A Theoretical and Empirical Investigation into the Equivalence of Graph Neural Networks and the Weisfeiler-Leman Algorithm

From the faculty of Mathematics, Physics, and Computer Science approved for the purpose of obtaining the academic degree of Bachelor of Sciences.

Eric Tillmann Bill

Supervision:
Prof. Dr. rer. nat. Christopher Morris

Informatik 6
RWTH Aachen University

Contents

1 Introduction

Yet to come!

2 Related Work

Yet to come!

3 Preliminaries

First, we introduce a couple of notions and definitions that will be used throughout this thesis. In particular, the definitions will play a crucial role in the theoretical part that follows. We start with some general notations, introduce a general graph definition, and familiarize the reader with the Weisfeiler-Leman algorithm. We will then introduce each framework independently, first the 1-WL+NN and then GNN. At the end, we will briefly introduce important properties of collections of functions computed by both methods.

3.1 General Notation

We first introduce a couple of notations and definitions that will be used throughout the thesis. With $[n]$, we denote the set $\{1, \dots, n\} \subset \mathbb{N}$ for any $n \in \mathbb{N}$ and with $\{\!\{\dots\}\!\}$ we denote a multiset which is formally defined as a 2 tuple (X, m) with X being a set of all unique elements and $m : X \to \mathbb{N}_{\geq 1}$ a mapping that maps every element in X to its number of occurrences in the multiset.

3.2 Graphs

A graph G is a 3-tuple $G := (V, E, l)$ that consists of the set of all nodes V, the set of all edges $E \subseteq V \times V$ and a label function $l : M \to \Sigma$ with M being either $V, V \cup E$ or E and $\Sigma \subset \mathbb{N}$ a finite alphabet. Moreover, let $\mathcal{G}$ be the set of all finite graphs. Note, that our definition of the label function allows for graphs with labels either only on the nodes, only on the edges, or on both nodes and edges. Sometimes the values assigned by l are called features, but this is usually only the case when Σ is multidimensional, which we do not cover in this thesis. In addition, although we have defined it this way, the labeling function is optional, and in cases where no labeling function is given, we add the trivial labeling function $f_1 : V(G) \to \{1\}$. Further, G can be either directed or undirected, depending on the definition of E, where $E \subseteq \{(v, u) \mid v, u \in V\}$ defines a directed and $E \subseteq \{(v, u), (u, v) \mid v, u \in V, v \neq u\}$ such that for every $(v, u) \in E$ also $(u, v) \in E$ defines an undirected graph. Additionally, we will use the notation $V(G)$ and $E(G)$ to denote the set of nodes of G and the set of edges of G respectively, as well as l_G to denote the label function of G. With $\mathcal{N}(v)$ for $v \in V(G)$ we denote the set of neighbors of v with $\mathcal{N}(v) := \{u \mid (u, v) \in E(G)\}$.

A coloring of a Graph G is a function $C : V(G) \to \mathbb{N}$ that assigns each node in the graph a color (here a positive integer). Further, a coloring C induces a partition P on the set of nodes, for which we define C^{-1} being the function that maps each color $c \in \mathbb{N}$ to its class of nodes with $C^{-1}(c) = \{v \in V(G) \mid C(v) = c\}$. In addition, we define $h_{G,C}$ as the histogram of graph G with coloring C, that maps every color in the image of C under $V(G)$ to the number of occurrences. In detail, $\forall c \in \mathbb{N} : h_{G,C}(c) := |\{v \in V(G) \mid C(v) = c\}| = |C^{-1}(c)|$

Permutation-invariance and -equivariance

We use S_n to denote the symmetric group over the elements $[n]$ for any $n > 0$. S_n consists of all permutations over these elements. Let G be a graph with $V(G) = [n]$, applying a permutation $\pi \in S_n$ on G, is defined as $G_\pi := \pi \cdot G$ where $V(G_\pi) = \{\pi(1), \dots, \pi(n)\}$ and $E(G_\pi) = \{(\pi(v), \pi(u)) \mid (v, u) \in E(G)\}$. We will now introduce two key concepts for classifying functions on graphs.

Definition 1 (Permutation Invariant)**.** Let $f : \mathcal{G} \to \mathcal{X}$ be an arbitrary function and let $V(G) = [n]$ for some $n \in \mathbb{N}$. The function f is *permutation-invariant* if and only if for all $G \in \mathcal{G}$ where $n_G := \mid V(G) \mid$ and for every $\pi \in S_{n_G}$: $f(G) = f(\pi \cdot G)$.

Definition 2 (Permuation Equivariant)**.** Let $f : \mathcal{G} \to \mathcal{X}$ be an arbitrary function and let $V(G) = [n]$ for some $n \in \mathbb{N}$. The function f is *permuation-equivariant* if and only if for all $G \in \mathcal{G}$ where $n_G := \mid V(G) \mid$ and for every $\pi \in S_{n_G}$: $f(G) = \pi^{-1} \cdot f(\pi \cdot G)$.

3.3 Weisfeiler and Leman Algorithm

The Weisfeiler-Leman algorithm consists of two main parts, first the coloring algorithm and second the graph isomorphism test. We will introduce them in this section.

Add the continues variant of the 1-WL by adding HASH functiions Morris: Faster kernel for graphs with continuous attributes via hashing

The Weisfeiler-Leman graph coloring algorithm

The 1-WL algorithm computes a node coloring of its input graph in each iteration. A color for a node is computed using only the coloring of its neighbors and the node itself. The algorithm will continue as long as it has not converged, and returns the final coloring of the graph.

Definition 3 (1-WL Algorithm)**.** Let $G = (V, E, l)$ be a graph, then in each iteration i, the 1-WL computes a node coloring $C_i : V(G) \to \mathbb{N}$. In iteration $i = 0$, the initial coloring is $C_0 = l$ or if l is non-existing $\forall v \in V(G) : C_0(v) = c$ for an arbitrary constant $c \in \mathbb{N}$. For $i > 0$, the algorithm assigns a color to $v \in V(G)$ as follows:

$$C_i(v) = \mathsf{RELABEL}(C_{i-1}(v), \{\!\{C_{i-1}(u) \mid u \in \mathcal{N}(v)\}\!\}),$$

where RELABEL injectively maps the above pair to a unique, previously not used, natural number. Although this is not a formal restriction by the inventors, we further require the function to always map to the next minimal natural number. Thereby we can contain the size of the codomain of each coloring for all iterations. The algorithm terminates when the number of colors between two iterations does not change, meaning the algorithm terminates after iteration i if the following condition is satisfied:

$$\forall v, w \in V(G) : C_i(v) = C_i(w) \iff C_{i+1}(v) = C_{i+1}(w).$$

Upon terminating we define $C_\infty := C_i$ as the stable coloring, such that 1-WL$(G) := C_\infty$.

The colorings computed in each iteration always converge to the final one, such that the algorithm always terminates. In more detail, **?** showed that it always holds after at most $|V(G)|$ iterations. For an illustration of this coloring algorithm, see Figure 2. Moreover, based on the work of **?** about efficient refinement strategies, **?** proved that the stable coloring C_∞ can be computed in time $\mathcal{O}(|V(G)| + |E(G)| \cdot \log |V(G)|)$.

The Weisfeiler-Leman Graph Isomorphism Test

Definition 4 (1-WL Isomorphism Test)**.** To determine if two graphs $G, H \in \mathcal{G}$ are non-isomorphic ($G \ncong H$), one applies the 1-WL coloring algorithm on both graphs "in parallel" and checks after each iteration if the occurrences of each color are equal, else the algorithm would terminate and conclude non-isomorphic. Formally, the algorithm concludes non-isomorphic in iteration i if there exists a color c such that:

$$|\{v \in V(G) \mid c = C_i(v)\}| \neq |\{v \in V(H) \mid c = C_i(v)\}|.$$

Note that this test is only sound and not complete for the *graph isomorphism problem*. Counterexamples where the algorithm fails to distinguish non-isomorphic graphs can be easily constructed, see Figure 1 which was discovered and proven by **?**.

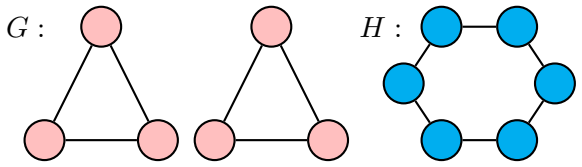


Figure 1: An example of two graphs G and H that are non-isomorphic but cannot be distinguished by the 1-WL isomorphism test.

3.4 1-WL+NN

As seen in the previous section, the 1-WL algorithm is quite powerful in identifying substructures. With the 1-WL+NN framework, we define functions that utilize this structural information to derive further application-specific insights. We do this by combining well-known machine learning techniques and the algorithm.

Definition 5 (1-WL+NN)**.** We say the function $\mathcal{B} : \mathcal{G} \to \mathbb{R}^m$ is computable by 1-WL+NN, if it can be compromised as $\mathcal{B}(\cdot) = \text{MLP} \circ f_{\text{enc}} \circ \text{1-WL}(\cdot)$, where f_{enc} is a permutation invariant encoding function that maps graph-colorings to fixed-sized vectors, and MLP is a multilayer perceptron.

As a concrete example of a collection of functions computable by 1-WL+NN we will introduce the collection $\mathfrak{B}_k$ that is parametrized by $k \in \mathbb{N}_{\geq 1}$. All functions $\mathcal{B} \in \mathfrak{B}_k$ use the *counting-encoding* function f_{count} as their encoding function, and are constrained in their domain to only work over a subset $\mathcal{X}$ of $\mathcal{G}$. We will define this particular encoding function in the following:

Definition 6 (Counting Encoding Functions)**.** For $k \in \mathbb{N}_{\geq 1}$, let

$$\mathcal{X} = \{G \in \mathcal{G} \mid \forall x \in V(G) \cup E(G) : l_G(x) \leq k\} \subset \mathcal{G}$$

be the set of all graphs, where the label alphabet Σ of the respective label function l is bounded with $\Sigma \subseteq [k]$. We define the *counting-encoding* function $f_{\text{count}} : \text{1-WL}(\mathcal{X}) \to \mathbb{N}^K$ as the function that maps a graph coloring C_∞ of a graph $G \in \mathcal{X}$ to a vector $v \in \mathbb{N}^K$ such that the c.th component of v is equal to the number of occurrences of the color c in the coloring C_∞. More formally, for $G \in \mathcal{X}$ let C_∞ be the final coloring upon the termination of the 1-WL algorithm on G and h_{G,C_∞} the respective color histogram. Then f_{count} maps C_∞ to a vector $v \in \mathbb{N}^K$, such that for all $c \in [K] : v_c = h_{G,C_\infty}(c)$, where v_c denotes the c.th component of the vector v. Important to note, due to the bounded label alphabet Σ of all graphs $G \in \mathcal{X}$ by the parameter k, there exists a minimal K for the codomain $\mathbb{N}^K$ of f_{count}, such that f_{count} is well-defined on all graphs $G \in \mathcal{X}$.

To illustrate how this encoding function works and why we coined it *counting-encoding*, we will quickly introduce an example graph G. In Figure 2, we give a visual representation of G and its stable coloring after applying the 1-WL algorithm to it. The *counting-encoding* function f_{count} counts through all colors $i \in [K]$ and sets each i.th component of the output vector to the number of occurrences in the final coloring. Therefore, the respective color histogram

$h_{G,C_\infty} = \{\!\{2,2,3,4\}\!\}$ of G is being mapped to $v \in \mathbb{N}^K$ with $v = (0,2,1,1,0,\ldots,0)^T$, since color 2 appears two times, while color 3 and 4 occur only once. All other components of v are set to 0.

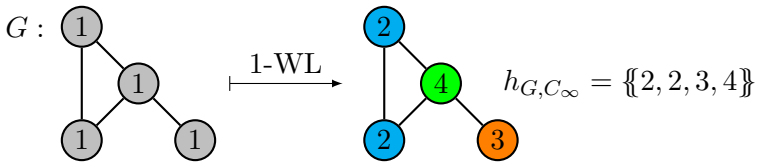


Figure 2: An example of the final coloring computed by applying the 1-WL algorithm on the graph G. The graph G consists of 4 nodes with all their labels being initially set to 1. Note that each label corresponds to a color, which we have also plotted for illustration purposes.

3.5 Graph Neural Networks (Message Passing)

A Graph Neural Network (GNN) is a composition of multiple layers, where each layer computes a new feature for each node and edge. The GNN layer thus technically obtains a new graph that is structurally identical to the previous one, but contains new feature information. After an input graph has been passed through all layers, there can be an additional final function, aggregating the computed information into a fixed size output. With this, it is possible to apply a GNN to every graph, regardless of its size, as the "computation" will only take place on the nodes and edges of the graph.

Note that in the following we will restrict the definition to only consider node features, however, one can easily extend it to also include edge features.

Definition 7 (Graph Neural Network)**.** Let $G = (V, E, l)$ be an arbitrary graph. A Graph Neural Network (GNN) is a composition of multiple layers where each layer t is represented by a function $f^{(t)}$ that works over the set of nodes $V(G)$. To begin with, we need a function $f^{(0)} : V(G) \to \mathbb{R}^{1\times d}$ that is consistent with l, that translates all labels into a vector representation. Further, for every $t > 0$, $f^{(t)}$ is of the format:

$$f^{(t)}(v) = f_{\text{merge}}^{W_{1,t}}(f^{(t-1)}(v),\ f_{\text{agg}}^{W_{2,t}}(\{\!\{f^{(t-1)}(w) \mid w \in \mathcal{N}(v)\}\!\})),$$

where $f_{\text{merge}}^{W_{1,t}}$ and $f_{\text{agg}}^{W_{2,t}}$ are arbitrary differentiable functions with $W_{1,t}$ and $W_{2,t}$ their respective parameters. Additionally, $f_{\text{agg}}^{W_{2,t}}$ has to be permuation-invariant.

Depending on the objective, whether the GNN is tasked with a graph or a node task, the last layer differs. In the case of graph tasks, we add a permutation-invariant aggregation function to the end, here called READOUT, that aggregates over every node and computes a fixed-size output vector for the entire graph, e.g. a label for graph classification. In order to ensure that we can train the GNN in an end-to-end fashion, we require READOUT to be also differentiable. Let $\mathcal{A}$ be an instance of the described GNN framework. Further, let $K \in \mathbb{N}$ be the number of layers of the GNN, $\mathcal{G}$ the set of all graphs, $\mathcal{Y}$ the task-specific output set (e.g. labels of a classification task), then the overall function computed by $\mathcal{A}$ is:

$$\mathcal{A} : \mathcal{G} \to \mathcal{Y} : x \mapsto \text{READOUT} \circ f^{(K)} \circ \ldots \circ f^{(0)}(x),$$

if $\mathcal{A}$ is configured for a graph task, otherwise:

$$\mathcal{A} : \mathcal{G} \to \mathcal{Y} : x \mapsto f^{(K)} \circ \ldots \circ f^{(0)}(x).$$

Note that, as we require all aggregation functions to be permutation-invariant, the total composition $\mathcal{A}$ is permutation-invariant, and with similar reasoning, it is also differentiable. This enables us to train $\mathcal{A}$ like any other machine learning method in an end-to-end fashion, regardless of the underlying encoding used for graphs. This definition and use of notation are inspired by **?** and **?**.

To demonstrate what kind of functions are typically used, we provide functions used by **?** for a node classification:

$$f_{\text{merge}}^{W_{1,t}}(v) = \text{ReLU}(W_{\text{merge}} \cdot \mathsf{concat}(f^{(t-1)}(v),\ f_{\text{agg}}^{W_{2,t}}(v)))$$
$$f_{\text{agg}}^{W_{2,t}}(v) = \mathsf{MAX}(\{\text{ReLU}(W_{\text{pool}} \cdot f^{(t-1)}(u) + b) \mid u \in \mathcal{N}(v)\})$$

where ReLU is a non-linear element wise activation function, MAX the element-wise max operator; W_{merge}, W_{pool} are trainable matrices, b a trainable vector and concat the concatenation function.

3.6 Important for later

In this section, we introduce a formal definition of multilayer perceptron as it is required in a later proof, as well as the $\simeq_{1\text{WL}}$ relation. Additionally, two very important properties for collections of functions.

Definition 8 (Multilayer Perceptron)**.** Multilayer perceptrons are a class of functions from $\mathbb{R}^n$ to $\mathbb{R}^m$, with $n, m \in \mathbb{N}$. In this thesis, we define a multilayer perceptron as a finite sequence, such that a multilayer perceptron MLP is defined as $\text{MLP} := (\text{MLP})_{i \in [k]}$ where k is the number of layers. For every $i \in [k]$, the i.th layer of the MLP is the i.th item in the finite sequence $(\text{MLP})_i$. Further, all layers are recursively defined as:

$$(\text{MLP})_1(v) := v$$
$$(\text{MLP})_{i+1}(v) := \sigma(W_i \cdot (\text{MLP})_i(v) + b_i), \quad \forall i \in [k-1]$$

where σ is an element wise activation function, W_i is the weight matrix and b_i the bias vector of layer i. Note, that for each W_i, the succeeding W_{i+1} must have the same number of columns as W_i has rows, in order to be well-defined. Similarly, for every layer i, W_i and b_i have to have the same number of rows. Following this definition, when applying a MLP on input $v \in \mathbb{R}^n$ it is $\text{MLP}(v) := (\text{MLP})_k(v)$.

Definition 9 (1-WL Relation)**.** For any graphs G, H we will denote $G \simeq_{1\text{WL}} H$ if the 1-WL isomorphism test can not distinguish both graphs. Note that due to the soundness of this algorithm, if $G \not\simeq_{1\text{WL}} H$, we always can conclude that $G \not\simeq H$.

The $\simeq_{1\text{WL}}$ relation can further be classified as an equivalence relation, as it is reflexive, symmetric and transitive. With this, we introduce a notation of its equivalence classes. Let $G \in \mathcal{G}$, then we denote with $\mathcal{G}/\simeq_{1\text{WL}}(G) := \{G' \in \mathcal{G} \mid G \simeq_{1\text{WL}} G'\}$ its equivalence class.

Definition 10 (1-WL-Discriminating)**.** Let $\mathcal{C}$ be a collection of permutation invariant functions from $\mathcal{X}$ to $\mathbb{R}$. We say $\mathcal{C}$ is **1-WL-Discriminating** if for all graphs $G_1, G_2 \in \mathcal{X}$ for which the 1-WL isomorphism test concludes non-isomorphic ($G_1 \not\simeq_{1\text{WL}} G_2$), there exists a function $h \in \mathcal{C}$ such that $f(G_1) \neq f(G_2)$.

Definition 11 (GNN-Approximating)**.** Let $\mathcal{C}$ be a collection of permutation invariant functions from $\mathcal{X}$ to $\mathbb{R}$. We say $\mathcal{C}$ is **GNN-Approximating** if for all permutation-invariant functions $\mathcal{A}$ computed by a GNN, and for all $\epsilon \in \mathbb{R}$ with $\epsilon > 0$, there exists $h_{\mathcal{A},\epsilon} \in \mathcal{C}$ such that $\|\mathcal{A} - h_{\mathcal{A},\epsilon}\|_\infty := \sup_{G \in \mathcal{X}} |f(G) - h_{\mathcal{A},\epsilon}(G)| < \epsilon$

4 Theoretical Connection

This section is the main part of our theoretical investigation of the two frameworks. We will present 4 intriguing theorems, which will be proven separately afterwards. These results will form the basis for the empirical part that follows. The first two theorems will establish an equivalence between the two frameworks when the set of graphs is finite. The last two theorems will go one step further and establish a connection for continuous graph features and prove a somewhat weaker connection between them.

Throughout the first two theorems we will concentrate on a finite collection of finite graphs which we will denote with $\mathcal{X} \subset \mathcal{G}$.

Theorem 12 (Finite Case: "1-WL+NN $\subseteq$ GNN")**.** Let $\mathcal{C}$ be a collection of functions from $\mathcal{X}$ to $\mathbb{R}$ computable by GNNs, then $\mathcal{C}$ is also computable by 1-WL+NN.

Theorem 13 (Finite Case: "GNN $\subseteq$ 1-WL+NN")**.** Let $\mathcal{C}$ be a collection of functions from $\mathcal{X}$ to $\mathbb{R}$ computable by 1-WL+NN, then $\mathcal{C}$ is also computable by GNNs.

With these, we showed the equivalence between both frameworks such that every function computed by 1-WL+NN is also computable by a GNN, and vice versa.

Notice that, we didn't leverage any constraints on the format of graphs throughout the theorems and their corresponding proves, but rather kept it general. In order to investigate the relation between both frameworks for continuous features spaces, we will first introduce an encoding of graphs that will be used throughout the proof of both the following theorems.

Definition 14. Let K be an arbitrary continuous body, we decode graphs with n nodes as a matrix $G \in K^{n \times n}$, where $G_{i,i}$ decode the node labels for $i \in [n]$, and $G_{i,j}$ with $i \neq j \in [n]$ decode edge connectivity and corresponding edge features. We say that there is an edge between node i and j if and only if $G_{i,j} \neq 0$. Further, if G encodes an undirected graph, G is symmetric.

Theorem 15 (General Case: "1-WL+NN $\subseteq$ GNN")**.** Let $\mathcal{C}$ be a collection of functions from $K^{n \times n}$ to $\mathbb{R}$ that is GNN-Approximating, than $\mathcal{C}$ is also 1-WL-Discriminating.

Theorem 16 (General Case: "GNN $\subseteq$ 1-WL+NN")**.** Let $\mathcal{C}$ be a collection of functions from $K^{n \times n}$ to $\mathbb{R}$ that is 1-WL-Discriminating, then $\mathcal{C}$ is also GNN-Approximating.

4.1 Proof of Theorem 12

We will prove Theorem 12 by introducing a couple of small lemmas, which combined prove the theorem. In detail, in Lemma 17 we show the existence of a collection computed by 1-WL+NN that is 1-WL-Discriminating. In Lemmas 18 to 20 we derive properties of 1-WL+NN functions we will use throughout Lemmas 21 to 23 with which we prove the theorem. We took great inspiration for Lemmas 21 to 23 from the proof presented in section 3.1 in the work of **?**.

Lemma 17. There exists a collection $\mathcal{C}$ of functions from $\mathcal{X}$ to $\mathbb{R}$ computable by 1-WL+NN that is 1-WL-Discriminating.

Proof. We consider the collection $\mathfrak{B}_k$ (Definition 6) of functions from $\mathcal{X}$ to $\mathbb{R}$ computed by 1-WL+NN, where we choose k as follows:

$$k := \max(\{l_G(v) \mid G \in \mathcal{X}, v \in V(G)\}),$$

the largest label of any node of any graph in $\mathcal{X}$. Note that we can compute k, since $\mathcal{X}$ is finite.

Let $G_1, G_2 \in \mathcal{X}$ such that the 1-WL isomorphism test concludes non-isomorphic ($G_1 \not\simeq_{1\text{WL}} G_2$). Let C_1, C_2 be the final coloring computed by the 1-WL algorithm when applied on G_1, G_2 respectively. Due to $G_1 \not\simeq_{1\text{WL}} G_2$, there exists a color $c \in \mathbb{N}$ such that $h_{G_1,C_1}(c) \neq h_{G_2,C_2}(c)$. If we now consider as multilayer perceptron $\text{MLP}_c : \mathbb{N}^K \to \mathbb{R}, v \mapsto W \cdot v$ with $W \in \mathbb{N}^{1\times K}$ such that $W_{1,c} := 1$ and $W_{1,i} := 0$ for all $i \in [K] \setminus \{c\}$, we can construct $\mathcal{B}$ as $\mathcal{B}(\cdot) := \text{MLP} \circ f_{\text{count}} \circ \text{1-WL}(\cdot)$, such that $\mathcal{B} \in \mathfrak{B}_k$ (K is a constant introduced by f_{count}). Then $\mathcal{B}(G_i) := h_{G,C_i}(c)$ for $i \in \{1, 2\}$, such that we can conclude $\mathcal{B}(G_1) \neq \mathcal{B}(G_2)$, and since G_1, G_2 were arbitrary chosen, we can conclude the proof. □

Lemma 18 (1-WL+NN Equivalence)**.** Let $\mathcal{C}$ be a collection of functions computable by 1-WL+NN, then for every function $\mathcal{B} \in \mathcal{C}$ and every pair of graphs $G_1, G_2 \in \mathcal{X}$: if $G_1 \simeq_{1\text{WL}} G_2$ than $\mathcal{B}(G_1) = \mathcal{B}(G_2)$.

Proof. Let $\mathcal{C}$ be a collection of functions computed by 1-WL+NN. Let $\mathcal{B}$ be an arbitrary function in $\mathcal{C}$, then $\mathcal{B}$ is comprised as follows: $\mathcal{B}(\cdot) = \text{MLP} \circ f_{\text{enc}} \circ \text{1-WL}(\cdot)$. Let $G_1, G_2 \in \mathcal{X}$ be arbitrary graphs with $G_1 \simeq_{1\text{WL}} G_2$, then by definition of the relation $\simeq_{1\text{WL}}$ we know that $\text{1-WL}(G_1) = \text{1-WL}(G_2)$. With this the equivalence follows immediatly. □

Lemma 19 (1-WL+NN Permuation Invariance)**.** Let $\mathcal{C}$ be a collection of functions computable by 1-WL+NN, then every function $\mathcal{B} \in \mathcal{C}$ is permutation-invariant.

Proof. Let $G_1, G_2 \in \mathcal{X}$ be arbitrary graphs with $G_1 \simeq G_2$ and $\mathcal{B}$ an arbitrary function computable by 1-WL+NN. Since the 1-WL algorithm is sound, we know that $G_1 \simeq G_2$ implies $G_1 \simeq_{1\text{WL}} G_2$. Using Lemma 18, we can therefore conclude that: $\mathcal{B}(G_1) = \mathcal{B}(G_2)$. □

Lemma 20 (1-WL+NN Composition)**.** Let $\mathcal{C}$ be a collection of functions computable by 1-WL+NN. Further, let $h_1, \dots h_n \in \mathcal{C}$ and $\text{MLP}^\bullet$ an multilayer perceptron, than the function $\mathcal{B}$ composed of $\mathcal{B}(\cdot) := \text{MLP}(h_1(\cdot), \dots, h_n(\cdot))$ is also computable by 1-WL+NN.

Proof. Assume the above and let $f_1, \dots, f_n$ be the encoding functions, as well as $\text{MLP}_1, \dots, \text{MLP}_n$ be the multilayer perceptrons used by $h_1, \dots h_n$ respectively. The idea of this proof is, we construct an encoding function f^* that maps a coloring C_∞ to a concatenation of the vectors obtained when applying each encoding function f_i individually. Additionally, we construct a multilayer perceptron MLP^* that takes in this concatenation of vectors and simulates all $\text{MLP}_1, \dots, \text{MLP}_n$ simultaneously on their respective section of the encoding vector of f^*, and applies afterwards the given $\text{MLP}^\bullet$ on the concatenation of the output of all MLP_i's. See Figure 3 for a sketch of the proof idea. A complete proof can be found in the Appendix, as this proof is very technical and not that interesting.

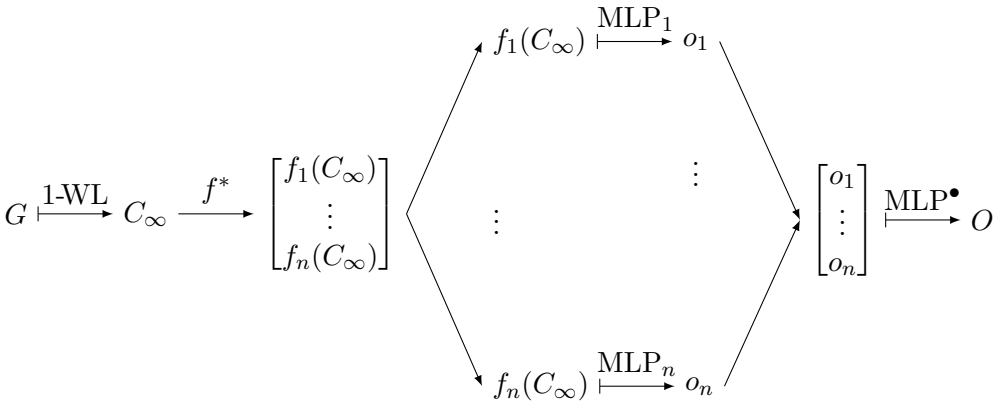


Figure 3: Proof idea for Lemma 20, how the constructed functions f^* and MLP^* will work on input G. Here we denote with C_∞ the final coloring computed by the 1-WL algorithm when applied on G. Further, we let o_i be the output computed by MLP_i on input $f_i(C_\infty)$.

□

Lemma 21. Let $\mathcal{C}$ be a collection of functions from $\mathcal{X}$ to $\mathbb{R}$ computable by 1-WL+NN that is 1-WL-Discriminating. Then for all $G \in \mathcal{X}$, there exists a function h_G from $\mathcal{X}$ to $\mathbb{R}$ computable by 1-WL+NN, such that for all $G^* \in \mathcal{X} : h_G(G^*) = 0$ if and only if $G \simeq_{1\text{WL}} G^*$.

Proof. For any $G_1, G_2 \in \mathcal{X}$ with $G_1 \not\simeq_{1\text{WL}} G_2$ let $f_{G_1,G_2} \in \mathcal{C}$ be the function distinguishing them, with $f_{G_1,G_2}(G_1) \neq f_{G_1,G_2}(G_2)$. We define the function $\overline{f}_{G_1,G_2}$ working over $\mathcal{X}$ as follows:

$$\begin{aligned} \overline{f}_{G_1,G_2}(\cdot) &= |f_{G_1,G_2}(\cdot) - f_{G_1,G_2}(G_1)| \\ &= \max(f_{G_1,G_2}(\cdot) - f_{G_1,G_2}(G_1)) + \max(f_{G_1,G_2}(G_1) - f_{G_1,G_2}(\cdot)) \end{aligned} \tag{0.1}$$

Note, that in the formula above "$f_{G_1,G_2}(G_1)$" is a fixed constant and the resulting function $\overline{f}_{G_1,G_2}$ is non-negative. Let $G_1 \in \mathcal{X}$ now be fixed, we will construct the function h_{G_1} with the desired properties as follows:

$$h_{G_1}(x) = \sum_{G_2 \in \mathcal{X},\ G_1 \not\simeq_{1\text{WL}} G_2} \overline{f}_{G_1,G_2}(x).$$

Since $\mathcal{X}$ is finite, the sum is finite and therefore well-defined. Next, we will prove that for a fixed graph $G_1 \in \mathcal{X}$, the function h_{G_1} is correct on input $G^* \in \mathcal{X}$:

1. If $G_1 \simeq_{1\text{WL}} G^*$, then for every function $\overline{f}_{G_1,G_2}$ of the sum with $G_1 \not\simeq_{1\text{WL}} G_2$, we know, using Lemma 18, that $\overline{f}_{G_1,G_2}(G^*)$ is equal to $\overline{f}_{G_1,G_2}(G_1)$ which is by definition 0, such that $h_{G_1}(G^*) = 0$.
2. If $G_1 \not\simeq_{1\text{WL}} G^*$, then $\overline{f}_{G_1,G^*}(G^*)$ is a summand of the overall sum, and since $\overline{f}_{G_1,G^*}(G^*) > 0$, we can conclude $h_{G_1}(G*) > 0$ due to the non-negativity of each function $\overline{f}$.

This function can be encoded in an MLP by replacing the max terms of the last line in Equation 0.1 by the activation function ReLU. Therefore, we can conclude with Lemma 20 that for every graph G, h_G is also 1-WL+NN computable. □

Lemma 22. Let $\mathcal{C}$ be a collection of functions from $\mathcal{X}$ to $\mathbb{R}$ computable by 1-WL+NN so that for all $G \in \mathcal{X}$, there exists $h_G \in \mathcal{C}$ satisfying $h_G(G^*) = 0$ if and only if $G \simeq_{1\text{WL}} G^*$ for all $G^* \in \mathcal{X}$. Then for every $G \in \mathcal{X}$, there exists a function φ_G computable by 1-WL+NN such that for all $G^* \in \mathcal{X}$: $\varphi_G(G^*) = \mathbb{1}_{G \simeq_{1\text{WL}} G^*}$.

Proof. Assuming the above. Due to $\mathcal{X}$ being finite, we can define for every graph G the constant:

$$\delta_G := \frac{1}{2} \min_{G^* \in \mathcal{X}, G \not\simeq_{1\text{WL}} G^*} |h_G(G^*)| > 0.$$

With this constant, we can use a so-called "bump" function working from $\mathbb{R}$ to $\mathbb{R}$ that will be similar to the indicator function. We define this function for parameter $a \in \mathbb{R}$ with $a > 0$ as:

$$\psi_a(x) := \max(\frac{x}{a} - 1,\ 0) + \max(\frac{x}{a} + 1,\ 0) - 2 \cdot \max(\frac{x}{a},\ 0).$$

The interesting property of ψ_a is that it maps every value x to 0, except when x is being drawn from the interval $(-a, a)$. In particular, it maps x to 1 if and only if x is equal to 0. See Figure 5 in the Appendix for a plot of the relevant part of this function with exemplary values for a.

We use these properties to define for every graph $G \in \mathcal{X}$ the function $\varphi_G(\cdot) := \psi_{\delta_G}(h_G(\cdot))$. We will quickly demonstrate that this function is equal to the indicator function, for this let G be fixed and G^*, an arbitrary graph from $\mathcal{X}$, the input:

1. If $G \simeq_{1\text{WL}} G^*$, then $h_G(G^*) = 0$ resulting in $\varphi_G(G^*) = \psi_{\delta_G}(0) = 1$.
2. If $G \not\simeq_{1\text{WL}} G^*$ then $h_G(G^*) \neq 0$, such that $|h_G(G^*)| > \delta_G$ resulting in $\varphi_G(G^*) = 0$.

Note that we can encode φ_G via a single MLP layer, where δ_G is a constant and the max operator is replaced by the non-linear activation function ReLU. With Lemma 20 we can therefore conclude that φ_G is computable by 1-WL+NN for every graph $G \in \mathcal{X}$. □

Lemma 23. Let $\mathcal{C}$ be a collection of functions from $\mathcal{X}$ to $\mathbb{R}$ computable by 1-WL+NN so that for all $G \in \mathcal{X}$, there exists $\varphi_G \in \mathcal{C}$ satisfying $\forall G^* \in \mathcal{X} : \varphi_G(G^*) = \mathbb{1}_{G \simeq_{1\text{WL}} G^*}$, then every permutation invariant function computable by a GNN is also computable by 1-WL+NN.

Proof. Assume the above. For any permutation invariant function $\mathcal{A}$ computed by an GNN that works over $\mathcal{X}$ to $\mathbb{R}$, we show that it can be decomposed as follows for any $G^* \in \mathcal{X}$ as input:

$$\begin{aligned}
\mathcal{A}(G^*) &= \Big(\frac{1}{|\mathcal{X}/\!\simeq_{1\text{WL}}(G^*)|} \sum_{G \in \mathcal{X}} \mathbb{1}_{G \simeq_{1\text{WL}} G^*} \Big) \cdot \mathcal{A}(G^*) \\
&= \frac{1}{|\mathcal{X}/\!\simeq_{1\text{WL}}(G^*)|} \sum_{G \in \mathcal{X}} \mathcal{A}(G) \cdot \mathbb{1}_{G \simeq_{1\text{WL}} G^*} \\
&= \sum_{G \in \mathcal{X}} \frac{\mathcal{A}(G)}{|\mathcal{X}/\!\simeq_{1\text{WL}}(G)|} \cdot \varphi_G(G^*) \qquad (0.2)
\end{aligned}$$

with $\mathcal{X}/\!\simeq_{1\text{WL}}(G^*)$ we denote the set of all graphs G over $\mathcal{X}$ that are equivalent to G^* according to the $\simeq_{1\text{WL}}$ relation.

Since $\mathcal{A}$ is permutation-invariant, and GNNs are at most as good as the 1-WL algorithm in distinguishing non-isomorphic graphs, we can use the fact that for every graph $G, H \in \mathcal{X}$ with $G \simeq_{1\text{WL}} H$: $\mathcal{A}(G) = \mathcal{A}(H)$. Therefore, we can decompose $\mathcal{A}$ as stated in Equation 0.2 in a single MLP layer with $\frac{\mathcal{A}(G)}{|\mathcal{X}/\!\simeq_{1\text{WL}}(G)|}$ being constants and $\varphi_G \in \mathcal{C}$ encoding the indicator function. Combined with the Lemma 20, we can conclude that $\mathcal{A}$ is computable by 1-WL+NN. Important to note, we can only do this since $\mathcal{X}$ is finite, making the overall sum finite and the cardinality of $\mathcal{X}/\!\simeq_{1\text{WL}}(G)$ well-defined for all graphs. □

4.2 Proof of Theorem 13

In this section we will shortly prove the converse direction. However, the proof is not as interesting as the previous one, as it is mainly just a consequence of Definition 7 of GNNs.

Proof of Theorem 13. Let $\mathcal{C}$ be a collection of functions from $\mathcal{X}$ to $\mathbb{R}$ computable by 1-WL+NN. For any $\mathcal{B} \in \mathcal{C}$ we know, by using Lemma 19 that $\mathcal{B}$ is permutation invariant such that it qualifies as a READOUT function for GNNs. By constructing a GNN with 0 layers and $\mathsf{READOUT}(\cdot) := \mathcal{B}(\cdot)$, we obtain a function $\mathcal{A}$ that is computed by this GNN. Moreover, $\mathcal{A}$ is equivalent to $\mathcal{B}$, proving that every function in $\mathcal{C}$ is computable by GNNs. □

4.3 Proof of Theorem 15

We took inspiration for this proof from the work of **?**.

Proof of Theorem 15. Let $\mathcal{C}$ be a collection of continues functions from $K^{n\times n}$ to $\mathbb{R}$ that is GNN approximating. Let $G_1, G_2 \in K^{n\times n}$ with $G_1 \not\simeq_{1\text{WL}} G_2$, then we define the function f_{G_1} on input $G^* \in \mathcal{X}$:

$$f_{G_1}(G^*) = \min_{\pi\in S_n} d(G_1,\ \pi^T G^* \pi),$$

where $d(\cdot,\ \cdot)$ is the euclidean distance between two matrices. Since f_{G_1} is permutation invariant, we can construct a GNN with 0 layers and f_{G_1} as its READOUT function, thereby constructing a GNN computing f_{G_1} and consequently showing that the function is GNN computable. We choose $\epsilon := \frac{1}{2} \cdot f_{G_1}(G_2) > 0$, then there exists a function $h_\epsilon \in \mathcal{C}$ approximating f_{G_1} within ϵ accuracy. With this, we have a function $h_\epsilon \in \mathcal{C}$ that can distinguish G_1 and G_2, since:

$$\begin{aligned} |h_\epsilon(G_1) - h_\epsilon(G_2)| &> |(f_{G_1}(G_1) - \epsilon) - (f_{G_1}(G_2) + \epsilon)| \\ &= |f_{G_1}(G_2) - 2\cdot\epsilon| && \text{by definition } f_{G_1}(G_1) = 0 \\ &= |f_{G_1}(G_2) - 2\cdot\frac{1}{2}\cdot f_{G_1}(G_2)| = 0. \end{aligned}$$

□

4.4 Proof of Theorem 16

OPEN FOR NOW: Can 1-WL+NN even compute continuous functions? The *wl* algorithm does not seem to work well over continuous graph features.

Lemma 24. Let $\mathcal{C}$ be a collection of continuous functions from $\mathcal{G}$ to $\mathbb{R}$ computable by 1-WL+NN. If $\mathcal{C}$ is 1-WL-Discriminating, then for any chosen graph $G^* \in \mathcal{G}$ and any $\epsilon \in \mathbb{R}$ with $\epsilon >$, the function h_{G^*} from $\mathcal{G}$ to $\mathbb{R}$ with the following properties:

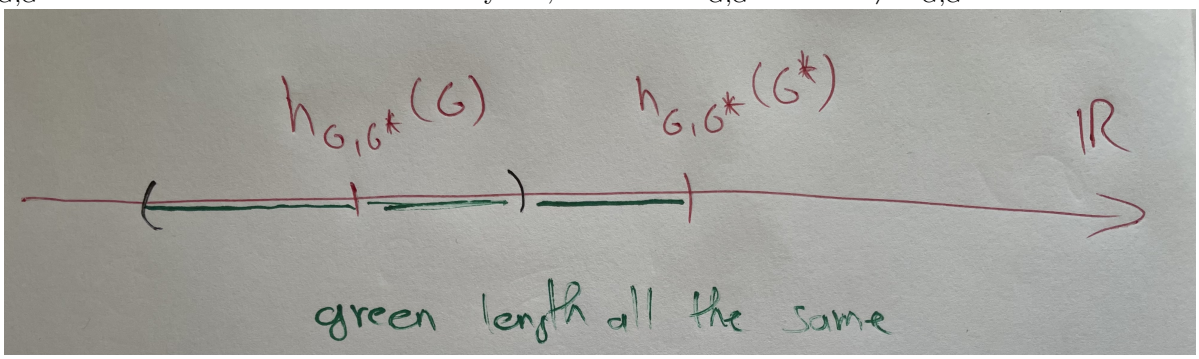
1. For all $G \in \mathcal{G} : h_{G^*}(G) \geq 0$.
2. For all $G \in \mathcal{G}$ with $G \simeq_{1\text{WL}} G^* : h_{G^*}(G) = 0$.
3. There exists a constant $\delta_{G^*} > 0$, such that for all $G \in \mathcal{G}$ with $h_{G^*}(G) < \delta_{G^*}$ there exists a graph $G' \in \mathcal{X}/\simeq_{1\text{WL}}(G)$ in the equivalence class of G such that $\|G' - G^*\|_2 < \epsilon$

One can interpret this function h_{G^*} as a kind of loss function, yielding no loss on input G if G is not distinguishable by the 1-WL algorithm from G^* ($G \simeq_{1\text{WL}} G^*$), only a small loss then G is close to G^* (the loss is upper bounded by δ_{G^*}), and a large loss otherwise.

Proof. Let $G^* \in \mathcal{G}$ be fixed and $\epsilon > 0$ be given. Since $\mathcal{C}$ is 1-WL-Discriminating, we know that for every $G \in \mathcal{G}$ with $G^* \not\simeq_{1\text{WL}} G$, there exists a function $h_{G,G^*} \in \mathcal{C}$ such that $h_{G,G^*}(G) \neq h_{G,G^*}(G^*)$. We use this property and construct for each $G \in \mathcal{G}$ with $G^* \not\simeq_{1\text{WL}} G$ a set A_{G,G^*} as follows:

$$A_{G,G^*} := \{G' \in \mathcal{G} \mid h_{G,G^*}(G') \in (h_{G,G^*}(G) \pm \frac{|h_{G,G^*}(G) - h_{G,G^*}(G^*)|}{2})\}.$$

One can use the illustration below for a better understanding, when a graph G' is contained in A_{G,G^*} and when not. Also one can easily see, that $G \in A_{G,G^*}$ and $G^* \notin A_{G,G^*}$.



Furthermore, by assumption h_{G,G^*} is continuous, A_{G,G^*} is an open set. For every $G \in \mathcal{G}$ with $G \simeq_{1\text{WL}} G^*$ we define A_{G,G^*} as follows:

$$A_{G,G^*} := \{G' \in \mathcal{G} \mid \|G' - G\|_2 < \epsilon\},$$

where $\|\cdot\|_2$ is the l_p norm with $p = 2$.

Thus $\{A_{G,G^*}\}_{G \in \mathcal{G}}$ is an open cover of $\mathcal{G}$. Since $\mathcal{G}$ is compact, there exists a finite subset $\mathcal{G}_0$ such that $\{A_{G,G^*}\}_{G \in \mathcal{G}_0}$ also covers $\mathcal{G}$. Hence, $\forall G \in \mathcal{G} \exists G_0 \in \mathcal{G}_0 : G \in A_{G_0,G^*}$.

We define the desired function h_{G^*} on input $G \in \mathcal{G}$ as follows:

$$h_{G^*}(G) = \sum_{G_0 \in \mathcal{G}_0 \setminus \mathcal{X}/\simeq_{1\text{WL}}(G^*)} \overline{h}_{G_0,G^*},$$

where we define $\overline{h}_{G_0,G^*}$ almost exactly the same as in the previous proof:

$$\begin{aligned}\overline{h}_{G_0,G^*}(\cdot) &= |h_{G_0,G^*}(\cdot) - h_{G_0,G^*}(G^*)| \\ &= \max(h_{G_0,G^*}(\cdot) - h_{G_0,G^*}(G^*)) + \max(h_{G_0,G^*}(G^*) - h_{G_0,G^*}(\cdot))\end{aligned}$$

We will shortly proof, that h_{G^*} fulfills the desired properties:

1. By construction, any $\overline{h}_{G_0,G^*}$ is non-negative, such that the sum over these functions is also non-negative.
2. If $G \simeq_{1\text{WL}} G^*$, using Lemma 18 we know that $h_{G^*}(G) = h_{G^*}(G^*)$, and by definition $h_{G^*}(G^*) = 0$, such that we can conclude $h_{G^*}(G) = 0$.
3. Let $\delta_{G^*} := \frac{1}{2}\min_{G_0 \in \mathcal{G} \setminus \mathcal{X}/\simeq_{1\text{WL}}(G^*)} |h_{G_0,G^*}(G_0) - h_{G_0,G^*}(G^*)|$. Prove by contraposition, assume that for every graph $G' \in \mathcal{X}/\simeq_{1\text{WL}}(G)$ in the equivalence class of G, $\|G' - G^*\|_2 \geq \epsilon$. Hence, $G \notin \bigcup_{G' \in \mathcal{X}/\simeq_{1\text{WL}}(G)} A_{G',G^*}$ (not in the union of l_2 balls of size ϵ around graphs of the equivalence class of G^*). However, since $\{A_{G,G^*}\}_{G \in \mathcal{G}_0}$ is a cover of $\mathcal{G}$, there must exist a $G_0 \in \mathcal{G}_0$ with $G_0 \not\simeq_{1\text{WL}} G^*$ such that $G \in A_{G_0,G^*}$. Thus, by definition of A_{G_0,G^*} we know that $h_{G_0,G^*}(G) \in (h_{G_0,G^*}(G_0) \pm \frac{|h_{G_0,G^*}(G_0) - h_{G_0,G^*}(G^*)|}{2})$, which when reformulated states:

$$|h_{G_0,G^*}(G) - h_{G_0,G^*}(G_0)| < \frac{1}{2}|h_{G_0,G_0}(G) - h_{G_0,G^*}(G^*)| \tag{0.3}$$

With this property we can prove $\overline{h}_{G_0,G^*}(G) \leq \delta_{G^*}$:

$$\begin{aligned}\overline{h}_{G_0,G^*}(G) &= |h_{G_0,G^*}(G) - h_{G_0,G^*}(G^*)| \\ &\geq |h_{G_0,G^*}(G_0) - h_{G_0,G^*}(G^*)| - |h_{G_0,G^*}(G) - h_{G_0,G^*}(G_0)| &(0.4)\\ &\geq \frac{1}{2}|h_{G_0,G^*}(G_0) - h_{G_0,G^*}(G^*)| &(0.5)\\ &\geq \frac{1}{2}\min_{G_0 \in \mathcal{G} \setminus \mathcal{X}/\simeq_{1\text{WL}}(G^*)} |h_{G_0,G^*}(G_0) - h_{G_0,G^*}(G^*)| =: \delta_{G^*}\end{aligned}$$

To understand these inequalities, it helps to visualize them, see therefor Figure 4. Further, by using Equation 0.3, we know that the blue segment is always small than half of the green segment. Then one can easily see that for the inequality used in Equation 0.4, that the green minus the blue segment is always smaller than the yellow segment. And for Equation 0.5, since we know that the blue segment is at most half the green segment, we can lower bound their difference by half the green segment.

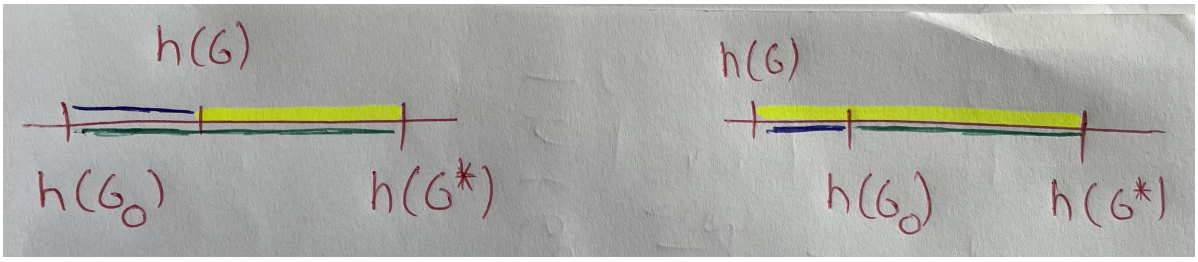


Figure 4: An illustration to better understand the above proof. Since we are using the absolute function to measure the distance, we do not need to illustrate all possible case scenarios, since they are symmetric to one of those already shown. In the table below, you find a legend, explaining which color represents which term.

Color	Term
yellow	$\lvert h_{G_0,G^*}(G) - h_{G_0,G^*}(G^*)\rvert$
green	$\lvert h_{G_0,G^*}(G_0) - h_{G_0,G^*}(G^*)\rvert$
blue	$\lvert h_{G_0,G^*}(G) - h_{G_0,G^*}(G_0)\rvert$

□

Lemma 25. Ich hab angst.

Proof. Let $\mathcal{A}$ be a continuous function from $\mathcal{G}$ to $\mathbb{R}$ computable by a GNN. Since $\mathcal{G}$ is compact, this implies that $\mathcal{A}$ is even uniformaly continuous on $\mathcal{G}$. This implies that $\forall \epsilon > 0 \exists \delta > 0$ such that $\forall G_1, G_2 \in \mathcal{G}$, if $\|G_1 - G_2|\_2 < \delta$, then $|f(G_1) - f(G_2)| < \epsilon$. Further, since $\mathcal{A}$ is GNN computable, we know that $\forall G_1, G_2 \in \mathcal{G}$, if $G' \in \mathcal{X}/\simeq_{1\text{WL}}(G_1)$ with $\|G' - G_2\|\_2 < \delta$ exists, than $|f(G_1) - f(G_2)| < \epsilon$.

Somehow talk about the not being better at distinguishin graphs!

Given any $G \in \mathcal{G}$. By assumption, we know that $h_G \in \mathcal{C}$ exists. We define the set $B_{G,a} := \{G' \in \mathcal{G} \mid h_G(G') \in [0, a)\}$, as the set of graphs that are mapped into the open interval $[0, a)$ by h_G. Then, by definition of h_G, there exists a constant δ_G such that:

$$B_{G,\delta_G} \subseteq \bigcup_{G' \in \mathcal{G},\ G' \simeq_{1\text{WL}} G} \{G'' \in \mathcal{G} \mid \|G'' - G'\|\_2 < r\}.$$

Hence, we know that $\{B_{G,a}\}_{G \in \mathcal{G}}$ is an oper cover of $\mathcal{G}$. Because $\mathcal{G}$ is compact, there exists a finite subset $G_0 \subseteq \mathcal{G}$ such that $\{B_{G_0,a}\}_{G_0 \in \mathcal{G}_0}$ also covers $\mathcal{G}$.

For each $G_0 \in \mathcal{G}_0$, we construct the function φ_{G_0} on $\mathcal{G}$ as follows:

$$\varphi_{G_0}(\cdot) := \max(\delta_{G_0} - h_{G_0}(\cdot),\ 0).$$

The function has two important properties, for once it is non-negative and for any $G \in \mathcal{G}$: $\varphi_{G_0}(G) = 0$ if and only if $G \in B_{G_0,\delta_{G_0}}$.

Further, using φ_{G_0} we construct for each $G_0 \in \mathcal{G}_0$ the function φ_{G_0} on $\mathcal{G}$ as follows:

$$\psi_{G_0}(\cdot) := \frac{\varphi_{G_0(\cdot)}}{\sum_{G' \in G_0} \varphi_{G'}(\cdot)},$$

which is well-defined, since $\{B_{G_0,a}\}_{G_0 \in \mathcal{G}_0}$ is a cover of $\mathcal{G}$, we can conclude that for any input graph G on $\psi_{G_0}(\cdot)$ there exists a $G_0 \in \mathcal{G}_0$ such that $\varphi_{G_0}(G) > 0$ making the denominator not 0.

□

5 Empirical Investigation

Yet to come!

6 Appendix

6.1 Figures and graphs

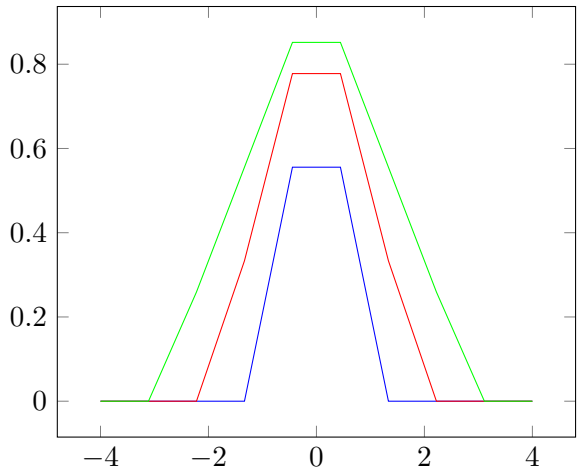


Figure 5: Illustration of the so-called "bump" function $\psi_a(x)$ used in the proof of Lemma 22. Here the colors of the displayed functions correspond to the parameter a set to $a := 1$ in blue, $a := 2$ in red and $a := 3$ in green.

6.2 Proofs

Proof of Lemma 20. Let $\mathcal{C}$ be a collection of functions computed by 1-WL+NN, $h_1, \ldots, h_n \in \mathcal{C}$, and $\text{MLP}^\bullet$ a multilayer perceptron. Further, let $f_1, \ldots, f_n$ be the encoding functions, as well as $\text{MLP}_1, \ldots, \text{MLP}_n$ be the multilayer perceptrons used by $h_1, \ldots h_n$ respectively. As outlined above, we will now construct f^* and MLP^*, such that for all graphs $G \in \mathcal{X}$:

$$\text{MLP}^\bullet(h_1(G), \ldots, h_n(G)) = \text{MLP}^* \circ f^* \circ \text{1-WL}(G)$$

such that we can conclude that the composition of multiple functions computable by 1-WL+NN, is in fact also 1-WL+NN computable.

We define the new encoding function f^* to work as follows on input C_∞:

$$f^*(C_\infty) := \mathsf{concat}(\begin{bmatrix} f_1(C_\infty) \\ \vdots \\ f_n(C_\infty) \end{bmatrix}),$$

where conccat is the concatenation functions, concatenating all encoding vectors to one single vector.

Using the decomposition introduced in Definition 8, we can decompose each MLP_i at layer $j > 1$ as follows: $(\text{MLP}_i)_j(v) := \sigma(W_j^i \cdot (\text{MLP}_i)_{j-1}(v) + b_j^i)$. Using this notation we construct

MLP* as follows:

$$
\begin{aligned}
&(\text{MLP}^*)_1(v) := v \\
&(\text{MLP}^*)_{j+1}(v) := \sigma(W_j^* \cdot (\text{MLP}^*)_j(v) + \mathsf{concact}(\begin{bmatrix} b_j^1 \\ \vdots \\ b_j^n \end{bmatrix})) && ,\forall j \in [k] \\
&(\text{MLP}^*)_{j+k+1}(v) := (\text{MLP}^\bullet)_{j+1}(v) && ,\forall j \in [k^\bullet - 1]
\end{aligned}
$$

where k is the maximum number of layers of the set of MLP_i's, $k^\bullet$ is the number of layers of the given MLP$^\bullet$ and σ an element wise activation function. Thereby, we define in the first equation line, that the start of the sequence is the input, with the second line, we construct the "simultaneous" execution of the MLP_i's, and in the last equation line, we add the layers of the given MLP$^\bullet$ to the end. Further, we define the weight matrix W_j^* as follows:

$$
W_j^* := \begin{bmatrix} W_j^1 & 0 & \dots & 0 \\ 0 & W_j^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & W_j^n \end{bmatrix},
$$

such that we build a new matrix where each individual weight matrix is placed along the diagonal. Here we denote with 0, zero matrices with the correct dimensions, such that W_j^* is well-defined. Important to note, should for an MLP_i, W_j^i not exist, because it has less than j layers, we use for W_j^i the identity matrix I_m where m is the dimension of the output computed by MLP_i.

Then $\mathcal{B}(\cdot) := \text{MLP}^* \circ f^* \circ \text{1-WL}(\cdot)$ is 1-WL+NN computable and equavalent to $\text{MLP}^\bullet(h_1, \dots, h_n)$ □

fine tune.