

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

[22] R. Liao, R. Urtasun, and R. Zemel. A PAC-bayesian approach to generalization bounds for graph neural networks. In *ICLR*, 2021.

[23] L. M. Lovász. Large networks and graph limits. In *volume 60 of Colloquium Publications*, 2012. doi: 10.1090/coll/060.

[24] L. M. Lovász and B. Szegedy. Szemerédi's lemma for the analyst. *GAFA Geometric And Functional Analysis*, 17:252–270, 2007.

[25] S. Maskey, R. Levie, Y. Lee, and G. Kutyniok. Generalization analysis of message passing neural networks on large random graphs. In *NeurIPS*. Curran Associates, Inc., 2022.

[26] S. Maskey, R. Levie, and G. Kutyniok. Transferability of graph neural networks: An extended graphon approach. *Applied and Computational Harmonic Analysis*, 63:48–83, 2023. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2022.11.008.

[27] O. Méndez-Lucio, M. Ahmad, E. A. del Rio-Chanona, and J. K. Wegner. A geometric deep learning approach to predict binding conformations of bioactive molecules. *Nature Machine Intelligence*, 3:1033–1039, 2021.

[28] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4602–4609, Jul. 2019. doi: 10.1609/aaai.v33i01.33014602.

[29] C. Morris, F. Geerts, J. Tönshoff, and M. Grohe. Wl meet vc. In *ICML*. PMLR, 2023.

[30] L. Ruiz, L. F. O. Chamon, and A. Ribeiro. Graphon signal processing. *IEEE Transactions on Signal Processing*, 69:4961–4976, 2021.

[31] F. Scarselli, A. C. Tsoi, and M. Hagenbuchner. The vapnik–chervonenkis dimension of graph and recursive neural networks. *Neural Networks*, 108:248–259, 2018.

[32] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. doi: 10.1017/CBO9781107298019.

[33] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991. doi: 10.1017/CBO9780511813658.

[34] K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.