GENERALIZATION BOUNDS FOR MESSAGE-PASSING GNNS ON DIRECTED GRAPHS WITH EDGE FEATURES

Anonymous authors
Paper under double-blind review

ABSTRACT

We establish generalization bounds for message-passing graph neural networks on directed graphs with multi-dimensional node and edge features, a setting that, to our knowledge, is not covered by existing generalization analyses. Messages are computed by Lipschitz functions of the triple (destination feature, edge feature, source feature), and each layer may include a global readout term enabling long-range interaction. We introduce a family of pseudo-metrics on graphs of fixed order n, defined by matching unrolling trees and vanishing on permutations by construction. Every network in our class is Lipschitz with respect to these metrics with an explicit constant, and bounding the covering number yields generalization bounds via the robustness framework of Xu & Mannor (2012). The bounds hold for any learning algorithm returning a hypothesis in the class and require no prior over model parameters. Using the data-dependent theorem of Kawaguchi et al. (2022), the dependence on the covering number C improves from $\sqrt{C}$ to $\sqrt{|\mathcal{T}_S| \ln C}$, where $|\mathcal{T}_S|$ counts only the covering classes occupied by the training sample. For the general architecture, the Lipschitz constant grows with n and with depth; for a simplified subclass without global aggregation or destination-dependent messages, it becomes independent of n.

1 INTRODUCTION

Graph-structured data arises naturally across various disciplines. In social networks, nodes represent individuals and edges encode relationships (Easley & Kleinberg, 2010); in financial systems, nodes represent financial institutions and directed edges represent interbank obligations whose magnitudes determine systemic risk (Eisenberg & Noe, 2001; Gonon et al., 2026a); in chemistry, nodes model atoms and edges encode chemical bonds with associated bond-type and distance features (Gilmer et al., 2017). The omnipresence of such data has made machine learning on graphs an active research area. Message-passing graph neural networks (MPNNs), introduced by Gori et al. (2005), Scarselli et al. (2009), Merkwirth & Lengauer (2005) and Gilmer et al. (2017), have emerged as the dominant paradigm for machine learning on graphs (Corso et al., 2024).

Despite their practical success, the theoretical foundations of MPNNs remain an active area of research with many open questions (Morris et al., 2024). Progress has been made on *expressivity*—the ability of a GNN to distinguish non-isomorphic graphs or to approximate functions on graphs—through the connection to the Weisfeiler–Leman (WL) isomorphism test (Weisfeiler & Leman, 1968; Xu et al., 2018; Morris et al., 2019). The *generalization* properties of GNNs, i.e. the ability of a trained model to make reliable predictions on unseen graphs drawn from the same distribution as the train data, are less understood. The generalization gap—the difference between empirical training loss and expected test loss—is fundamental to any application of GNNs, yet most existing theoretical results are restricted to simple architectures or yield bounds too loose to be informative in practice.

A major limitation of the existing generalization theory is the treatment of edge information. Most existing bounds apply to graphs with either no edge features or scalar adjacency weights, and typically assume undirected graphs. The importance of edge features and edge directionality in GNNs, however, is well documented. The MPNN framework of Gilmer et al. (2017) explicitly accommodates edge features in the message function and demonstrated their value for quantum chemistry. The Graph Network (GN) framework of Battaglia et al. (2018) likewise supports rich edge represen-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tations. For node classification on heterophilic graphs, edge directionality was shown to substantially improve performance (Rossi et al., 2024), and the edged WL test (Yue et al., 2024) formalizes the resulting expressivity gains from edge features. In scene-graph generation, directed edges encode the type and direction of relationships between objects (Herzig et al., 2018) and in financial networks, directed edges represent the precise bilateral debt obligations between institutions, which are essential inputs for computing systemic risk (Gonon et al., 2026a). Despite this motivation, no generalization theory exists for GNNs that support directed graphs with multi-dimensional edge features.

The GNN architectures we study directly address this gap. Abbreviating $h_v := h_v^{(l-1)}$, in each layer l these models feature a message-passing update of the form

$$h_k^{(l)} = \rho^{(l)}\left(h_k, \sum_{i\in\mathcal{N}(k)} \phi^{(l)}(h_k, e_{i,k}, h_i)\,, \sum_{j\in[n]} \alpha^{(l)}\left(h_j, \sum_{i\in\mathcal{N}(j)} \psi^{(l)}(h_j, e_{i,j}, h_i)\right)\right),$$

where node features $h_v^{(l)} \in \mathbb{R}^{d^l}$ and edge features $e_{u,v} \in \mathbb{R}^{d'}$ are both arbitrary dimensional and graphs are directed. This ACR (Aggregate–Combine–Readout) structure (Barceló et al., 2020) includes, compared to the standard MPNN of Gilmer et al. (2017), an additional global-aggregation term via readout function $\alpha^{(l)}$ and aggregation function $\psi^{(l)}$ that enables long-range interaction between nodes. When using feedforward neural networks as functions $\rho^{(l)}, \phi^{(l)}, \alpha^{(l)}, \psi^{(l)}$, this GNN type, called "extended PENN" (permutation-equivariant neural network), was shown to outperform simple message-passing GNNs on certain specialized tasks (Herzig et al., 2018; Gonon et al., 2026a). It is moreover universal, in the sense of approximating measurable permutation-equivariant or permutation-invariant functions on graphs when graphs are augmented with random node features (Gonon et al., 2026b). The present paper provides the complementary generalization theory: generalization bounds for this type of GNN operating on directed graphs with multi-dimensional node and edge features.

Related Work The earliest generalization results bound the Vapnik–Chervonenkis (VC) dimension. Scarselli et al. (2018) obtained bounds for a very specific class of GNNs, closer to recurrent neural networks, with shared weights between layers and based on Pfaffian functions. D'Inverno et al. (2025) extend this analysis to modern MPNNs with Pfaffian message-passing functions. Morris et al. (2023) establish a connection between the VC dimension of a specific sum-aggregation MPNN class and the 1-WL algorithm; however, the analysis is restricted to boolean node features and a simple MPNN architecture. Garg et al. (2020) derive generalization bounds for simple MPNNs via the Rademacher complexity, extending techniques for feedforward and recurrent neural networks, noting possible extensions to GNNs with message passing weighted by scalar edge features. Similarly, Lv (2021) derived bounds for GCNs with one hidden layer. Liao et al. (2020) establish PAC-Bayesian generalization bounds for multilayer GCNs and MPNNs, depending on the spectral norm of the weight matrices and the maximum node degree. Those bounds were subsequently improved by Sun & Lin (2024). However, PAC-Bayesian bounds require specifying a prior over model parameters.

The work most directly related to ours uses covering numbers and robustness theory. Maskey et al. (2022) lift MPNNs to graphons and derive generalization bounds in a graphon framework. Levie (2023) establishes a graphon-signal analysis of GNNs and proves Lipschitz continuity of MPNNs with respect to the cut distance, yielding generalization bounds via compactness. However, graphons are symmetric objects, so this framework is restricted to undirected graphs. Moreover, lifting GNNs to graphons requires order-normalized aggregation of the form $\frac{1}{n}\sum_{v\in\mathcal{N}(u)}(\cdot)$, which is incompatible with the widely used sum aggregation, and edge features cannot be naturally incorporated. Rauchwerger et al. (2024) employ the DIDM-mover's distance (based on iterated degree measures) for a joint expressivity, universality, and generalization analysis; their framework likewise applies to undirected attributed graphs with normalized aggregation and does not accommodate multi-dimensional edge features. Pseudo-metrics on unrolled computation trees were introduced by Chuang & Jegelka (2022), whose Tree Mover's Distance aligns subtrees by hierarchical optimal transport and yields stability and generalization guarantees for MPNNs. Building on that line, Vasileiou et al. (2024) introduce the *forest distance*—a tree-unrolling pseudo-metric for undirected, node-attributed MPNNs with sum or mean aggregation—and derive fine-grained generalization bounds via the Xu–Mannor robustness theorem (Xu & Mannor, 2012). Our pseudo-metrics

differ in how trees are compared: rather than transporting mass between subtrees, we minimize over edge-preserving bijections of the full unrolling tree, which is what allows edge directions and multi-dimensional edge features to be carried through the Lipschitz estimates. We also consider larger classes of GNNs.

2 PRELIMINARIES

Notation Let $\mathbb{N} = \{1, 2, \ldots\}$ denote the natural numbers and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. We write $[N] := \{1, \ldots, N\}$. Let $\|\cdot\|$ denote the Euclidean norm. For a tensor or vector containing only zeros we write $\mathbf{0}$. Unless specified otherwise the dimensions will be clear from context. Given two metric spaces $(\mathcal{X}, d_{\mathcal{X}}), (\mathcal{Y}, d_{\mathcal{Y}})$ a function $f : \mathcal{X} \to \mathcal{Y}$ is Lipschitz continuous if there exists a constant $K > 0$ such that $d_{\mathcal{Y}}(f(x), f(y)) \leq K d_{\mathcal{X}}(x, y)$ for all $x, y \in \mathcal{X}$. Finally, recall the following standard notion of the covering number (Xu & Mannor, 2012). For a metric space $(\mathcal{X}, d)$ and $U \subset \mathcal{X}$ we say $\hat{U} \subset \mathcal{X}$ is an ε-cover of U, if for all $u \in U$ there exists $\hat{u} \in \hat{U}$ such that $d(u, \hat{u}) \leq \varepsilon$. The ε-covering number of U is $\mathcal{C}(U, d, \varepsilon) = \min\{|\hat{U}| \mid \hat{U} \text{ is an } \varepsilon\text{-cover of } U\}$.

Domain of graphs Let $\mathcal{G}_n$ be the set of directed, featured graphs with $n \in \mathbb{N}$ nodes whose node and edge features are bounded by $M > 0$. For any graph $g \in \mathcal{G}_n$, we denote the node set by $V(g) \subset \mathbb{N}$ (without loss of generality, we usually take $V(g) = [n]$) and the edge set by $E(g) \subset V(g) \times V(g)$. Each node $v \in V(g)$ has a $d \in \mathbb{N}$-dimensional feature vector $h_v \in [-M, M]^d \subset \mathbb{R}^d$, and each edge (u, v) has a $d' \in \mathbb{N}$-dimensional feature vector $e_{u,v} \in [-M, M]^{d'} \subset \mathbb{R}^{d'}$. We define the neighborhood $\mathcal{N}(i)$ of a node $i \in [n]$ of graph g as $\{j \in [n] \mid (j, i) \in E(g)\}$. If several graphs are involved, we may indicate the graph by a subscript, writing $\mathcal{N}_g$ for clarity; when the unrolling tree $\mathcal{T}$ is involved as well (Section 3), we use $\mathcal{N}_g$ versus $\mathcal{N}_{\mathcal{T}}$ to specify whether a neighborhood is taken with respect to the graph structure or the (tree) structure.

Since we will only consider GNNs for which it is irrelevant whether two nodes are unconnected or connected by a "zero-edge", we may represent a graph $g \in \mathcal{G}_n$ directly by its node and edge features,

$$w^g := \left(h_1^g, \ldots, h_n^g,\ (e_{i,j}^g)_{i,j \in [n]}\right) \in [-M, M]^p, \qquad p := nd + n^2 d',$$

where $e_{i,j}^g$ is set to $\mathbf{0}$ if no edge from i to j exists in g; every point of $[-M, M]^p$ is w^g for some $g \in \mathcal{G}_n$, so $g \mapsto w^g$ identifies $\mathcal{G}_n$ with the raw feature cube $[-M, M]^p$.

Let S_n denote the symmetric group of all permutations of $[n]$. For $\sigma \in S_n$ and $g \in \mathcal{G}_n$, we define the permuted graph $\sigma(g) \in \mathcal{G}_n$ by permuting the node- and edge-feature blocks of w^g according to σ,

$$h_i^{\sigma(g)} := h_{\sigma^{-1}(i)}^g, \qquad e_{i,j}^{\sigma(g)} := e_{\sigma^{-1}(i),\, \sigma^{-1}(j)}^g, \qquad i, j \in [n].$$

Graph neural networks We consider graph neural networks of the following form. Let $g \in \mathcal{G}_n$ and set $h_v^{(0)} = h_v$ for all $v \in [n]$, with $d^{(0)} = d$. Let $L \in \mathbb{N}$ and $d^{(l)}, d_1^{(l)}, d_2^{(l)}, d_3^{(l)} \in \mathbb{N}, l = 1, \ldots, L$. For each $l \in [L]$, define the maps

$$\phi^{(l)} : \mathbb{R}^{2d^{(l-1)} + d'} \to \mathbb{R}^{d_1^{(l)}}, \qquad \psi^{(l)} : \mathbb{R}^{2d^{(l-1)} + d'} \to \mathbb{R}^{d_2^{(l)}},$$
$$\alpha^{(l)} : \mathbb{R}^{d^{(l-1)} + d_2^{(l)}} \to \mathbb{R}^{d_3^{(l)}}, \qquad \rho^{(l)} : \mathbb{R}^{d^{(l-1)} + d_1^{(l)} + d_3^{(l)}} \to \mathbb{R}^{d^{(l)}}.$$

Abbreviating $h_k := h_k^{(l-1)}$, in each message-passing step $l = 1, \ldots, L$, an L-layer graph neural network updates the representation of each node $k \in [n]$ by

$$h_k^{(l)} = \rho^{(l)}\left(h_k,\ \sum_{i \in \mathcal{N}(k)} \phi^{(l)}(h_k, e_{i,k}, h_i)\,,\ \sum_{j \in [n]} \alpha^{(l)}\left(h_j,\ \sum_{i \in \mathcal{N}(j)} \psi^{(l)}(h_j, e_{i,j}, h_i)\right)\right). \quad (1)$$

We compute a permutation-invariant graph representation h_g from the normalized sum of node representations via

$$h_g := \tau\left(\frac{1}{|V(g)|} \sum_{k \in V(g)} h_k^{(L)}\right), \quad (2)$$

where $\tau : \mathbb{R}^{d^{(L)}} \to \mathbb{R}^{d_{\text{out}}}$ is the readout function.

This architecture is a specific subclass of ACR-GNNs with sum aggregation. With suitable choices of the functions ρ, α, ϕ, ψ, and τ, it encompasses many commonly used GNN architectures. In particular, if ρ, α, ϕ, and ψ are chosen as feedforward neural networks, (1) coincides with the extended PENN architecture discussed in Gonon et al. (2026b), which becomes a universal approximator of permutation-equivariant functions when graphs are augmented with random node features. For our generalization analysis, we focus on the subset of L-layer GNNs, $L \in \mathbb{N}$, in which all components are Lipschitz continuous with a common constant $K > 0$, and the functions $\phi^{(l)}$ and $\psi^{(l)}$, $l \in [L]$, are required to return zero whenever the input corresponds to a "zero-edge."

$$\mathcal{H}_{L,K} := \left\{ f : g \mapsto h_g \;\middle|\; \begin{array}{r} h_g \text{ is computed by (1) and (2),} \\ \rho^{(l)}, \alpha^{(l)}, \phi^{(l)}, \psi^{(l)},\ l \in [L], \text{ and } \tau \text{ are } K\text{-Lipschitz,} \\ \phi^{(l)}(\cdot, \mathbf{0}, \cdot) = \psi^{(l)}(\cdot, \mathbf{0}, \cdot) = \mathbf{0} \text{ for all } l \in [L] \end{array} \right\} \tag{3}$$

Since every K-Lipschitz function is also $\max(K,1)$-Lipschitz and $\mathcal{H}_{L,K} \subseteq \mathcal{H}_{L,\max(K,1)}$, we assume throughout that $K \geq 1$. All bounds below hold for $K < 1$ as well, with K replaced by $\max(K,1)$. Requiring the output to be zero on zero-edges is a natural condition. For example, most edge-sensitive GNNs are defined only for scalar edge features (weighted graphs) and take the form $\phi(h_j, e_{i,j}, h_i) = \bar{\phi}(h_j, h_i)\, e_{i,j}$ for some function $\bar{\phi}$, such that consequently $\phi(\cdot, \mathbf{0}, \cdot) = \mathbf{0}$. With bounded edge features, a simple example of a Lipschitz function that vanishes on zero-edges is $\phi(h_j, e_{i,j}, h_i) = \|e_{i,j}\| \cdot \sigma\big(\bar{\phi}(h_j, e_{i,j}, h_i)\big)$, where $\bar{\phi}$ is any Lipschitz function and σ is a Lipschitz, bounded activation function.

Supervised learning Let $\mathcal{G}$ be a set of graphs, $\mathcal{Y}$ a set of labels, and $\mathcal{X}$ a set of embeddings. Let $\mathcal{H} \subseteq \mathcal{X}^{\mathcal{G}}$ be a hypothesis class, i.e., a set of mappings from $\mathcal{G}$ to $\mathcal{X}$. Define $\mathcal{Z} := \mathcal{G} \times \mathcal{Y}$ as the space of graph–label pairs, equipped with a probability distribution μ. We write $(G, Y) \sim \mu$ to indicate that (G, Y) is distributed on $\mathcal{Z}$ according to μ; the uppercase G denotes this random graph, as opposed to a deterministic graph $g \in \mathcal{G}$. Fix a sample size $k \in \mathbb{N}$ and let $\mathcal{S} \sim \mu^k$ denote a collection of k independent and identically distributed random variables on $\mathcal{Z}$. We call $\mathcal{S} = ((G^{(1)}, Y^{(1)}), \ldots, (G^{(k)}, Y^{(k)}))$ a *random sample* of size k, and any realization $s = ((g^{(1)}, y^{(1)}), \ldots, (g^{(k)}, y^{(k)})) \in \mathcal{Z}^k$ a sample of size k. A graph learning algorithm $\mathcal{A} : \mathcal{Z}^k \to \mathcal{H}$ assigns to each sample $s \in \mathcal{Z}^k$ a hypothesis $\mathcal{A}_s \in \mathcal{H}$. Let $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function, assumed to be bounded by some $M_\ell > 0$. The expected and empirical errors of $\mathcal{A}_s$ are defined as

$$\ell_{\exp}(\mathcal{A}_s) := \mathbb{E}_{(G,Y)\sim\mu}[\ell(\mathcal{A}_s(G), Y)], \qquad \ell_{\text{emp}}(\mathcal{A}_s) := \frac{1}{|s|} \sum_{(g,y)\in s} \ell(\mathcal{A}_s(g), y).$$

The quantity $|\ell_{\exp}(\mathcal{A}_s) - \ell_{\text{emp}}(\mathcal{A}_s)|$ is the *generalization gap*. It measures the difference between the empirical loss on the sample and the true expected loss on $\mathcal{Z}$. A small generalization gap indicates that the learned hypothesis generalizes well from the observed data to unseen graphs. When the sample is the random sample $\mathcal{S} = ((G^{(1)}, Y^{(1)}), \ldots, (G^{(k)}, Y^{(k)}))$, we write $\mathcal{A}_{\mathcal{S}}$ for the (random) hypothesis and $|\ell_{\exp}(\mathcal{A}_{\mathcal{S}}) - \ell_{\text{emp}}(\mathcal{A}_{\mathcal{S}})|$ for the corresponding generalization gap, which is a random variable that we aim to bound in a probabilistic sense.

3 Generalization Bounds for Permutation-Invariant GNNs

In this section, we introduce a family of pseudo-metrics for directed, featured graphs in $\mathcal{G}_n$, show Lipschitz continuity of GNNs with respect to these metrics, and derive upper bounds for the covering number of $\mathcal{G}_n$. Then we derive generalization bounds via the notion of robustness for single and multi-layer GNNs.

3.1 The Pseudo-Metrics

The pseudo-metrics we propose satisfy two key properties. (1) If for two graphs one is a permutation of the other, their distance should be zero. This is a very natural assumption. Since the nodes of a graph have no natural ordering, it should not make a difference whether we consider a graph or

one of its permutations. In fact, a trivial graph pseudo-metric is defined by $d(g^{(1)}, g^{(2)}) = 0$ if, for two graphs $g^{(1)}, g^{(2)}$, one is a permutation of the other, and $d(g^{(1)}, g^{(2)}) = \infty$ otherwise. This is a rather coarse metric, and we aim to establish a metric that is meaningful also for graphs that are not permutations of one another. Since we investigate graph neural networks, we would like to obtain a pseudo-metric for which the following holds. (2) If the distance between two graphs is small, then their representations computed by a GNN are close as well. Such a distance is important to obtain bounds on the generalization gap, as we will see later. To define such a metric, we first introduce the unrolling tree of a graph.

Definition 1. Let $g \in \mathcal{G}_n$ and $L \in \mathbb{N}_0$. The unrolling tree of g at level L, denoted by $\mathcal{T}_L(g)$, is a directed, featured graph whose nodes are arranged in levels. It is defined recursively as follows.

- On level 0, $\mathcal{T}_0(g)$ contains copies of the nodes $1, \dots, n$ of g, together with their node features $h_1, \dots, h_n$. We refer to these nodes as the *roots* of the unrolling tree.
- For $L \in \mathbb{N}$, the tree $\mathcal{T}_L(g)$ is constructed from $\mathcal{T}_{L-1}(g)$ as follows. For each node on level $L-1$ (a "parent"), we add n nodes on level L (its "children"), which are copies of the nodes $1, \dots, n$ of g with their corresponding features $h_1, \dots, h_n$. Each parent is connected to each of its children by a directed edge from the child to the parent. If parent and child correspond to nodes i and j of the original graph g, respectively, the edge is assigned the feature $e_{j,i}$ from g, or $\mathbf{0}$ if no edge from j to i exists in g.

We will define the distance between two graphs as the distance between their unrolling trees. To do so, we first introduce several additional ingredients.

Definition 2.

- The set of roots of an unrolling tree $\mathcal{T}$ is denoted by $\mathrm{roots}(\mathcal{T})$.
- The function $\mathrm{par} : V(\mathcal{T}) \to V(\mathcal{T}) \cup \{0\}$ returns the parent of a node, and 0 for roots.
- For a node $i \in V(\mathcal{T})$, the set of its children is denoted by $\mathrm{ch}(i)$, with $\mathrm{ch}(0) = [n]$.
- The map $\mathrm{orig} : V(\mathcal{T}) \to V(g)$ assigns to each node of the unrolling tree the corresponding node of the original graph.
- For two trees $\mathcal{T}^{(1)}$ and $\mathcal{T}^{(2)}$, a bijection $\chi : V(\mathcal{T}^{(1)}) \to V(\mathcal{T}^{(2)})$ is called *edge-preserving* if, for all $i, j \in V(\mathcal{T}^{(1)})$,

$$(i,j) \in E(\mathcal{T}^{(1)}) \iff (\chi(i), \chi(j)) \in E(\mathcal{T}^{(2)}).$$

Remark 3 (Graph features versus tree features). We use the same symbols h and e for the features of g and the features of $\mathcal{T}_L(g)$: for a tree node $i \in V(\mathcal{T}_L(g))$ we set $h_i := h_{\mathrm{orig}(i)}$, and for a tree edge (i,j) we set $e_{i,j} := e_{\mathrm{orig}(i),\mathrm{orig}(j)}$ (or $\mathbf{0}$, if there is no such edge in g). We keep this slight abuse of notation—rather than introducing separate tree-feature symbols—because in every formula that follows, the index alone (a graph node $i \in [n]$, or a tree node $i \in V(\mathcal{T}_L(g)) \setminus [n]$) already disambiguates which is meant, so new symbols would only lengthen formulas without adding information.

Based on unrolling trees, we define the family of pseudo-metrics between two graphs as follows.

Definition 4. Let $L \in \mathbb{N}_0$ and $g^{(1)}, g^{(2)} \in \mathcal{G}_n$. For $i, j \in [n]$, let $h_i^{g^{(1)}}, e_{i,j}^{g^{(1)}}$ and $h_i^{g^{(2)}}, e_{i,j}^{g^{(2)}}$ denote the node and edge features of $g^{(1)}$ and $g^{(2)}$, respectively. We define

$$d_L(g^{(1)}, g^{(2)}) := \min_{\substack{\chi : V(\mathcal{T}_L(g^{(1)})) \to V(\mathcal{T}_L(g^{(2)})) \\ \text{edge-preserving bijection}}} \sum_{\substack{i \in V(\mathcal{T}_L(g^{(1)})) \\ j = \mathrm{par}(i)}} \left\| (h_j^{g^{(1)}},\, e_{i,j}^{g^{(1)}},\, h_i^{g^{(1)}}) - (h_{\chi(j)}^{g^{(2)}},\, e_{\chi(i),\chi(j)}^{g^{(2)}},\, h_{\chi(i)}^{g^{(2)}}) \right\|, \tag{4}$$

where we set $\chi(0) = 0$, $h_0^{g^{(1)}} = h_0^{g^{(2)}} = \mathbf{0}$, and $e_{\cdot,0}^{g^{(1)}} = e_{\cdot,0}^{g^{(2)}} = \mathbf{0}$.

Note that d_L is defined, and Lemmas 5 and 7 hold verbatim, for directed featured graphs with arbitrary feature dimension and unbounded features; M is used only for the covering numbers. Further note that an edge-preserving bijection between two trees $\mathcal{T}_L(g^{(1)})$ and $\mathcal{T}_L(g^{(2)})$ has a very

rigid structure. The n roots of $\mathcal{T}_L(g^{(1)})$ can only be mapped to the n roots of $\mathcal{T}_L(g^{(2)})$. Likewise, the n children of any parent node $i \in V(\mathcal{T}_L(g^{(1)}))$ can only be mapped to the n children of $\chi(i)$. Consequently, for $L \in \mathbb{N}$, an edge-preserving map from $\mathcal{T}_L(g^{(1)})$ to $\mathcal{T}_L(g^{(2)})$ can be described by a collection of $|V(\mathcal{T}_{L-1}(g^{(1)}))| + 1$ permutations from S_n. For example, when $L = 1$, the map χ can be represented by $n + 1$ permutations $\sigma_0, \sigma_1, \ldots, \sigma_n \in S_n$:

$$d_1(g^{(1)}, g^{(2)}) = \min_{\substack{\chi: V(\mathcal{T}_1(g^{(1)})) \to V(\mathcal{T}_1(g^{(2)})) \\ \text{edge-preserving bijection}}} \sum_{\substack{i \in V(\mathcal{T}_1(g^{(1)})) \\ j = \mathrm{par}(i)}} \left\| (h_j^{g^{(1)}}, e_{i,j}^{g^{(1)}}, h_i^{g^{(1)}}) - (h_{\chi(j)}^{g^{(2)}}, e_{\chi(i),\chi(j)}^{g^{(2)}}, h_{\chi(i)}^{g^{(2)}}) \right\|$$

$$= \min_{\sigma_0, \sigma_1, \ldots, \sigma_n \in S_n} \sum_{j \in [n]} \left[\left\| h_j^{g^{(1)}} - h_{\sigma_0(j)}^{g^{(2)}} \right\| + \sum_{i \in [n]} \left\| (h_j^{g^{(1)}}, e_{i,j}^{g^{(1)}}, h_i^{g^{(1)}}) - (h_{\sigma_0(j)}^{g^{(2)}}, e_{\sigma_j(i),\sigma_0(j)}^{g^{(2)}}, h_{\sigma_j(i)}^{g^{(2)}}) \right\| \right].$$

We record that d_L is indeed a pseudo-metric, see Lemma 20 and, as mentioned above, the distance between a graph and any of its permutations is zero, see Lemma 21.

3.2 Lipschitz Continuity

By specifying the GNN class, we can derive explicit Lipschitz constants and hence bound the difference between node representations computed by GNNs with the graph distance. We start the analysis by focusing on one update step. In what follows we abbreviate $\tilde{K} := K_\rho(1 + K_\phi + K_\alpha + K_\alpha K_\psi)$, the constant produced by one update step.

Lemma 5. *Let $L \in \mathbb{N}$ and $g^{(1)}, g^{(2)} \in \mathcal{G}_n$. Let ρ, α, ϕ, ψ be update functions as in (1) of some $f \in \mathcal{H}_{1,K}$ and $K_\rho, K_\alpha, K_\phi, K_\psi \in (0, K]$ be their Lipschitz constants, respectively. Let $k \in V(\mathcal{T}_{L-1}(g^{(1)}))$ and $\bar{k} = \chi(k) \in V(\mathcal{T}_{L-1}(g^{(2)}))$, where χ is a distance-minimizing edge-preserving map for $d_L(g^{(1)}, g^{(2)})$. Let $h_k^{g^{(1)*}}$ and $h_{\bar{k}}^{g^{(2)*}}$ be the representations obtained by one update step (1) applied at k and at $\bar{k}$ respectively. Then we have*

$$\|h_k^{g^{(1)*}} - h_{\bar{k}}^{g^{(2)*}}\| \leq (K_\rho + K_\rho K_\phi + K_\rho K_\alpha + K_\rho K_\alpha K_\psi)\, d_L(g^{(1)}, g^{(2)}) = \tilde{K} d_L(g^{(1)}, g^{(2)}),$$

and

$$\sum_{k \in V(\mathcal{T}_{L-1}(g^{(1)}))} \|h_k^{g^{(1)*}} - h_{\bar{k}}^{g^{(2)*}}\| \leq n\tilde{K} d_L(g^{(1)}, g^{(2)}).$$

The previous lemma allows us to bound the difference of node representations computed in the update step of a 1-layer GNN from $\mathcal{H}_{1,K}$. We can use this result and bound the difference between the graph representations produced by a readout function applied to those node representations.

Proposition 6. *Let $g^{(1)}, g^{(2)} \in \mathcal{G}_n$. For $K \geq 1$, let $K \geq K_\tau > 0$ and τ be a K_τ-Lipschitz readout function as in (2) for some $f \in \mathcal{H}_{1,K}$. With $h^{g^{(1)*}} = f(g^{(1)})$ and $h^{g^{(2)*}} = f(g^{(2)})$, we obtain*

$$\|h^{g^{(1)*}} - h^{g^{(2)*}}\| \leq K_\tau\, \tilde{K}\, d_1(g^{(1)}, g^{(2)}) \leq 4K^4 d_1(g^{(1)}, g^{(2)}).$$

Furthermore, we can establish the following inequality that relates the d_L distance of two graphs to the d_{L-1} distance of the graphs after their nodes have been updated by the update step of a GNN.

Lemma 7. *Let $g^{(1)}, g^{(2)} \in \mathcal{G}_n$ and let $g^{(1)*}, g^{(2)*}$ be the directed featured graphs on n nodes obtained after updating the node features of $g^{(1)}$ and $g^{(2)}$ by one update step of some $f \in \mathcal{H}_{1,K}$. Then, for any $L \in \mathbb{N}$,*

$$d_{L-1}(g^{(1)*}, g^{(2)*}) \leq 12\, n^2 K^3\, d_L(g^{(1)}, g^{(2)}).$$

Since we can bound the effect of one update step, we can inductively bound the effect of multiple update steps. Doing so we obtain a multi-layer version of Proposition 6 that bounds the difference of graph representations obtained from a multi-layer GNN by the d_L distance of the input graphs.

Proposition 8. *Let $g^{(1)}, g^{(2)} \in \mathcal{G}_n$, and let $f \in \mathcal{H}_{L,K}$ for some $L \in \mathbb{N}$ and $K \geq 1$. Then for $c(K, L, n) = 4^L 3^{L-1} n^{2(L-1)} K^{3L+1}$,*

$$\|f(g^{(1)}) - f(g^{(2)})\| \leq c \cdot d_L(g^{(1)}, g^{(2)}).$$

3.3 Covering Number

We can relate the graph pseudo-metric d_L to the Euclidean distance on the raw feature vectors $w^g \in [-M, M]^p$, $p = nd + n^2d'$, introduced in the preliminaries. Since every point of the feature cube is the feature vector of an actual graph, this relation lets us transfer covering numbers of a Euclidean hypercube to $(\mathcal{G}_n, d_L)$.

Proposition 9. *Let $g^{(1)}, g^{(2)} \in \mathcal{G}_n$ and $L \in \mathbb{N}$. Set*

$$T_L := \begin{cases} 2, & L = 1, \\ 2 + |V(\mathcal{T}_{L-2}(g^{(1)}))|, & L \geq 2, \end{cases}$$

which is independent of the particular $g^{(1)} \in \mathcal{G}_n$. Then

$$d_L(g^{(1)}, g^{(2)}) \;\leq\; \sqrt{2}\, T_L\, n^{3/2}\, \|w^{g^{(1)}} - w^{g^{(2)}}\|_2.$$

We can therefore construct a covering of $(\mathcal{G}_n, d_L)$ directly from a covering of the feature cube $[-M, M]^p$.

Lemma 10. *Let $L \in \mathbb{N}$, $\varepsilon > 0$, and T_L, $p = nd + n^2d'$ as in Proposition 9. Then*

$$\mathcal{C}(\mathcal{G}_n, d_L, \varepsilon) \leq \mathcal{C}\Big([-M, M]^p,\ \|\cdot\|_2,\ \tfrac{\varepsilon}{\sqrt{2}\, T_L\, n^{3/2}}\Big).$$

Finally, we bound the covering number of a bounded Euclidean hypercube.

Lemma 11. *Let $I := [-M, M]$ and $p \in \mathbb{N}$. For $\varepsilon > 0$,*

$$\mathcal{C}(I^p, \|\cdot\|_2, \varepsilon) \leq (M\sqrt{p}/\varepsilon + 1)^p.$$

3.4 Generalization Bounds for Single-Layer GNNs

The results established so far, together with the notion of *robustness* introduced by Xu & Mannor (2012) and presented in Appendix B, allow us to derive generalization bounds for 1-layer GNNs. Recall that for any $n \in \mathbb{N}$, the set $\mathcal{G}_n$ consists of directed, featured graphs with n nodes, node- and edge-feature dimensions $d, d' \in \mathbb{N}$, and feature bound $M > 0$.

Theorem 12. *Let d_1 denote the pseudo-metric on $\mathcal{G}_n$ from Definition 4. Let $\mathcal{X} = \mathbb{R}$ (so the readout dimension is $d_{\text{out}} = 1$), $\mathcal{Y} \subseteq [0, 1]$ and $d_{\mathcal{X}} = d_{\mathcal{Y}} = |\cdot - \cdot|$. For some $M_\ell > 0$, let $x \wedge y := \min(x, y)$ and the loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be defined by*

$$\ell(x, y) := |x - y| \wedge M_\ell.$$

Let $K_{\mathcal{H}} \geq 1$ and let $\mathcal{H}_{1,K_{\mathcal{H}}} \subset \mathcal{X}^{\mathcal{G}_n}$ be the set of 1-layer GNNs whose components are $K_{\mathcal{H}}$-Lipschitz functions defined in (3).

Let $k \in \mathbb{N}$, $\mathcal{A} : \mathcal{Z}^k \to \mathcal{H}_{1,K_{\mathcal{H}}}$ be any learning algorithm, and $\mathcal{S}$ be a random sample of size k of $\mathcal{Z}$. Then, for any $\delta, \varepsilon > 0$, with probability at least $1 - \delta$

$$|\ell_{\exp}(\mathcal{A}_{\mathcal{S}}) - \ell_{\text{emp}}(\mathcal{A}_{\mathcal{S}})| \leq 2\varepsilon + M_\ell \sqrt{\frac{2C\ln(2) + 2\ln(1/\delta)}{k}},$$

where

$$C = \mathcal{C}(\mathcal{G}_n, d_1, \varepsilon/(8K_{\mathcal{H}}^4)) \cdot \mathcal{C}(\mathcal{Y}, |\cdot|, \varepsilon/2).$$

By bounding the covering number of $(\mathcal{G}_n, d_1)$, Theorem 12 yields explicit upper bounds on the generalization gap.

Corollary 13. *In the same setting as in Theorem 12, let $p := nd + n^2d'$. Then for any $\delta, \varepsilon > 0$, with probability at least $1 - \delta$,*

$$|\ell_{\exp}(\mathcal{A}_{\mathcal{S}}) - \ell_{\text{emp}}(\mathcal{A}_{\mathcal{S}})| \leq 2\varepsilon + M_\ell \sqrt{\frac{2\lceil 1/\varepsilon \rceil \left((16\sqrt{2}\, M\, n^{3/2} \sqrt{p}\, K_{\mathcal{H}}^4)/\varepsilon + 1\right)^p \ln(2) + 2\ln(1/\delta)}{k}}.$$

Next, we present an alternative version of this bound that improves the exponential dependence on n, d and d'.

3.5 Improved Generalization Bounds for Single-Layer GNNs

The dependence on C in the bound of Theorem 12 is of order $\sqrt{C}$, which can be problematic when the covering number is very large or scales poorly. In our setting, for instance, the term C, which scales like $(c/\varepsilon)^{nd+n^2d'}$ for a constant c (see Corollary 13), becomes enormous even for small graph sizes. It is therefore desirable to avoid such exponential scaling. The following proposition by Kawaguchi et al. (2022) shows that the dependence on C can be reduced to order $\ln(C)$ by bounding the generalization gap only on the relevant side and incorporating certain data-dependent refinements.

Proposition 14 (Theorem 1, Kawaguchi et al. (2022))**.** *Let $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function that is upper bounded by some $M_\ell > 0$. If the learning algorithm $\mathcal{A} : \mathcal{Z}^k \to \mathcal{H}$ is $(C, \epsilon(\cdot))$-robust for $C \in \mathbb{N}$ (with partition $\{A_i\}_{i=1}^C$) and $\epsilon : \mathcal{Z}^k \to \mathbb{R}_+$ and $\mathcal{S}$ is a random sample of size $k \in \mathbb{N}$ of $\mathcal{Z}$, then for any $\delta > 0$, with probability at least $1 - \delta$*

$$\ell_{\exp}(\mathcal{A}_\mathcal{S}) \le \ell_{\mathrm{emp}}(\mathcal{A}_\mathcal{S}) \;+\; \epsilon(\mathcal{S}) \;+\; \zeta(\mathcal{A}_\mathcal{S}) \left((\sqrt{2}+1) \sqrt{\frac{|\mathcal{T}_\mathcal{S}| \ln(2C/\delta)}{k}} \;+\; \frac{2|\mathcal{T}_\mathcal{S}| \ln(2C/\delta)}{k} \right),$$

where

$$\mathcal{I}_i^\mathcal{S} := \{\, j \in [k] : Z^{(j)} \in A_i \,\}, \quad \zeta(\mathcal{A}_\mathcal{S}) := \max_{(g,y)\in\mathcal{Z}} \ell(\mathcal{A}_\mathcal{S}(g), y), \quad \mathcal{T}_\mathcal{S} := \{\, i \in [C] : |\mathcal{I}_i^\mathcal{S}| \ge 1 \,\}.$$

Here $|\mathcal{T}_\mathcal{S}|$ is the number of classes that actually appear in the sample $\mathcal{S}$, so it is always at most k. It can be far smaller when numerous classes have very small probabilities and are only sparsely populated. Kawaguchi et al. (2022) provide a theoretical analysis and empirical evidence that this situation is common. In particular, under assumptions on the class-probabilities $\{p_i\}_{i=1}^C$, the quantity $|\mathcal{T}_\mathcal{S}|$ does not grow with C but increases only on the order of $\ln(k)$; see Proposition 3 in Kawaguchi et al. (2022). These assumptions are only realistic when the feature distribution is discrete or supported only on a low-dimensional subset of the domain. In this case

$$|\mathcal{T}_\mathcal{S}| \ll C \quad \text{and} \quad |\mathcal{T}_\mathcal{S}| \ll k.$$

For a distribution with a density on $[-M, M]^p$, where in the regime $k \ll C$ almost every sample point occupies a covering class of its own, $|\mathcal{T}_\mathcal{S}| \approx k$. The refinement is therefore informative precisely when the data are concentrated, rather than in complete generality.

Additionally, compared to Proposition 23, M_ℓ is replaced by $\zeta(\mathcal{A}_\mathcal{S})$. This is another refinement because M_ℓ, which bounds the worst-case maximum loss over the entire hypothesis class $M_\ell \ge \max_{(g,y)\in\mathcal{Z}} \max_{h\in\mathcal{H}} \ell(h(g), y)$, is replaced by the maximal loss that is actually attained on $\mathcal{Z}$ by the hypothesis returned by the learning algorithm.

With these refinements, Proposition 14 yields a potentially stronger version of Theorem 12.

Theorem 15. *In the same setting as in Theorem 12, for any $\delta, \varepsilon > 0$, with probability at least $1 - \delta$*

$$\ell_{\exp}(\mathcal{A}_\mathcal{S}) \le \ell_{\mathrm{emp}}(\mathcal{A}_\mathcal{S}) \;+\; 2\varepsilon \;+\; \zeta(\mathcal{A}_\mathcal{S}) \left((\sqrt{2}+1) \sqrt{\frac{|\mathcal{T}_\mathcal{S}| \ln(2C/\delta)}{k}} \;+\; \frac{2|\mathcal{T}_\mathcal{S}| \ln(2C/\delta)}{k} \right),$$

where

$$C = \mathcal{C}(\mathcal{G}_n, d_1, \varepsilon/(8K_\mathcal{H}^4)) \cdot \mathcal{C}(\mathcal{Y}, |\cdot|, \varepsilon/2).$$

3.6 Multi-Layer GNNs

The multi-layer case can be treated similarly to the single-layer setting by using the Lipschitz constant from Proposition 8.

Theorem 16. *Let d_L denote the pseudo-metric on $\mathcal{G}_n$ from Definition 4. Let $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} \subseteq [0, 1]$ and $d_\mathcal{X} = d_\mathcal{Y} = |\cdot|$. For some $M_\ell > 0$, let the loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be defined by*

$$\ell(x, y) := |x - y| \wedge M_\ell.$$

Let $K_\mathcal{H} \ge 1$ and let $\mathcal{H}_{L,K_\mathcal{H}} \subset \mathcal{X}^{\mathcal{G}_n}$ be the set of L-layer GNNs whose components are $K_\mathcal{H}$-Lipschitz functions defined in (3).

Let $k \in \mathbb{N}$, $\mathcal{A} : \mathcal{Z}^k \to \mathcal{H}_{L,K_\mathcal{H}}$ be any learning algorithm, and $\mathcal{S}$ be a random sample of size k of $\mathcal{Z}$. Then, for any $\delta, \varepsilon > 0$, with probability at least $1 - \delta$,

$$|\ell_{\exp}(\mathcal{A}_\mathcal{S}) - \ell_{\mathrm{emp}}(\mathcal{A}_\mathcal{S})| \leq 2\varepsilon + M_\ell \sqrt{\frac{2C\ln(2) + 2\ln(1/\delta)}{k}},$$

where

$$C = \mathcal{C}\left(\mathcal{G}_n,\, d_L,\, \frac{\varepsilon}{2 \cdot 4^L 3^{L-1} K_\mathcal{H}^{3L+1} n^{2(L-1)}}\right) \cdot \mathcal{C}(\mathcal{Y}, |\cdot|, \varepsilon/2).$$

Theorem 17. *In the same setting as in Theorem 16, for any $\delta, \varepsilon > 0$, with probability at least $1 - \delta$*

$$\ell_{\exp}(\mathcal{A}_\mathcal{S}) \leq \ell_{\mathrm{emp}}(\mathcal{A}_\mathcal{S}) \;+\; 2\varepsilon \;+\; \zeta(\mathcal{A}_\mathcal{S}) \left((\sqrt{2}+1) \sqrt{\frac{|\mathcal{T}_\mathcal{S}| \ln(2C/\delta)}{k}} \;+\; \frac{2|\mathcal{T}_\mathcal{S}| \ln(2C/\delta)}{k} \right),$$

where

$$C = \mathcal{C}\left(\mathcal{G}_n,\, d_L,\, \frac{\varepsilon}{2 \cdot 4^L 3^{L-1} K_\mathcal{H}^{3L+1} n^{2(L-1)}}\right) \cdot \mathcal{C}\left(\mathcal{Y}, |\cdot|, \frac{\varepsilon}{2}\right).$$

To close this section, we discuss how depth and graph size propagate into the bounds. Using that T_L is of order n^{L-1} and $\sqrt{p}$ of order n, the calculations presented in Remark 29 yield for $n \geq 2, M \geq \varepsilon > 0$,

$$\ln C \leq p \left[\left(3L - \tfrac{1}{2}\right) \ln n + (3L+1) \ln K_\mathcal{H} + L \ln 12 + \ln \tfrac{1}{\varepsilon} + \ln\left(4M\sqrt{d+d'}\right) \right] + \ln \left\lceil \tfrac{1}{\varepsilon} \right\rceil . \quad (5)$$

Two distinct effects are visible. The depth enters $\ln C$ only linearly, so deep networks are not a dominant obstruction. The graph size n on the other hand enters through the exponent $p = nd + n^2 d'$, making $\ln C$ of order $L \cdot n^2 \ln n$. This is what separates the two bounds. In Theorem 16 the quantity C/k must be small, which demands a sample size exponential in p and is out of reach already for moderate n. In the data-dependent bound of Theorem 17 only $|\mathcal{T}_\mathcal{S}| \ln(2C/\delta)/k$ must be small, which grows linearly in the number of raw features p and in the depth. The data-dependent refinement thus converts a requirement exponential in p into one linear in p; whether it is usable in practice rests on $|\mathcal{T}_\mathcal{S}|$, discussed in Section 3.5. The dependence on n in the Lipschitz constant itself can be removed by simplifying the architecture, as presented in Appendix D.

4 Conclusion

We have developed a generalization theory for GNNs operating on directed graphs with multi-dimensional node and edge features, addressing a setting previously not covered in the literature.

For the general class $\mathcal{H}_{L,K}$ of ACR-GNNs, we introduced a family of graph pseudo-metrics d_L that respect permutation symmetry by construction, established explicit Lipschitz constants for GNNs in $\mathcal{H}_{L,K}$ with respect to d_L for both the single- and multi-layer case, and bounded the covering number of $(\mathcal{G}_n, d_L)$ by that of a bounded Euclidean feature cube. Combining these with the robustness framework of Xu & Mannor (2012) yields generalization bounds that hold for any learning algorithm returning a hypothesis in $\mathcal{H}_{L,K}$, without requiring a prior over the model parameters. A further improvement is achieved through the data-dependent theorem of Kawaguchi et al. (2022), which replaces the $\sqrt{C}$ factor of the Xu–Mannor bound by $\sqrt{|T_\mathcal{S}| \ln C}$, where $|\mathcal{T}_\mathcal{S}|$ counts only the covering classes actually occupied by the training sample. Since C enters only logarithmically, the resulting bound remains meaningful even when C is large, as it typically is for our graph covering numbers. Whether it is informative then depends on $|\mathcal{T}_\mathcal{S}|$, which grows as $O(\ln k)$ under the distributional assumptions of Kawaguchi et al. (2022) but is of order k for feature distributions with a density. Separately, for the simplified class $\mathcal{H}^*_{L,K}$, which drops the global aggregation and lets messages depend only on the edge and source-node features, the same arguments yield a Lipschitz constant independent of the graph size n, though the covering number retains its dependence on n (Appendix D).

Interesting directions for further research include tighter bounds on the covering number, e.g. via WL equivalence class counts for a discrete feature space, an extension from fixed to bounded order graphs, for instance by restricting $\rho(\mathbf{0}, \cdot, \cdot) = \alpha(\mathbf{0}, \cdot) = \mathbf{0}$ and using an unnormalized sum in (2), and an empirical study of $|\mathcal{T}_\mathcal{S}|$ on relevant graph datasets.

AI USE STATEMENT

In this work, we used generative AI tools to assist with polishing the manuscript, in the writing of proofs, to provide feedback on our research methodology, and to interpret results. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, to propose or refine hypotheses, to formulate mathematical claims, or to provide critical ingredients for proving mathematical claims. Generating synthetic data sets, implementing methods, assisting with translation, cleaning and reformatting datasets, and supporting qualitative and thematic data analysis are not applicable to this work. We have reviewed all AI-assisted work. Large language models were instructed to find typos and check the proofs in the appendix and proposed corrections to several steps, including mis-indexed summations and an incorrectly stated multiset identity in the proofs of Lemma 5, Lemma 7, Lemma 18, and Corollary 31; each proposed correction was checked by hand before adoption, and several were rejected. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

This is a theoretical paper and involves no datasets, code or experiments. Every result stated in the main text is restated in full and proved in Appendix C. The standing conventions under which the constants are derived are collected in Remark 27, and the auxiliary facts about unrolling trees and the robustness framework on which the proofs rely are stated and proved in Appendix A and Appendix B.

REFERENCES

Pablo Barceló, Egor V Kostylev, Mikael Monet, Jorge Pérez, Juan Reutter, and Juan-Pablo Silva. The logical expressiveness of graph neural networks. In *8th International Conference on Learning Representations (ICLR 2020)*, 2020.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

Ching-Yao Chuang and Stefanie Jegelka. Tree mover's distance: Bridging graph metrics and stability of graph neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2944–2957, 2022.

Gabriele Corso, Hannes Stark, Stefanie Jegelka, Tommi Jaakkola, and Regina Barzilay. Graph neural networks. *Nature Reviews Methods Primers*, 4(1):17, 2024.

Giuseppe Alessio D'Inverno, Monica Bianchini, and Franco Scarselli. VC dimension of Graph Neural Networks with Pfaffian activation functions. *Neural Networks*, 182:106924, 2025.

David Easley and Jon Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*, volume 1. Cambridge University Press, 2010.

Larry Eisenberg and Thomas H. Noe. Systemic risk in financial systems. *Management Science*, 47 (2):236–249, 2001. ISSN 00251909, 15265501. URL http://www.jstor.org/stable/2661572.

Vikas Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks. In *International conference on machine learning*, pp. 3419–3430. PMLR, 2020.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

Lukas Gonon, Thilo Meyer-Brandis, and Niklas Weber. Computing systemic risk measures with graph neural networks. *SIAM Journal on Financial Mathematics*, 17(2):565–619, 2026a. doi: 10.1137/24M1697402. URL https://doi.org/10.1137/24M1697402.

Lukas Gonon, Thilo Meyer-Brandis, and Niklas Weber. Universal graph neural networks. *arXiv preprint arXiv:2607.26699*, 2026b.

Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. In *Proceedings. IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pp. 729–734. IEEE, 2005.

Roei Herzig, Moshiko Raboh, Gal Chechik, Jonathan Berant, and Amir Globerson. Mapping images to scene graphs with permutation-invariant structured prediction. *Advances in Neural Information Processing Systems*, 31, 2018.

Kenji Kawaguchi, Zhun Deng, Kyle Luh, and Jiaoyang Huang. Robustness implies generalization via data-dependent generalization bounds. In *International conference on machine learning*, pp. 10866–10894. PMLR, 2022.

Ron Levie. A graphon-signal analysis of graph neural networks. *Advances in Neural Information Processing Systems*, 36:64482–64525, 2023.

Renjie Liao, Raquel Urtasun, and Richard Zemel. A PAC-Bayesian approach to generalization bounds for graph neural networks. *arXiv preprint arXiv:2012.07690*, 2020.

Shaogao Lv. Generalization bounds for graph convolutional neural networks via Rademacher complexity. *arXiv preprint arXiv:2102.10234*, 2021.

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

Christian Merkwirth and Thomas Lengauer. Automatic generation of complementary descriptors with molecular graph networks. *Journal of chemical information and modeling*, 45(5):1159–1168, 2005.

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

Christopher Morris, Floris Geerts, Jan Tönshoff, and Martin Grohe. WL meet VC. In *International conference on machine learning*, pp. 25275–25302. PMLR, 2023.

Christopher Morris, Nadav Dym, Haggai Maron, Ismail Ilkan Ceylan, Fabrizio Frasca, Ron Levie, Derek Lim, Michael M Bronstein, Martin Grohe, and Stefanie Jegelka. Future directions in foundations of graph machine learning. *CoRR*, 2024.

Levi Rauchwerger, Stefanie Jegelka, and Ron Levie. Generalization, expressivity, and universality of graph neural networks on attributed graphs. *arXiv preprint arXiv:2411.05464*, 2024.

Emanuele Rossi, Bertrand Charpentier, Francesco Di Giovanni, Fabrizio Frasca, Stephan Günnemann, and Michael M. Bronstein. Edge directionality improves learning on heterophilic graphs. In Soledad Villar and Benjamin Chamberlain (eds.), *Proceedings of the Second Learning on Graphs Conference*, volume 231 of *Proceedings of Machine Learning Research*, pp. 25:1–25:27. PMLR, 27–30 Nov 2024. URL https://proceedings.mlr.press/v231/rossi24a.html.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. Computational capabilities of graph neural networks. *IEEE Transactions on Neural Networks*, 20 (1):81–102, 2009. doi: 10.1109/TNN.2008.2005141.

Franco Scarselli, Ah Chung Tsoi, and Markus Hagenbuchner. The Vapnik–Chervonenkis dimension of graph and recursive neural networks. *Neural Networks*, 108:248–259, 2018.

Tan Sun and Junhong Lin. PAC-Bayesian adversarially robust generalization bounds for graph neural network. *arXiv preprint arXiv:2402.04038*, 2024.

Antonis Vasileiou, Ben Finkelshtein, Floris Geerts, Ron Levie, and Christopher Morris. Covered forest: Fine-grained generalization analysis of graph neural networks. *arXiv preprint arXiv:2412.07106*, 2024.

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

Huan Xu and Shie Mannor. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Xiao Yue, Bo Liu, Feng Zhang, and Guangzhi Qu. Edged Weisfeiler-Lehman algorithm. In *International Conference on Artificial Neural Networks*, pp. 93–109. Springer, 2024.

A FUNDAMENTALS OF THE PSEUDO-METRICS

This appendix contains additional results surrounding the pseudo-metrics, some of which will be referenced throughout.

Lemma 18 (Children equal neighbors under the zero-edge convention)**.** *Let $g \in \mathcal{G}_n$, $L \in \mathbb{N}$, and let $k' \in V(\mathcal{T}_L(g))$ be a node not on level L, with $\mathrm{orig}(k') = k \in [n]$. For any function ψ with $\psi(\cdot, \mathbf{0}, \cdot) = \mathbf{0}$,*

$$\sum_{i \in \mathcal{N}_g(k)} \psi(h_k, e_{i,k}, h_i) = \sum_{i \in \mathrm{ch}(k')} \psi(h_{k'}, e_{i,k'}, h_i).$$

Consequently, updating the representation of $k \in V(g)$ by one step of a GNN from $\mathcal{H}_{1,\cdot}$ (via (1), with $\mathcal{N}(k)$ read as the graph neighborhood $\mathcal{N}_g(k)$) is equivalent to updating k' within the unrolling tree $\mathcal{T}_L(g)$ (via the same formula (1), with the local aggregation's $\mathcal{N}(k')$ read as the tree-children $\mathrm{ch}(k')$):

$$\begin{aligned}&\rho\left(h^g_k, \sum_{i \in \mathcal{N}_g(k)} \phi(h^g_k, e^g_{i,k}, h^g_i), \sum_{j \in [n]} \alpha\left(h^g_j, \sum_{i \in \mathcal{N}_g(j)} \psi(h^g_j, e^g_{i,j}, h^g_i)\right)\right)\\ =&\rho\left(h^g_{k'}, \sum_{i \in \mathrm{ch}(k')} \phi(h^g_{k'}, e^g_{i,k'}, h^g_i), \sum_{j \in [n]} \alpha\left(h^g_j, \sum_{i \in \mathcal{N}_g(j)} \psi(h^g_j, e^g_{i,j}, h^g_i)\right)\right).\end{aligned}$$

Moreover, the global term's summation may equivalently be taken over $[n] = V(g)$ (i.e. always the roots) or localized *to $\mathrm{ch}(\mathrm{par}(k'))$, the copies of the roots that are children of $\mathrm{par}(k')$:*

$$\sum_{j \in [n]} \alpha\left(h^g_j, \sum_{i \in \mathcal{N}_g(j)} \psi(h^g_j, e^g_{i,j}, h^g_i)\right) = \sum_{j \in \mathrm{ch}(\mathrm{par}(k'))} \alpha\left(h^g_j, \sum_{i \in \mathrm{ch}(j)} \psi(h^g_j, e^g_{i,j}, h^g_i)\right).$$

Proof. Since g has n nodes, $\mathrm{ch}(k')$ consists of one copy of every original node $1, \ldots, n$, and orig restricts to a bijection $\mathrm{ch}(k') \to [n]$ under which the edge feature $e_{i,k'}$ of a child i equals $e_{\mathrm{orig}(i),k}$ of g (or $\mathbf{0}$, if that graph edge is absent). Hence, writing $u := \mathrm{orig}(i)$,

$$\begin{aligned}\sum_{i \in \mathrm{ch}(k')} \psi(h_{k'}, e_{i,k'}, h_i) &= \sum_{u \in [n]} \psi(h_k, e_{u,k}, h_u)\\ &= \sum_{u \in \mathcal{N}_g(k)} \psi(h_k, e_{u,k}, h_u) + \sum_{u \notin \mathcal{N}_g(k)} \underbrace{\psi(h_k, \mathbf{0}, h_u)}_{=\mathbf{0}}\\ &= \sum_{u \in \mathcal{N}_g(k)} \psi(h_k, e_{u,k}, h_u).\end{aligned}$$

Applying this identity with ϕ in place of ψ gives the local-term equality; applying it with ψ inside every summand of the global term (using that $\mathcal{N}_g(j)$ does not depend on which tree-copy of j is used) gives the remaining equality. For the localization claim, each $p \in V(\mathcal{T}_{L-2}(g)) \cup \{0\}$ has exactly n children, and orig restricts to a bijection $\mathrm{ch}(p) \to [n]$; taking $p = \mathrm{par}(k')$ shows that summing the (already node-indexed, tree-copy-independent) summand over $\mathrm{ch}(p)$ or over $[n]$ gives the same value. □

The following elementary observation about edge-preserving bijections will be used repeatedly.

Lemma 19. *Let $L \in \mathbb{N}$ and $g^{(1)}, g^{(2)} \in \mathcal{G}_n$. Every edge-preserving bijection $\chi : V(\mathcal{T}_L(g^{(1)})) \to V(\mathcal{T}_L(g^{(2)}))$ maps level-l nodes to level-l nodes for all $0 \le l \le L$, satisfies $\mathrm{par}(\chi(i)) = \chi(\mathrm{par}(i))$, and restricts to an edge-preserving bijection $V(\mathcal{T}_l(g^{(1)})) \to V(\mathcal{T}_l(g^{(2)}))$ for every $l \le L$. In particular it maps the n roots of $\mathcal{T}_L(g^{(1)})$ onto the n roots of $\mathcal{T}_L(g^{(2)})$, and for every internal node p it maps the children of p bijectively onto the children of $\chi(p)$.*

Proof. In $\mathcal{T}_L$ every non-root node has exactly one out-edge (to its parent) and roots have none, so the roots are exactly the nodes of out-degree 0; since χ preserves edges in both directions, it maps roots

to roots. A node lies on level l iff its distance to a root along parent-edges is l, a quantity preserved by an edge-preserving bijection, giving level preservation and $\mathrm{par} \circ \chi = \chi \circ \mathrm{par}$. The restriction to $V(\mathcal{T}_l(g^{(1)}))$ is then a bijection onto $V(\mathcal{T}_l(g^{(2)}))$ that still preserves edges. □

Lemma 20. *For every $L \in \mathbb{N}_0$, d_L is a pseudo-metric on $\mathcal{G}_n$; that is, $d_L(g^{(1)}, g^{(1)}) = 0$, $d_L(g^{(1)}, g^{(2)}) = d_L(g^{(2)}, g^{(1)})$, and $d_L(g^{(1)}, g^{(3)}) \leq d_L(g^{(1)}, g^{(2)}) + d_L(g^{(2)}, g^{(3)})$ for all $g^{(1)}, g^{(2)}, g^{(3)} \in \mathcal{G}_n$.*

Proof. Non-negativity and $d_L(g^{(1)}, g^{(1)}) = 0$ (take $\chi = \mathrm{id}$) are immediate. For symmetry, if χ is an edge-preserving bijection $V(\mathcal{T}_L(g^{(1)})) \to V(\mathcal{T}_L(g^{(2)}))$ then χ^{-1} is one in the opposite direction, and reindexing the sum in (4) by $i' = \chi(i)$ shows the two minimization problems have the same value. For the triangle inequality, let χ_{12} and χ_{23} be minimizers for $d_L(g^{(1)}, g^{(2)})$ and $d_L(g^{(2)}, g^{(3)})$; then $\chi_{23} \circ \chi_{12}$ is an edge-preserving bijection $V(\mathcal{T}_L(g^{(1)})) \to V(\mathcal{T}_L(g^{(3)}))$, and applying the triangle inequality of $\|\cdot\|$ termwise gives $d_L(g^{(1)}, g^{(3)}) \leq d_L(g^{(1)}, g^{(2)}) + d_L(g^{(2)}, g^{(3)})$. The same argument applies verbatim to $\tilde{d}_L$ (Appendix D). □

Lemma 21. *Let $g \in \mathcal{G}_n$ and $\sigma \in S_n$. Then, for any $L \in \mathbb{N}_0$,*

$$d_L(g, \sigma(g)) = 0.$$

Proof. Write $g' := \sigma(g)$; recall from the definition of $\sigma(g)$ in the preliminaries that $h^{g'}_{\sigma(i)} = h^g_i$ and $e^{g'}_{\sigma(j),\sigma(i)} = e^g_{j,i}$ for all $i, j \in [n]$. Let $\tilde{\chi} : V(\mathcal{T}_L(g)) \to V(\mathcal{T}_L(g'))$ be the edge-preserving bijection all of whose component permutations equal σ. Then, for every summand of (4) indexed by $(i, j = \mathrm{par}(i))$,

$$(h^g_j, e^g_{i,j}, h^g_i) = (h^{g'}_{\tilde{\chi}(j)}, e^{g'}_{\tilde{\chi}(i),\tilde{\chi}(j)}, h^{g'}_{\tilde{\chi}(i)}),$$

so $\tilde{\chi}$ makes the minimized sum in (4) equal to 0, and since $d_L \geq 0$ this proves $d_L(g, g') = 0$. □

B GENERALIZATION BOUNDS VIA ROBUSTNESS

Here we present the notion of robustness for learning algorithms, following Xu & Mannor (2012), from which we will obtain generalization bounds. Recall the supervised learning setting in which $\mathcal{G}$ is a set of graphs, $\mathcal{Y}$ a set of labels, $\mathcal{X}$ a set of embeddings, $\mathcal{Z} := \mathcal{G} \times \mathcal{Y}$, and $\mathcal{H} \subseteq \mathcal{X}^{\mathcal{G}}$ a hypothesis class.

Definition 22 (Definition 2, Xu & Mannor (2012))**.** A (graph) learning algorithm $\mathcal{A} : \mathcal{Z}^k \to \mathcal{H}$ is $(C, \epsilon(\cdot))$-robust for $C \in \mathbb{N}$ and $\epsilon : \mathcal{Z}^k \to \mathbb{R}_+$ if the following holds.

$\mathcal{Z}$ can be partitioned into C disjoint sets $A_i, i \in [C]$, such that for all $s \in \mathcal{Z}^k$ the following holds. For all $(g, y) \in s$, all $(g', y') \in \mathcal{Z}$ and all $i \in [C]$: if $(g, y), (g', y') \in A_i$, then

$$|\ell(\mathcal{A}_s(g), y) - \ell(\mathcal{A}_s(g'), y')| \leq \epsilon(s),$$

where $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ is a loss function and $\mathcal{A}_s = \mathcal{A}(s)$.

Following Xu & Mannor (2012), the robustness of a learning algorithm yields an upper bound on the generalization gap.

Proposition 23 (Theorem 1, Xu & Mannor (2012))**.** *Let $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function bounded above by some $M_\ell > 0$. If a learning algorithm $\mathcal{A} : \mathcal{Z}^k \to \mathcal{H}$ is $(C, \epsilon(\cdot))$-robust for some $C \in \mathbb{N}$ and $\epsilon : \mathcal{Z}^k \to \mathbb{R}_+$, and if $\mathcal{S}$ is a random sample of size $k \in \mathbb{N}$ from $\mathcal{Z}$, then for any $\delta > 0$, with probability at least $1 - \delta$,*

$$\left|\ell_{\exp}(\mathcal{A}_{\mathcal{S}}) - \ell_{\text{emp}}(\mathcal{A}_{\mathcal{S}})\right| \leq \epsilon(\mathcal{S}) + M_\ell \sqrt{\frac{2C \ln(2) + 2\ln(1/\delta)}{k}}.$$

Therefore, in order to bound the generalization gap it suffices to establish robustness of the learning algorithm. As the next proposition shows, this can be achieved for *any* learning algorithm by ensuring suitable continuity properties of the loss function ℓ and the hypothesis class $\mathcal{H}$.

Proposition 24 (Corollary 12, Vasileiou et al. (2024))**.** *Let $(\mathcal{G}, d_{\mathcal{G}})$, $(\mathcal{X}, d_{\mathcal{X}})$, and $(\mathcal{Y}, d_{\mathcal{Y}})$ be pseudo-metric spaces, and let $\mathcal{H}$ be a class of $K_{\mathcal{H}} > 0$ Lipschitz hypotheses from $(\mathcal{G}, d_{\mathcal{G}})$ to $(\mathcal{X}, d_{\mathcal{X}})$. Define $d_\infty : (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}_+$ by*

$$d_\infty((x, y), (x', y')) := \max\{d_{\mathcal{X}}(x, x'), d_{\mathcal{Y}}(y, y')\}.$$

If $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ is a $K_\ell > 0$ Lipschitz loss function with respect to d_∞, then for any $\varepsilon > 0$, every learning algorithm $\mathcal{A} : \mathcal{Z}^k \to \mathcal{H}$ is $(C, \epsilon(\cdot))$-robust with

$$C = \mathcal{C}(\mathcal{G}, d_{\mathcal{G}}, \varepsilon/(2K_{\mathcal{H}})) \cdot \mathcal{C}(\mathcal{Y}, d_{\mathcal{Y}}, \varepsilon/2), \qquad \text{and} \qquad \epsilon(s) \equiv K_\ell\, \varepsilon.$$

In the previous result the loss function must fulfill some conditions. We argue that this is not a strong imposition. As the next proposition shows, any capped norm is a valid loss function, since it is both bounded and Lipschitz continuous with respect to d_∞.

Proposition 25. *Let $d_{\text{out}} \in \mathbb{N}, M_\ell > 0$, $\mathcal{X} = \mathbb{R}^{d_{\text{out}}}$, and $\mathcal{Y} \subseteq \mathbb{R}^{d_{\text{out}}}$. Let $\|\cdot\|$ be any norm on $\mathbb{R}^{d_{\text{out}}}$ and $d_{\mathcal{X}} = d_{\mathcal{Y}}$ be the metric induced by $\|\cdot\|$. Then the function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$*

$$\ell(x, y) := \|x - y\| \wedge M_\ell$$

is 2-Lipschitz with respect to d_∞.

Proof. Since $\|\cdot\|$ is a norm, for any $u, v \in \mathcal{X}$

$$|\|u\| - \|v\|| \leq \|u - v\| \quad \text{and} \quad \|u + v\| \leq \|u\| + \|v\|.$$

Let $(x, y), (x', y') \in \mathcal{X} \times \mathcal{Y}$.

Case 1: If both $\|x - y\|, \|x' - y'\| \leq M_\ell$,

$$\begin{aligned}|\ell(x, y) - \ell(x', y')| &= |\|x - y\| - \|x' - y'\|| \leq \|x - y - x' + y'\| \leq \|x - x'\| + \|y' - y\| \\ &\leq 2\max\{\|x - x'\|, \|y - y'\|\}.\end{aligned}$$

Case 2: If both $\|x - y\|, \|x' - y'\| \geq M_\ell$, then

$$|\ell(x, y) - \ell(x', y')| = M_\ell - M_\ell = 0 \leq 2\max\{\|x - x'\|, \|y - y'\|\}.$$

Case 3: If $\|x - y\| \leq M_\ell$ and $\|x' - y'\| \geq M_\ell$, then $\ell(x, y) = \|x - y\|$, $\ell(x', y') = M_\ell$, and $M_\ell \leq \|x' - y'\|$, so

$$\begin{aligned}|\ell(x, y) - \ell(x', y')| &= M_\ell - \|x - y\| \leq \|x' - y'\| - \|x - y\| \leq \big|\|x - y\| - \|x' - y'\|\big| \\ &\leq \|x - y - x' + y'\| \leq \|x - x'\| + \|y - y'\| \leq 2\max\{\|x - x'\|, \|y - y'\|\}.\end{aligned}$$

Case 4: If $\|x - y\| \geq M_\ell$ and $\|x' - y'\| \leq M_\ell$, the roles of (x, y) and (x', y') are exchanged and the bound follows from Case 3 by symmetry. □

Remark 26. In the case of one dimensional labels in $[0, 1]$, i.e., for $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = [0, 1]$, another natural loss function is $\ell(x, y) := |\sigma(x) - y|$, where σ is the 1-Lipschitz sigmoid function. This loss is also 2-Lipschitz with respect to d_∞ and bounded by $M_\ell = 1$:

$$\begin{aligned}|\ell(x, y) - \ell(x', y')| &= ||\sigma(x) - y| - |\sigma(x') - y'|| \leq |\sigma(x) - \sigma(x') + y' - y| \\ &\leq |\sigma(x) - \sigma(x')| + |y - y'| \leq |x - x'| + |y - y'| \\ &\leq 2\max\{|x - x'|, |y - y'|\}.\end{aligned}$$

C PROOFS

This section contains proofs of the results in the main body of the paper.

Remark 27. Throughout the paper we adopt the following two standing conventions.

(i) $\mathbf{K \geq 1}$. As established in Section 3, we assume $K \geq 1$ whenever a Lipschitz constant K of a hypothesis class is fixed. In particular the component constants satisfy $K_\rho, K_\alpha, K_\phi, K_\psi \leq K$ with $K \geq 1$.

(ii) **Zero-edge irrelevance.** We only consider GNNs for which it is irrelevant whether two nodes are unconnected or connected by a "zero-edge" (edge feature $\mathbf{0}$). This is exactly what the conditions $\phi^{(l)}(\cdot, \mathbf{0}, \cdot) = \psi^{(l)}(\cdot, \mathbf{0}, \cdot) = \mathbf{0}$ encode, and it is what makes the raw-feature representation w^g and the unrolling tree (with zero-filled edges for absent edges) faithful descriptions of the GNN's computation.

Lemma 5. *Let $L \in \mathbb{N}$ and $g^{(1)}, g^{(2)} \in \mathcal{G}_n$. Let ρ, α, ϕ, ψ be update functions as in (1) of some $f \in \mathcal{H}_{1,K}$ and $K_\rho, K_\alpha, K_\phi, K_\psi \in (0, K]$ be their Lipschitz constants, respectively. Let $k \in V(\mathcal{T}_{L-1}(g^{(1)}))$ and $\bar{k} = \chi(k) \in V(\mathcal{T}_{L-1}(g^{(2)}))$, where χ is a distance-minimizing edge-preserving map for $d_L(g^{(1)}, g^{(2)})$. Let $h_k^{g^{(1)*}}$ and $h_{\bar{k}}^{g^{(2)*}}$ be the representations obtained by one update step (1) applied at k and at $\bar{k}$ respectively. Then we have*

$$\|h_k^{g^{(1)*}} - h_{\bar{k}}^{g^{(2)*}}\| \leq (K_\rho + K_\rho K_\phi + K_\rho K_\alpha + K_\rho K_\alpha K_\psi)\, d_L(g^{(1)}, g^{(2)}) = \tilde{K} d_L(g^{(1)}, g^{(2)}),$$

and

$$\sum_{k \in V(\mathcal{T}_{L-1}(g^{(1)}))} \|h_k^{g^{(1)*}} - h_{\bar{k}}^{g^{(2)*}}\| \leq n\tilde{K} d_L(g^{(1)}, g^{(2)}).$$

Proof. For $i, j \in V(\mathcal{T}_L(g^{(1)}))$ denote $\bar{i} = \chi(i)$ and $\bar{j} = \chi(j)$. By the zero-edge convention (Remark 27) and Lemma 18, for any tree node the message sum $\sum_{i \in \mathcal{N}(\cdot)} \phi$ over graph in-neighbors equals the sum over all n children of that node in the unrolling tree; the same holds for ψ. Throughout this proof we therefore read $\mathcal{N}(\cdot)$ as ranging over the (tree) children, and Lemma 19 lets us pair the children of k with those of $\bar{k}$ through χ, so that reindexing the $g^{(2)}$-sums by χ is legitimate.

$$\begin{aligned}
&\|h_k^{g^{(1)*}} - h_{\bar{k}}^{g^{(2)*}}\| \\
&\leq K_\rho \left\| \left(h_k^{g^{(1)}}, \sum_{i \in \mathcal{N}(k)} \phi(h_k^{g^{(1)}}, e_{i,k}^{g^{(1)}}, h_i^{g^{(1)}}), \sum_{j \in [n]} \alpha\left(h_j^{g^{(1)}}, \sum_{i \in \mathcal{N}(j)} \psi(h_j^{g^{(1)}}, e_{i,j}^{g^{(1)}}, h_i^{g^{(1)}}) \right) \right) \right. \\
&\qquad \left. - \left(h_{\bar{k}}^{g^{(2)}}, \sum_{i \in \mathcal{N}(\bar{k})} \phi(h_{\bar{k}}^{g^{(2)}}, e_{i,\bar{k}}^{g^{(2)}}, h_i^{g^{(2)}}), \sum_{j \in [n]} \alpha\left(h_j^{g^{(2)}}, \sum_{i \in \mathcal{N}(j)} \psi(h_j^{g^{(2)}}, e_{i,j}^{g^{(2)}}, h_i^{g^{(2)}}) \right) \right) \right\| \\
&\leq K_\rho \left[\underbrace{\left\| h_k^{g^{(1)}} - h_{\bar{k}}^{g^{(2)}} \right\|}_{\leq \|(\cdot,\cdot,h_k^{g^{(1)}}) - (\cdot,\cdot,h_{\bar{k}}^{g^{(2)}})\| \leq d_L(g^{(1)}, g^{(2)})} + \left\| \sum_{i \in \mathcal{N}(k)} \phi(h_k^{g^{(1)}}, e_{i,k}^{g^{(1)}}, h_i^{g^{(1)}}) - \sum_{i \in \mathcal{N}(\bar{k})} \phi(h_{\bar{k}}^{g^{(2)}}, e_{i,\bar{k}}^{g^{(2)}}, h_i^{g^{(2)}}) \right\| \right. \\
&\qquad \left. + \left\| \sum_{j \in [n]} \alpha\left(h_j^{g^{(1)}}, \sum_{i \in \mathcal{N}(j)} \psi(h_j^{g^{(1)}}, e_{i,j}^{g^{(1)}}, h_i^{g^{(1)}}) \right) - \sum_{j \in [n]} \alpha\left(h_j^{g^{(2)}}, \sum_{i \in \mathcal{N}(j)} \psi(h_j^{g^{(2)}}, e_{i,j}^{g^{(2)}}, h_i^{g^{(2)}}) \right) \right\| \right] \\
&\leq K_\rho \left[d_L(g^{(1)}, g^{(2)}) + K_\phi \underbrace{\sum_{i \in \mathcal{N}(k)} \left\| (h_k^{g^{(1)}}, e_{i,k}^{g^{(1)}}, h_i^{g^{(1)}}) - (h_{\bar{k}}^{g^{(2)}}, e_{\bar{i},\bar{k}}^{g^{(2)}}, h_{\bar{i}}^{g^{(2)}}) \right\|}_{\leq d_L(g^{(1)}, g^{(2)})} \right.
\end{aligned}$$

$$
\begin{aligned}
&\quad + \sum_{\substack{j\in[n]\\ \bar{j}=\chi(j)}} \left\| \alpha\left(h_j^{g^{(1)}}, \sum_{i\in\mathcal{N}(j)} \psi(h_j^{g^{(1)}}, e_{i,j}^{g^{(1)}}, h_i^{g^{(1)}})\right) - \alpha\left(h_{\bar{j}}^{g^{(2)}}, \sum_{\bar{i}\in\mathcal{N}(\bar{j})} \psi(h_{\bar{j}}^{g^{(2)}}, e_{\bar{i},\bar{j}}^{g^{(2)}}, h_{\bar{i}}^{g^{(2)}})\right)\right\| \Bigg] \\
&\leq K_\rho \Bigg[d_L(g^{(1)}, g^{(2)}) + K_\phi d_L(g^{(1)}, g^{(2)}) \\
&\quad + K_\alpha \sum_{j\in[n]} \left[\left\| h_j^{g^{(1)}} - h_{\bar{j}}^{g^{(2)}} \right\| + \left\| \sum_{i\in\mathcal{N}(j)} \psi(h_j^{g^{(1)}}, e_{i,j}^{g^{(1)}}, h_i^{g^{(1)}}) - \sum_{\bar{i}\in\mathcal{N}(\bar{j})} \psi(h_{\bar{j}}^{g^{(2)}}, e_{\bar{i},\bar{j}}^{g^{(2)}}, h_{\bar{i}}^{g^{(2)}}) \right\| \right] \Bigg] \\
&\leq K_\rho \Bigg[d_L(g^{(1)}, g^{(2)}) + K_\phi d_L(g^{(1)}, g^{(2)}) + K_\alpha \underbrace{\sum_{j\in[n]} \left\| h_j^{g^{(1)}} - h_{\bar{j}}^{g^{(2)}} \right\|}_{\leq d_L(g^{(1)}, g^{(2)})} \\
&\quad + K_\alpha K_\psi \underbrace{\sum_{j\in[n]} \sum_{i\in\mathcal{N}(j)} \left\| (h_j^{g^{(1)}}, e_{i,j}^{g^{(1)}}, h_i^{g^{(1)}}) - (h_{\bar{j}}^{g^{(2)}}, e_{\bar{i},\bar{j}}^{g^{(2)}}, h_{\bar{i}}^{g^{(2)}}) \right\|}_{\leq d_L(g^{(1)}, g^{(2)})} \Bigg] \\
&\leq K_\rho \left[d_L(g^{(1)}, g^{(2)}) + K_\phi d_L(g^{(1)}, g^{(2)}) + K_\alpha d_L(g^{(1)}, g^{(2)}) + K_\alpha K_\psi d_L(g^{(1)}, g^{(2)}) \right] \leq \tilde{K} d_L(g^{(1)}, g^{(2)}).
\end{aligned}
$$

This proves the first statement of the lemma. For the second statement, recall from Lemma 18 that the global term of any internal node k may be summed over $\mathrm{ch}(\mathrm{par}(k))$ (the copies of all original nodes that are the children of $\mathrm{par}(k)$) instead of over the roots $[n]$, i.e.

$$
\sum_{j\in[n]} \alpha\left(h_j^{g^{(1)}}, \sum_{i\in\mathcal{N}(j)} \psi(h_j^{g^{(1)}}, e_{i,j}^{g^{(1)}}, h_i^{g^{(1)}})\right) = \sum_{j\in\mathrm{ch}(\mathrm{par}(k))} \alpha\left(h_j^{g^{(1)}}, \sum_{i\in\mathcal{N}(j)} \psi(h_j^{g^{(1)}}, e_{i,j}^{g^{(1)}}, h_i^{g^{(1)}})\right).
$$

We obtain

$$
\begin{aligned}
&\sum_{k\in V(\mathcal{T}_{L-1}(g^{(1)}))} \|h_k^{g^{(1)*}} - h_{\bar{k}}^{g^{(2)*}}\| \\
&\leq K_\rho \sum_{k\in V(\mathcal{T}_{L-1}(g^{(1)}))} \Bigg\| \left(h_k^{g^{(1)}}, \sum_{i\in\mathcal{N}(k)} \phi(h_k^{g^{(1)}}, e_{i,k}^{g^{(1)}}, h_i^{g^{(1)}}), \sum_{j\in[n]} \alpha\left(h_j^{g^{(1)}}, \sum_{i\in\mathcal{N}(j)} \psi(h_j^{g^{(1)}}, e_{i,j}^{g^{(1)}}, h_i^{g^{(1)}})\right)\right) \\
&\quad - \left(h_{\bar{k}}^{g^{(2)}}, \sum_{i\in\mathcal{N}(\bar{k})} \phi(h_{\bar{k}}^{g^{(2)}}, e_{i,\bar{k}}^{g^{(2)}}, h_i^{g^{(2)}}), \sum_{j\in[n]} \alpha\left(h_j^{g^{(2)}}, \sum_{i\in\mathcal{N}(j)} \psi(h_j^{g^{(2)}}, e_{i,j}^{g^{(2)}}, h_i^{g^{(2)}})\right)\right) \Bigg\| \\
&\leq K_\rho \Bigg[\sum_{k\in V(\mathcal{T}_{L-1}(g^{(1)}))} \underbrace{\left\| h_k^{g^{(1)}} - h_{\bar{k}}^{g^{(2)}} \right\|}_{\leq \|(\cdot,\cdot,h_k^{g^{(1)}}) - (\cdot,\cdot,h_{\bar{k}}^{g^{(2)}})\|} \\
&\quad + \sum_{k\in V(\mathcal{T}_{L-1}(g^{(1)}))} \left\| \sum_{i\in\mathcal{N}(k)} \phi(h_k^{g^{(1)}}, e_{i,k}^{g^{(1)}}, h_i^{g^{(1)}}) - \sum_{i\in\mathcal{N}(\bar{k})} \phi(h_{\bar{k}}^{g^{(2)}}, e_{i,\bar{k}}^{g^{(2)}}, h_i^{g^{(2)}}) \right\| \\
&\quad + \sum_{k\in V(\mathcal{T}_{L-1}(g^{(1)}))} \Bigg\| \sum_{j\in\mathrm{ch}(\mathrm{par}(k))} \alpha\left(h_j^{g^{(1)}}, \sum_{i\in\mathcal{N}(j)} \psi(h_j^{g^{(1)}}, e_{i,j}^{g^{(1)}}, h_i^{g^{(1)}})\right) \\
&\qquad\qquad - \sum_{j\in\mathrm{ch}(\mathrm{par}(\bar{k}))} \alpha\left(h_j^{g^{(2)}}, \sum_{i\in\mathcal{N}(j)} \psi(h_j^{g^{(2)}}, e_{i,j}^{g^{(2)}}, h_i^{g^{(2)}})\right) \Bigg\| \Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&\leq K_\rho \Bigg[d_L(g^{(1)}, g^{(2)}) + K_\phi \underbrace{\sum_{k \in V(\mathcal{T}_{L-1}(g^{(1)}))} \sum_{i \in \mathcal{N}(k)} \left\| (h_k^{g^{(1)}}, e_{i,k}^{g^{(1)}}, h_i^{g^{(1)}}) - (h_{\bar{k}}^{g^{(2)}}, e_{\bar{i},\bar{k}}^{g^{(2)}}, h_{\bar{i}}^{g^{(2)}}) \right\|}_{\leq d_L(g^{(1)}, g^{(2)})} \\
&\quad + \sum_{k \in V(\mathcal{T}_{L-1}(g^{(1)}))} \sum_{j \in \mathrm{ch}(\mathrm{par}(k))} \left\| \alpha\left(h_j^{g^{(1)}}, \sum_{i \in \mathcal{N}(j)} \psi(h_j^{g^{(1)}}, e_{i,j}^{g^{(1)}}, h_i^{g^{(1)}}) \right) - \alpha\left(h_{\bar{j}}^{g^{(2)}}, \sum_{\bar{i} \in \mathcal{N}(\bar{j})} \psi(h_{\bar{j}}^{g^{(2)}}, e_{\bar{i},\bar{j}}^{g^{(2)}}, h_{\bar{i}}^{g^{(2)}}) \right) \right\| \Bigg] \\
&\leq K_\rho \Big[d_L(g^{(1)}, g^{(2)}) + K_\phi d_L(g^{(1)}, g^{(2)}) \\
&\quad + K_\alpha \sum_{k \in V(\mathcal{T}_{L-1}(g^{(1)}))} \sum_{j \in \mathrm{ch}(\mathrm{par}(k))} \left[\left\| h_j^{g^{(1)}} - h_{\bar{j}}^{g^{(2)}} \right\| + \left\| \sum_{i \in \mathcal{N}(j)} \psi(h_j^{g^{(1)}}, e_{i,j}^{g^{(1)}}, h_i^{g^{(1)}}) - \sum_{i \in \mathcal{N}(\bar{j})} \psi(h_{\bar{j}}^{g^{(2)}}, e_{\bar{i},\bar{j}}^{g^{(2)}}, h_{\bar{i}}^{g^{(2)}}) \right\| \right] \Big] .
\end{aligned}
$$

Since each $p \in V(\mathcal{T}_{L-2}(g^{(1)})) \cup \{0\}$ is the parent of exactly n nodes in $V(\mathcal{T}_{L-1}(g^{(1)}))$, we observe that

$$
\sum_{k \in V(\mathcal{T}_{L-1}(g^{(1)}))} \sum_{j \in \mathrm{ch}(\mathrm{par}(k))} = \sum_{k \in V(\mathcal{T}_{L-1}(g^{(1)}))} \sum_{p = \mathrm{par}(k)} \sum_{j \in \mathrm{ch}(p)} = n \sum_{p \in V(\mathcal{T}_{L-2}(g^{(1)})) \cup \{0\}} \sum_{j \in \mathrm{ch}(p)} = n \sum_{j \in V(\mathcal{T}_{L-1}(g^{(1)}))} .
$$

With this,

$$
\begin{aligned}
&K_\rho \Big[d_L(g^{(1)}, g^{(2)}) + K_\phi d_L(g^{(1)}, g^{(2)}) \\
&\quad + K_\alpha \sum_{k \in V(\mathcal{T}_{L-1}(g^{(1)}))} \sum_{j \in \mathrm{ch}(\mathrm{par}(k))} \left[\left\| h_j^{g^{(1)}} - h_{\bar{j}}^{g^{(2)}} \right\| + \left\| \sum_{i \in \mathcal{N}(j)} \psi(h_j^{g^{(1)}}, e_{i,j}^{g^{(1)}}, h_i^{g^{(1)}}) - \sum_{i \in \mathcal{N}(\bar{j})} \psi(h_{\bar{j}}^{g^{(2)}}, e_{\bar{i},\bar{j}}^{g^{(2)}}, h_{\bar{i}}^{g^{(2)}}) \right\| \right] \Big] \\
&= K_\rho \Big[d_L(g^{(1)}, g^{(2)}) + K_\phi d_L(g^{(1)}, g^{(2)}) \\
&\quad + nK_\alpha \sum_{j \in V(\mathcal{T}_{L-1}(g^{(1)}))} \left[\left\| h_j^{g^{(1)}} - h_{\bar{j}}^{g^{(2)}} \right\| + \left\| \sum_{i \in \mathcal{N}(j)} \psi(h_j^{g^{(1)}}, e_{i,j}^{g^{(1)}}, h_i^{g^{(1)}}) - \sum_{i \in \mathcal{N}(\bar{j})} \psi(h_{\bar{j}}^{g^{(2)}}, e_{\bar{i},\bar{j}}^{g^{(2)}}, h_{\bar{i}}^{g^{(2)}}) \right\| \right] \Big] \\
&\leq K_\rho \Big[d_L(g^{(1)}, g^{(2)}) + K_\phi d_L(g^{(1)}, g^{(2)}) + nK_\alpha d_L(g^{(1)}, g^{(2)}) \\
&\quad + nK_\alpha K_\psi \sum_{j \in V(\mathcal{T}_{L-1}(g^{(1)}))} \sum_{i \in \mathcal{N}(j)} \left\| (h_j^{g^{(1)}}, e_{i,j}^{g^{(1)}}, h_i^{g^{(1)}}) - (h_{\bar{j}}^{g^{(2)}}, e_{\bar{i},\bar{j}}^{g^{(2)}}, h_{\bar{i}}^{g^{(2)}}) \right\| \Big] \\
&\leq K_\rho \Big[d_L(g^{(1)}, g^{(2)}) + K_\phi d_L(g^{(1)}, g^{(2)}) + nK_\alpha d_L(g^{(1)}, g^{(2)}) + nK_\alpha K_\psi d_L(g^{(1)}, g^{(2)}) \Big] \leq n\tilde{K} d_L(g^{(1)}, g^{(2)}).
\end{aligned}
$$

□

Proposition 6. *Let $g^{(1)}, g^{(2)} \in \mathcal{G}_n$. For $K \geq 1$, let $K \geq K_\tau > 0$ and τ be a K_τ-Lipschitz readout function as in (2) for some $f \in \mathcal{H}_{1,K}$. With $h^{g^{(1)*}} = f(g^{(1)})$ and $h^{g^{(2)*}} = f(g^{(2)})$, we obtain*

$$
\|h^{g^{(1)*}} - h^{g^{(2)*}}\| \leq K_\tau \, \tilde{K} \, d_1(g^{(1)}, g^{(2)}) \leq 4K^4 d_1(g^{(1)}, g^{(2)}).
$$

Proof. Let $h_i^{g^{(1)*}}, h_i^{g^{(2)*}}, i \in [n]$ be the node representation of $g^{(1)}$ and $g^{(2)}$ after the update step of f. Let χ be a distance-minimizing edge-preserving map of $d_1(g^{(1)}, g^{(2)})$. Then

$$
\begin{aligned}
\left\| h^{g^{(1)*}} - h^{g^{(2)*}} \right\| &= \left\| \tau\left(\frac{1}{n} \sum_{k \in [n]} h_k^{g^{(1)*}} \right) - \tau\left(\frac{1}{n} \sum_{k \in [n]} h_k^{g^{(2)*}} \right) \right\| \\
&\leq K_\tau \frac{1}{n} \sum_{\substack{k \in [n] \\ \bar{k} = \chi(k)}} \left\| h_k^{g^{(1)*}} - h_{\bar{k}}^{g^{(2)*}} \right\| \leq K_\tau \tilde{K} d_1(g^{(1)}, g^{(2)}),
\end{aligned}
$$

where, in the last step, we applied Lemma 5. □

Lemma 7. *Let $g^{(1)}, g^{(2)} \in \mathcal{G}_n$ and let $g^{(1)*}, g^{(2)*}$ be the directed featured graphs on n nodes obtained after updating the node features of $g^{(1)}$ and $g^{(2)}$ by one update step of some $f \in \mathcal{H}_{1,K}$. Then, for any $L \in \mathbb{N}$,*

$$d_{L-1}(g^{(1)*}, g^{(2)*}) \leq 12\, n^2 K^3\, d_L(g^{(1)}, g^{(2)}).$$

Proof. Let $\tilde{\chi}$ be a distance-minimizing edge-preserving map of $d_L(g^{(1)}, g^{(2)})$. To simplify notation let $\Theta_0 = \mathrm{roots}(\mathcal{T}_{L-1}(g^{(1)*}))$ and $\Theta = V(\mathcal{T}_{L-1}(g^{(1)*})) \setminus \mathrm{roots}(\mathcal{T}_{L-1}(g^{(1)*}))$.

$$\begin{aligned}
&d_{L-1}(g^{(1)*}, g^{(2)*}) \\
&= \min_{\substack{\chi: V(\mathcal{T}_{L-1}(g^{(1)*})) \to V(\mathcal{T}_{L-1}(g^{(2)*})) \\ \text{edge-preserving bijection}}} \sum_{\substack{i \in V(\mathcal{T}_{L-1}(g^{(1)*})) \\ j = \mathrm{par}(i)}} \left\| (h_j^{g^{(1)*}}, e_{i,j}^{g^{(1)*}}, h_i^{g^{(1)*}}) - (h_{\chi(j)}^{g^{(2)*}}, e_{\chi(i),\chi(j)}^{g^{(2)*}}, h_{\chi(i)}^{g^{(2)*}}) \right\| \\
&= \min_{\substack{\chi: V(\mathcal{T}_{L-1}(g^{(1)*})) \to V(\mathcal{T}_{L-1}(g^{(2)*})) \\ \text{edge-preserving bijection}}} \left\{ \sum_{i \in \Theta_0} \left\| h_i^{g^{(1)*}} - h_{\chi(i)}^{g^{(2)*}} \right\| + \sum_{\substack{i \in \Theta \\ j = \mathrm{par}(i)}} \left\| (h_j^{g^{(1)*}}, e_{i,j}^{g^{(1)}}, h_i^{g^{(1)*}}) - (h_{\chi(j)}^{g^{(2)*}}, e_{\chi(i),\chi(j)}^{g^{(2)}}, h_{\chi(i)}^{g^{(2)*}}) \right\| \right\} \\
&\leq \sum_{i \in \Theta_0} \left\| h_i^{g^{(1)*}} - h_{\tilde{\chi}(i)}^{g^{(2)*}} \right\| + \sum_{\substack{i \in \Theta \\ j = \mathrm{par}(i)}} \left\| (h_j^{g^{(1)*}}, e_{i,j}^{g^{(1)}}, h_i^{g^{(1)*}}) - (h_{\tilde{\chi}(j)}^{g^{(2)*}}, e_{\tilde{\chi}(i),\tilde{\chi}(j)}^{g^{(2)}}, h_{\tilde{\chi}(i)}^{g^{(2)*}}) \right\| \\
&\leq \sum_{i \in \Theta_0} \left\| h_i^{g^{(1)*}} - h_{\tilde{\chi}(i)}^{g^{(2)*}} \right\| + \sum_{\substack{i \in \Theta \\ j = \mathrm{par}(i)}} \left\| h_j^{g^{(1)*}} - h_{\tilde{\chi}(j)}^{g^{(2)*}} \right\| + \underbrace{\sum_{\substack{i \in \Theta \\ j = \mathrm{par}(i)}} \left\| e_{i,j}^{g^{(1)}} - e_{\tilde{\chi}(i),\tilde{\chi}(j)}^{g^{(2)}} \right\|}_{\leq d_L(g^{(1)}, g^{(2)})} + \sum_{i \in \Theta} \left\| h_i^{g^{(1)*}} - h_{\tilde{\chi}(i)}^{g^{(2)*}} \right\| \\
&\leq \sum_{i \in V(\mathcal{T}_{L-1}(g^{(1)}))} \left\| h_i^{g^{(1)*}} - h_{\tilde{\chi}(i)}^{g^{(2)*}} \right\| + n \sum_{j \in V(\mathcal{T}_{L-1}(g^{(1)}))} \left\| h_j^{g^{(1)*}} - h_{\tilde{\chi}(j)}^{g^{(2)*}} \right\| + d_L(g^{(1)}, g^{(2)}) \\
&\leq n\tilde{K} d_L(g^{(1)}, g^{(2)}) + n^2 \tilde{K} d_L(g^{(1)}, g^{(2)}) + d_L(g^{(1)}, g^{(2)}) \leq 12\, n^2 K^3 d_L(g^{(1)}, g^{(2)}),
\end{aligned}$$

where the last step uses $\tilde{K} \leq 4K^3$ and $K \geq 1$ (so $1 \leq 4K^3 \leq 4n^2K^3$ and $n\tilde{K} \leq 4nK^3 \leq 4n^2K^3$; Remark 27). Here we used that Θ contains no roots, so $\mathrm{par}(j) \in V(\mathcal{T}_{L-2}(g^{(1)}))$ for every $j \in \Theta$, and that the multiset $\{\, \mathrm{par}(j) \mid j \in \Theta \,\}$ contains each element of $V(\mathcal{T}_{L-2}(g^{(1)}))$ exactly n times. Hence the parent-feature sum satisfies $\sum_{j \in \Theta} \|h_{\mathrm{par}(j)}^{g^{(1)*}} - \cdots\| = n \sum_{p \in V(\mathcal{T}_{L-2}(g^{(1)}))} \|h_p^{g^{(1)*}} - \cdots\| \leq n \sum_{i \in V(\mathcal{T}_{L-1}(g^{(1)}))} \|h_i^{g^{(1)*}} - \cdots\|$, using $V(\mathcal{T}_{L-2}(g^{(1)})) \subset V(\mathcal{T}_{L-1}(g^{(1)}))$. Lemma 5 was applied in the penultimate step. □

Proposition 8. *Let $g^{(1)}, g^{(2)} \in \mathcal{G}_n$, and let $f \in \mathcal{H}_{L,K}$ for some $L \in \mathbb{N}$ and $K \geq 1$. Then for $c(K, L, n) = 4^L 3^{L-1} n^{2(L-1)} K^{3L+1}$,*

$$\|f(g^{(1)}) - f(g^{(2)})\| \leq c \cdot d_L(g^{(1)}, g^{(2)}).$$

Proof. Denote by $g_l^{(1)}, g_l^{(2)}$ the graphs with updated node features after the l-th update step of f, so $g_0^{(1)} := g^{(1)}$, $g_0^{(2)} := g^{(2)}$. By Proposition 6 and since $\tilde{K} \leq 4K^3$,

$$\|f(g^{(1)}) - f(g^{(2)})\| \leq K_\tau \tilde{K} d_1(g_{L-1}^{(1)}, g_{L-1}^{(2)}) \leq 4K^4 d_1(g_{L-1}^{(1)}, g_{L-1}^{(2)}).$$

We apply Lemma 7 iteratively and obtain

$$\begin{aligned}
4K^4 d_1(g_{L-1}^{(1)}, g_{L-1}^{(2)}) &\leq 4K^4 (12n^2K^3)\, d_2(g_{L-2}^{(1)}, g_{L-2}^{(2)}) \\
&\vdots \\
&\leq 4K^4 (12n^2K^3)^{L-1} d_L(g^{(1)}, g^{(2)})
\end{aligned}$$

$$= 4K^4\, 12^{L-1} n^{2(L-1)} K^{3(L-1)} d_L(g^{(1)}, g^{(2)})$$
$$= 4^L 3^{L-1} n^{2(L-1)} K^{3L+1} d_L(g^{(1)}, g^{(2)}),$$

where the last equality uses $4 \cdot 12^{L-1} = 4^L 3^{L-1}$. Choosing $c(K, L, n) = 4^L 3^{L-1} n^{2(L-1)} K^{3L+1}$ yields the desired result. □

Proposition 9. *Let $g^{(1)}, g^{(2)} \in \mathcal{G}_n$ and $L \in \mathbb{N}$. Set*

$$T_L := \begin{cases} 2, & L = 1, \\ 2 + |V(\mathcal{T}_{L-2}(g^{(1)}))|, & L \geq 2, \end{cases}$$

which is independent of the particular $g^{(1)} \in \mathcal{G}_n$. Then

$$d_L(g^{(1)}, g^{(2)}) \;\leq\; \sqrt{2}\, T_L\, n^{3/2}\, \|w^{g^{(1)}} - w^{g^{(2)}}\|_2.$$

Proof. Choosing the edge-preserving bijection χ whose component permutations are all the identity (which fixes each original node) upper-bounds the minimum in (4):

$$d_L(g^{(1)}, g^{(2)}) \leq \sum_{\substack{i \in V(\mathcal{T}_L(g^{(1)})) \\ j = \mathrm{par}(i)}} \left\| (h_j^{g^{(1)}}, e_{i,j}^{g^{(1)}}, h_i^{g^{(1)}}) - (h_j^{g^{(2)}}, e_{i,j}^{g^{(2)}}, h_i^{g^{(2)}}) \right\|.$$

Split off the n roots (for which $j = 0$, $h_0 = \mathbf{0}$, $e_{i,0} = \mathbf{0}$) and reindex the remaining terms by the original labels $(u, v) := (\mathrm{orig}(j), \mathrm{orig}(i))$ of parent and child. Each internal node has exactly one child per original node, and among the $|V(\mathcal{T}_{L-1}(g^{(1)}))|$ internal nodes each original node occurs as $\mathrm{orig}(j)$ exactly $|V(\mathcal{T}_{L-1}(g^{(1)}))|/n = 1 + |V(\mathcal{T}_{L-2}(g^{(1)}))|$ times (with the convention $|V(\mathcal{T}_{-1})| = 0$). Hence every pair $(u, v) \in [n]^2$ contributes $1 + |V(\mathcal{T}_{L-2}(g^{(1)}))|$ identical terms, and

$$\begin{aligned} d_L(g^{(1)}, g^{(2)}) &\leq \sum_{v \in [n]} \|h_v^{g^{(1)}} - h_v^{g^{(2)}}\| \\ &\quad + \big(1 + |V(\mathcal{T}_{L-2}(g^{(1)}))|\big) \sum_{u,v \in [n]} \left\| (h_u^{g^{(1)}}, e_{v,u}^{g^{(1)}}, h_v^{g^{(1)}}) - (h_u^{g^{(2)}}, e_{v,u}^{g^{(2)}}, h_v^{g^{(2)}}) \right\| \\ &\leq T_L \sum_{u,v \in [n]} \left\| (h_u^{g^{(1)}}, e_{v,u}^{g^{(1)}}, h_v^{g^{(1)}}) - (h_u^{g^{(2)}}, e_{v,u}^{g^{(2)}}, h_v^{g^{(2)}}) \right\|, \end{aligned}$$

where the last step bounds each root term $\|h_v^{g^{(1)}} - h_v^{g^{(2)}}\|$ by the $(u, v) = (v, v)$ summand of the triple sum. Applying the Cauchy–Schwarz inequality to the n^2 summands we get

$$\begin{aligned} &\leq T_L \sum_{u,v \in [n]} \left\| (h_u^{g^{(1)}}, e_{v,u}^{g^{(1)}}, h_v^{g^{(1)}}) - (h_u^{g^{(2)}}, e_{v,u}^{g^{(2)}}, h_v^{g^{(2)}}) \right\| \\ &= T_L \sum_{u,v \in [n]} 1 \cdot \left\| (h_u^{g^{(1)}}, e_{v,u}^{g^{(1)}}, h_v^{g^{(1)}}) - (h_u^{g^{(2)}}, e_{v,u}^{g^{(2)}}, h_v^{g^{(2)}}) \right\| \\ &\leq T_L \sqrt{\sum_{u,v \in [n]} 1^2} \cdot \sqrt{\sum_{u,v \in [n]} \left\| (h_u^{g^{(1)}}, e_{v,u}^{g^{(1)}}, h_v^{g^{(1)}}) - (h_u^{g^{(2)}}, e_{v,u}^{g^{(2)}}, h_v^{g^{(2)}}) \right\|^2} \end{aligned}$$

and using

$$\begin{aligned} &\sum_{u,v \in [n]} \left\| (h_u^{g^{(1)}}, e_{v,u}^{g^{(1)}}, h_v^{g^{(1)}}) - (h_u^{g^{(2)}}, e_{v,u}^{g^{(2)}}, h_v^{g^{(2)}}) \right\|^2 \\ &= 2n \sum_{u \in [n]} \|h_u^{g^{(1)}} - h_u^{g^{(2)}}\|^2 + \sum_{u,v \in [n]} \|e_{v,u}^{g^{(1)}} - e_{v,u}^{g^{(2)}}\|^2 \leq 2n\, \|w^{g^{(1)}} - w^{g^{(2)}}\|_2^2, \end{aligned}$$

(each node feature appears $2n$ times, each edge feature once, and $2n \geq 1$) we obtain

$$d_L(g^{(1)}, g^{(2)}) \leq T_L\, n \sqrt{2n}\, \|w^{g^{(1)}} - w^{g^{(2)}}\|_2 = \sqrt{2}\, T_L\, n^{3/2}\, \|w^{g^{(1)}} - w^{g^{(2)}}\|_2.$$ □

Lemma 10. *Let $L \in \mathbb{N}$, $\varepsilon > 0$, and T_L, $p = nd + n^2 d'$ as in Proposition 9. Then*

$$\mathcal{C}(\mathcal{G}_n, d_L, \varepsilon) \leq \mathcal{C}\Big([-M, M]^p,\ \|\cdot\|_2,\ \tfrac{\varepsilon}{\sqrt{2}\, T_L\, n^{3/2}}\Big).$$

Proof. Put $r := \varepsilon/(\sqrt{2}\, T_L\, n^{3/2})$ and $N := \mathcal{C}([-M, M]^p, \|\cdot\|_2, r)$, and let $w_1, \ldots, w_N \in [-M, M]^p$ be an r-cover. Each w_m is the feature vector of some graph $g_m \in \mathcal{G}_n$. For $g \in \mathcal{G}_n$ choose m with $\|w^g - w_m\|_2 \leq r$; then by Proposition 9,

$$d_L(g, g_m) \leq \sqrt{2}\, T_L\, n^{3/2} \|w^g - w_m\|_2 \leq \varepsilon.$$

Hence $g_1, \ldots, g_N$ is an ε-cover of $(\mathcal{G}_n, d_L)$. □

Lemma 11. *Let $I := [-M, M]$ and $p \in \mathbb{N}$. For $\varepsilon > 0$,*

$$\mathcal{C}(I^p, \|\cdot\|_2, \varepsilon) \leq (M\sqrt{p}/\varepsilon + 1)^p.$$

Proof. Partition I^p into axis-aligned subcubes of side length $2\varepsilon/\sqrt{p}$; this requires $\lceil M\sqrt{p}/\varepsilon \rceil$ subcubes per axis, hence $N := \lceil M\sqrt{p}/\varepsilon \rceil^p$ subcubes with centers $c_1, \ldots, c_N$. Each subcube has diameter $\sqrt{\sum_{i \in [p]} (2\varepsilon/\sqrt{p})^2} = 2\varepsilon$, so every point of a subcube lies within ε of its center. Thus the ε-balls around $c_1, \ldots, c_N$ cover I^p, and $\mathcal{C}(I^p, \|\cdot\|_2, \varepsilon) \leq \lceil M\sqrt{p}/\varepsilon \rceil^p \leq (M\sqrt{p}/\varepsilon + 1)^p$. □

Theorem 12. *Let d_1 denote the pseudo-metric on $\mathcal{G}_n$ from Definition 4. Let $\mathcal{X} = \mathbb{R}$ (so the readout dimension is $d_{\text{out}} = 1$), $\mathcal{Y} \subseteq [0, 1]$ and $d_{\mathcal{X}} = d_{\mathcal{Y}} = |\cdot|$. For some $M_\ell > 0$, let the loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be defined by*

$$\ell(x, y) := |x - y| \wedge M_\ell.$$

Let $K_{\mathcal{H}} \geq 1$ and let $\mathcal{H}_{1,K_{\mathcal{H}}} \subset \mathcal{X}^{\mathcal{G}_n}$ be the set of 1-layer GNNs whose components are $K_{\mathcal{H}}$-Lipschitz functions defined in (3).

Let $k \in \mathbb{N}$, $\mathcal{A} : \mathcal{Z}^k \to \mathcal{H}_{1,K_{\mathcal{H}}}$ be any learning algorithm, and $\mathcal{S}$ be a random sample of size k of $\mathcal{Z}$. Then, for any $\delta, \varepsilon > 0$, with probability at least $1 - \delta$

$$|\ell_{\exp}(\mathcal{A}_{\mathcal{S}}) - \ell_{\text{emp}}(\mathcal{A}_{\mathcal{S}})| \leq 2\varepsilon + M_\ell \sqrt{\frac{2C \ln(2) + 2\ln(1/\delta)}{k}},$$

where

$$C = \mathcal{C}(\mathcal{G}_n, d_1, \varepsilon/(8K_{\mathcal{H}}^4)) \cdot \mathcal{C}(\mathcal{Y}, |\cdot|, \varepsilon/2).$$

Proof. Without loss of generality assume $\mathcal{Y} = [0, 1]$; the covering numbers below are therefore written for $[0, 1]$. We note that by Proposition 25 the loss function ℓ is 2-Lipschitz continuous with respect to d_∞. By Proposition 6 all GNNs in $\mathcal{H}_{1,K_{\mathcal{H}}}$ are Lipschitz continuous with Lipschitz constant $4K_{\mathcal{H}}^4$. Therefore, by Proposition 24, for any $\varepsilon > 0$ $\mathcal{A}$ is

$$(\mathcal{C}(\mathcal{G}_n, d_1, \varepsilon/(8K_{\mathcal{H}}^4)) \cdot \mathcal{C}([0, 1], |\cdot|, \varepsilon/2), 2\epsilon(\cdot))\text{-robust},$$

where $\epsilon(\cdot) \equiv \varepsilon$. Finally, we apply Proposition 23 and obtain that for a random sample $\mathcal{S}$ of size $k \in \mathbb{N}$ of $\mathcal{Z}$, for any $\delta > 0$, with probability at least $1 - \delta$

$$|\ell_{\exp}(\mathcal{A}_{\mathcal{S}}) - \ell_{\text{emp}}(\mathcal{A}_{\mathcal{S}})| \leq 2\varepsilon + M_\ell \sqrt{\frac{2C \ln(2) + 2\ln(1/\delta)}{k}},$$

where

$$C = \mathcal{C}(\mathcal{G}_n, d_1, \varepsilon/(8K_{\mathcal{H}}^4)) \cdot \mathcal{C}([0, 1], |\cdot|, \varepsilon/2).$$

□

Remark 28. Theorem 12 also holds for $\mathcal{X} = \mathbb{R}^{d_{\text{out}}}$ and $\mathcal{Y} \subseteq [-M_{\mathcal{Y}}, M_{\mathcal{Y}}]^{d_{\text{out}}}$ for some $M_{\mathcal{Y}} > 0$ and $d_{\text{out}} \in \mathbb{N}$. Moreover, the loss function ℓ may be any norm $\|\cdot\|$ on $\mathbb{R}^{d_{\text{out}}}$ capped at $M_\ell > 0$, i.e., $\ell(x, y) = \|x - y\| \wedge M_\ell$, provided that $d_{\mathcal{X}} = d_{\mathcal{Y}}$ are the metrics induced by this norm; see Proposition 25. Note, however, that the choice of norm and the value of $M_{\mathcal{Y}}$ affect the covering number $\mathcal{C}(\mathcal{Y}, d_{\mathcal{Y}}, \varepsilon/2)$.

Corollary 13. *In the same setting as in Theorem 12, let $p := nd + n^2 d'$. Then for any $\delta, \varepsilon > 0$, with probability at least $1 - \delta$,*

$$|\ell_{\exp}(\mathcal{A}_{\mathcal{S}}) - \ell_{\mathrm{emp}}(\mathcal{A}_{\mathcal{S}})| \le 2\varepsilon + M_\ell \sqrt{\frac{2\lceil 1/\varepsilon \rceil \left((16\sqrt{2}\, M\, n^{3/2} \sqrt{p}\, K_{\mathcal{H}}^4)/\varepsilon + 1\right)^p \ln(2) + 2\ln(1/\delta)}{k}}.$$

Proof. In Theorem 12 we replace the two covering numbers by explicit upper bounds. For the graphs, Lemma 10 with $L = 1$ (so $T_1 = 2$) followed by Lemma 11 gives, at radius $\varepsilon/(8K_{\mathcal{H}}^4)$,

$$\mathcal{C}\Big(\mathcal{G}_n, d_1, \tfrac{\varepsilon}{8K_{\mathcal{H}}^4}\Big) \le \mathcal{C}\Big([-M, M]^p, \|\cdot\|_2, \tfrac{\varepsilon/(8K_{\mathcal{H}}^4)}{2\sqrt{2}\, n^{3/2}}\Big) \le \left(\frac{16\sqrt{2}\, M\, n^{3/2} \sqrt{p}\, K_{\mathcal{H}}^4}{\varepsilon} + 1\right)^p.$$

For the labels, $\mathcal{C}([0,1], |\cdot|, \varepsilon/2) = \lceil 1/\varepsilon \rceil$. Substituting both into Theorem 12 and using that the loss is 2-Lipschitz (so the leading term is 2ε) yields the claim. □

Theorem 15. *In the same setting as in Theorem 12, for any $\delta, \varepsilon > 0$, with probability at least $1 - \delta$*

$$\ell_{\exp}(\mathcal{A}_{\mathcal{S}}) \le \ell_{\mathrm{emp}}(\mathcal{A}_{\mathcal{S}}) \;+\; 2\varepsilon \;+\; \zeta(\mathcal{A}_{\mathcal{S}}) \left((\sqrt{2}+1) \sqrt{\frac{|\mathcal{T}_{\mathcal{S}}| \ln(2C/\delta)}{k}} \;+\; \frac{2|\mathcal{T}_{\mathcal{S}}| \ln(2C/\delta)}{k} \right),$$

where

$$C = \mathcal{C}(\mathcal{G}_n, d_1, \varepsilon/(8K_{\mathcal{H}}^4)) \cdot \mathcal{C}(\mathcal{Y}, |\cdot|, \varepsilon/2).$$

Proof. Exactly as in the proof of Theorem 12, the loss is 2-Lipschitz (Proposition 25) and every $f \in \mathcal{H}_{1,K_{\mathcal{H}}}$ is $4K_{\mathcal{H}}^4$-Lipschitz (Proposition 6), so by Proposition 24 the algorithm $\mathcal{A}$ is $(C, 2\varepsilon)$-robust with C as stated. Here the constant robustness function has value $\epsilon(\mathcal{S}) = 2\varepsilon$. Applying the data-dependent bound of Proposition 14 (in place of Proposition 23) to a random sample $\mathcal{S}$ of size k yields the claim. □

Theorem 16. *Let d_L denote the pseudo-metric on $\mathcal{G}_n$ from Definition 4. Let $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} \subseteq [0,1]$ and $d_{\mathcal{X}} = d_{\mathcal{Y}} = |\cdot|$. For some $M_\ell > 0$, let the loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be defined by*

$$\ell(x, y) := |x - y| \wedge M_\ell.$$

Let $K_{\mathcal{H}} \ge 1$ and let $\mathcal{H}_{L,K_{\mathcal{H}}} \subset \mathcal{X}^{\mathcal{G}_n}$ be the set of L-layer GNNs whose components are $K_{\mathcal{H}}$-Lipschitz functions defined in (3).

Let $k \in \mathbb{N}$, $\mathcal{A} : \mathcal{Z}^k \to \mathcal{H}_{L,K_{\mathcal{H}}}$ be any learning algorithm, and $\mathcal{S}$ be a random sample of size k of $\mathcal{Z}$. Then, for any $\delta, \varepsilon > 0$, with probability at least $1 - \delta$,

$$|\ell_{\exp}(\mathcal{A}_{\mathcal{S}}) - \ell_{\mathrm{emp}}(\mathcal{A}_{\mathcal{S}})| \le 2\varepsilon + M_\ell \sqrt{\frac{2C\ln(2) + 2\ln(1/\delta)}{k}},$$

where

$$C = \mathcal{C}\left(\mathcal{G}_n,\, d_L,\, \frac{\varepsilon}{2 \cdot 4^L 3^{L-1} K_{\mathcal{H}}^{3L+1} n^{2(L-1)}}\right) \cdot \mathcal{C}(\mathcal{Y}, |\cdot|, \varepsilon/2).$$

Proof. Without loss of generality assume $\mathcal{Y} = [0,1]$; the covering numbers below are therefore written for $[0,1]$. By Proposition 25, the loss ℓ is bounded by M_ℓ and 2-Lipschitz with respect to d_∞. By Proposition 8, every $f \in \mathcal{H}_{L,K_{\mathcal{H}}}$ satisfies

$$\|f(g^{(1)}) - f(g^{(2)})\| \le 4^L 3^{L-1} K_{\mathcal{H}}^{3L+1} n^{2(L-1)}\, d_L(g^{(1)}, g^{(2)}), \qquad g^{(1)}, g^{(2)} \in \mathcal{G}_n.$$

Therefore, by Proposition 24, for any $\varepsilon > 0$ the algorithm $\mathcal{A}$ is

$$\left(\mathcal{C}\left(\mathcal{G}_n,\, d_L,\, \frac{\varepsilon}{2 \cdot 4^L 3^{L-1} K_{\mathcal{H}}^{3L+1} n^{2(L-1)}}\right) \cdot \mathcal{C}\Big([0,1], |\cdot|, \frac{\varepsilon}{2}\Big),\; 2\varepsilon\right)\text{-robust.}$$

Applying Proposition 23 to a random sample $\mathcal{S}$ of size k yields the claim. □

Theorem 17. *In the same setting as in Theorem 16, for any $\delta, \varepsilon > 0$, with probability at least $1-\delta$*

$$\ell_{\exp}(\mathcal{A}_{\mathcal{S}}) \le \ell_{\mathrm{emp}}(\mathcal{A}_{\mathcal{S}}) \;+\; 2\varepsilon \;+\; \zeta(\mathcal{A}_{\mathcal{S}})\left((\sqrt{2}+1)\sqrt{\frac{|\mathcal{T}_{\mathcal{S}}|\ln(2C/\delta)}{k}} \;+\; \frac{2|\mathcal{T}_{\mathcal{S}}|\ln(2C/\delta)}{k}\right),$$

where

$$C = \mathcal{C}\left(\mathcal{G}_n,\, d_L,\, \frac{\varepsilon}{2\cdot 4^L 3^{L-1} K_{\mathcal{H}}^{3L+1} n^{2(L-1)}}\right)\cdot \mathcal{C}\left(\mathcal{Y}, |\cdot|, \frac{\varepsilon}{2}\right).$$

Proof. As in the proof of Theorem 16, the loss ℓ is bounded by M_ℓ and 2-Lipschitz with respect to d_∞ (Proposition 25), and by Proposition 8 every $f \in \mathcal{H}_{L,K_{\mathcal{H}}}$ is $4^L 3^{L-1} K_{\mathcal{H}}^{3L+1} n^{2(L-1)}$-Lipschitz with respect to d_L. By Proposition 24 the algorithm $\mathcal{A}$ is therefore $(C, \epsilon(\cdot))$-robust with C as stated and $\epsilon(\cdot) \equiv 2\varepsilon$. Applying the data-dependent bound of Proposition 14 (in place of Proposition 23) to a random sample $\mathcal{S}$ of size k yields the claim. □

Remark 29. This remark derives an explicit upper bound on $\ln C$, valid for all $n \ge 2, M \ge \varepsilon > 0$, from which the dependence on depth and graph size quoted in Section 3.6 can be read off. Let $p := nd + n^2 d'$, recall $c(K_{\mathcal{H}}, L, n) = 4^L 3^{L-1} K_{\mathcal{H}}^{3L+1} n^{2(L-1)}$, and combine Lemma 10 with Lemma 11 at the radius $\varepsilon/(2c(K_{\mathcal{H}}, L, n))$ required by Theorem 16. Then, using $\mathcal{C}(\mathcal{Y}, |\cdot|, \varepsilon/2) \le \mathcal{C}([0,1], |\cdot|, \varepsilon/2) = \lceil 1/\varepsilon \rceil$,

$$C \;\le\; \left\lceil \tfrac{1}{\varepsilon} \right\rceil (B+1)^p, \qquad B \;:=\; \frac{2\sqrt{2}\, M \sqrt{p}\, T_L\, n^{3/2}\, c(K_{\mathcal{H}}, L, n)}{\varepsilon}.$$

For $n \ge 2, d \ge 1, L \ge 1, K_{\mathcal{H}} \ge 1$, we have $\sqrt{p} \ge \sqrt{nd} \ge \sqrt{2}$, $T_L \ge 2$, $n^{3/2} \ge 2\sqrt{2}$, and $c(K_{\mathcal{H}}, L, n) \ge 4$, such that

$$B \ge \frac{2\sqrt{2}\cdot M \cdot \sqrt{2} \cdot 2 \cdot 2\sqrt{2} \cdot 4}{\varepsilon} \ge \frac{M}{\varepsilon}$$

which is ≥ 1 for $\varepsilon \le M$. Since T_L is of order n^{L-1} and $\sqrt{p}$ of order n, the powers of n in B collect as $(L-1) + \frac{3}{2} + 2(L-1) + 1 = 3L - \frac{1}{2}$, and

$$\begin{aligned}\ln B &= \ln\left(2\sqrt{2}\, M\right) + \tfrac{1}{2}\ln p + \ln T_L + \tfrac{3}{2}\ln n + \ln c(K_{\mathcal{H}}, L, n) + \ln \tfrac{1}{\varepsilon} \\ &= \left(3L - \tfrac{1}{2}\right)\ln n + (3L+1)\ln K_{\mathcal{H}} + L\ln 12 + \ln\tfrac{1}{\varepsilon} + \underbrace{\ln \tfrac{2\sqrt{2}M}{3} + \tfrac{1}{2}\ln\tfrac{p}{n^2} + \ln\tfrac{T_L}{n^{L-1}}}_{\le \ln\left(2M\sqrt{d+d'}\right) \text{ for } n \ge 2}.\end{aligned}$$

For bounding the last three terms we used that

$$\begin{aligned}&\ln \tfrac{2\sqrt{2}M}{3} \le \ln M, \\ &\tfrac{1}{2}\ln\tfrac{p}{n^2} = \ln\left(\sqrt{\tfrac{nd+n^2d'}{n^2}}\right) \le \ln\left(\sqrt{\tfrac{n^2d+n^2d'}{n^2}}\right) = \ln\left(\sqrt{d+d'}\right)\end{aligned}$$

and that for $n \ge 2$,

$$\begin{aligned}\frac{T_L}{n^{L-1}} = \frac{2+n+\cdots+n^{L-1}}{n^{L-1}} &= \frac{2}{n^{L-1}} + \sum_{j=0}^{L-2} n^{-j} \le \frac{2}{2^{L-1}} + \sum_{j=0}^{L-2} 2^{-j} \\ &= 2^{-(L-2)} + \frac{1-(\frac{1}{2})^{L-1}}{1-\frac{1}{2}} = 2^{-(L-2)} + \left(2 - 2^{-(L-2)}\right) = 2.\end{aligned}$$

Since $B \ge 1$ it holds

$$\ln(B+1) \le \ln(2B) = \ln B + \ln 2$$

and

$$\ln 2 + \ln\left(2M\sqrt{d+d'}\right) = \ln\left(4M\sqrt{d+d'}\right),$$

such that altogether

$$\ln C \le p\left[\left(3L - \tfrac{1}{2}\right)\ln n + (3L+1)\ln K_{\mathcal{H}} + L\ln 12 + \ln\tfrac{1}{\varepsilon} + \ln\left(4M\sqrt{d+d'}\right)\right] + \ln\left\lceil \tfrac{1}{\varepsilon}\right\rceil.$$

D Generalization Bounds for Permutation-Invariant Multi-Layer GNNs with Simple Message-Passing

We consider the simplified class $\mathcal{H}^*_{L,K}$, where we remove the global-pooling component given by α and ψ in the message passing and replace the tuple $(h_j, e_{i,j}, h_i)$ with $(e_{i,j}, h_i)$ such that the message now only depends on the edge feature and the feature of the source node, but not anymore on the destination node of the edge. In each message-passing step, the update rule for $\mathcal{H}^*_{L,K}$ is

$$h_k^{(l)} = \rho^{(l)}\left(h_k^{(l-1)}, \sum_{i \in \mathcal{N}(k)} \phi^{(l)}\left(e_{i,k}, h_i^{(l-1)}\right)\right), \qquad l \in [L], \tag{6}$$

followed by the readout (2). Formally,

$$\mathcal{H}^*_{L,K} := \left\{ f : g \mapsto h_g \;\middle|\; \begin{array}{r} h_g \text{ is computed by (6) and (2),} \\ \rho^{(l)},\ \phi^{(l)},\ l \in [L], \text{ and } \tau \text{ are } K\text{-Lipschitz,} \\ \phi^{(l)}(\mathbf{0}, \cdot) = \mathbf{0} \text{ for all } l \in [L]. \end{array} \right\} \tag{7}$$

With this simpler message passing we can also simplify the graph pseudo-metric with respect to which such GNNs are Lipschitz continuous.

Definition 30. Let $L \in \mathbb{N}_0$ and $g^{(1)}, g^{(2)} \in \mathcal{G}_n$. For $i, j \in [n]$, let $h_i^{g^{(1)}}, e_{i,j}^{g^{(1)}}$ and $h_i^{g^{(2)}}, e_{i,j}^{g^{(2)}}$ denote the node and edge features of $g^{(1)}$ and $g^{(2)}$, respectively. We define

$$\tilde{d}_L(g^{(1)}, g^{(2)}) := \min_{\substack{\chi : V(\mathcal{T}_L(g^{(1)})) \to V(\mathcal{T}_L(g^{(2)})) \\ \text{edge-preserving bijection}}} \sum_{\substack{i \in V(\mathcal{T}_L(g^{(1)})) \\ j = \mathrm{par}(i)}} \left\| (e_{i,j}^{g^{(1)}}, h_i^{g^{(1)}}) - (e_{\chi(i),\chi(j)}^{g^{(2)}}, h_{\chi(i)}^{g^{(2)}}) \right\|, \tag{8}$$

where we set $\chi(0) = 0$, $h_0^{g^{(1)}} = h_0^{g^{(2)}} = \mathbf{0}$, and $e_{\cdot,0}^{g^{(1)}} = e_{\cdot,0}^{g^{(2)}} = \mathbf{0}$.

Under these modifications, the dependence on n disappears from the Lipschitz constant of GNNs, as we can see in the following corollaries. First we present a version of Lemma 5 for the simplified message-passing architecture.

Corollary 31. *Let $L \in \mathbb{N}$ and $g^{(1)}, g^{(2)} \in \mathcal{G}_n$. Let ρ and ϕ be update functions as in (6) of some $f \in \mathcal{H}^*_{1,K}$ that are $K_\rho, K_\phi \in (0, K]$ Lipschitz, respectively. Let $k \in V(\mathcal{T}_{L-1}(g^{(1)}))$ and $\bar{k} = \chi(k) \in V(\mathcal{T}_{L-1}(g^{(2)}))$, where χ is a distance-minimizing edge-preserving map for $\tilde{d}_L(g^{(1)}, g^{(2)})$. With $h_k^{g^{(1)*}}$ and $h_{\bar{k}}^{g^{(2)*}}$ the node representations after one update step of f, we have*

$$\|h_k^{g^{(1)*}} - h_{\bar{k}}^{g^{(2)*}}\| \le \tilde{K}^* \, \tilde{d}_L(g^{(1)}, g^{(2)}),$$

and

$$\sum_{k \in V(\mathcal{T}_{L-1}(g^{(1)}))} \|h_k^{g^{(1)*}} - h_{\bar{k}}^{g^{(2)*}}\| \le \tilde{K}^* \, \tilde{d}_L(g^{(1)}, g^{(2)}),$$

where $\tilde{K}^* := K_\rho + K_\rho K_\phi$.

Proof. The proof follows the same reasoning as the proof of Lemma 5 adapted to the simpler GNN architecture. As there, the zero-edge convention (Remark 27) and Lemma 18 let us read $\mathcal{N}(\cdot)$ as ranging over the n tree children and pair the children of k with those of $\bar{k}$ through χ (Lemma 19); we write $\bar{i} = \chi(i)$. By K_ρ-Lipschitz continuity of ρ and K_ϕ-Lipschitz continuity of ϕ,

$$\|h_k^{g^{(1)*}} - h_{\bar{k}}^{g^{(2)*}}\| \le K_\rho \left[\|h_k^{g^{(1)}} - h_{\bar{k}}^{g^{(2)}}\| + K_\phi \sum_{i \in \mathcal{N}(k)} \|(e_{i,k}^{g^{(1)}}, h_i^{g^{(1)}}) - (e_{\bar{i},\bar{k}}^{g^{(2)}}, h_{\bar{i}}^{g^{(2)}})\| \right].$$

Since χ minimizes $\tilde{d}_L$, the node-feature discrepancy satisfies $\|h_k^{g^{(1)}} - h_{\bar{k}}^{g^{(2)}}\| \le \tilde{d}_L(g^{(1)}, g^{(2)})$ (it is a sub-sum of the minimized expression, obtained by selecting the term with child index $i = k$

and parent $j = \mathrm{par}(k)$, where $j = 0$ when k is a root, in which case the summand reduces to $\|(\mathbf{0}, h_k^{g^{(1)}}) - (\mathbf{0}, h_{\chi(k)}^{g^{(2)}})\| = \|h_k^{g^{(1)}} - h_{\chi(k)}^{g^{(2)}}\|$).

Similarly, the sum over neighbors satisfies $\sum_{i\in\mathcal{N}(k)} \|(e_{i,k}^{g^{(1)}}, h_i^{g^{(1)}}) - (e_{\tilde{i},\tilde{k}}^{g^{(2)}}, h_{\tilde{i}}^{g^{(2)}})\| \le \tilde{d}_L(g^{(1)}, g^{(2)})$. Combining gives the first bound.

For the second bound, summing over all $k \in V(\mathcal{T}_{L-1}(g^{(1)}))$:

$$\begin{aligned}\sum_{k\in V(\mathcal{T}_{L-1}(g^{(1)}))} \|h_k^{g^{(1)*}} - h_{\tilde{k}}^{g^{(2)*}}\| &\le K_\rho \underbrace{\sum_k \|h_k^{g^{(1)}} - h_{\tilde{k}}^{g^{(2)}}\|}_{\le \tilde{d}_L(g^{(1)}, g^{(2)})} \\ &\quad + K_\rho K_\phi \underbrace{\sum_k \sum_{i\in\mathcal{N}(k)} \|(e_{i,k}^{g^{(1)}}, h_i^{g^{(1)}}) - (e_{\tilde{i},\tilde{k}}^{g^{(2)}}, h_{\tilde{i}}^{g^{(2)}})\|}_{\le \tilde{d}_L(g^{(1)}, g^{(2)})} \\ &\le \tilde{K}^*\, \tilde{d}_L(g^{(1)}, g^{(2)}).\end{aligned}$$

Each of the two sums is individually bounded by $\tilde{d}_L(g^{(1)}, g^{(2)})$, being a sub-sum of the minimized expression defining it. Crucially, no factor of n arises for this simplified GNN type. □

Similarly, we can also formulate a version of Lemma 7 for the simpler architecture.

Corollary 32. *Let $g^{(1)}, g^{(2)} \in \mathcal{G}_n$ and let $g^{(1)*}, g^{(2)*}$ be the featured graphs on n nodes obtained after updating the node features of $g^{(1)}$ and $g^{(2)}$ by one update step of some $f \in \mathcal{H}^*_{1,K}$. Then, for any $L \in \mathbb{N}$,*

$$\tilde{d}_{L-1}(g^{(1)*}, g^{(2)*}) \le 4K^2\, \tilde{d}_L(g^{(1)}, g^{(2)}).$$

Proof. Let $\tilde{\chi}$ be a distance-minimizing edge-preserving map for $\tilde{d}_L(g^{(1)}, g^{(2)})$. Since edge features are unchanged ($e^{g^{(1)*}} = e^{g^{(1)}}$, $e^{g^{(2)*}} = e^{g^{(2)}}$),

$$\begin{aligned}\tilde{d}_{L-1}(g^{(1)*}, g^{(2)*}) &\le \sum_{\substack{i\in V(\mathcal{T}_{L-1}(g^{(1)}))\\ j=\mathrm{par}(i)}} \|(e_{i,j}^{g^{(1)}},\, h_i^{g^{(1)*}}) - (e_{\tilde{\chi}(i),\tilde{\chi}(j)}^{g^{(2)}},\, h_{\tilde{\chi}(i)}^{g^{(2)*}})\| \\ &\le \underbrace{\sum_{i,\, j=\mathrm{par}(i)} \|e_{i,j}^{g^{(1)}} - e_{\tilde{\chi}(i),\tilde{\chi}(j)}^{g^{(2)}}\|}_{\le \tilde{d}_L(g^{(1)}, g^{(2)})} + \underbrace{\sum_{i,\, j=\mathrm{par}(i)} \|h_i^{g^{(1)*}} - h_{\tilde{\chi}(i)}^{g^{(2)*}}\|}_{\le \tilde{K}^*\tilde{d}_L(g^{(1)}, g^{(2)})} \\ &\le (1 + \tilde{K}^*)\, \tilde{d}_L(g^{(1)}, g^{(2)}) \le 4K^2\, \tilde{d}_L(g^{(1)}, g^{(2)}),\end{aligned}$$

where the first sum is a sub-sum of the minimized expression defining $\tilde{d}_L$, the second sum is bounded by Corollary 31, and the last step uses $\tilde{K}^* \le 2K^2$ together with $1 \le 2K^2$ (as $K \ge 1$). □

Combining the previous corollaries yields the following Lipschitz constant for the simplified message-passing architecture, which—in contrast to the general multi-layer bound of Proposition 8—is independent of n.

Proposition 33. *Let $g^{(1)}, g^{(2)} \in \mathcal{G}_n$ and $f \in \mathcal{H}^*_{L,K}$ for some $L \in \mathbb{N}$ and $K \ge 1$. Then for $c(K, L) = 2^{2L-1}K^{2L+1}$,*

$$\|f(g^{(1)}) - f(g^{(2)})\| \le c \cdot \tilde{d}_L(g^{(1)}, g^{(2)}).$$

Proof. First, note that Proposition 6 holds for the simplified architecture if $\tilde{K}$ is replaced by $\tilde{K}^*$. This is easy to see; we simply invoke Corollary 31 instead of Lemma 5 in the proof.

Denote by $g_l^{(1)}, g_l^{(2)}$ the graphs after the l-th update step of f, so $g_0^{(1)} := g^{(1)}$, $g_0^{(2)} := g^{(2)}$. By (the simplified version of) Proposition 6, and using $K_\tau, K_\rho, K_\phi \le K$,

$$\|f(g^{(1)}) - f(g^{(2)})\| \le K_\tau\, \tilde{K}^* \tilde{d}_1(g_{L-1}^{(1)}, g_{L-1}^{(2)}) \le 2K^3 \tilde{d}_1(g_{L-1}^{(1)}, g_{L-1}^{(2)}).$$

Applying Corollary 32 iteratively $(L-1)$ times yields

$$\begin{aligned} 2K^3\tilde{d}_1(g_{L-1}^{(1)}, g_{L-1}^{(2)}) &\le 2K^3(4K^2)^{L-1}\tilde{d}_L(g^{(1)}, g^{(2)}) \\ &= 2K^3\,4^{L-1}K^{2(L-1)}\tilde{d}_L(g^{(1)}, g^{(2)}) \\ &= 2^{2L-1}K^{2L+1}\tilde{d}_L(g^{(1)}, g^{(2)}). \end{aligned}$$

Choosing $c(K, L) = 2^{2L-1}K^{2L+1}$ yields the result. □

In contrast to Proposition 8, this bound is *independent of* n. This shows that with the described simplifications of the message-passing architecture the dependence on n in the Lipschitz bound vanishes making the bounds more similar to those obtained in Vasileiou et al. (2024), besides our more general message-passing structure and our accounting for edge features.

The same reasoning used to bound the covering number of $(\mathcal{G}_n, d_L)$ in Section 3 applies to $\tilde{d}_L$; the only difference is that the feature triple $(h_j, e_{i,j}, h_i)$ is replaced by the pair $(e_{i,j}, h_i)$, so a node feature contributes to only one (rather than two) of the two endpoints of an edge term.

Corollary 34. *Let $L \in \mathbb{N}$, $\varepsilon > 0$, and $p = nd + n^2d'$. With T_L as in Proposition 9 (applied to $g^{(1)}$'s unrolling tree),*

$$\tilde{d}_L(g^{(1)}, g^{(2)}) \le T_L\, n^{3/2}\, \|w^{g^{(1)}} - w^{g^{(2)}}\|_2, \qquad g^{(1)}, g^{(2)} \in \mathcal{G}_n,$$

and consequently

$$\mathcal{C}(\mathcal{G}_n, \tilde{d}_L, \varepsilon) \le \left(\frac{M\sqrt{p}\,T_L\, n^{3/2}}{\varepsilon} + 1\right)^p.$$

Proof. Exactly as in the proof of Proposition 9, bound $\tilde{d}_L(g^{(1)}, g^{(2)})$ via the identity bijection and reindex by $(u, v) := (\mathrm{orig}(j), \mathrm{orig}(i))$ to get $\tilde{d}_L(g^{(1)}, g^{(2)}) \le T_L \sum_{u,v\in[n]} \|(e_{v,u}^{g^{(1)}}, h_v^{g^{(1)}}) - (e_{v,u}^{g^{(2)}}, h_v^{g^{(2)}})\|$. Now each pair (u, v) contributes only the single node feature h_v (rather than both h_u, h_v), so summing squared norms over $u, v \in [n]$ gives exactly $n \sum_{v\in[n]} \|h_v^{g^{(1)}} - h_v^{g^{(2)}}\|^2 + \sum_{u,v\in[n]} \|e_{v,u}^{g^{(1)}} - e_{v,u}^{g^{(2)}}\|^2 \le n\, \|w^{g^{(1)}} - w^{g^{(2)}}\|_2^2$ (using $n \ge 1$ and that the two sums add up exactly to $\|w^{g^{(1)}} - w^{g^{(2)}}\|_2^2$). Cauchy–Schwarz over the n^2 terms then gives $\tilde{d}_L(g^{(1)}, g^{(2)}) \le T_L\, n\sqrt{n}\, \|w^{g^{(1)}} - w^{g^{(2)}}\|_2 = T_L n^{3/2} \|w^{g^{(1)}} - w^{g^{(2)}}\|_2$ (one factor of $\sqrt{2}$ fewer than for d_L in Proposition 9, since only one, not two, node features appear per edge term). The covering-number bound then follows exactly as in Lemma 10 together with Lemma 11. □

Removing the dependence on n from the Lipschitz constant does not remove it from the bound. The covering-number estimates of Lemmata 10 and 11 (respectively Corollary 34) still involve $T_L = 2+|V(\mathcal{T}_{L-2}(g))|$, which is of order n^{L-1}. Repeating the computation in Remark 29 with $c(K, L) = 2^{2L-1}K^{2L+1}$ in place of $c(K_{\mathcal{H}}, L, n)$, the powers of n in the base collect as $L + \frac{3}{2}$ rather than $3L - \frac{1}{2}$, since the Lipschitz constant now contributes none. As this base is raised to the exponent $p = nd + n^2d'$, the growth of $\ln C$ is again linear in L and of order $n^2 L \ln n$: the simplified architecture reduces the coefficient of L in the exponent from 3 to 1, but not the order.