A. Vector Space of Invariant and Equivariant Linear Graph Operators

Denote with $\Gamma(k)$ the set of all partitions of $\{1,\dots,k\}$. Bell's number $\mathrm{Bell}(k)$ represents the cardinality of $\Gamma(k)$. Given a partition $\gamma \in \Gamma(k)$, we say that multi-index $\mathbf{a} \in \{1,\dots,n\}^k$ *complies* with γ if, for any $i,j \in \{1,\dots,n\}$, we have that $\mathbf{a}_i = \mathbf{a}_j$ if and only if i,j belong to the same set in γ.

For every $\gamma \in \Gamma(k)$, tensor $\mathbf{I}_\gamma \in \mathcal{T}^k$ is defined as $(\mathbf{I}_\gamma)_{\mathbf{a}} = 1$ for any $\mathbf{a}$ complying with γ, and 0 otherwise. With the scalar tensor product

$$\mathbf{I} \cdot \mathbf{T} := \sum_{\mathbf{a} \in \{1,\dots,n\}^k} \mathbf{I}_{\mathbf{a}} \mathbf{T}_{\mathbf{a}} \quad \in \mathbb{R},$$

for any $\mathbf{I}, \mathbf{T} \in \mathcal{T}^k$. We obtain that basis $\{\mathbf{I}_\gamma\}_{\gamma \in \Gamma(k)}$ is orthogonal, because the locations in which the tensors are non-null are disjoints.

Similarly, for every $\gamma \in \Gamma(k+l)$, tensor $\mathbf{E}_\gamma \in \mathcal{T}^{k+l}$ is defined as $(\mathbf{E}_\gamma)_{\mathbf{a}} = 1$ for any $\mathbf{a}$ complying with γ, and 0 otherwise. Here the tensor product $\mathbf{E} \cdot \mathbf{T}$ between $\mathbf{E} \in \mathcal{T}^{k+l}$ and $\mathbf{T} \in \mathcal{T}^k$ is defined as

$$\mathbf{E} \cdot \mathbf{T} := \sum_{\mathbf{b} \in \{1,\dots,n\}^l} \sum_{\mathbf{a} \in \{1,\dots,n\}^k} \mathbf{E}_{\mathbf{b},\mathbf{a}} \mathbf{T}_{\mathbf{a}} \quad \in \mathcal{T}^l.$$

Again, for different γ, γ', the basis elements $\mathbf{E}_\gamma$ and $\mathbf{E}_{\gamma'}$ are orthogonal.

It is not rare that graphs come with both node and edge attributes, say F_{node}- and F_{edge}-dimensional, respectively. In this case one can create a the Cartesian product space of dimension $F = F_{node} + F_{edge}$, and represent any n-node graph g as a tensor $\mathbf{A}_g \in \mathcal{T}^2 \times \mathbb{R}^F$. Accordingly, an equivariant map is function $f : \mathcal{T}^k \times \mathbb{R}^F \to \mathcal{T}^l \times \mathbb{R}^{F'}$, so that $f(\pi \star \mathbf{T}) = \pi \star f(\mathbf{T})$ for every $\pi \in S_n$, and where π is now acting on all components, but the last one. Similarly, we extend the definition of invariant maps.

We refer the reader to the original paper for a detailed description (Maron et al., 2019b).

B. Proofs

B.1. Proof of Lemma 1

The proof employs the functions $f^{\otimes}$ in the form

$$f^{\otimes}(g) = \sum_{s=1}^{S} H_{\sum k_s} \Big[\rho_e \left(F^{(s,1)}_{2,k_{s,1}}(\mathbf{A}_g; \boldsymbol{\theta}_{s,1}) \right) \otimes \dots \\ \dots \otimes \rho_e \left(F^{(s,T)}_{2,k_{s,T}}(\mathbf{A}_g; \boldsymbol{\theta}_{s,1}) \right); \boldsymbol{\theta}_H \Big] + b$$

with functions $\{H^{(s)}_i\}$ and $\{F^{(s,t)}_{2,j}\}$ linear invariant and equivariant functions as defined in (2), but without the bias terms. We denote with $\mathcal{N}^{\otimes}(\rho_e)$ the set of such functions letting S, T vary in $\mathbb{N}$, $b \in \mathbb{R}$, and for any $(k_{1,1}, \boldsymbol{\theta}_{s,t}, \boldsymbol{\theta}_H) \in \mathcal{W}$. We also denote with $\mathcal{N}(\rho_e)$ the restriction of $\mathcal{N}^{\otimes}(\rho_e)$ to $T = 1$, which is contained in the closure under finite sums of set $\mathcal{F}(\mathrm{id}, \rho_e)$, where $\mathrm{id} : \mathbb{R} \to \mathbb{R}$ is the identity function.

1) Keriven & Peyré (2019) showed that $\mathcal{N}^{\otimes}(\sigma)$, with σ the sigmoid activation, separates $\mathcal{G}$ [Lem. 2] and that $\mathcal{N}(\rho_e)$ is dense in $\mathcal{N}^{\otimes}(\sigma)$ for any squashing function ρ_e [Lem. 3]. Since $\mathcal{N}(\rho_e) \subseteq \mathcal{F}(\mathrm{id}, \rho_e)$, then $\mathcal{F}(\mathrm{id}, \rho_e, \mathcal{W})$ is dense on $\mathcal{N}^{\otimes}(\rho_e)$, and it derives that $\mathcal{F}(\mathrm{id}, \rho_e, \mathcal{W})$ separates points of $\mathcal{G}$, as well. We conclude that for any pair of distinct graphs $g_1 \neq g_2$ in $\mathcal{G}$, there is a function $\tilde{f} \in \mathcal{F}(\mathrm{id}, \rho_e, \mathcal{W})$, such that $\tilde{f}(g_1) \neq \tilde{f}(g_2)$.

2) Notice that, for any $a, b \in \mathbb{R}$, $f_{a,b}(\cdot) := a\tilde{f}(\cdot) + b \in \mathcal{F}(\mathrm{id}, \rho_e, \mathcal{W})$, and when $a \neq 0$ we also have $f_{a,b}(g_1) \neq f_{a,b}(g_2)$; therefore, we can push $f_{a,b}(g_1)$ and $f_{a,b}(g_2)$ to any desired location. Let $\delta := \tilde{f}(g_2) - \tilde{f}(g_1) > 0$ and, without loss on generality, assume that $\rho_i(0) \neq \rho_i(1)$ (in fact, ρ_i is non-constant). With the choice $a = \frac{1}{\delta}$ and $b = -\frac{\tilde{f}(g_2)}{\delta}$, we obtain that $f_{a,b}(g_2) = 0$ and $f_{a,b}(g_1) = 1$ and $\rho_i(f_{a,b}(g_2)) \neq \rho_i(f_{a,b}(g_1))$, which proves that for any $g_1 \neq g_2$ there exists a function $f \in \mathcal{F}(\rho_i, \rho_e, \mathcal{W})$ such that $f(g_1) \neq f(g_2)$.

B.2. Proof of Theorem 1

1) Since the $\delta(g_1, g_2) := (\psi(g_1) - \psi(g_2))^2 \geq 0$, $\delta(g_1, g_2) = \delta(g_2, g_1)$ and $\delta(g_1, g_1) = 0$ for any $g_1, g_2 \in \mathcal{G}$, then the same properties hold also for $d_P(g_1, g_2) = \sqrt{\mathbb{E}[\delta(g_1, g_2)]}$.

2) The Cauchy-Schwarz inequality

$$|\mathbb{E}[X_1 X_2]|^2 \leq \mathbb{E}[X_1^2]\mathbb{E}[X_2^2], \tag{19}$$

holds for any pair of random variables X_1, X_2; in fact, notice that

$$0 \leq \frac{1}{2}\,\mathbb{E}\left[\left(\frac{X}{\sqrt{\mathbb{E}[X^2]}} - \frac{Y}{\sqrt{\mathbb{E}[Y^2]}}\right)^2\right] \\ = 1 - \frac{\mathbb{E}\,[XY]}{\sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]}};$$

3) By (19), we have

$$\mathbb{E}[(X_1 + X_2)^2] \leq \mathbb{E}[X_1^2] + 2\sqrt{\mathbb{E}[X_1^2]\mathbb{E}[X_2^2]} + \mathbb{E}[X_2^2] \\ = \left(\sqrt{\mathbb{E}[X_1^2]} + \sqrt{\mathbb{E}[X_2^2]}\right)^2. \tag{20}$$

The triangular inequality follows from the choice $X_1 = \psi(g_1; \mathbf{w}) - \psi(g_3; \mathbf{w})$ and $X_2 = \psi(g_3; \mathbf{w}) - \psi(g_2; \mathbf{w})$.

4) Finally, the identifiability property (5) is proved by the following Lemma 2.

Lemma 2. *Under Assumptions (A1) and (A2), we have that for any pair of graphs $g_1, g_2 \in \mathcal{G}$*

$$g_1 = g_2 \iff d_P(g_1, g_2) = 0.$$

Proof. Denote with $\mathcal{W}_k$ the parameter set $\mathcal{W}$ in which the hidden tensor order is fixed to k.

1) Assumption (A1) enables Lemma 1, therefore for any pair $g_1, g_2 \in \mathcal{G}$ with $g_1 \neq g_2$ there exists a parameter configuration $\tilde{\mathbf{w}} \in \mathcal{W}$ so that $\psi(g_1, \tilde{\mathbf{w}}) \neq \psi(g_2, \tilde{\mathbf{w}})$.

2) Again from Assumption (A1), for any $k \in \mathbb{N}$, we have that $\psi(g; \cdot) : \mathcal{W} \to \mathbb{R}$ limited to $\mathcal{W}_k$, is continuous, as it is the composition of linear operators and continuous activations ρ_i, ρ_e. This holds in particular for $\tilde{k}$, the hidden tensor order associated with $\tilde{\mathbf{w}}$. Therefore, there is a neighbourhood $U_{g_1,g_2}(\tilde{\mathbf{w}})$ of $\tilde{\mathbf{w}}$ such that

$$|\psi(g_1, \mathbf{w}) - \psi(g_2, \mathbf{w})| \geq \frac{\varepsilon_{g_1,g_2}}{2}, \qquad \forall \mathbf{w} \in U(\tilde{\mathbf{w}}),$$

with $\varepsilon_{g_1,g_2} = |\psi(g_1; \tilde{\mathbf{w}}) - \psi(g_2; \tilde{\mathbf{w}})| > 0$.

3) Assumption (A2) ensures that $\mathrm{supp}(P) = \mathcal{W}$, and that $P(U_{g_1,g_2}(\tilde{\mathbf{w}}))$ is strictly positive, independently on the choice of graphs g_1, g_2. We conclude that for any pair g_1, g_2 of distinct graphs,

$$d_P(g_1, g_2) \geq P(U_{g_1,g_2}(\tilde{\mathbf{w}})) \frac{\varepsilon_{g_1,g_2}}{2} > 0.$$

□

C. Limiting the Order of the Hidden Tensor: Weighted GRNF

Allowing order k to grow indefinitely might result in an infeasible computation load. In the following section, we show how to cope with this problem by defining a d_P and κ_P over a distribution P, while sampling parameter $\mathbf{w}$ from a different and more convenient one, $\overline{P}$.

Limiting k to be less or equal than k^* results in distance $d_P(\cdot, \cdot)$ which is not metric, in general, and one can build practical counterexamples. Consider a P with $\mathrm{supp}(P) = \mathcal{W}$, and assume p to be the marginal probability mass function associated to k. Consider also a nonempty subset $\mathcal{K} \subseteq \mathbb{N}$ and define the probability function $\overline{P}$ with marginal probability mass function

$$\bar{p}(k) = \frac{p(k)}{P(\mathcal{K})}, \qquad k \in \mathcal{K}, \qquad \text{and } 0 \text{ otherwise}$$

where $P(\mathcal{K})$ is the normalizing factor $\sum_{l \in \mathcal{K}} p(l)$. We obtain that approximating κ_P and d_P^2 by sampling the parameters $\mathbf{w}$ from $\overline{P}$, and considering the following modified GRNF

$$\overline{\mathbf{z}}(\,\cdot\,; \mathbf{W}) := \sqrt{P(\mathcal{K})}\, \mathbf{z}(\,\cdot\,; \mathbf{W}), \tag{21}$$

yields a practical alternative, as shown in the following lemma. We call $\overline{\mathbf{z}}$ in (21) *bounded-order GRNF* to distinguish it from the *plain* GRNF (10).

Lemma 3. *Consider the bounded-order GRNF (21). If ρ_i is bounded by a constant C_{ρ_i}, then*

$$\mathbb{E}\left[|\overline{\mathbf{z}}(g_1; \mathbf{W}) - \overline{\mathbf{z}}(g_2; \mathbf{W})|_2^2\right] \begin{cases} \geq d_P(g_1, g_2)^2 - (1 - \mathbb{P}_p(\mathcal{K}))\, 4C_{\rho_i}^2 \\ \leq d_P(g_1, g_2)^2. \end{cases}$$

Proof. **1)** Let us start with a generic random variable X.

$$\begin{aligned} \mathbb{E}_{\overline{p}}[X] &= \sum_{k=1}^{\infty} \overline{p}(k) X = \sum_{k \in \mathcal{K}} \overline{p}(k) X = \frac{1}{\mathbb{P}_p(\mathcal{K})} \sum_{k \in \mathcal{K}} p(k) X \\ &= \frac{1}{\mathbb{P}_p(\mathcal{K})} \sum_{k=1}^{\infty} p(k) X - \sum_{k \notin \mathcal{K}} p(k) X \\ &= \frac{1}{\mathbb{P}_p(\mathcal{K})} \mathbb{E}_p[X] - \frac{1}{\mathbb{P}_p(\mathcal{K})} \sum_{k \notin \mathcal{K}} p(k) X. \end{aligned}$$

2) Substituting $X = (\mathbf{z}(g_1; \mathbf{w}) - \mathbf{z}(g_2; \mathbf{w}))^2$, and then taking the expectation with respect to the joint $\overline{P}$, we get $\mathbb{E}_{\overline{P}}[X] \leq d_P(g_1, g_2)^2$. Finally, being ρ_i bounded by a constant C_{ρ_i}, we get $X \leq 4C_{\rho_i}^2$, hence the thesis.

□

We stress that, despite the hidden orders are sampled from the bounded-order distribution $\overline{p}$, the result relates to distance $d_P(\cdot, \cdot)$, which is with respect to the original distribution P with $\mathrm{supp}(P) = \mathcal{W}$.

As we can see, completely avoiding certain hidden tensor orders k comes with the price of biased estimations, which does not ensure convergence in (13) and (11). One can obtain consistent approximations (13) and (11) while mitigating the computational and memory burden by selecting probability distribution $\overline{P}$, which down-weights large hidden-tensor orders maintaining $\mathrm{supp}(\overline{P}) = \mathcal{W}$. We can also make a step further and let the entire distribution P vary. We result in the Weighted GRNF defined in (14). This type of embedding is a generalization of both the plain GRNF (10) and the bounded-order GRNF (21).

We prove that a generalized version of Theorem 2 that applies to the weighted GRNF.

Theorem 3. *Consider a distribution $\overline{P}$ over $\mathcal{W}$, with $\mathrm{supp}(\overline{P}) = \mathcal{W}$. If there exists a positive constant $\overline{C_{\mathcal{G}}}$ such that the fourth momentum*

$$\mathbb{E}_{\mathbf{w}\sim\overline{P}}\left[\frac{p(\mathbf{w})^2}{\overline{p}(\mathbf{w})^2}\psi(g;\mathbf{w})^4\right] < \overline{C_{\mathcal{G}}}$$

for any choice of $g \in \mathcal{G}$, then for any value $\varepsilon > 0$ and $\delta \in (0,1)$, when $M \geq \frac{16\overline{C_{\mathcal{G}}}}{\delta\,\varepsilon^2}$ we have

$$\mathbb{P}\left(\left||\overline{\mathbf{z}}(g_1) - \overline{\mathbf{z}}(g_2)|_2^2 - d_P(g_1,g_2)^2\right| \geq \varepsilon\right) \leq \delta.$$

Proof. Following the rationale of the proof of Theorem 2, we prove that $\mathbb{E}[|\overline{\mathbf{z}}(g_1) - \overline{\mathbf{z}}(g_2)|_2^2] = d_P(g_1,g_2)^2$ and that $\mathbb{E}[|\overline{\mathbf{z}}(g_1) - \overline{\mathbf{z}}(g_2)|_2^2]$ scales as $O(M^{-1})$. Finally, we apply the Chebyshev's inequality.

1) Denote with $\Delta(\mathbf{w}) = |\psi(g_1;\mathbf{w}) - \psi(g_2;\mathbf{w})|_2^2$.

$$\begin{aligned}
&\mathbb{E}_{\overline{P}^M}\left[|\overline{\mathbf{z}}(g_1) - \overline{\mathbf{z}}(g_2)|_2^2\right] = \\
&\quad= \sum_{m=1}^{M}\mathbb{E}_{\overline{P}}\left[\frac{1}{M}\frac{p(\mathbf{w}_m)}{\overline{p}(\mathbf{w}_m)}\Delta(\mathbf{w}_m)\right] \\
&\quad= \frac{1}{M}\sum_{m=1}^{M}\int_{\mathcal{W}}\frac{p(\mathbf{w})}{\overline{p}(\mathbf{w})}\,\Delta(\mathbf{w})\,\mathrm{d}P(\mathbf{w}) \\
&\quad= \int_{\mathcal{W}}\overline{p}(\mathbf{w})\,\frac{p(\mathbf{w})}{\overline{p}(\mathbf{w})}\,\Delta(\mathbf{w})\,\mathrm{d}\mathbf{w} \\
&\quad= \mathbb{E}_P\left[\Delta(\mathbf{w})\right] = d_P(g_1,g_2)^2
\end{aligned}$$

This holds thanks to the fact that $\overline{p}(\mathbf{w}) \neq 0$ for every $\mathbf{w} \in \mathcal{W}$, otherwise we would end up with a result similar to Lemma 3.

2) The variance can be bound in the same manner of (12), obtaining

$$\begin{aligned}
&\mathrm{Var}\left[|\overline{\mathbf{z}}(g_1) - \overline{\mathbf{z}}(g_2)|_2^2\right] \\
&\quad= \frac{1}{M}\mathrm{Var}\left[\frac{p(\mathbf{w})}{\overline{p}(\mathbf{w})}\left(\psi(g_1;\mathbf{w}) - \psi(g_2;\mathbf{w})\right)^2\right] \\
&\quad\leq \frac{16\,\overline{C_{\mathcal{G}}}}{M}
\end{aligned}$$

3) Chebyshev's inequality gives us the bound

$$\mathbb{P}_{\overline{P}^M}\left(\left||\overline{\mathbf{z}}(g_1) - \overline{\mathbf{z}}(g_2)|_2^2 - d_P(g_1,g_2)\right| \geq \varepsilon\right) \leq \frac{16\,\overline{C_{\mathcal{G}}}}{M\,\varepsilon^2}$$

from which the thesis follows.

□

D. Computational Complexity

Let us consider M random features with hidden tensor of order k. The computational complexity of (10) is given by:

- $\mathrm{Bell}(2+k)$ operations of the form $\mathbf{E}_\gamma\mathbf{T}$, each with cost $O(n^{2+k}\,F\,F_h)$;
- then, we perform M linear combinations $\sum_\gamma \theta_\gamma \mathbf{I}_\gamma T$ and addition of the bias term, each with cost $O((2\mathrm{Bell}(2+k)+1)n^k\,F_h)$;
- in order to compute $H_k(\mathbf{T})$ (2) for each of the M features, we perform $\mathrm{Bell}(k)$ operations of the form $\mathbf{I}_\gamma T$, which scale as $O(n^k \cdot F_h)$. Considering also the linear combination $\sum_\gamma \theta_\gamma \mathbf{I}_\gamma T$ and the bias term θ', we have $O(\mathrm{Bell}(k)\,(n^k\,F_h + 1) + 1)$.

The total computational complexity for creating a graph representation is:

$$\begin{aligned}
&O(\mathrm{Bell}(2+k)n^{2+k}\,F\,F_h) \\
&\qquad + O(M(2\mathrm{Bell}(2+k)+1)n^k\,F_h) \\
&\qquad\quad + O(M\mathrm{Bell}(k)\,(n^k\,F_h+1) + M).
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
&O\big(\mathrm{Bell}(2+k)\,n^{2+k}\,F\,F_h \\
&\qquad + M\,n^k\,F_h\,(\mathrm{Bell}(2+k) + \mathrm{Bell}(k))\big).
\end{aligned}$$

E. Implementation Details

GRNF implementation A PyTorch (Paszke et al., 2019) implementation of GRNF is available at the following link `https://github.com/dzambon/graph-random-neural-features` and adopts the efficient version for $k = 1, 2$ described in (Maron et al., 2019b). When not specified, $\rho_e(x) = \max\{0, x\}$ is the rectified linear unit, $\rho_i(x) = \tanh(x)$ is the hyperbolic tangent, $F_h = 4$ features in the hidden tensor, the probability of having order $k = 1$ and $k = 2$ in the hidden tensor is $2/3$ and $1/3$, respectively, and the weights θ_F, θ_H drawn from a standard Gaussian distribution. The provided implementation can run on ordinary laptops.

Replicability of the experiments The source code for running all the synthetic experiments is available at the GRNF repository. All the other experiments are performed with the framework provided by Errica et al. (2020) at the repository `https://github.com/diningphil/gnn-comparison`. All competitor models considered in our study are set up with the hyper-parameters suggested in (Errica et al., 2020).