Rademacher Meets Colors: More Expressivity, *but at What Cost*?

Martin Carrasco[*]
University of Fribourg
martin.carrrascocastaneda@unifr.ch

Caio F. Deberaldini Netto[*]
Johns Hopkins University
cnetto1@jh.edu

Vahan A. Martirosyan[*]
Université Paris-Saclay
vahan.martirosyan@centralesupelec.fr

Aneeqa Mehrab[*]
University of Ferrara
aneeqa.mehrab@unife.it

Ehimare Okoyomon[*]
Technical University of Munich
e.okoyomon@tum.de

Caterina Graziani
University of Siena
caterina.graziani2@unisi.it

Abstract

The expressive power of graph neural networks (GNNs) is typically understood through their correspondence with graph isomorphism tests such as the Weisfeiler–Leman (WL) hierarchy. While more expressive GNNs can distinguish a richer set of graphs, they are also observed to suffer from higher generalization error. This work provides a theoretical explanation for this trade-off by linking expressivity and generalization through the lens of coloring algorithms. Specifically, we show that the number of equivalence classes induced by WL colorings directly bounds the GNN's Rademacher complexity – a key data-dependent measure of generalization. Our analysis reveals that greater expressivity leads to higher complexity and thus weaker generalization guarantees. Furthermore, we prove that the Rademacher complexity is stable under perturbations in the color counts across different samples, ensuring robustness to sampling variability across datasets. Importantly, our framework is not restricted to message-passing GNNs or 1-WL, but extends to arbitrary GNN architectures and expressivity measures that partition graphs into equivalence classes. These results unify the study of expressivity and generalization in GNNs, providing a principled understanding of why increasing expressive power often comes at the cost of generalization.

1 Introduction

Graph Neural Networks (GNNs) [26, 11] have shown great success in learning tasks across many domains such as social networks, knowledge graphs, and chemistry [33]. This empirical success has sparked a growing interest in understanding the theoretical capabilities of GNNs, leading to the characterization of their expressive power – a measure of the model's ability to discriminate non-isomorphic graphs. One fundamental insight into GNNs' expressivity is the relationship between these model classes and the Weisfeiler-Leman (WL) graph isomorphism tests. Previous research from Xu et al. [30] and Morris et al. [22] established that message-passing GNNs (MPGNN) are at most as powerful as the 1-dimensional WL test (1-WL), highlighting a fundamental limitation in

[*]These authors contributed equally to this work.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: New Perspectives in Advancing Graph Machine Learning.

the expressivity of these model architectures. Since then, this relationship has been extended to a range of more expressive GNN variants, [27, 7, 4, 24, 32, 5, 14, 3, 1], each one endowed with the corresponding WL test. While expressivity is a meaningful and active focus of GNN research, it offers limited insight into the fundamental issue of an architecture's ability to generalize to graphs outside of its training set.

Recent efforts have begun to connect this characterization of expressivity to generalization theory. Morris et al. [23] provided the first direct connection between the GNN expressivity and the VC dimension by showing that the VC dimension is tightly related to the number of graphs that can be distinguished by 1-WL. For GNNs with piecewise activation functions and in settings where an upper bound on the graphs' order (number of vertices) is known, they proved that the VC dimension equals the maximum number of pairwise 1-WL-distinguishable graphs, while for graphs with bounded individual color complexity, they derived bounds of $\mathcal{O}(P \log(puP))$, where P is the number of parameters, u is the number of node colors, and p is the number of pieces of the activation functions. D'Inverno et al. [8] extended these VC dimension analyses to GNNs with Pfaffian activation functions (such as tanh, sigmoid), providing bounds that also depend on the maximum number of node colors per graph. Further related work connecting expressivity and generalization is provided in Appendix A.

Other contributions have investigated using Rademacher Complexity as a more refined and data-dependent approach to bounding generalization. Garg et al. [10] provided the first Rademacher complexity bounds for message-passing GNNs, explicitly accounting for the local permutation invariance of GNNs. Their bounds are tighter than existing VC dimension guarantees, but depend solely on the parameters of the given GNN architecture, leaving the connection between the WL color distributions and the Rademacher complexity unexplored.

Our work directly addresses this gap by using coloring algorithms to relate the expressive power of GNNs to their Rademacher complexity, thus providing a theoretical justification for the observed trade-off between expressive power and generalization performance. The paper's main contributions are summarized as follows:

1. We derive a novel upper bound on the empirical Rademacher complexity of GNNs in terms of the number of equivalence classes induced by the graph coloring function. Moreover, the bound is tight under the assumption that all the classes have the same cardinality.
2. Our results are general, applying not only to the 1-WL algorithm or a particular GNN class, but to any GNN architecture together with its associated coloring function.
3. Last, we establish stability guarantees to show that our bounds remain reliable even when a *similar* sample set is used to calculate the empirical Rademacher complexity.

This connection unifies expressivity and generalization, extending the previous analyses to arbitrary GNN architectures, and provides a more comprehensive view of the performance of GNN models.

The paper is organized as follows. Section 2 introduces the necessary notation and preliminaries. Section 3 presents our main theoretical results, establishing a connection between expressivity and generalization via coloring functions, including a stability analysis (Section 3.1) and generalization to arbitrary coloring schemes (Section 3.2). Section 4 discusses limitations and outlines directions for future work, while Section 5 draws preliminary conclusions. Further related work, definitions, technical material, and proofs can be found in Appendix A and Sections B to E.

2 Notation and Preliminaries

For $n \geq 1$, let $[n] := \{1, 2, \ldots, n\}$. We use $\{\!\{\ldots\}\!\}$ to denote multisets, i.e., the generalization of sets allowing for multiple instances of each of their elements.

Graphs. A graph $G = (V, E)$ is a pair with finite set of vertices or nodes V and edges $E \subseteq \{\{u, v\} \subseteq V \mid u \neq v\}$. For ease of notation, we denote the edge $\{u, v\}$ in E by (u, v) or (v, u). If not otherwise stated, we set $n := |V|$, and the graph is of order n. Let $\mathcal{N}(v)$ be the *neighborhood* of a node $v \in \mathcal{V}$, i.e. the set of all nodes adjacent to v, and $d(v)$ the *degree* of a node $v \in \mathcal{V}$, i.e., the number of neighbors $|\mathcal{N}(v)|$. An **attributed graph** $G = (V, E, \alpha)$ is a triple with a graph (V, E) and node-attribute function $\alpha : V \to A$, where A is a a finite subset of $\mathbb{R}^d$, for some $d > 0$. We consider the space of finite, simple, undirected, attributed graphs, denoted by $\mathcal{G}$.

Graph Neural Networks. Message-passing GNNs (MPGNNs) learn real-valued vectors, called *embeddings*, for each node by iteratively updating their features based on aggregated information from their neighbors. Specifically, the embedding $h_\ell(v)$ for node v at layer ℓ is computed as:

$$h_\ell(v) = \texttt{COMBINE}^{(\ell)}\left(h_{\ell-1}(v), \texttt{AGGREGATE}^{(\ell)}\left(\{\!\{h_{\ell-1}(u)\}\!\}_{u\in\mathcal{N}(v)}\right)\right). \tag{1}$$

After L layers, we obtain the final node embedding, which we denote as $h_L(v)$. For graph-level tasks, these are aggregated into a graph representation $h_L(G)$. The full architectural details are in Appendix B.

Expressivity. The expressive power of MPGNNs is studied via their capability to distinguish non-isomorphic graphs. Research has shown that MPGNNs are at most as powerful as the Weisfeiler-Leman (1-WL) test, a well-known isomorphism heuristic [30, 22, 29]. The 1-WL test works by partitioning the nodes of a graph into equivalence classes, where equivalent nodes are assigned the same color based on their neighborhood structure. At each iteration ℓ, the color $c_\ell(v)$ of a node v is updated by hashing its previous color with the multiset of its neighbors' colors:

$$c_\ell(v) = \texttt{HASH}\Big(c_{\ell-1}(v), \{\!\{c_{\ell-1}(u)\}\!\}_{u\in\mathcal{N}(v)}\Big). \tag{2}$$

This process continues until the partitioning is stable. We denote by $c(v)$ the color of node v at convergence of the partitions, that is, $c(v) := c_L(v)$ where L is the first iteration after which the partition no longer changes. The full algorithm is described in Appendix B.

The 1-WL induces an equivalence relation $\overset{WL}{\sim}$ on nodes, such that $u \overset{WL}{\sim} v \Leftrightarrow c(u) = c(v)$.

We define the color of a graph G, or *color histogram* of G, the multiset of colors of its nodes:

$$c(G) = \{\!\{c(v)\}\!\}_{v\in V}. \tag{3}$$

The set of graph colors is denoted by $\mathcal{GC}$. To test whether two graphs are isomorphic, 1-WL is applied to both graphs. If the colors of the two graphs differ, i.e., the graphs have a different number of nodes with the same color, the graphs are non-isomorphic. If the colors are the same, the algorithm is inconclusive, meaning that the two graphs may be, but are *not guaranteed*, isomorphic. More concretely, given two graphs $G = (V, E)$ and $G' = (V', E')$ we can define the equivalence relation induced by 1-WL on graphs as:

$$G \overset{\text{WL}}{\equiv} G' \Leftrightarrow c(G) = c(G') \Leftrightarrow \{\!\{c(v)\}\!\}_{v\in V} = \{\!\{c(u)\}\!\}_{u\in V'}. \tag{4}$$

Remark 2.1. Comparing Eq. (1) and Eq. (2) reveals that they share the same structure: both the updates rely on combining a node's features (or color), with the features (or colors) of its neighbours. WL test has been proved to be an upper bound for GNNs' expressivity [30, 22]. This means that if two nodes have the same color, they must also have the same embedding:

$$c(u) = c(v) \implies h_L(u) = h_L(v). \tag{5}$$

The converse holds when $\texttt{COMBINE}^{(\ell)}$ and $\texttt{AGGREGATE}^{(\ell)}$ (see Eq. 1) are injective functions [30] and for graph level tasks this requires that $\texttt{READOUT}^{(\ell)}$ is also injective.

Throughout the paper, we adopt this correspondence between MPGNNs and 1-WL as a running example to support the exposition and enhance readability. Nonetheless, the presented results hold in general for arbitrary GNN architectures along with the coloring test that upper bounds their expressive power (see Section 3.2).

Generalization. We briefly review the main definitions in the theory of Rademacher complexity. We invite the reader to consult Mohri et al. [21] for a comprehensive treatment of the topic. Let $S = \{(G_i, y_i)\}_{i\in[m]} \sim \mathcal{D}^m$ be a dataset composed of m i.i.d. samples which we assume are drawn from an underlying distribution $\mathcal{D}$ on $\mathcal{G}\times\mathcal{Y}$, with $\mathcal{Y} = \{-1, +1\}$. Sometimes we subsume the set $\mathcal{Y}$, writing $S = \{G_1, \dots G_m\}$. For any fixed GNN architecture, we assume a hypothesis class

$$\mathcal{F} = \{f \colon \mathcal{G} \to [-1, 1]\}$$

of possible graph-level functions that can be learned by this GNN. In the following, $f(G;\Theta)$ is the output of a function parametrized by Θ under a fixed GNN architecture.

Given a loss function ℓ that measures the prediction error, we define for each $f \in \mathcal{F}$ the empirical and true (or population) risk, respectively, by

$$L_S(f) = \frac{1}{m}\sum_{j=1}^{m} \ell(f(G_j;\Theta), y_j), \quad L(f) = \mathbb{E}_{(G,y)\sim\mathcal{D}}[\ell(f(G;\Theta), y)]. \tag{6}$$

The generalization error is defined as the difference between the true and empirical risk, and it is bounded by the model complexity. To quantify complexity, let $\sigma_1, \ldots, \sigma_m$ be independent Rademacher variables and define the *empirical* Rademacher complexity

$$\mathcal{R}_S(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{j=1}^{m} \sigma_j f(G_j;\Theta) \right], \tag{7}$$

with *population* counterpart

$$\mathcal{R}_m(\mathcal{F}) = \mathbb{E}_{S\sim\mathcal{D}^m}\left[\mathcal{R}_S(\mathcal{F})\right]. \tag{8}$$

Rademacher complexity measures how well a function class $\mathcal{F}$ can correlate with random noise. High Rademacher complexity indicates that there exists a function in $\mathcal{F}$ that is potentially "overfitting" the labels. The Rademacher complexity can be used to bound the generalization error of a hypothesis class, as formalized in the next result.

Lemma 2.2 (Mohri et al. [Theorem 3.3])**.** *For any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $f \in \mathcal{F}$ and any loss function ℓ:*

$$L(f) \leq L_S(f) + 2\mathcal{R}_S(\ell \circ \mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}, \tag{9}$$

where $\ell \circ \mathcal{F}$ denotes the standard function composition, i.e., $\ell \circ \mathcal{F} := \{\ell(f(G;\Theta), y) \mid f \in \mathcal{F}\}$.

Moreover, if the loss function $\ell\colon [-1,1]^2 \to \mathbb{R}$ is Lipschitz with constant γ (relative to any norm-induced metric), then $\mathcal{R}_S(\ell \circ \mathcal{F}) \leq \gamma\,\mathcal{R}_S(\mathcal{F})$. This result is known as Talagrand's contraction lemma.

Combining Lemma 2.2 and Talagrand's lemma, we claim that it suffices to bound the empirical Rademacher complexity $\mathcal{R}_S(\mathcal{F})$ to bound the generalization error of the class $\mathcal{F}$.

Proposition 2.3. *Let ℓ be a Lipschitz loss function, of constant γ. For any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $f \in \mathcal{F}$:*

$$L(f) \leq L_S(f) + 2\gamma\,\mathcal{R}_S(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}. \tag{10}$$

Some examples of such Lipschitz loss functions in the context of (graph) classification are the logistic loss (*log loss*), the cross-entropy (CE) when applied to the output of a softmax layer [19] or to the output of a logistic function when its input is bounded (check Appendix E.1), and a margin loss [10]. Particularly for this work, since our hypothesis class is $\mathcal{F} = \{f\colon \mathcal{G} \to [-1,1]\}$, we can either use directly the latter loss with an activation function that gives outputs in the interval $[-1,1]$ (e.g., *tanh*), or combine the GNN's output with a linear transformation $[-1,1] \to [0,1]$ and use one of the other two loss functions (see Appendix E.2). The next section explores the natural connection between Rademacher complexity and expressivity.

3 Rademacher Meets Colors

The connection between expressivity and generalization can be drawn by relating the coloring algorithm characterizing a GNN's expressive power to the Rademacher complexity of its hypothesis class. Coloring algorithms *partition* the sample $S = \{G_1, \ldots, G_m\}$ in p disjoint sets $I_1, \ldots, I_p$, where each I_j is an equivalence class containing all graphs with the same color c_j[1]. This imposes structural constraints on the function class: any function f implementable by the architecture must be constant over equivalence classes. As a consequence, this limits

[1]We use the terms (graph) colors and equivalence classes interchangeably.

its possibility to overfit arbitrary labels, since not all labelings are compatible with the partitioning. For ease of presentation, we first introduce our results in the familiar setting of message-passing GNNs and their connection to the 1-WL coloring algorithm. This framework, however, extends beyond the familiar WL setting; we refer the reader to Section 3.2 for further details.

First, we extend Definition B.1 to describe how the graph-level output $f(G;\Theta)$ is computed, thereby specifying the hypothesis class $\mathcal{F}$ under consideration. Let $h_L(G) \in \mathbb{R}^d$ be the global embedding of the graph G, obtained by combining the node embeddings using a `READOUT` function such as *sum*, *max*, or *mean*. Then, the GNN output $f(G;\Theta)$ is computed by applying an activation function (e.g., the hyperbolic tangent) $\psi(\cdot)$ to the linearly transformed graph embedding:

$$f(G;\Theta) = \psi(\boldsymbol{\beta}^\intercal h_L(G)) \in [-1,1], \tag{11}$$

where $\boldsymbol{\beta} \in \mathbb{R}^d$ is a trainable parameter. The following result bounds the Rademacher complexity of message-passing GNNs in terms of the number of graph colors p.

Proposition 3.1. *Let $S = \{G_1, \ldots, G_m\}$ be a sample of m graphs, partitioned into p disjoint sets $\{I_1, \ldots, I_p\}$ by a coloring function. Let $\mathcal{F}$ be a class of functions whose output $f(G;\Theta)$ is the same on each graph of a fixed class $G \in I_j$. The empirical Rademacher complexity of $\mathcal{F}$ on S is bounded by:*

$$\mathcal{R}_S(\mathcal{F}) \leq \frac{\sup_\Theta L(\Theta)\sqrt{p}}{m} \tag{12}$$

where $L(\Theta) = \sqrt{\sum_{i=1}^m f(G_i;\Theta)^2}$ is the ℓ_2-norm of the function's outputs over the sample S.

The proof can be found in the Appendix D.1.

The previous result holds for a general class of functions $\mathcal{F} = \{f : \mathcal{G} \to \mathbb{R}\}$ where the output respects, for example, the 1-WL equivalence. Usually, when studying Rademacher complexity, the focus is restricted to functions with bounded outputs. Without such an assumption, the Rademacher complexity may become infinite, in which case the resulting generalization bounds are meaningless. The following corollary explores the case where f maps to the interval $[-1,1]$.

Corollary 3.2. *Under the assumptions of Proposition 3.1 and in the special case where every function $f \in \mathcal{F}$ maps to $[-1,1]$, the empirical Rademacher complexity is bounded by:*

$$\mathcal{R}_S(\mathcal{F}) \leq \sqrt{\frac{p}{m}} \tag{13}$$

Proof. The result follows from Prop. 3.1, noting that if the output space is $[-1,1]$, then $\sup_\Theta \sqrt{\sum_{i=1}^m f(G_i;\Theta)^2} = \sqrt{m}$. □

The bound in Corollary 3.2 scales as $\sqrt{p/m}$, where m is the number of graphs in the sample and p is the number of equivalence classes, e.g. those produced by 1-WL. Intuitively, p measures the diversity of the graphs within the sample S with respect to the 1-WL test: the more equivalence classes there are, the more heterogeneous the dataset appears to the GNN. For example, the smallest possible bound is met in the extreme case when $p = 1$, namely when the graphs are indistinguishable under the 1-WL – for instance, when all graphs are regular and have the same order. As p grows, the dataset becomes more complex, and the bound increases accordingly, reflecting a higher risk of overfitting. Moreover, since any MPGNN is at most as expressive as the 1-WL test, this result provides a unifying upper bound for a broad family of architectures and it extends beyond that (see Section 3.2). Finally, the dependence on $1/\sqrt{m}$ matches standard learning-theoretic intuition: increasing the sample size tightens the bound regardless of architectural expressivity.

Under the assumption that all the equivalence classes have the same cardinality, the bound of Corollary 3.2 is proven to be tight and asymptotically correct:

Proposition 3.3 (Uniform partitioning assumption)**.** *Let $S = \{G_1, \ldots, G_m\}$ be a sample of m graphs, partitioned by a coloring function into p disjoint sets $\{I_1, \ldots, I_p\}$ of the same cardinality. Let $\mathcal{F}$ be a class of functions whose output $f(G;\Theta)$ is a constant value $f_j(\Theta)$ for each graph in a particular I_j, and the sign of $f_j(\Theta)$ is arbitrary for each partition, i.e., $sign(f_j(\Theta)) \in \{-1,1\}$, $\forall j$. Then, the empirical Rademacher complexity of $\mathcal{F}$ on S is bounded by:*

$$\mathcal{R}_S(\mathcal{F}) \geq \sqrt{\frac{p}{2m}}.$$

The proof can be found in Appendix D.2.

Proposition 3.3 shows that the bound of Corollary 3.2 is tight under uniform partitioning, i.e., when all color classes have the same cardinality. In this case, the empirical Rademacher complexity admits both upper and lower bounds of the same order, that is, $\mathcal{R}_S(\mathcal{F}) = O(\sqrt{p/m})$.

Last, leveraging Corollary 3.2, we can improve the Dudley entropy integral bound (see Theorem C.2) on the empirical Rademacher complexity [2] by incorporating the number of graph colors in the inequality. Let $\mathcal{N}(\mathcal{F}_{|S}, \epsilon, ||\cdot||_2)$ be the covering number of $\mathcal{F}_{|S}$ [2] at radius ϵ under $||\cdot||_2$ (the ℓ_2 norm). Then:

Proposition 3.4. *Let $S = \{G_1, \ldots, G_m\}$ be a sample of m graphs, partitioned into p disjoint sets $\{I_1, \ldots, I_p\}$ by a coloring function. Let $\mathcal{F}$ be a class of functions $f : \mathcal{G} \to [-1, 1]$ whose output is the same on each graph of a fixed I_j. Assume $\mathbf{0} \in \mathcal{F}$. The empirical Rademacher complexity of $\mathcal{F}$ on S is bounded by:*

$$\mathcal{R}_S(\mathcal{F}) \leq \inf_{\alpha>0} \left(\frac{4\alpha\sqrt{p}}{m} + \frac{12}{m} \int_{\alpha}^{\sqrt{m}} \sqrt{\log \mathcal{N}(\mathcal{F}_{|S}, \epsilon, ||\cdot||_2)} d\epsilon \right) \tag{14}$$

where the bound is reduced due to the p-dimensional structure of the output space $\mathcal{F}_{|S}$.

The proof can be found in Appendix D.3 and is very similar to the proof of Lemma A.5 from [2].

Relative to prior work [2], the first term in the bound is tighten from $1/\sqrt{m}$ to $\sqrt{p}/m$, yielding a concrete improvement and an explicit characterization in terms of graph colors. Furthermore, this bound applies generally to all GNN architectures but can be further refined for a specific function class by bounding the covering number of $\mathcal{F}_{|S}$. We refer the interested reader to Garg et al. [10] for an example of such a covering number bound for message-passing GNNs.

3.1 Stability of Rademacher Complexity under color perturbation

The previous results bound the Rademacher complexity of a class of functions $\mathcal{F}$ on a fixed sample S. However, we are also interested in how the change in samples affects the complexity of $\mathcal{F}$. We show our Rademacher complexity bounds remain meaningful under noisy perturbations, and in particular that $\mathcal{R}_S(\mathcal{F})$ is Lipschitz-continuous in the underlying color counts. Concretely, if each color's count shifts by at most ϵ_j, then the resulting change in the empirical Rademacher complexity scales only linearly in ϵ_j.

Proposition 3.5. *Let*

$$S = \{G_1, \ldots, G_m\} \quad \textit{and} \quad S' = \{G'_1, \ldots, G'_m\}$$

be two samples of size m. Applying a coloring procedure to both samples yields two sets of colors, and we denote by $\mathcal{GC}$ their union. Suppose that for every color $c_j \in \mathcal{GC}$, the number of graphs with color c_j in the two samples differs by at most ϵ_j:

$$|\mu_j(S) - \mu_j(S')| \leq \epsilon_j. \tag{15}$$

Then the empirical Rademacher complexities of $\mathcal{F}$ on these two samples satisfy:

$$|\mathcal{R}_S(\mathcal{F}) - \mathcal{R}_{S'}(\mathcal{F})| \leq \sum_{c_j \in \mathcal{GC}} \frac{\epsilon_j}{m} \tag{16}$$

The proof of Proposition 3.5 is deferred to Appendix D.5.

Proposition 3.5 guarantees that our generalization bounds are robust to domain shifts between datasets that are close in the color space. Bridging to the results from the previous section, as the number of colors p increases, the Rademacher complexity grows; however, this growth is smooth because the empirical Rademacher complexity is Lipschitz-continuous with respect to color perturbations.

[2] The notation $\mathcal{F}_{|S}$ denotes the *restriction* of the class of functions $\mathcal{F}$ to functions with domain in S.

3.2 Extension to arbitrary coloring functions

In this work, we connect expressivity and generalization through the lens of coloring algorithms. Our analysis is not restricted to 1-WL, but applies to any pair $(\mathcal{A}, T)$ where $\mathcal{A}$ is a GNN architecture and T is the coloring algorithm that bounds its expressive power. With a small abuse of notation we write $\mathcal{A} \sqsubseteq T$. For example, MPGNN $\sqsubseteq$ 1-WL. Given such a pair $(\mathcal{A}, T)$, the coloring algorithm T partitions the space of graphs into p_T equivalence classes, such that two graphs with the same color belong to the same partition and get the same output from the architecture $\mathcal{A}$. All results presented in the previous section for (MPGNN, 1-WL) immediately extend to this general case by replacing p with p_T. This observation enables us to compare the generalization abilities of architectures with different expressive power. Indeed, if $T \sqsubseteq S$, then $p_T \leq p_S$, which implies that the bound on the Rademacher complexity is larger for S (see Fig. 1 for a visual representation of the comparison between 1-WL and k-WL). This confirms the common belief that expressivity comes *at the cost of* generalization power. The key observation is that this holds for *any expressivity measurement* which relies on partitioning onto equivalence classes. As a consequence, more expressive GNNs (e.g., k-GNNs [22], CW networks [4], Subgraph GNNs [31] or Path GNNs [12]) correspond to a larger p, leading to higher Rademacher complexity bounds and thus to an increased risk of overfitting.

4 Limitations and Future Work

Our theoretical contributions focus on graph-level binary classification, which is consistent with the current state of the art in generalization theory [28], and aligns well with the most frequently encountered scenario in GNN benchmarking. Nonetheless, future work includes extending our generalization bounds to cover a wider range of graph learning tasks, including multi-class evaluation, regression, and node-level bounds. Given known extensions of Rademacher complexity to multi-class classification and bounded regression problems [16, 21], we anticipate adapting those techniques to broaden the theoretical scope of our results.

Another natural extension of our framework would be to replace the discrete partitioning induced by WL colorings with pseudometric-based notions of structural similarity, in the spirit of Maskey et al. [20]. This would enable finer-grained expressivity analyses and potentially tighter generalization bounds. In parallel, we aim to empirically investigate how well these Rademacher-based bounds apply in practice. We will conduct systematic studies across several GNN function classes and benchmark tasks to verify how the number of distinct colors influences GNN generalization performance.

Lastly, our current analysis bounds the empirical Rademacher complexity without making assumptions on the underlying graph distribution. As a future direction, we aim to study how the *true* Rademacher behaves when graphs are sampled from known probabilistic models, such as Random Graph Models (RGMs), and to further extend this analysis to general limiting objects such as graphons [18], which generate these models. For instance, a particularly interesting future direction is to analyze how graph size affects the number of distinct colors in samples from graphons, and to investigate the asymptotic behavior of our theory in this setting.

5 Conclusions

This study leverages Rademacher complexity to draw a direct connection between the generalization ability of GNNs and their expressivity. The resulting bounds depend on the number of partitions induced by an arbitrary coloring algorithm, making the framework broadly applicable across different architectures and expressivity measures. We show that more expressivity comes *at the cost* of a higher upper bound on the models' generalization error, implying that less expressive models are at a *lower risk* of overfitting than more expressive ones. Moreover, for any fixed coloring algorithm, we show that the change in Rademacher complexity between two samples scales linearly with the difference in their color multiplicities. Ultimately, this work highlights an inherent interplay between expressivity and generalization, motivating further analysis of the generalization power of expressive methods.

Author Contributions. All authors contributed to the final version of this paper. Specific contributions [3] in what follows.
Martin Carrasco: Conceptualization; Visualization; Writing; Review, and Editing.
Caio F. Deberaldini Netto: Conceptualization; Formal analysis (Prop. E.1, Prop. E.2); Writing; Review and Editing.
Vahan A. Martirosyan: Conceptualization; Formal analysis (Prop. 3.1, Cor. 3.2, Prop. 3.4); Review and Editing.
Aneeqa Mehrab: Contributed to the Writing.
Ehimare Okoyomon: Conceptualization; Writing; Review and Editing.
Caterina Graziani: Conceptualization; Supervision; Formal analysis (Prop. 3.3, Prop. 3.5); Writing; Review and Editing.

Acknowledgments and Disclosure of Funding

We are deeply grateful to Alessandro Micheli, who contributed from the early stages of this project and provided the original stability result that inspired our extension to graphs. We also thank Franco Scarselli for his continuous support and precious feedback. Last, we would like to acknowledge the 2025 London Geometry and Machine Learning Summer School (LOGML), where this research project started. In particular, we would like to express our gratitude to the members of the organizing committee: Vincenzo Marco De Luca, Massimiliano Esposito, Simone Foti, Valentina Giunchiglia, Daniel Platt, Pragya Singh, Arne Wolf, and Zhengang Zhong.

References

[1] Ralph Abboud, Radoslav Dimitrov, and Ismail Ilkan Ceylan. Shortest path networks for graph property prediction. In *Learning on graphs conference*, pages 5–1. PMLR, 2022. 2

[2] Peter L Bartlett, Dylan J Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 6240–6249, 2017. 6, 12, 17

[3] Beatrice Bevilacqua, Fabrizio Frasca, Derek Lim, Balasubramaniam Srinivasan, Chen Cai, Gopinath Balamurugan, Michael M Bronstein, and Haggai Maron. Equivariant subgraph aggregation networks. *arXiv preprint arXiv:2110.02910*, 2021. 2

[4] Cristian Bodnar, Fabrizio Frasca, Nina Otter, Yuguang Wang, Pietro Lio, Guido F Montufar, and Michael Bronstein. Weisfeiler and lehman go cellular: Cw networks. *Advances in neural information processing systems*, 34:2625–2640, 2021. 2, 7

[5] Cristian Bodnar, Fabrizio Frasca, Yuguang Wang, Nina Otter, Guido F Montufar, Pietro Lio, and Michael Bronstein. Weisfeiler and lehman go topological: Message passing simplicial networks. In *International conference on machine learning*, pages 1026–1037. PMLR, 2021. 2

[6] Jan Böker. Graph similarity and homomorphism densities. *arXiv preprint arXiv:2104.14213*, 2021. 11

[7] Giorgos Bouritsas, Fabrizio Frasca, Stefanos Zafeiriou, and Michael M Bronstein. Improving graph neural network expressivity via subgraph isomorphism counting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):657–668, 2022. 2

[8] Giuseppe Alessio D'Inverno, Monica Bianchini, and Franco Scarselli. Vc dimension of graph neural networks with pfaffian activation functions. *Neural Networks*, 182:106924, 2025. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2024.106924. URL https://www.sciencedirect.com/science/article/pii/S0893608024008530. 2

[9] Billy J Franks, Christopher Morris, Ameya Velingker, and Floris Geerts. Weisfeiler-leman at the margin: When more expressivity matters. *arXiv preprint arXiv:2402.07568*, 2024. 11

[3] Inspired by the Contributor Role Taxonomy at https://credit.niso.org.

[10] Vikas Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks. In *International conference on machine learning*, pages 3419–3430. PMLR, 2020. 2, 4, 6

[11] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Message passing neural networks. In *Machine learning meets quantum physics*, pages 199–214. Springer, 2020. 1

[12] Caterina Graziani, Tamara Drucks, Fabian Jogl, Monica Bianchini, Franco Scarselli, T Gartner, et al. The expressive power of path-based graph neural networks. *Proceedings of the 41st International Conference on Machine Learning*, 235:16226–16249, 2024. 7

[13] Uffe Haagerup. The best constants in the Khintchine inequality. *Studia Mathematica*, 70(3): 231–283, 1981. 15

[14] Lecheng Kong, Yixin Chen, and Muhan Zhang. Geodesic graph neural network for efficient graph representation learning. *Advances in neural information processing systems*, 35:5896–5909, 2022. 2

[15] Ron Levie. A graphon-signal analysis of graph neural networks. *Advances in Neural Information Processing Systems*, 36:64482–64525, 2023. 11

[16] Jian Li, Yong Liu, Rong Yin, Hua Zhang, Lizhong Ding, and Weiping Wang. Multi-class learning: From theory to algorithm. *Advances in Neural Information Processing Systems*, 31, 2018. 7

[17] Shouheng Li, Floris Geerts, Dongwoo Kim, and Qing Wang. Towards bridging generalization and expressivity of graph neural networks. *arXiv preprint arXiv:2410.10051*, 2024. 11

[18] László Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012. 7

[19] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pages 23803–23828. pmlr, 2023. 4

[20] Sohir Maskey, Raffaele Paolino, Fabian Jogl, Gitta Kutyniok, and Johannes F Lutzeyer. Graph representational learning: When does more expressivity hurt generalization? *arXiv preprint arXiv:2505.11298*, 2025. 7, 11

[21] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT Press, 2012. 3, 4, 7

[22] 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, pages 4602–4609, 2019. 1, 3, 7

[23] Christopher Morris, Floris Geerts, Jan Tönshoff, and Martin Grohe. WL meet VC. In *International conference on machine learning*, pages 25275–25302. PMLR, 2023. 2

[24] Raffaele Paolino, Sohir Maskey, Pascal Welke, and Gitta Kutyniok. Weisfeiler and leman go loopy: A new hierarchy for graph representational learning. *Advances in Neural Information Processing Systems*, 37:120780–120831, 2024. 2

[25] Levi Rauchwerger, Stefanie Jegelka, and Ron Levie. Generalization, expressivity, and universality of graph neural networks on attributed graphs. *arXiv preprint arXiv:2411.05464*, 2024. 11

[26] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008. 1

[27] Erik Thiede, Wenda Zhou, and Risi Kondor. Autobahn: Automorphism-based graph neural nets. *Advances in Neural Information Processing Systems*, 34:29922–29934, 2021. 2

[28] Antonis Vasileiou, Stefanie Jegelka, Ron Levie, and Christopher Morris. Survey on generalization theory for graph neural networks. *arXiv preprint arXiv:2503.15650*, 2025. 7, 11

[29] Boris Weisfeiler and AA Lehman. A Reduction of a Graph to a Canonical Form and an Algebra arising during this Reduction. In *Nauchno-Technicheskaya Informatsia*, pages 2(9):12—-16, 1968. 3

[30] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018. 1, 3

[31] Bohang Zhang, Guhao Feng, Yiheng Du, Di He, and Liwei Wang. A complete expressiveness hierarchy for subgraph gnns via subgraph weisfeiler-lehman tests. In *International Conference on Machine Learning*, pages 41019–41077. PMLR, 2023. 7

[32] Muhan Zhang and Pan Li. Nested graph neural networks. *Advances in Neural Information Processing Systems*, 34:15734–15747, 2021. 2

[33] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81, 2020. 1

A Additional Related works

The fundamental trade-off between expressivity and generalization in GNNs is attracting increasing attention within the community. A recent work by Maskey et al. [20] demonstrated that more expressive GNNs may have worse generalization capabilities, unless their increased complexity is balanced by sufficiently large training sets or reduced structural distance between training and test graphs. Their analysis introduces pseudo-metrics that capture structural similarity and reveal when expressivity hurts generalization. The case when more expressive power affects generalization is further refined by Franks et al. [9]. The authors propose using *partial concepts* to derive bounds of VC dimension independent of the length of the embedding vector, d. Additionally, they show that for certain classes of graphs there are tighter lower bounds, thus confirming that more expressivity is not always worse. In a more general approach, [25] establishes a bound on the generalization error independentof both the data and the parameter of *any* MPNN. However, this approach eliminates the nuances that exist in different datasets. Depending on the task at hand, a more fine-grained analysis requires taking into account the distribution of graphs, for instance, by using the construction of graphons such as in [15]. Nonetheless, that approach has its own drawbacks, since under the used metrics, sparse graphs converge to the empty graph, which hinders seamless adoption in our context. Vasileiou et al. [28] make use of previous results on generalization, robustness and expressivity are collapsed under a single framework. This relies on a new pseudo-metric termed *Forest Distance*, inspired by *Tree Distance* [6]. Nevertheless, the bounds are not data dependent and while vertex-attributed graphs are considered, only discrete attributes are assumed. Additionally, it only works if the aggregation method for graphs is mean pooling. Meanwhile, Li et al. [17] proposed the notion of a k-variance margin-based generalization bound, defining the structural quality of graph embeddings in terms of their expressive power.

Collectively, these findings align with our work, illustrating a more nuanced relationship between model expressivity and generalization, and they are not restricted to message-passing GNNs, as is also the case for our study. Our work differs in that it provides a theoretical analysis grounded in Rademacher complexity, using coloring-based partitioning as a formal lens to characterize expressivity.

B Formal Definitions

Definition B.1 (MPGNN)**.** Let $G = (V, E, \alpha)$ be an attributed graph. We denote by $h_\ell(v)$ the embedding of node v at layer ℓ. The embeddings are initialized with $h_0(v) \in \mathbf{R}^d$ in a way which is *consistent* with $\alpha(v)$, namely, $h_0(v) = h_0(u)$ iff $\alpha(v) = \alpha(u)$. The GNN propagation scheme for iteration $\ell \in [L],\ \ell > 0$ is defined as:

$$h_\ell(v) = \texttt{COMBINE}^{(\ell)}\left(h_{\ell-1}(v), \texttt{AGGREGATE}^{(\ell)}\left(\{\!\{h_{\ell-1}(u)\}\!\}_{u\in\mathcal{N}(v)}\right)\right). \tag{17}$$

where $\texttt{AGGREGATE}^{(\ell)}$ and $\texttt{COMBINE}^{(\ell)}$ are differentiable parameterized functions, e.g. neural networks, and $\texttt{AGGREGATE}^{(\ell)}$ is permutation invariant over multisets.

Once all L layers have been applied, we obtain a final embedding for each node, $h_L(v)$. In the case of graph-level tasks, e.g., graph classification, a readout layer then compresses these node embeddings into a graph embedding $h_L(G)$:

$$h_L(G) = \texttt{READOUT}(\{\!\{h_L(v)\}\!\}_{v\in V}), \tag{18}$$

where $\texttt{READOUT}$ can be a differentiable parameterized function.

Definition B.2 (1-WL)**.** Let $G = (V, E, \alpha)$ be an attributed graph with $\alpha : V \to A \subset \mathbb{R}^d$, and $\mathcal{C}$ be a discrete set of colors. We begin by seeding each node with an initial color $\mathbf{c}_0(v) = \texttt{HASH}_0(\alpha(v))$, where $\texttt{HASH}_0 : A \to \mathcal{C}$ is an injective function mapping each node attribute to a color. For each iteration $\ell \in [L]$, the 1-WL algorithm updates every node color by

$$c_\ell(v) = \texttt{HASH}\Big(c_{\ell-1}(v), \{\!\{c_{\ell-1}(u)\}\!\}_{u\in\mathcal{N}(v)}\Big), \tag{19}$$

where $\texttt{HASH}$ is any injective map that encodes the pair consisting of the node's previous color and the multiset of its neighbours' colors as a new element of $\mathcal{C}$.
The algorithm terminates with a stable coloring when the partitioning does not change between iterations. We denote by $c(v)$ the color of node v at convergence of the partitions, that is, $c(v) := c_L(v)$ where L is the first iteration after which the partition no longer changes.

C Covering Numbers and Dudley's Integral

In this appendix, we introduce some theoretical tools used throughout the paper and in the proofs. Our goal is to keep the presentation self-contained, highlighting only the results needed in later proofs. In particular, we recall basic notions on covering numbers and we include the Dudley entropy integral, which is instrumental in deriving bounds on the Rademacher complexity.

C.1 Covering Numbers

We begin with the definition of a covering number, which quantifies the "size" of a function class in a given pseudometric space.

Definition C.1 (Covering Number)**.** Let $(\mathcal{X}, d)$ be a pseudometric space and let $\mathcal{F}$ be a subset of $\mathcal{X}$. For any $\varepsilon > 0$, the internal ε-covering number of $\mathcal{F}$, denoted $\mathcal{N}(\mathcal{F}, \varepsilon, d)$, is the minimum cardinality of a set $C \subseteq \mathcal{F}$ such that for every $f \in \mathcal{F}$, there exists some $f_c \in C$ with $d(f, f_c) \leq \varepsilon$. Unless otherwise specified, all covering numbers in this paper are internal.

C.2 Dudley's Entropy Integral

Dudley's entropy integral bounds the Rademacher complexity of a function class using its covering numbers. We consider the restriction of $\mathcal{F}$ to a sample $S = \{x_1, x_2, \ldots, x_m\}$, which is the set of vectors: $\mathcal{F}_{|S} = \{(f(x_1), \ldots, f(x_m)) | f \in \mathcal{F}\} \subseteq \mathcal{R}^m$. The covering number in the theorem is computed with respect to the standard Euclidean norm $||\cdot||_2$ on $\mathcal{R}^m$.

Theorem C.2 (Dudley's Entropy Integral Bound [2])**.** *Let $\mathcal{F}$ be a real-valued function class taking values in $[0, 1]$, and assume that $\mathbf{0} \in \mathcal{F}$. Then*

$$\mathcal{R}_S(\mathcal{F}) \leq \inf_{\alpha > 0} \left\{ \frac{4\alpha}{\sqrt{m}} + \frac{12}{m} \int_{\alpha}^{\sqrt{m}} \sqrt{\log \mathcal{N}(\mathcal{F}_{|S}, \epsilon, ||\cdot||_2)} \mathrm{d}\epsilon \right\}. \tag{20}$$

In Proposition 3.4, we derive a tighter bound using the p-dimensional structure of $\mathcal{F}_{|S}$.

D Proofs

Proposition D.1. *Let $S = \{G_1, \ldots, G_m\}$ be a sample of m graphs, partitioned into p disjoint sets $\{I_1, \ldots, I_p\}$ by a coloring function. Let $\mathcal{F}$ be a class of functions whose output $f(G; \Theta)$ is the same on each graph of a fixed class $G \in I_j$. The empirical Rademacher complexity of $\mathcal{F}$ on S is bounded by:*

$$\mathcal{R}_S(\mathcal{F}) \leq \frac{\sup_\Theta L(\Theta)\sqrt{p}}{m}$$

where $L(\Theta) = \sqrt{\sum_{i=1}^m f(G_i; \Theta)^2}$ is the L2-norm of the function's outputs over the sample S.

Proof. The proof proceeds by first reorganizing the sum by graph colour, then applying the Cauchy-Schwarz inequality to separate the function-dependent norm from the Rademacher variables, and finally using Jensen's inequality.

First, we write the definition of the empirical Rademacher complexity and group the sum over the p partitions

$$I_j := \{i \in [m] : c(G_i) = c_j\}. \tag{21}$$

Let $f_j(\Theta)$ be the constant output for any graph in partition I_j.

$$\begin{aligned} \mathcal{R}_S(\mathcal{F}) &= \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_\Theta \frac{1}{m} \sum_{i=1}^m \sigma_i f(G_i; \Theta) \right] \\ &= \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_\Theta \sum_{j=1}^p f_j(\Theta) \sum_{i \in I_j} \sigma_i \right]. \end{aligned} \tag{22}$$

Let $Z_j = \sum_{i \in I_j} \sigma_i$. The inner sum is $\sum_{j=1}^{p} f_j(\Theta) Z_j$. We apply the Cauchy-Schwarz inequality to this sum over j:

$$\sum_{j=1}^{p} f_j(\Theta) Z_j = \sum_{j=1}^{p} \Big(f_j(\Theta) \sqrt{|I_j|} \Big) \Big(\frac{Z_j}{\sqrt{|I_j|}} \Big) \leq \sqrt{\sum_{j=1}^{p} f_j(\Theta)^2 |I_j|} \cdot \sqrt{\sum_{j=1}^{p} \frac{Z_j^2}{|I_j|}}$$

The first term on the right is precisely the L2-norm $L(\Theta)$, since $\sum_{j=1}^{p} f_j(\Theta)^2 |I_j| = \sum_{i=1}^{m} f(G_i; \Theta)^2 = L(\Theta)^2$. Substituting this back into the main expression gives:

$$\begin{aligned}
\mathcal{R}_S(\mathcal{F}) &\leq \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\Theta} \left(L(\Theta) \cdot \sqrt{\sum_{j=1}^{p} \frac{Z_j^2}{|I_j|}} \right) \right] \\
&= \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\left(\sup_{\Theta} L(\Theta) \right) \cdot \sqrt{\sum_{j=1}^{p} \frac{Z_j^2}{|I_j|}} \right] \\
&= \frac{\sup_{\Theta} L(\Theta)}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sqrt{\sum_{j=1}^{p} \frac{(\sum_{i \in I_j} \sigma_i)^2}{|I_j|}} \right]
\end{aligned}$$

The second line follows because the term involving the Rademacher variables σ_i does not depend on Θ, allowing us to separate the supremum. The third line follows because $\sup_{\Theta} L(\Theta)$ is a constant with respect to the expectation over σ.

Next, we apply Jensen's inequality to the expectation. Since the square root function is concave, $\mathbb{E}[\sqrt{X}] \leq \sqrt{\mathbb{E}[X]}$.

$$\mathcal{R}_S(\mathcal{F}) \leq \frac{\sup_{\Theta} L(\Theta)}{m} \sqrt{\mathbb{E}_{\boldsymbol{\sigma}} \left[\sum_{j=1}^{p} \frac{(\sum_{i \in I_j} \sigma_i)^2}{|I_j|} \right]}$$

Finally, we evaluate the expectation inside the square root. By the linearity of expectation and the fact that σ_i are independent random variables with $\mathbb{E}[\sigma_i] = 0$ and $\mathbb{E}[\sigma_i^2] = 1$, we have:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}} \left[\sum_{j=1}^{p} \frac{(\sum_{i \in I_j} \sigma_i)^2}{|I_j|} \right] &= \sum_{j=1}^{p} \frac{\mathbb{E}_{\boldsymbol{\sigma}}[(\sum_{i \in I_j} \sigma_i)^2]}{|I_j|} \\
&= \sum_{j=1}^{p} \frac{\sum_{i \in I_j} \mathbb{E}[\sigma_i^2] + \sum_{i \neq k} \mathbb{E}[\sigma_i \sigma_k]}{|I_j|} \\
&= \sum_{j=1}^{p} \frac{\sum_{i \in I_j} \mathbb{E}[\sigma_i^2]}{|I_j|} \\
&= \sum_{j=1}^{p} \frac{|I_j|}{|I_j|} \\
&= p
\end{aligned}$$

Substituting this result back gives the final bound:

$$\mathcal{R}_S(\mathcal{F}) \leq \frac{\sup_{\Theta} L(\Theta) \sqrt{p}}{m}$$

□

Proposition D.2 (Uniform cardinality assumption)**.** *Let $S = \{G_1, \ldots, G_m\}$ be a sample of m graphs, partitioned by a coloring function into p disjoint sets $\{I_1, \ldots, I_p\}$ of the same cardinality. Let $\mathcal{F}$ be a class of functions whose output $f(G; \Theta)$ is a constant value $f_j(\Theta)$ for each graph in a*

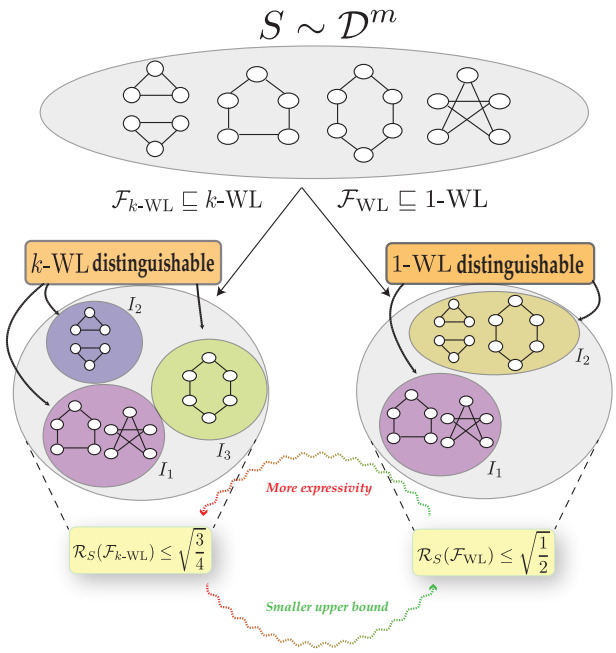

Figure 1: Two function classes $\mathcal{F}_{k\text{-WL}}$ and $\mathcal{F}_{\text{WL}}$, constrained respectively by 1-WL and k-WL expressivity, induce different partitions of a sample S. The more expressive function class $\mathcal{F}_{k\text{-WL}}$ *distinguishes* more graphs, leading to a finer partition of the sample and a larger number of equivalence classes. Since Rademacher complexity depends on the number of partitions of the input space, the coarser partition induced by $\mathcal{F}_{\text{WL}}$ yields a tighter upper bound on $\mathcal{R}_S$.

particular I_j, and the sign of $f_j(\Theta)$ is arbitrary for each partition, i.e., $sign(f_j(\Theta)) \in \{-1, 1\}$, $\forall j$. Then, the empirical Rademacher complexity of $\mathcal{F}$ on S is bounded by:

$$\mathcal{R}_S(\mathcal{F}) \geq \sqrt{\frac{p}{2m}}.$$

Proof. First, we write the definition of the empirical Rademacher complexity and group the sum over the p partitions

$$I_j := \{i \in [m] : c(G_i) = c_j\}. \tag{23}$$

Let $f_j(\Theta)$ be the constant output for any graph in partition I_j.

$$\begin{aligned}\mathcal{R}_S(\mathcal{F}) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\Theta} \frac{1}{m}\sum_{i=1}^{m} \sigma_i f(G_i;\Theta)\right] \\ &= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\Theta} \sum_{j=1}^{p} f_j(\Theta)\sum_{i\in I_j}\sigma_i\right]\end{aligned} \tag{24}$$

The sup is obtained for $f_j(\Theta) = sign(\sum_{i\in I_j}\sigma_i)$. Then,

$$\frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\Theta} \sum_{j=1}^{p} f_j(\Theta)\sum_{i\in I_j}\sigma_i\right] = \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{j=1}^{p}|\sum_{i\in I_j}\sigma_i|\right] \tag{25}$$

By the linearity of expectation and the fact that σ_i are independent random variables with $\mathbb{E}[\sigma_i] = 0$ and $\mathbb{E}[\sigma_i^2] = 1$, we have:

$$\frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{j=1}^{p}|\sum_{i\in I_j}\sigma_i|\right] = \frac{1}{m}\sum_{j=1}^{p}\mathbb{E}_{\boldsymbol{\sigma}}\left[|\sum_{i\in I_j}\sigma_i|\right] \tag{26}$$

By Khintchine's inequality we get the following bound on the expected value [13]:

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[|\sum_{i\in I_j}\sigma_i|\right] \geq \sqrt{\frac{\mu_j}{2}} \tag{27}$$

Then, substituting back into Eq. (26):

$$\frac{1}{m}\sum_{j=1}^{p}\mathbb{E}_{\boldsymbol{\sigma}}\left[|\sum_{i\in I_j}\sigma_i|\right] \geq \frac{1}{m}\sum_{j=1}^{p}\sqrt{\frac{\mu_j}{2}}. \tag{28}$$

In the hypothesis that all the classes have the same cardinality, $\mu_j = m/p$ we conclude that:

$$\begin{aligned}\mathcal{R}_S(\mathcal{F}) &\geq \frac{1}{m}\sum_{j=1}^{p}\sqrt{\frac{\mu_j}{2}} \\ &= \frac{1}{m}\sum_{j=1}^{p}\sqrt{\frac{m}{2p}} \\ &= \frac{p}{m}\sqrt{\frac{m}{2p}} \\ &= \sqrt{\frac{p}{2m}}.\end{aligned}$$

□

Proposition D.3 (Rademacher complexity bound under Partition Structure)**.** *Let $S = \{G_1, \ldots, G_m\}$ be a sample of m graphs, partitioned into p disjoint sets $\{I_1, \ldots, I_p\}$ by a coloring function. Let $\mathcal{F}$ be a class of functions $f : \mathcal{G} \to [-1, 1]$ whose output is the same on each graph of a fixed I_j. Assume $\mathbf{0} \in \mathcal{F}$. The empirical Rademacher complexity of $\mathcal{F}$ on S is bounded by:*

$$\mathcal{R}_S(\mathcal{F}) \leq \inf_{\alpha>0}\left(\frac{4\alpha\sqrt{p}}{m} + \frac{12}{m}\int_{\alpha}^{\sqrt{m}}\sqrt{\log\mathcal{N}(\mathcal{F}_{|S}, \epsilon, ||\cdot||_2)}d\epsilon\right)$$

where the bound is reduced due to the p-dimensional structure of the output space $\mathcal{F}_{|S}$.

Proof. The proof uses the Dudley entropy integral, adapting the bound to the p-dimensional structure of the function class $\mathcal{F}$.

Let us define a sequence of scales $\epsilon_k = \sqrt{m}\cdot 2^{-(k-1)}$ for $k \geq 1$. For each k, let V_k be a minimal ϵ_k-cover of the set of output vectors $\mathcal{F}_{|S}$ with respect to the ℓ_2-norm, so $|V_k| = \mathcal{N}(\mathcal{F}_{|S}, \epsilon_k, ||\cdot||_2)$. Note that $\mathcal{F}_{|S}$ lies within the unit hypercube $[-1, 1]^m$ therefore the cover is finite for $\epsilon_k > 0$.

For any function $f \in \mathcal{F}$, let $\boldsymbol{f}_{|S}$ be its corresponding vector in $\mathbb{R}^m$. Let $\boldsymbol{v}^k[\boldsymbol{f}]$ be a vector in V_k such that $||\boldsymbol{f}_{|S} - \boldsymbol{v}^k[\boldsymbol{f}]||_2 \leq \epsilon_k$. We decompose the vector $\boldsymbol{f}_{|S}$ using a telescoping sum:

$$\boldsymbol{f}_{|S} = (\boldsymbol{f}_{|S} - \boldsymbol{v}^N[\boldsymbol{f}]) + \sum_{k=1}^{N-1}(\boldsymbol{v}^k[\boldsymbol{f}] - \boldsymbol{v}^{k+1}[\boldsymbol{f}]) + \boldsymbol{v}^1[\boldsymbol{f}]$$

The quantity to bound is $m\mathcal{R}_S(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}}\sup_{f\in\mathcal{F}}\langle\boldsymbol{\sigma}, \boldsymbol{f}_{|S}\rangle$. Substituting the decomposition and using the triangle inequality for suprema gives:

$$\begin{aligned}m\mathcal{R}_S(\mathcal{F}) \leq \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\langle\boldsymbol{\sigma}, \boldsymbol{f}_{|S} - \boldsymbol{v}^N[\boldsymbol{f}]\rangle\right] &+ \sum_{k=1}^{N-1}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\langle\boldsymbol{\sigma}, \boldsymbol{v}^k[\boldsymbol{f}] - \boldsymbol{v}^{k+1}[\boldsymbol{f}]\rangle\right] \\ &+ \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\langle\boldsymbol{\sigma}, \boldsymbol{v}^1[\boldsymbol{f}]\rangle\right]\end{aligned}$$

For the last term, note the scale $\epsilon_1 = \sqrt{m}$. Since $f(G_i) \in [-1, 1]$, we have $||\boldsymbol{f}_{|S}||_2^2 = \sum_{i=1}^{m} f(G_i)^2 \leq m$, which means $||\boldsymbol{f}_{|S}||_2 \leq \sqrt{m}$. The zero vector $\mathbf{0}$ is in $\mathcal{F}_{|S}$ because $\mathbf{0} \in \mathcal{F}$,

and for any $\boldsymbol{f}_{|S}$, $||\boldsymbol{f}_{|S} - \mathbf{0}||_2 \leq \sqrt{m} = \epsilon_1$. Therefore, $V_1 = \{\mathbf{0}\}$ is a valid ϵ_1-cover. We can choose $\boldsymbol{v}^1[\boldsymbol{f}] = \mathbf{0}$ for all $f \in \mathcal{F}$, causing the last term to be zero.

The first term is the expected supremum over the set of residual vectors, $\mathcal{G}_N = \{\boldsymbol{f}_{|S} - \boldsymbol{v}^N[\boldsymbol{f}] : f \in \mathcal{F}\}$, which is precisely $m \cdot \mathcal{R}_S(\mathcal{G}_N)$.

To bound the Rademacher complexity, $\mathcal{R}_S(\mathcal{G}_N)$, we can leverage two key properties of this class of residual functions. First, $\mathcal{G}_N$ inherits the partition structure from $\mathcal{F}$, which means $(f(G_i) - v_i^N[f]) = (f(G_j) - v_j^N[f])$ whenever $f(G_i) = f(G_j)$.

Given these properties, we can apply Proposition 3.1:

$$\mathcal{R}_S(\mathcal{G}_N) \leq \frac{\sup_{\boldsymbol{g} \in \mathcal{G}_N} ||\boldsymbol{g}||_2 \cdot \sqrt{p}}{m}$$

By definition of the cover V_N, we have $\sup_{g \in \mathcal{G}_N} ||\boldsymbol{g}||_2 \leq \epsilon_N$. Substituting this in:

$$\mathcal{R}_S(\mathcal{G}_N) \leq \frac{\epsilon_N \sqrt{p}}{m}$$

Therefore, the term we need to bound is $m \cdot \mathcal{R}_S(\mathcal{G}_N)$, which gives:

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f \in \mathcal{F}} \langle \boldsymbol{\sigma}, \boldsymbol{f}_{|S} - \boldsymbol{v}^N[\boldsymbol{f}] \rangle\right] = m \cdot \mathcal{R}_S(\mathcal{G}_N) \leq \epsilon_N \sqrt{p}$$

For each $k \in \{1, \ldots, N-1\}$, we bound $\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f \in \mathcal{F}} \langle \boldsymbol{\sigma}, \boldsymbol{v}^k[\boldsymbol{f}] - \boldsymbol{v}^{k+1}[\boldsymbol{f}] \rangle\right]$. Let $W_k = \{\boldsymbol{v}^k[\boldsymbol{f}] - \boldsymbol{v}^{k+1}[\boldsymbol{f}] : f \in \mathcal{F}\}$. The expression is the Rademacher complexity of this finite set, $\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{w} \in W_k} \langle \boldsymbol{\sigma}, \boldsymbol{w} \rangle\right]$. The size of this set is $|W_k| \leq |V_k| \cdot |V_{k+1}|$. Since the covering numbers are monotonic ($\epsilon_k > \epsilon_{k+1} \implies |V_k| \leq |V_{k+1}|$), we have $|W_k| \leq |V_{k+1}|^2$. The norm of any element $w \in W_k$ is bounded by the triangle inequality:

$$||\boldsymbol{w}||_2 = ||\boldsymbol{v}^k[\boldsymbol{f}] - \boldsymbol{v}^{k+1}[\boldsymbol{f}]||_2 \leq ||\boldsymbol{v}^k[\boldsymbol{f}] - \boldsymbol{f}_{|S}||_2 + ||\boldsymbol{f}_{|S} - \boldsymbol{v}^{k+1}[\boldsymbol{f}]||_2 \leq \epsilon_k + \epsilon_{k+1}$$

Since $\epsilon_k = 2\epsilon_{k+1}$, the norm is bounded by $3\epsilon_{k+1}$. Applying Massart's Lemma:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{w \in W_k} \langle \sigma, w \rangle\right] &\leq (3\epsilon_{k+1}) \cdot \sqrt{2 \log |W_k|} \leq 3\epsilon_{k+1}\sqrt{2\log(|V_{k+1}|^2)} \\ &= 3\epsilon_{k+1}\sqrt{4 \log |V_{k+1}|} = 6\epsilon_{k+1}\sqrt{\log \mathcal{N}(\mathcal{F}_{|S}, \epsilon_{k+1}, ||\cdot||_2)}\end{aligned}$$

Combining the bounds on the residual and chain links yields:

$$\begin{aligned}m\mathcal{R}_S(\mathcal{F}) &\leq \epsilon_N\sqrt{p} + \sum_{k=1}^{N-1} 6\epsilon_{k+1}\sqrt{\log \mathcal{N}(\mathcal{F}_{|S}, \epsilon_{k+1}, ||\cdot||_2)} \\ &\leq \epsilon_N\sqrt{p} + 12\sum_{k=1}^{N-1}(\epsilon_k - \epsilon_{k+1})\sqrt{\log \mathcal{N}(\mathcal{F}_{|S}, \epsilon_k, ||\cdot||_2)}\end{aligned}$$

Using the standard step of bounding the sum with an integral and the monotonicity of the covering number:

$$m\mathcal{R}_S(\mathcal{F}) \leq \epsilon_N\sqrt{p} + 12\int_{\epsilon_{N+1}}^{\sqrt{m}} \sqrt{\log \mathcal{N}(\mathcal{F}_{|S}, \epsilon, ||\cdot||_2)}d\epsilon$$

For any given $\alpha > 0$, we choose N to be the largest integer such that $\epsilon_{N+1} > \alpha$. This implies $\epsilon_{N+2} \leq \alpha$. From the definition of the scales, we have $\epsilon_N = 2\epsilon_{N+1} = 4\epsilon_{N+2}$. This gives the bound $\epsilon_N \leq 4\alpha$. The lower limit of the integral, ϵ_N, is greater than α. Therefore, the integral is over a smaller domain than $[\alpha, \sqrt{m}]$, so we can bound it:

$$\epsilon_N\sqrt{p} + 12\int_{\epsilon_N}^{\sqrt{m}} \sqrt{\log \mathcal{N}(\ldots)}d\epsilon \leq 4\alpha\sqrt{p} + 12\int_{\alpha}^{\sqrt{m}} \sqrt{\log \mathcal{N}(\ldots)}d\epsilon$$

Substituting these into the main inequality:

$$m\mathcal{R}_S(\mathcal{F}) \leq 4\alpha\sqrt{p} + 12\int_{\alpha}^{\sqrt{m}} \sqrt{\log \mathcal{N}(\mathcal{F}_{|S}, \epsilon, ||\cdot||_2)}d\epsilon$$

Dividing by m gives the bound for our chosen α. This improves upon the classical bound [2] replacing $4\alpha/\sqrt{m}$ with $4\alpha\sqrt{p}/m$. As this holds for any $\alpha > 0$, we may take the infimum to find the tightest bound:

$$\mathcal{R}_S(\mathcal{F}) \leq \inf_{\alpha>0}\left(\frac{4\alpha\sqrt{p}}{m} + \frac{12}{m}\int_{\alpha}^{\sqrt{m}} \sqrt{\log \mathcal{N}(\mathcal{F}_{|S}, \epsilon, ||\cdot||_2)}d\epsilon\right)$$

This completes the proof. □

Lemma D.4. *Let X be a nonempty set and let $f, g : X \to \mathbb{R}$ be two real–valued functions. Then*

$$\big|\sup_{x\in X} f(x) \;-\; \sup_{x\in X} g(x)\big| \leq \sup_{x\in X}\big|f(x) - g(x)\big|. \tag{29}$$

We are now ready to prove Proposition 3.5.

Proposition D.5. *Let*

$$S = \{G_1, \ldots, G_m\} \quad \textit{and} \quad S' = \{G'_1, \ldots, G'_m\}$$

be two samples of size m. Applying a coloring procedure to both samples yields two sets of colours, $\mathcal{GC}(S)$ and $\mathcal{GC}(S')$ respectively. We denote by $\mathcal{GC}$ their union.
Suppose that for every colour $c_j \in \mathcal{GC}$, the number of graphs with colour c_j in the two samples differs by at most ϵ_j:

$$|\mu_j(S) - \mu_j(S')| \leq \epsilon_j. \tag{30}$$

Then the empirical Rademacher complexities of $\mathcal{F}$ on these two samples satisfy:

$$|\mathcal{R}_S(\mathcal{F}) - \mathcal{R}_{S'}(\mathcal{F})| \leq \sum_{c_j\in\mathcal{GC}} \frac{\epsilon_j}{m} \tag{31}$$

Proof of Proposition 3.5. From Eq. (24), the empirical Rademacher complexity can be expressed in terms of graph colours as:

$$\mathcal{R}_S(\mathcal{F}) := \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i f(G_i;\Theta)\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{c_j\in\mathcal{GC}(S)} f_j(\Theta)\sum_{i\in I_j(S)}\sigma_i\right], \tag{32}$$

where $\mathcal{GC}(S)$ is the set of colours appearing in S, $I_j(S)$ is the set of indices of graphs with colour c_j in S, and $f_j(\Theta)$ is the constant output for any graph in partition I_j. An analogous expression holds for S'.

To compare the empirical Rademacher Complexity of two different samples S and S', we require the definition to be invariant under permutations of the graph indices (otherwise, the Rademacher complexity of the same sample could change simply by reordering its elements). To ensure this invariance, we rewrite $\sum_{i\in I_j(S)}\sigma_i$ in terms of colours and their multiplicity, that is $\mu_j := |I_j|$:

$$\sum_{i\in I_j(S)}\sigma_i = \sum_{i=1}^{\mu_j(S)}\sigma_{j,i}$$

where for every color c_j, the sequence $(\sigma_{j,i})_{i\geq 1}$ is shared across samples (with each sample using only the first $\mu_j(S)$ terms, depending on its multiplicity). Additionally,we re-index Eq.32 over the union of colours $\mathcal{GC}$.

$$|\mathcal{R}_S(\mathcal{F}) - \mathcal{R}_{S'}(\mathcal{F})| \leq \mathbb{E}_{\boldsymbol{\sigma}}\left|\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{c_j\in\mathcal{GC}} f_j(\Theta)\sum_{i=1}^{\mu_j(S)}\sigma_{j,i} - \sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{c_j\in\mathcal{GC}} f_j(\Theta)\sum_{i=1}^{\mu_j(S')}\sigma_{j,i}\right|. \tag{33}$$

Note that if a colour c occurs only in one sample, its multiplicity is zero in the other, so the corresponding contribution vanishes. Now by Lemma D.4, we can upper-bound the difference of suprema by the supremum of the differences, yielding:

$$\begin{aligned}|\mathcal{R}_S(\mathcal{F}) - \mathcal{R}_{S'}(\mathcal{F})| &\leq \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{c_j\in\mathcal{GC}} f_j(\Theta)\sum_{i=1}^{\mu_j(S)}\sigma_{j,i} - \frac{1}{m}\sum_{c_j\in\mathcal{GC}} f_j(\Theta)\sum_{i=1}^{\mu_j(S')}\sigma_{j,i}\right|\right] \\ &\leq \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{c_j\in\mathcal{GC}}|f_j(\Theta)|\left|\sum_{i=1}^{\mu_j(S)}\sigma_{c,i} - \sum_{i=1}^{\mu_c(S')}\sigma_{j,i}\right|\right].\end{aligned} \tag{34}$$

Now, set $\min\ \mu_j := \min(\mu_j(S), \mu_j(S'))$ and $\max\ \mu_j := \max(\mu_j(S), \mu_j(S'))$ and, given that the Rademacher sequence $(\sigma_{j,i})_{i\geq 1}$ is shared across samples, we can write:

$$\left|\sum_{i=1}^{\mu_j(S)}\sigma_{j,i} - \sum_{i=1}^{\mu_j(S')}\sigma_{j,i}\right| = \left|\sum_{i=min\,\mu_j+1}^{max\,\mu_j}\sigma_i\right| \leq |\mu_j(S) - \mu_j(S')|$$

Hence, the bound in Eq.34 can be rewritten in terms of color multiplicities, independently of the σ_i's:

$$|\mathcal{R}_S(\mathcal{F}) - \mathcal{R}_{S'}(\mathcal{F})| \leq \sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{c_j\in\mathcal{GC}}|f_j(\Theta)|\left|\mu_j - \mu'_j\right|.$$

The final bound is obtained using the fact that $\sup_{f\in\mathcal{F}}|f(G)| \leq 1$:

$$|\mathcal{R}_S(\mathcal{F}) - \mathcal{R}_{S'}(\mathcal{F})| \leq \frac{1}{m}\sum_{c_j\in\mathcal{GC}}\left|\mu_j - \mu'_j\right| = \sum_{c_j\in\mathcal{GC}}\frac{\epsilon_j}{m}.$$

□

E Lipschitz continuity of loss functions

Proposition E.1. *Let ℓ_{CE} be the cross-entropy loss function. Moreover, let $f(G_i) = \psi(\boldsymbol{\beta}^\top\phi(G_i))$ be a GNN output, where $\psi(\cdot)$ is the logistic activation function (i.e., $\psi\colon\mathbb{R}\to[0,1]$), and $\phi(\cdot)$ the GNN's final representation. Assume that $\|\phi\|_\infty \leq b_\phi$ and $\|\boldsymbol{\beta}\|_1 \leq B_\beta$, for constants $b, B_\beta > 0$. Therefore, ℓ_{CE} is Lipschitz continuous.*

Proof. The cross-entropy loss is defined as:

$$\ell_{CE}(f(G), y) = -\sum_{i=1}^{m}\left[y_i\log(f(G_i)) + (1-y_i)\log(1-f(G_i))\right].$$

The partial derivative of ℓ_{CE} with respect to $z_i = \boldsymbol{\beta}^T\phi(G_i)$ is

$$\begin{aligned}\frac{\partial\ell_{CE}}{\partial z_i} &= -\left[y_i\frac{\psi'(z_i)}{\psi(z_i)} - (1-y_i)\frac{\psi'(z_i)}{1-\psi(z_i)}\right] \\ &= -\left[y_i\frac{\psi(z_i)(1-\psi(z_i))}{\psi(z_i)} - (1-y_i)\frac{\psi(z_i)(1-\psi(z_i))}{1-\psi(z_i)}\right] \\ &= -\left[y_i(1-\psi(z_i)) - (1-y_i)\psi(z_i)\right] \\ &= \psi(z_i) - y_i \qquad (*)\end{aligned}$$

Since $\phi(G_i)$ is bounded by b_ϕ in L_∞-norm, and $\|\boldsymbol{\beta}\|_1 \leq B_\beta$, we have

$$|z_i| = |\boldsymbol{\beta}^T\phi(G_i)| \leq \sum_{j=1}^{d}|\beta_j||\phi_j(G_i)| \leq \|\phi\|_\infty\|\boldsymbol{\beta}\|_1 \leq b_\phi B_\beta.$$

Thus, $z_i \in [-b_\phi B_\beta, b_\phi B_\beta]$, and the sigmoid function $\psi(z)$ satisfies:

$$\psi(-b_\phi B_\beta) \leq \psi(z_i) \leq \psi(b_\phi B_\beta),$$

for all G_i.

Then, we have that

$$(^*) : |\psi(z_i) - y_i| \leq \max\{|\psi(b_\phi B_\beta)|, |1 - \psi(b_\phi B_\beta)|\},$$

since $y_i \in \{0, 1\}$.

The derivative of the loss with respect to z_i is bounded; therefore, $\ell_{CE}(f(G_i), y_i)$ is Lipschitz continuous in $\phi(G_i)$.

□

Proposition E.2. *Assume the conditions from Prop. E.1 hold. Moreover, assume that the activation function is a sigmoid, i.e., $\psi\colon \mathbb{R} \to [a, b]$, for $a, b \in \mathbb{R}$ and $a < b$. In addition, assume that its derivative is bounded, $|\psi'(x)| \leq C$, for $C > 0$. Analogously, let ℓ_{CE} be the cross-entropy loss function, and define $g\colon [a, b] \to [0, 1]$, $g(x) = \frac{x-a}{b-a}$. Therefore, $\ell_{CE}(g \circ f(G_i))$ is Lipschitz continuous.*

Proof. Using the cross-entropy loss definition shown before, we have:

$$\ell_{CE}(g \circ f(G), y) = -\sum_{i=1}^{m} \left[y_i \log(f(g \circ G_i)) + (1 - y_i) \log(1 - g \circ f(G_i)) \right].$$

Analogously, the partial derivative of ℓ_{CE} with respect to $z_i = \beta^\top \phi(G_i)$ is

$$\begin{aligned}
\frac{\partial \ell_{CE}}{\partial z_i} &= -\left[y_i \frac{1}{g(\psi(z_i))} - (1 - y_i) \frac{1}{1 - g(\psi(z_i))} \right] \cdot \frac{1}{g'(\psi(z_i))} \cdot \psi'(z_i) \\
&= -\left[y_i \frac{1}{g(\psi(z_i))} - (1 - y_i) \frac{1}{1 - g(\psi(z_i))} \right] \cdot \frac{1}{(b - a)} \cdot \psi'(z_i) \\
&= \left[\frac{g(\psi(z_i)) - y_i}{g(\psi(z_i))(1 - g(\psi(z_i)))} \right] \cdot \frac{C}{(b - a)} \qquad (^*)
\end{aligned}$$

Since $|z_i| \leq b_\phi B_\beta$, then $a < \psi(-b_\phi B_\beta) \leq \psi(z_i) \leq \psi(b_\phi B_\beta) < b$. Hence, we have that $0 < g(-b_\phi B_\beta) \leq g(\psi(z_i)) \leq g(b_\phi B_\beta) < 1$, for all G_i, and

$$(^*) : \left| \frac{g(\psi(z_i)) - y_i}{g(\psi(z_i))(1 - g(\psi(z_i)))} \right| \cdot \frac{C}{(b - a)} \leq \frac{C}{(b - a)} \cdot \max\left\{ \frac{1}{|g(\psi(b_\phi B_\beta))|}, \frac{1}{|1 - g(\psi(b_\phi B_\beta))|} \right\}.$$

Again, because the derivative of the loss with respect to z_i is bounded, we have that $\ell_{CE}(g \circ f(G_i), y_i)$ is Lipschitz continuous in $\phi(G_i)$. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction regarding the relation between expressivity and generalization match the content of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.

- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 4 discusses the limitations of the work and the analysis performed.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: For each theoretical result, the full set of assumptions is stated explicitly, with complete proofs deferred to the appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: No experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: No experiments.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not

including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: No experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: No experiments

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: No experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Yes, we respect the NeurIPS code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There are no societal impacts we are aware of.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:To the best of our knowledge there is not high risk of miss-use of this work.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: No existing assets used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets released.

    Guidelines:

    - The answer NA means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We did not use human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We did not use human participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were used for text editing or formatting purposes.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.