ON THE EFFECT OF INPUT PERTURBATIONS FOR GRAPH NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

The expressive power of a message passing graph neural network (MPGNN) depends on its architecture and the input node attributes. In this work, we study how this interplay is affected by input perturbations. First, perturbations of node attributes may act as noise and hinder predictive power. But, perturbations can also aid expressiveness, by making nodes more identifiable. Recent works show that unique node IDs are necessary to represent certain functions with MPGNNs. Our results relate properties of the noise, smoothness of the model and the geometry of the input graphs and task. In particular, we take the perspective of lower bounding smoothness for achieving discrimination: how much output variation is needed for exploiting random node IDs, or for retaining discriminability? Our theoretical results imply constraints on the model for exploiting random node IDs, and, conversely, insights into the tolerance of a given model class for retaining discrimination with perturbations of node attributes.

1 INTRODUCTION

Graph Neural Networks (GNNs) (Scarselli et al., 2008) are the state-of-the-art representation learning algorithms on graphs. Important variants of GNNs include Graph Attention Networks (GATs) Veličković et al. (2018) and Graph Convolutional Networks (GCNs) Kipf & Welling (2017). They have shown to be quite successful in practice, e.g., in physics Shlomi et al. (2020); Battaglia et al. (2016), predicting drug interations Lim et al. (2019), protein design Strokach et al. (2020); Ingraham et al. (2021), studying molecules Jin et al. (2018); Duvenaud et al. (2015), quantum chemistry Gilmer et al. (2017), and drug discovery Sun et al. (2020) (see Wu et al. (2020) for a survey on GNNs, and Zhou et al. (2020) for a review on applications).

Despite having numerous successful empirical results, it is observed that GNNs are sensitive to perturbation in both graph structure and node attributes (see e.g., Zugner & Günnemann (2019)), and they are vulnerable to over-smoothing: having deep networks results in poor representations (see e.g., Kipf & Welling (2017); Oono & Suzuki (2019)). Intuitively, deeper networks allow to fit more smooth functions to the given data, along with having hidden representations depending on large neighborhoods. These issues have been addressed in several papers, and subsequently, people also tried to provide a consistent theory for these observation.

The discrimination power or the expressive power of GNNs is also another well-studied topic, and many works notice that having similar initial node attributes at different nodes may affect the power of network to distinguish distinct graphs, as well as counting subgraphs, while node identifiers (IDs), which are typically randomly chosen, can resolve the issues. But how much smoothness/non-smoothness of the model can affect the benefits of those (random) node IDs? Here by the smoothness of the model we means (lower/upper) bounds involving (perhaps higher-order) derivatives of the representation with respect to hidden representations, where here these derivatives assumed to exist.

The goal of this paper is to unify these observations/questions in one theory and relate the expressive power of GNN to the smoothness of the model and the sensitivity of it to the random perturbation of initial node attributes (or having (random) node IDs[1]). Our approach is restrict the class of functions

[1]Also, initial node attributes can be considered random in cases that are task independent (while quite being informative). For example, a GNN running on a social network can be initialized with individuals gender (as initial attributes) while the learning task is independent from the information.

that GNNs can approximate (i.e., the expressive power of GNNs), if the model is robust against random perturbations of initial attributes, and satisfy upper/lower bounds on its (perhaps higher-order) derivatives with respect to hidden representations.

Our main finding is a *quantitative* lower bound relating the following important notions about GNNs to each other: expressive power, geometry of the noiseless, smoothness of the model, and sensitivity to node attributes. In particular, we observe that non-smoothness representations are essential to express many graph classifiers with GNNs while still remaining robust against random perturbation of attributes. This scenario can also happen when node attributes are still informative, but independent of the learning task[2]. In addition, since the class of graph classifiers achieved with smooth and non-sensitive GNNs is completely restricted, we conclude that we should either be non-robust against random perturbations, or we should accept non-smooth representation to break the bound.

In short, in this paper, we make the following contributions:

- We present a theory explaining why restriction the model to smooth GNNs and still demanding robustness against random perturbations of node attributes can dramatically decrease the expressive power of the model. The result is a *quantitative* lower bounds which allows to observe the effects of the discussed notions on the representation in one mathematical equation, which we refer to it as an uncertainty principle for GNNs too.
- We introduce the notion of *discrepancy* of graph classifiers for studying GNNs, study its properties, and relate it to the smoothness, sensitivity, and the expressive power of GNNs.
- Our bound also applies to the expressive power of GNN with random (or task independent) node attributes, which can be of independent interest. In particular, we study how smoothness can affect the expressive power of GNN with random node attributes and the results suggest that while node IDs can improve the expressive power in general, with more smooth GNNs can approximate a restricted class of graph classifiers.

2 Related works

Theory of GNNs. The expressive power of GNNs is limited to the so-called Weisfeiler-Lehman (WL) graph isomorphism test (Xu et al., 2019), and under specific conditions, they are universal approximator (Scarselli et al., 2009) (see (Azizian & marc lelarge, 2021; Sato, 2020) for more on expressive power). Note that function approximation and graph isomorphism testing with GNNs are equivalent (Chen et al., 2019). Beyond graph isomorphism testing, GNNs are known to be unable to count substructures (see for instance (Chen et al., 2020b; You et al., 2019)). To improve the expressive power of GNNs, it is proposed to use random attributes (Sato et al., 2021) or coloring nodes (Dasoulas et al., 2020). We notice that just being able to represent a functions does not mean that the representation is achievable with smooth functions. For example, in the context of learning on sets, it is shown that imposing regularity conditions on the underlying functions results in limitations on representing functions (Wagstaff et al., 2019). In addition, GNNs suffer from the problem of over-smoothing (Oono & Suzuki, 2019). For example, with dropping edges is an instance of defeating against over-smoothing problem (Rong et al., 2020). Recently, generalization and stability bounds for GNNs have been also derived, e.g., see (Garg et al., 2020; Verma & Zhang, 2019).

We note that although GNNs with infinite width and node identifiers are Turing-complete (Loukas, 2020a), depth/width lower bounds show that GNNs should scale to several graph learning problems (Loukas, 2020b;a). These papers however does not relate the expressive power of GNNs to the smoothness and they also work only for discrete spaces. For the two-layer neural networks, the expressive power and smoothness are conjectured to satisfy a quantitative law of robustness (Bubeck et al., 2021).

Other related works. Adversarial attacks on graph representations have been well studied, see for instance (Bojchevski & Günnemann, 2019a) which is about edge attack on random walk methods. In general, GNNs are known to be sensitive to adversarial perturbations (Zugner & Günnemann, 2019), and heuristic solutions based on task and application exist, e.g., (Zhu et al., 2019; Jin et al., 2020b). For more information about adversarial attacks/defenses on graphs, see reviews (Jin et al., 2020a;

[2]For example, in social networks, node attributes while carry information about nodes, may be independent of a specific learning task and can be considered as noisy inputs.

Sun et al., 2018; Xu et al., 2020), and for solutions see (Zügner & Günnemann, 2020; Liu et al., 2020; Tang et al., 2020; Zügner & Günnemann, 2019; Dai et al., 2018; Bojchevski & Günnemann, 2019b; Geisler et al., 2020; Chen et al., 2020a).

3 Background, Motivation, and Problem Formulation

3.1 Graph Neural Networks (GNNs)

Let $\bar{\mathbb{G}}_n$ denote the set of all simple graphs on n vertices, and let $\mathbb{G}_n \subseteq \bar{\mathbb{G}}_n$ be an arbitrary subset. Assume that for each element of $\mathbb{G}_n$, the maximum degree is upper/lower bounded by $\Delta_{\max}, \Delta_{\min} \in \mathbb{N}$. Attached to a simple graph $\mathtt{G} = ([n], \mathcal{E}_{\mathtt{G}})$, there are initial attributes $\mathbf{x}_v \in \mathcal{X}$, $v \in [n]$, for nodes, and non-zero $\mathbf{e}_{(u,v)} \in \mathcal{X}$, $(u,v) \in \mathcal{E}_{\mathtt{G}}$, for edges (by setting zero elsewhere it can also be extended to all pairs), where $\mathcal{X} \subseteq \mathbb{R}^d$ is a compact convex set. Thus, a graph with initial attributes is just a tensor which lies in $\mathbb{R}^{n\times n\times d}$. The initial attributes profile is denoted by $\mathtt{init}(\mathtt{G}) = (\{\mathbf{x}_v\}_{v\in[n]}, \{\mathbf{e}_{(u,v)}\}_{(u,v)\in \mathtt{E}})$. A graph representation is a function $h(.;\theta) : \mathbb{R}^{n\times n\times d} \to \mathbb{R}$, that is further invariant under permutations of nodes, where θ denotes the set of all (learnable) parameters, and $\theta \in \Theta$ for some compact set Θ. A Message Passing Graph Neural Network (MPGNN) is a particular way to represent graphs with specific iterative learnable functions, and it is a special case of Graph Neural Networks (GNNs) (Gilmer et al., 2017). In this paper, we consider MPNNs as graph representation and we use the two terms GNN and MPGNN interchangeably.

We consider the most expressive GNN, i.e., Graph Isomorphism Network (GIN) in our formulation (Xu et al., 2019). However, our analysis and results are not restricted to this model and are applicable to other first order GNNs with minor changes. Let us consider the following general formulation. In a GNN with k iterations, at iteration $\ell \in [k]$, we compute

$$\mathbf{h}_v^{(\ell)} = \Phi_\ell\Big(\mathbf{h}_v^{(\ell-1)} + \sum_{u\in\mathcal{N}(v)} \Psi_\ell\big(\mathbf{h}_u^{(\ell-1)}, \mathbf{e}_{(u,v)}\big)\Big), \tag{1}$$

for each node $v \in [n]$, where $\mathcal{N}(v)$ denotes the set of neighbors of node v, and Ψ_ℓ, Φ_ℓ, $\ell \in [k]$, are learnable functions with appropriate dimensions, e.g, they can be modeled by Multi-Layer Perceptron (MLP), thanks to the universal approximation property (Hornik et al., 1989; Hornik, 1991). In particular, to update, it first aggregates on the neighbors of a node $v \in [n]$, and then combine the result with the current representation of v to achieve the updated representation. Throughout this paper, we assume that $\Psi_\ell, \Phi_\ell, \ell \in [k]$, have continuous derivatives. For example, if the activation function of MLPs is twice differentiable, then the assumption is satisfied.

To initialize, let $\mathbf{h}_v^{(0)} = \mathbf{x}_v$, for $v \in [n]$. Finally, a function is applied, called READOUT function, to achieve the final representation of the graph $\mathtt{G}$,

$$h(\mathtt{G}) = \Phi\Big(\sum_{v\in[n]} \mathbf{h}_v^{(k)}\Big), \tag{2}$$

where $\Phi : \mathbb{R}^{d_k} \to \mathbb{R}$ is a continuously differentiable real-valued function. To emphasize, we can also denote the final representation of a graph $\mathtt{G}$ with $h(\mathtt{G}, \mathtt{init}(\mathtt{G}); \theta)$, where θ denotes all parameters involving the representation. Note that some GNNs use the concatenation function as READOUT, but here we focus on a permutation-invariant function of graphs which allow to analyze the graph classification problem. It is known that this special READOUT function can model any permutation-invariant function as well (Zaheer et al., 2017).

3.2 Random Node Attributes Perturbation

In the most simple case, the initial node attributes may be extractable from $\mathtt{G}$ itself (e.g, degree, or from data attached to the graph). They can also be informative for the underlying learning task. The perturbed representation[3], which is now a random variable (since the perturbation is random), is denoted by $h_{\text{pert}}(\mathtt{G})$ for any graph with attributes $\mathtt{G} \in \mathbb{R}^{n\times n\times d}$, and is achieved by $\mathbf{x}_v \sim \mathcal{N}(\mu_v, \frac{\sigma_n^2}{d} I_d)$

[3]In order to facilitate the analysis we consider the Gaussian perturbation in the paper but the assumption can be easily replaced to the other other perturbations (e.g., integer perturbation).

and $\mathbf{e}_{(u,v)} \sim \mathcal{N}(\rho_{(u,v)}, \frac{\sigma_e^2}{d} I_d)$ where σ_n^2, σ_e^2 denote the variance of perturbations, and μ_v, $v \in [n]$, are graph/node dependent noiseless node attributes. Still, $h(\mathtt{G})$ denotes the representation without randomness, achieved by noiseless initial attributes. We also consider the following bounds for the noiseless initial attributes: $\|\mu_v\|_2 \leq B_\mu$ for all $v \in [n]$, and $1 \leq \|\rho_{u,v}\|_2 \leq B_\rho$ for all $(u,v) \in \mathcal{E}_{\mathtt{G}}$, and $\rho_{u,v} = 0$ for other pairs[4].

3.3 Graph Classifiers and Discrepancy

In this paper, a function $f : \mathbb{R}^{n\times n\times d} \to \mathbb{R}$ is called a graph function if and only if for any $\mathtt{G}_1, \mathtt{G}_2 \in \mathbb{R}^{n\times n\times d}$, we have $f(\mathtt{G}_1) = f(\mathtt{G}_2)$ whenever $\mathtt{G}_1 \simeq \mathtt{G}_2$ where $\simeq$ means the graph isomorphism (with respect to the attached initial attributes). A graph classifier is a graph function whose codomain is $\{\pm 1\}$. Let $\mathbb{P}$ denote an arbitrary but fixed probability measure on $\mathbb{R}^{n\times n\times d}$ which is supported on the class of graphs with initial attributes[5]. The given dataset of graphs can be considered as i.i.d. samples from $\mathbb{P}$. For a graph classifier f, let η_f^+, (and also η_f^-) denote the conditional probability measures given $f(\mathtt{G}) = 1$ (or $f(\mathtt{G}) = -1$), and we refer to them as the measures associated with f.

For any $\mathtt{G}, \mathtt{G}' \in \mathbb{R}^{n\times n\times d}$ with initial attributes define their $\lambda-$distance as

$$\mathrm{dist}(\mathtt{G}, \mathtt{G}'; \lambda) := \min_{\pi\in\mathcal{S}_n} \Big\{ \sum_{v\in[n]} \|\mu_v - \mu'_{\pi(v)}\|_2 + \lambda \sum_{\substack{u,v\in[n]\\ u<v}} \|\rho_{(u,v)} - \rho'_{(\pi(u),\pi(v))}\|_2 \Big\}, \tag{3}$$

where $\mathcal{S}_n$ denotes the set of permutations of $[n]$.

For each $\lambda > 0$ this defines a metric on the space of graphs with initial attributes. In particular, for simple graphs without different initial attributes this reduces to the edit distance on graphs: the minimum number of edges to be added/removed to construct one graph from the other (with respect to the graph isomorphism). Note that $\mathrm{dist}(\mathtt{G}_1, \mathtt{G}_2; \lambda) = 0$ if and only if $\mathtt{G}_1 \simeq \mathtt{G}_2$ and always $\mathrm{dist}(\mathtt{G}_1, \mathtt{G}_2; \lambda) \leq 2B_\mu n + \lambda B_\rho \Delta_{\max} n$ if $\|\mu_v^i\|_2 \leq B_\mu$ and $\|\rho_{(u,v)}^i\|_2 \leq B_\rho$ for all nodes and $i = 1, 2$.

Discrepancy of a graph classifier is an important quantity in our analysis. Roughly speaking, in the case of simple graphs without initial attributes, it is the average number of edges required to be removed/added to change the output of the classifier, and it is similar to the notion of margin in classical classification problems. First note that we have already defined a metric on the set of graphs with initial attributes, hence we can define the Wasserstein distance for probability measures on this space.

Definition 1. *For a graph classifier f, the discrepancy of f is defined as* $\mathrm{disc}(f) := W_1(\eta_f^+, \eta_f^-)$, *where W_1 denotes the Wasserstein distance, and η_f^+, η_f^- are two measures associated with f. In words, the discrepancy of a graph classifier is the optimal transportation cost of changing the binary label, for two measures induced by random choice of the graphs.*

For more technical information about optimal transport in this space, see Appendix A. Also, we compute the discrepancy of well-known graph classifiers in Section 5.

An important problem in deep learning and graph representations learning is to determine the approximation power of a model: the classifiers that can be approximated by some parameters. Typically, we say that a graph classifier can be approximated by a class of graph representations if and only if we can successfully classify the data with a threshold function applied on the representation. Accordingly, under the perturbation of input, the associated probability of error must be small. In what follows, we state this definition rigorously.

Definition 2. *A graph classifier f is called to be approximated under perturbations (with the error probability of at most δ) by a class of parameterized GNNs $h(.;\theta)$, $\theta \in \Theta$, if and only if*

$$\inf_{\theta\in\Theta} \inf_{\gamma\in\mathbb{R}} \mathbb{P}\Big(f(\mathtt{G}) \neq \mathbb{1}\{h_{pert}(\mathtt{G}) \geq \gamma\} \Big) \leq \delta, \tag{4}$$

where $\gamma \in \mathbb{R}$ is a threshold which possibly depends on θ, and $\mathbb{1}\{.\}$ is the sign function.

[4]Clearly, those bounds must hold uniformly on all observable graphs.

[5]We assume that given each graph, the resulting measure in absolutely continuous with respect to the Lebesgue measure, and we assume that each graph classifier is measurable in this space.

4 MAIN RESULT

Let $\mathtt{G}_n^*$ denote the set of observable graphs with initial attributes. Also, let us denote the gradient of the graph representation for $\mathtt{G} \in \mathtt{G}_n^*$ with respect to initial attributes by $\partial_v h(\mathtt{G})$, $\partial_{(u,v)} h(\mathtt{G})$ for all $v \in [n]$, $(u,v) \in \mathcal{E}_{\mathtt{G}}$ (similarly defined for $h_{\text{pert}}(\mathtt{G})$). The main result of this paper is summarized in the following theorem.

Theorem 1. *Consider a parametrized class of GNNs $h(.;\theta)$ which obey the following properties.*

- *[polynomial approximation] For any graph with initial attributes $\mathtt{G} \in \mathtt{G}_n^*$, there exists a polynomial of initial attributes[6], of degree at most k_{poly}, such that*

$$|h_{poly}(\mathtt{G}) - h(\mathtt{G})| \leq C_{poly}\sqrt{\text{var}(h_{pert}(\mathtt{G}))}. \tag{5}$$

  This polynomial may depend of $\mathtt{G}$, and the RHS, which can be interpreted as the normalized approximation error by polynomials is bounded. The variance of representation is just used to normalize the error in RHS.

- *[lower bound on average gradients[7]] For any $\mathtt{G} \in \mathtt{G}_n^*$, $\|\mathbb{E}[\partial_v h_{pert}(\mathtt{G})]\|_2 \geq L_{\min}$, $\|\mathbb{E}[\partial_{(u,v)} h_{pert}(\mathtt{G})]\|_2 \geq L_{\min}$ for all $v \in [n]$, $(u,v) \in \mathcal{E}_{\mathtt{G}}$.*

- *[upper bound on norm of representations] For any $\mathtt{G} \in \mathtt{G}_n^*$, $v' \in [n]$, and $\ell \in [k]$,*

$$\|\mathbf{h}_{v'}^{(\ell-1)}\|_2 \leq \frac{1}{n}\sum_{v\in[n]} L_h\|\mathbf{x}_v\|_2 + \frac{1}{|\mathcal{E}_{\mathtt{G}}|}\sum_{(u,v)\in\mathcal{E}_{\mathtt{G}}} L_h\|\mathbf{e}_{(u,v)}\|_2 + B_h. \tag{6}$$

  This means that when $\|\mathbf{x}_v\|_2, \|\mathbf{e}_{(u,v)}\|_2 = O(1)$, then $\|\mathbf{h}_{v'}^{(\ell-1)}\|_2 = O(2L_h + B_h)$.

- *[upper bound on gradients] For any $\mathtt{G} \in \mathtt{G}_n^*$, $\|\partial_v h(\mathtt{G})\|_2 \leq L_{\max}$, $\|\partial_{(u,v)} h(\mathtt{G})\|_2 \leq L_{\max}$ for all $v \in [n]$, $(u,v) \in \mathcal{E}_{\mathtt{G}}$.*

- *[smoothness for aggregation] For any $\mathtt{G} \in \mathtt{G}_n^*$,*

$$\|\Psi_\ell\big(\mathbf{h}_u^{(\ell-1)}, \mathbf{e}_{(u,v)}\big)\|_2 \leq L_\Psi\|\mathbf{h}_u^{(\ell-1)}\|_2 + L_\Psi\|\mathbf{e}_{(u,v)}\|_2. \tag{7}$$

Then, a graph classifier $f : \mathbb{G}_n^* \to \{\pm 1\}$ can be approximated by the class (with the error probability of at most δ), only if

$$\Big(\text{disc}(f) + 8.68\Big(1 + 2n\delta C_{poly}(1 + 3^{\frac{1}{2}k_{poly}}))^4(2B_\mu + \lambda B_\rho \Delta_{\max})\Big) \tag{8}$$

$$\times L_{\max}(1 + \frac{1}{\lambda}(1 + 2kL_\Psi(L_h(B_\mu + \sigma_n + B_\rho + \sigma_e) + B_h) + 2kL_\Psi(1 + \sigma_e))) \tag{9}$$

$$\geq \frac{0.17 L_{\min}\sigma_n\sqrt{n}\sqrt{(1 + \frac{\sigma_e^2}{2\sigma_n^2}\Delta_{\min})}}{\Big(1 + 2C_{poly}(1 + 3^{\frac{1}{2}k_{poly}})\Big)^2}. \tag{10}$$

Let us interpret the result and conditions. Theorem 1 is a *quantitative* bound restricting the expressive power of GNNs with the smoothness of the model, variance of perturbations, and the discrepancy of graph classifiers. A more general form of this theorem is also available in appendices, which does not involve lower bounds on the average gradients. We note that a class of GNNs can satisfy the conditions of Theorem 1, if we have lower/upper bounds on the gradient of graph representation, and also we can locally approximate the normalized versions of it with polynomials, which we interpret them as smoothness conditions. Note that the approximation by polynomials essentially follows from having bounds on higher-order derivatives. However, the converse is not true. Also, the lower bound on the average gradients, for small perturbations, is essentially the gradient without perturbations.

[6] For vectors, consider polynomials of entires.

[7] This condition can be relaxed to average of these quantities too, we just considered the most simple case here and the more general form of the theorem is available in appendices.

Remark 1. *A graph representation is called smooth (on observable graphs) if and only if* [8] *the bounds* $B_\mu, B_\rho, B_h, L_h, kL_\Psi, C_{poly}, k_{poly} = O(1)$*, and* $= O(1)$ *are satisfied. And consider* $\sigma = \sigma_n = \sigma_e = \Theta(1)$ *(bounded variance perturbation), and* $\lambda = \Theta(1)$*. In this case, the condition in the theorem can be written as*

$$\Big(\mathrm{disc}(f) + O(n\delta\Delta_{\max}) \Big) \gtrsim \frac{L_{\min}}{L_{\max}} \sigma \sqrt{n}. \tag{11}$$

We interpret $L_{\min}/L_{\max}$ *as a notion of inverse condition number for smooth GNNs. The above condition means that if a smooth GNN further satisfies the condition* $L_{\min}/L_{\max} = \Omega(1)$ *(i.e., having small condition number), and has small probability of error* $\delta \ll (n\Delta_{\max})^{-1}$*, and we can approximate the graph classifier* f *with it (under perturbations), then the discrepancy of classifier must be lower bounded by* $O(\sqrt{n})$*. However, we will see in next section that this is quite restricting, and the class of those classifiers is not so large. Note that discrepancy is an intrinsic value assigned to the classifier, and it independent for the representation.*

Let us also notice an important fact about the results: the conditions are obtained via a statistical formulation of the problem, hence the computational tractability is not considered here. Indeed, the bounds even hold for arbitrary large networks as long as the smoothness conditions are satisfied.

For better interpretation of the result, we introduce a new notion in the following subsection.

4.1 Separability of (Noiseless) Node Attributes

As one can expect, the effect of random perturbation to node attributes depends on how noiseless attributes are separated. In other words, if noiseless attributes are well-separated from each other, probably having random perturbation could not change the final representation much, and the model is more expressive (see Definition 2). In order to measure this separation, we define *separability* of the noiseless node attributes as the maximum change when we add an edge to a graph.

Definition 3. *The separability of noiseless node attributes* $(\mu_v)_{v\in[n]}$ *is defined as*

$$\mathrm{sep}(\mu) := \max_{\substack{\mathtt{G},\mathtt{G}'\in\mathbb{G}_n^* \\ \mathrm{dist}(\mathtt{G},\mathtt{G}';\lambda)=1}} \sum_{v\in[n]} \|\mu_v - \mu'_v\|_2. \tag{12}$$

In words, the separability of node attributes means how much adding an edge can change the initial node attributes. It is worth mentioning that the higher separability can make the learning task easier.

Example 1. *In the absence of noiseless node attributes* $\mathrm{sep}(\mu) = 0$*. This may happen when the node attributes are completely random.*

Example 2. *If the initial node attributes are the degrees of nodes (either one-hot encoded or uncoded) then* $\mathrm{sep}(\mu) = 2$*.*

Example 3. *If the initial node attributes are one-hot encoded, and adding an edge can only change the noiseless attributes of at most* T *nodes, then* $\mathrm{sep}(\mu) \le 2T$*. If* $T \ll n$ *then* $\mathrm{sep}(\mu) \ll n$*.*

In many reasonable scenarios, we can expect $\mathrm{sep}(\mu) = O(1)$ or at least $\mathrm{sep}(\mu) = o(n)$.

4.2 Uncertainty and GNNs

Let us assume there is no edge attribute. Note that from the definition of distance function between graphs, we can conclude (if $\lambda = 1$) for a given set of observable graphs and a graph classifier f,

$$\mathrm{disc}(f) \le \widetilde{\mathrm{disc}}(f) \times (\mathrm{sep}(\mu) + O(1)) \tag{13}$$

where $\widetilde{\mathrm{disc}}(f)$ is the discrepancy of f with respect to the edit-distance.

[8] We bound kL_Ψ because the linear approximations of the network accumulate, and if the output is bounded, it makes sense to have this bound too. Indeed, k in our model can be arbitrary large. Also, one can rescale Ψ_ℓ functions and change the mode a little bit to come up with other cases, and the assumption leads to no loss of generality, other than boundedness of output along with other conditions.

In this case, the result in Theorem 1 can be written as

$$\Big(\widetilde{\mathrm{disc}}(f)+O(1)\Big)\Big(\mathrm{sep}(\mu)+O(1)\Big)\gtrsim \frac{L_{\min}}{L_{\max}}\sigma\sqrt{n}. \tag{14}$$

This can be considered as an *uncertainty* principle for smooth GNNs: the product of the discrepancy of the target graph classifier (which measures how simple is that graph classifier to be approximated by a GNN), and the separability of the noiseless node attributes (which measures how noiseless node attributes are capable to help the classification) is lower bounded by $\sqrt{n}$. Additionally, if also $\mathrm{sep}(\mu)=O(1)$, as we argued before, we have a strong lower bound: $\widetilde{\mathrm{disc}}(f)\gtrsim\sqrt{n}$. As we will see later in Section 5, this lower bound completely restricts the class of functions GNN is capable to approximate and consequently if also gives lower bounds on the estimation error of estimating graph parameters. Note that here the new bound is with respect to the edit-distance.

Remark 2. *The above bound also shows that if we want to approximate a "complex" graph classifier ($\widetilde{\mathrm{disc}}(f)=O(1)$) then one of the following cases must happen:*

- $\mathrm{sep}(\mu)\gtrsim\sqrt{n}$. *In this case, the noiseless node attributes can be quite informative.*
- *The model is non-smooth: one of the smoothness conditions are violated.*
- *The model has large probability of error (at least $O(n^{-1})$).*
- *The model is not robust against random perturbation of node attributes: the resulting representation is sensitive.*

4.3 Multi-class Classification

In this part, we extend Theorem 1 to the multi-class graph classifiers. A function $f:\mathbb{G}_n^*\to[p]$ is called a multi-class graph classifier, where $p\in\mathbb{N}$. The case $p=2$ reduces to the aforementioned class of (binary) graph classifiers.

First we define approximation of multi-class graph classifiers with GNNs.

Definition 4. *A multi-class graph classifier $f:\mathbb{G}_n^*\to[p]$ can be approximated (with error probability of at most δ) by a class of parameterized GNNs $h(.;\theta)$, $\theta\in\Theta$, if and only if*

$$\inf_{\theta\in\Theta}\inf_{\zeta\in\mathcal{Z}}\sup_{\mathbb{P}\in\{\mathbb{P}_{p'}\otimes\mathcal{N}:p'\in[p]\}}\mathbb{P}\Big(f(\mathtt{G})\neq\zeta(h_{pert}(\mathtt{G}))\Big)\leq\delta, \tag{15}$$

where $\mathcal{Z}$ is the set of piecewise constant functions $\zeta:\mathbb{R}\to[p]$ with at most $p-1$ discontinuity points. Also, $\mathbb{P}_{p'}$ is the conditional probability measure on $f^{-1}(p')$, for each $p'\in[p]$, and $\mathcal{N}$ denotes the Gaussian probability measure corresponded to the perturbations (see 3.2 for more explanations).

Remark 3. *The motivation of this definition is that if $f(G)\approx h_{pert}(\mathtt{G})$ then there is confidence intervals allowing to infer $f(G)$ from $h_{pert}(\mathtt{G})$ with high probability.*

Let us define the discrepancy of multi-class graph classifiers as follows.

Definition 5. *For a multi-class graph classifier $f:\mathbb{G}_n^*\to[p]$, the discrepancy of f is defined as*

$$\mathrm{disc}(f):=\min_{\substack{p_1,p_2\in[p]\\ p_1\neq p_2}}\mathrm{disc}(f_{p_1,p_2}), \tag{16}$$

where $f_{p_1,p_2}(\mathtt{G}):f^{-1}(p_1)\cup f^{-1}(p_2)\to\{\pm1\}$ is defined as

$$f_{p_1,p_2}(\mathtt{G}):=\begin{cases}+1 & \textit{iff}\quad f(\mathtt{G})=p_1\\ -1 & \textit{iff}\quad f(\mathtt{G})=p_2\end{cases} \tag{17}$$

To define the discrepancy of multi-class graph classifiers, we consider all pairs of differently labeled graphs according to p_1,p_2, and take the minimum discrepancy of the corresponded graph classifier. Note that the set of observable graphs here is $f^{-1}(p_1)\cup f^{-1}(p_2)$ for fixed $p_1\neq p_2$.

The main result of this part is the following proposition which directly follows from Theorem 1.

Proposition 1. *Theorem 1 holds for multi-class graph classifiers $f:\mathbb{G}_n^*\to[p]$.*

Note that to satisfy the smoothness conditions for multi-class graph classifiers, we can normalize the resulting representations f_{p_1,p_2}, since Theorem 1 holds for each p_1,p_2.

Table 1: Approximation bounds for graph parameters

$f:\mathbb{G}_n^* \to [0,T)$			
Parameter	$\mathrm{disc}(f_p)$	Discrepancy & additive error	κ
Number of edges	$\frac{n\Delta_{\max}}{2p}+O(1)$	$\mathrm{disc}(f_p)=\epsilon_{\text{additive}}+O(1)$	1
Number of triangles	$\frac{n}{p}+O(1)$	$\mathrm{disc}(f_p)=3\epsilon_{\text{additive}}+O(1)$	3
Number of subgraphs H	$\frac{nm_H}{pn_H}+O(1)$	$\mathrm{disc}(f_p)=m_H\epsilon_{\text{additive}}+O(1)$	m_H
Diameter	$\frac{n-1}{p}+O(1)$	$\mathrm{disc}(f_p)=\epsilon_{\text{additive}}+O(1)$	1
Max-cut	$\frac{n\Delta_{\max}}{2p}+O(1)$	$\mathrm{disc}(f_p)=\epsilon_{\text{additive}}+O(1)$	1
Min-cut	$\frac{n\Delta_{\max}}{2p}+O(1)$	$\mathrm{disc}(f_p)=\epsilon_{\text{additive}}+O(1)$	1
Clique number	$\frac{\Delta_{\max}^2}{p}+O(1)$	$\mathrm{disc}(f_p)=\Delta_{\max}\epsilon_{\text{additive}}+O(1)$	$\Delta_{\max}$
Independence number	$\frac{\Delta_{\max}^2}{p}+O(1)$	$\mathrm{disc}(f_p)=\Delta_{\max}\epsilon_{\text{additive}}+O(1)$	$\Delta_{\max}$
Number of connected components	$\frac{n}{p}+O(1)$	$\mathrm{disc}(f_p)=\epsilon_{\text{additive}}+O(1)$	1

5 Applications of the Derived Bound

In this part, we give some applications of the proved bound by computing the discrepancy of a number of well-known graph functions. First consider a graph property as $f:\mathbb{G}_n^*\to[0,T)$, e.g., the number of edges, diameter, etc. To approximate it with a GNN, let's consider the discrepancy of its quantization to a $p-$level function:

$$f_p(\mathtt{G}):=p' \quad \text{iff} \quad \frac{T(p'-1)}{p}\le f(\mathtt{G})<\frac{Tp'}{p}, \tag{18}$$

for any $p'\in[p]$. If f as a multi-class graph classifier can be approximated by GNNs, then f_p can be approximated by GNNs, and the estimation error for f is at most $\epsilon_{\text{additive}}=\frac{T}{p}$. More precisely, this quantity is the expectation of the absolute value of the error $f-f_{\text{GNN}}$.

Hence, to achieve a lower bound for representing a graph parameter (function) from the main result, we need to compute the discrepancy of f_p. We expect that $\mathrm{disc}(f_p)$ is decreasing with p, meaning that having more resolution applies more complexity for representation. We will make this observation more clear in the following proposition and achieve bounds on the estimation error $\epsilon_{\text{additive}}$.

Proposition 2. *For any graph parameter in Table 1, there is an appropriate set of graphs and $T\in\mathbb{R}$ such that we have the provided bounds on the table on the discrepancy of their quantization. Also, the additive error of approximation is related to the discrepancy in the third column.*

Note that the number of subgraphs can be interpreted as induced or not necessarily induced, and the bounds are achieved for both cases. Also, n_H and m_H denote the number of nodes and the number of edges of a small graph H, and typically for counting tasks we have $n_H, m_H=O(1)$.

Now using the above table and the result on the approximation of the multi-class graph classifiers, we can derive the following bound on the estimation error of the above graph parameters.

Remark 4. *The proposition gives a quantitative bound based on the variance of perturbation, smoothness of the model, separability of noiseless attributes, and the additive estimation error of representation. In particular, under the assumptions in Section 4.2, we have the following necessary condition:*

$$\Big(\kappa\times\epsilon_{\text{additive}}+O(1)\Big)\Big(\mathrm{sep}(\mu)+O(1)\Big)\gtrsim\sqrt{n}. \tag{19}$$

We can immediately conclude the following proposition.

Proposition 3. *Consider the case that $\mathrm{sep}(\mu)=O(1)$. Then, for counting subgraph H (either induced or not necessarily induced) with GNNs, we have*

$$\epsilon_{\text{additive}}\gtrsim\sqrt{n}. \tag{20}$$

This means that roughly speaking smooth GNNs with node attributes perturbation are unable to count subgraphs with the additive estimation error $o(\sqrt{n})$.

The lower bounds in previous propositions suggest that smooth GNNs with node attributes perturbations are subject to high approximation error for the graph parameters in Table 1, and that is why we can argue that those graph representations are highly restricted to large error.

6 CONCLUSION

In this paper, we theoretically analyze GNNs with smooth representations, along with perturbed node attributes. It is observed that having smoothness highly restricts the model, and imposes several lower bounds on the approximation error for graph parameters. Our bound also allows to achieve compute a trade-off relating the expressive power of GNNs to the smoothness, separability of noiseless attributes, and the variance of perturbation.

REFERENCES

Waiss Azizian and marc lelarge. Expressive power of invariant and equivariant graph neural networks. In *International Conference on Learning Representations*, 2021.

Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems*, pp. 4502–4510, 2016.

Aleksandar Bojchevski and Stephan Günnemann. Adversarial attacks on node embeddings via graph poisoning. In *International Conference on Machine Learning*, pp. 695–704. PMLR, 2019a.

Aleksandar Bojchevski and Stephan Günnemann. Certifiable robustness to graph perturbations. *arXiv preprint arXiv:1910.14356*, 2019b.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Sébastien Bubeck, Yuanzhi Li, and Dheeraj M Nagaraj. A law of robustness for two-layers neural networks. In *Conference on Learning Theory*, pp. 804–820. PMLR, 2021.

Xu Chen, Ya Zhang, Ivor Tsang, and Yuangang Pan. Learning robust node representation on graphs. *arXiv preprint arXiv:2008.11416*, 2020a.

Zhengdao Chen, Soledad Villar, Lei Chen, and Joan Bruna. On the equivalence between graph isomorphism testing and function approximation with gnns. In *Advances in Neural Information Processing Systems*, pp. 15894–15902, 2019.

Zhengdao Chen, Lei Chen, Soledad Villar, and Joan Bruna. Can graph neural networks count substructures? *arXiv preprint arXiv:2002.04025*, 2020b.

Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. Adversarial attack on graph structured data. In *International conference on machine learning*, pp. 1115–1124. PMLR, 2018.

George Dasoulas, Ludovic Dos Santos, Kevin Scaman, and Aladin Virmaux. Coloring graph neural networks for node disambiguation. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 2126–2132. International Joint Conferences on Artificial Intelligence Organization, 7 2020.

David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pp. 2224–2232, 2015.

Vikas Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks. In *Int. Conference on Machine Learning (ICML)*, pp. 5204–5215. 2020.

Simon Geisler, Daniel Zügner, and Stephan Günnemann. Reliable graph neural networks via robust aggregation. *Advances in Neural Information Processing Systems*, 33, 2020.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1263–1272, 2017.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 1991.

Kurt Hornik, Maxwell Stinchcombe, Halbert White, et al. Multilayer feedforward networks are universal approximators. *Neural networks*, 1989.

John Ingraham, Vikas Kamur Garg, Regina Barzilay, and Tommi S Jaakkola. Generative models for graph-based protein design. 2021.

Wei Jin, Yaxin Li, Han Xu, Yiqi Wang, and Jiliang Tang. Adversarial attacks and defenses on graphs: A review and empirical study. *arXiv preprint arXiv:2003.00653*, 2020a.

Wei Jin, Yao Ma, Xiaorui Liu, Xianfeng Tang, Suhang Wang, and Jiliang Tang. Graph structure learning for robust graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 66–74, 2020b.

Wengong Jin, Kevin Yang, Regina Barzilay, and Tommi Jaakkola. Learning multimodal graph-to-graph translation for molecular optimization. *arXiv preprint arXiv:1812.01070*, 2018.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

FD Lesley and VI Rotar. Some remarks on lower bounds of chebyshev's type for half-lines. *JIPAM*, 4(5):96, 2003.

Jaechang Lim, Seongok Ryu, Kyubyong Park, Yo Joong Choe, Jiyeon Ham, and Woo Youn Kim. Predicting drug–target interaction using a novel graph neural network with 3d structure-embedded graph representation. *Journal of chemical information and modeling*, 59(9):3981–3988, 2019.

Yang Liu, Xianzhuo Xia, Liang Chen, Xiangnan He, Carl Yang, and Zibin Zheng. Certifiable robustness to discrete adversarial perturbations for factorization machines. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 419–428, 2020.

Andreas Loukas. What graph neural networks cannot learn: depth vs width. In *International Conference on Learning Representations*, 2020a.

Andreas Loukas. How hard is to distinguish graphs with graph neural networks? In *Advances in Neural Information Processing Systems*, 2020b.

Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. *arXiv preprint arXiv:1905.10947*, 2019.

Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. Dropedge: Towards deep graph convolutional networks on node classification. In *International Conference on Learning Representations*, 2020.

Ryoma Sato. A survey on the expressive power of graph neural networks. *arXiv preprint arXiv:2003.04078*, 2020.

Ryoma Sato, Makoto Yamada, and Hisashi Kashima. Random features strengthen graph neural networks. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pp. 333–341. SIAM, 2021.

F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. Computational capabilities of graph neural networks. *IEEE Transactions on Neural Networks*, 20(1):81–102, 2009.

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.

Jonathan Shlomi, Peter Battaglia, and Jean-Roch Vlimant. Graph neural networks in particle physics. *Machine Learning: Science and Technology*, 2(2):021001, 2020.

Alexey Strokach, David Becerra, Carles Corbi-Verge, Albert Perez-Riba, and Philip M Kim. Fast and flexible protein design using deep graph neural networks. *Cell Systems*, 11(4):402–411, 2020.

Lichao Sun, Yingtong Dou, Carl Yang, Ji Wang, Philip S Yu, Lifang He, and Bo Li. Adversarial attack and defense on graph data: A survey. *arXiv preprint arXiv:1812.10528*, 2018.

Mengying Sun, Sendong Zhao, Coryandar Gilvary, Olivier Elemento, Jiayu Zhou, and Fei Wang. Graph convolutional networks for computational drug development and discovery. *Briefings in bioinformatics*, 21(3):919–935, 2020.

Xianfeng Tang, Yandong Li, Yiwei Sun, Huaxiu Yao, Prasenjit Mitra, and Suhang Wang. Transferring robustness for graph neural network against poisoning attacks. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 600–608, 2020.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Saurabh Verma and Zhi-Li Zhang. Stability and generalization of graph convolutional neural networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1539–1548, 2019.

Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

Edward Wagstaff, Fabian Fuchs, Martin Engelcke, Ingmar Posner, and Michael A Osborne. On the limitations of representing functions on sets. In *International Conference on Machine Learning*, pp. 6487–6494, 2019.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 2020.

Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17(2):151–178, 2020.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Jiaxuan You, Rex Ying, and Jure Leskovec. Position-aware graph neural networks. In *International Conference on Machine Learning*, pp. 7134–7143. PMLR, 2019.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in neural information processing systems*, pp. 3391–3401, 2017.

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.

Dingyuan Zhu, Ziwei Zhang, Peng Cui, and Wenwu Zhu. Robust graph convolutional networks against adversarial attacks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1399–1407, 2019.

Daniel Zügner and Stephan Günnemann. Certifiable robustness and robust training for graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 246–256, 2019.

Daniel Zügner and Stephan Günnemann. Certifiable robustness of graph convolutional networks under structure perturbations. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1656–1665, 2020.

Daniel Zugner and Stephan Günnemann. Adversarial attacks on graph neural networks via meta learning. In *International Conference on Learning Representations*, 2019.

A Optimal Transportation Duality

Let $\mathbb{G}_n^* \subseteq \mathbb{R}^{n\times n\times d}$ denote the set of all graphs with initial attributes which is equipped with the $\lambda-$distance. Let $\mathcal{P}(\mathbb{G}_n^*)$ denote the set of all probability measures on the set of graphs with initial attributes that are absolutely continuous with respect to the Lebesgue measure. The Wasserstein distance between two probability measures $\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}(\mathbb{G}_n^*)$ is defined as

$$W_1(\mathbb{P}_1, \mathbb{P}_2) := \inf_{\mathbb{P}\in\mathcal{M}(\mathbb{P}_1,\mathbb{P}_2)} \mathbb{E}_{(\mathtt{G}_1,\mathtt{G}_2)\sim\mathbb{P}}[\mathrm{dist}(\mathtt{G}_1, \mathtt{G}_2; \lambda)], \tag{21}$$

where $\mathcal{M}(\mathbb{P}_1, \mathbb{P}_2)$ denotes the set of all couplings of $\mathbb{P}_1, \mathbb{P}_2$. Note that $(\mathcal{P}(\mathbb{G}_n^*), W_1)$ is a metric space. From the duality principle, we have the following alternative definition for the Wasserstein distance,

$$W_1(\mathbb{P}_1, \mathbb{P}_2) = \sup_{f\in\mathscr{F}_1} \Big\{\mathbb{E}_{\mathtt{G}\sim\mathbb{P}_1}[f(\mathtt{G})] - \mathbb{E}_{\mathtt{G}\sim\mathbb{P}_2}[f(\mathtt{G})]\Big\}, \tag{22}$$

where

$$\mathscr{F}_1 := \Big\{f_1 : \mathbb{G}_n^* \to \mathbb{R} : \sup_{\mathtt{G}\neq\mathtt{G}'} \frac{|f(\mathtt{G}) - f(\mathtt{G}')|}{\mathrm{dist}(\mathtt{G}, \mathtt{G}'; \lambda)} \leq 1\Big\}. \tag{23}$$

For more on optimal transport, see (Villani, 2003).

B Proof of Theorem 1

Since the proof of the theorem is long, let us decompose it into a few subsections.

B.1 A More General Result

Let us restate the main theorem (in a more general case) here.

Theorem 2. *Under the conditions of Theorem 1, except the lower bound on average gradients, a graph classifier $f : \mathbb{G}_n^* \to \{\pm 1\}$ can be approximated by the class (with the error probability of at most δ), only if*

$$(\mathrm{disc}(f) + 8.68\Big(1 + 2n\delta C_{poly}(1 + 3^{\frac{1}{2}k_{poly}}))^4(2B_\mu + \lambda B_\rho\Delta_{\max})) \tag{24}$$

$$\times L_{\max}(1 + \frac{1}{\lambda}(1 + 2kL_\Psi L_h(B_\mu + \sigma_n + B_\rho + \sigma_e) + 2kL_\Psi(1 + \sigma_e))) \tag{25}$$

$$\geq \frac{0.17}{\Big(1 + 2C_{poly}(1 + 3^{\frac{1}{2}k_{poly}})\Big)^2} \max\Big\{\inf_{\mathtt{G}\in\mathbb{G}_n^*} \sqrt{\mathrm{var}(h_{pert}(\tilde{\mathtt{G}}))}, \tag{26}$$

$$\mathbb{E}[\sqrt{\mathrm{var}(h_{pert}(\tilde{\mathtt{G}}))}] - 8.68\delta\Big(1 + 2C_{poly}(1 + 3^{\frac{1}{2}k_{poly}})\Big)^4 \sup_{\mathtt{G}\in\mathbb{G}_n^*} \sqrt{\mathrm{var}(h_{pert}(\tilde{\mathtt{G}}))}\Big\}, \tag{27}$$

Remark 5. *This result is more general that that of Theorem 1. In particular, the RHS involves the average variance, if δ is small enough, and this means that we are outperforming Theorem 1, which can be concluded from the worst-case infimum variance (details are below).*

We conclude the main result with the following proposition.

Proposition 4. *We have that*

$$\inf_{\mathtt{G}\in\mathbb{G}_n^*} \sqrt{\mathrm{var}(h_{pert}(\tilde{\mathtt{G}}))} \geq L_{\min}\sigma_n\sqrt{n}\sqrt{(1 + \frac{\sigma_e^2}{2\sigma_n^2}\Delta_{\min})}. \tag{28}$$

Proof. For any fixed $\mathtt{G} \in \mathbb{G}_n^*$, $h_{\text{pert}}(\mathtt{G})$ is a deterministic function of a number of Gaussian random variables: $\mathbf{x}_v = \mu_v + \mathbf{n}_v$ and $\mathbf{e}_{(u,v)} = \rho_{(u,v)} + \mathbf{n}_{(u,v)}$ such that $\mathbf{n}_v \sim \mathcal{N}(0, \frac{\sigma_n^2}{d} I_d)$ and $\mathbf{n}_{(u,v)} \sim$

$\mathcal{N}(0, \frac{\sigma_e^2}{d} I_d)$ for $v \in [n]$ and $(u, v) \in \mathcal{E}_{\mathtt{G}}$. From the theory of Hermit polynomials and Gaussian Hilbert spaces (O'Donnell, 2014), we know that

$$\mathrm{var}(h_{\text{pert}}(\tilde{\mathtt{G}})) \geq \sigma_n^2 \sum_{v \in [n]} \|\mathbb{E}[\partial_v h_{\text{pert}}(\mathtt{G})]\|_2^2 + \sigma_e^2 \sum_{(u,v) \in \mathcal{E}_{\mathtt{G}}} \|\mathbb{E}[\partial_{(u,v)} h_{\text{pert}}(\mathtt{G})]\|_2^2. \tag{29}$$

The RHS can be interpreted as the square norm of the average gradient around perturbations. Intuitively, if perturbation is small, this is essentially the gradient without perturbations. From the assumption, we simply get

$$\mathrm{var}(h_{\text{pert}}(\tilde{\mathtt{G}})) \geq n\sigma_n^2 L_{\min}^2 + |\mathcal{E}_{\mathtt{G}}|\sigma_e^2 L_{\min}^2 \geq n L_{\min}^2 (\sigma_n^2 + \frac{1}{2}\sigma_e^2 \Delta_{\min}). \tag{30}$$

We are done. □

B.2 Approximating the Graph Classifier

In this part, we approximate the given graph classifier f with an auxiliary classifier $\hat{f}$ which aggress with f with high probability, we then exploit properties of the new classifier to proceed the proof. Note that if the probability of error in Equation 4 is bounded for any graph in $\mathbb{G}_n^*$, instead of random graphs, then we can skip this part since the approximation is not required and all graphs satisfy the conditions below.

Consider is a graph classifier f which can be approximated by GNNs, meaning that there exist θ, γ such that

$$\mathbb{P}\Big(f(\mathtt{G}) \neq \mathbb{1}\{h_{\text{pert}}(\mathtt{G}) \geq \gamma\}\Big) \leq \delta. \tag{31}$$

Now consider the following set:

$$\mathcal{A} := \Big\{\tilde{\mathtt{G}} \in \mathbb{G}_n^* : \mathbb{P}(f(\mathtt{G}) \neq \mathbb{1}\{h_{\text{pert}}(\mathtt{G}) \geq \gamma\} | \mathtt{G} = \tilde{\mathtt{G}}) \leq \frac{1}{C_1}\Big\}, \tag{32}$$

where C_1 is an arbitrary fixed constant to be determined later. For simplicity, let us define

$$\mathbb{P}_{\tilde{\mathtt{G}}}(f(\mathtt{G}) \neq \mathbb{1}\{h_{\text{pert}}(\mathtt{G}) \geq \gamma\}) := \mathbb{P}(f(\mathtt{G}) \neq \mathbb{1}\{h_{\text{pert}}(\mathtt{G}) \geq \gamma\} | \mathtt{G} = \tilde{\mathtt{G}}). \tag{33}$$

Note that we have

$$\begin{aligned}
\mathbb{P}\Big(f(\mathtt{G}) \neq \mathbb{1}\{h_{\text{pert}}(\mathtt{G}) \geq \gamma\}\Big) &= \mathbb{P}\Big(f(\mathtt{G}) \neq \mathbb{1}\{h_{\text{pert}}(\mathtt{G}) \geq \gamma\} | \mathcal{A}\Big)\mathbb{P}(\mathcal{A}) && (34)\\
&\quad + \mathbb{P}\Big(f(\mathtt{G}) \neq \mathbb{1}\{h_{\text{pert}}(\mathtt{G}) \geq \gamma\} | \mathcal{A}^c\Big)\mathbb{P}(\mathcal{A}^c) && (35)\\
&\geq \mathbb{P}\Big(f(\mathtt{G}) \neq \mathbb{1}\{h_{\text{pert}}(\mathtt{G}) \geq \gamma\} | \mathcal{A}^c\Big)(1 - \mathbb{P}(\mathcal{A})) && (36)\\
&\geq \inf_{\tilde{\mathtt{G}} \in \mathcal{A}^c} \mathbb{P}_{\tilde{\mathtt{G}}}\Big(f(\mathtt{G}) \neq \mathbb{1}\{h_{\text{pert}}(\mathtt{G}) \geq \gamma\}\Big)(1 - \mathbb{P}(\mathcal{A})) && (37)\\
&\geq \frac{1}{C_1}(1 - \mathbb{P}(\mathcal{A})). && (38)
\end{aligned}$$

Thus, $\mathbb{P}(\mathcal{A}) \geq 1 - C_1\delta$.

Now, consider the function $\hat{f} : \mathcal{A} \subseteq \mathbb{G}_n^* \to \{\pm 1\}$ which is the restriction of f to the set $\mathcal{A}$. We write

$$\begin{aligned}
\mathrm{disc}(\hat{f}) &= W_1(\eta_{\hat{f}}^+, \eta_{\hat{f}}^-) && (39)\\
&\overset{(a)}{\leq} W_1(\eta_{\hat{f}}^+, \eta_f^+) + W_1(\eta_f^+, \eta_f^-) + W_1(\eta_f^-, \eta_{\hat{f}}^-) && (40)\\
&= \mathrm{disc}(f) + W_1(\eta_{\hat{f}}^+, \eta_f^+) + W_1(\eta_{\hat{f}}^-, \eta_f^-), && (41)
\end{aligned}$$

where (a) holds by the triangle inequality. Recall that $\mathbb{G}_n^*$ is the set of graphs with initial attributes. To bound the other terms, consider the following coupling between the measures $\eta_{\hat{f}}^\kappa, \eta_f^\kappa$ for $\kappa \in \{\pm\}$. Let η^κ denote the following probability measure on $\mathbb{G}_n^* \times \mathbb{G}_n^*$: first we choose an element $\mathtt{G}_1$ with respect to the probability measure η_f^κ, and then if it belongs to $\mathcal{A}$ then the second element $\mathtt{G}_2$ is

the same as $\mathtt{G}_1$. Otherwise, we draw a random element of $\mathcal{A}$ for $\mathtt{G}_2$, with respect to the probability measure $\eta^{\kappa}_{\hat{f}}$. This is clearly a coupling between $\eta^{\kappa}_{\hat{f}}, \eta^{\kappa}_{f}$, and so we conclude

$$W_1(\eta^{\kappa}_{\hat{f}}, \eta^{\kappa}_{f}) \leq \mathbb{E}_{\eta^{\kappa}}[\mathrm{dist}(\mathtt{G}_1, \mathtt{G}_2; \lambda)] \tag{42}$$

$$\leq (2B_\mu n + \lambda B_\rho \Delta_{\max} n)(1 - \mathbb{P}(\mathcal{A})) \tag{43}$$

$$\leq C_1 \delta n (2B_\mu + \lambda B_\rho \Delta_{\max}), \tag{44}$$

since the trivial bound $\mathrm{dist}(\mathtt{G}_1, \mathtt{G}_2; \lambda) \leq 2B_\mu n + \lambda B_\rho \Delta_{\max} n$ holds for all graphs.

Combining the above result with (41) shows that

$$\mathrm{disc}(\hat{f}) \leq \mathrm{disc}(f) + 2C_1 \delta n (2B_\mu + \lambda B_\rho \Delta_{\max}). \tag{45}$$

B.3 Duality Principle and Approximating GNNs

In this part, we use the optimal transportation duality, and then we approximate the GNN representation with a polynomial, along with we using mean-value theorem to bound the approximation error. This approximation allows us to compute useful bounds for the given graph representation.

Now let us consider associated measures of $\hat{f}$ and note that $\mathbb{E}[h_{\text{pert}}(\tilde{\mathtt{G}})] : \mathbb{G}_n^* \to \mathbb{R}$ is a function on the space $(\mathbb{G}_n^*, \mathrm{dist}(.,.;\lambda))$, thus

$$\mathrm{disc}(\hat{f}) = W_1(\eta^{+}_{\hat{f}}, \eta^{-}_{\hat{f}}) \tag{46}$$

$$= \sup_{f_1 \in \mathscr{F}_1} \Big\{ \mathbb{E}_{\mathtt{G} \sim \eta^{+}_{\hat{f}}}[f_1(\mathtt{G})] - \mathbb{E}_{\mathtt{G} \sim \eta^{-}_{\hat{f}}}[f_1(\mathtt{G})] \Big\} \tag{47}$$

$$\geq \frac{1}{\mathscr{L}(h)} \Big(\mathbb{E}_{\mathtt{G} \sim \eta^{+}_{\hat{f}}}[\mathbb{E}[h_{\text{pert}}(\mathtt{G})]] - \mathbb{E}_{\mathtt{G} \sim \eta^{-}_{\hat{f}}}[\mathbb{E}[h_{\text{pert}}(\mathtt{G})]] \Big), \tag{48}$$

where

$$\mathscr{L}(h) := \sup_{\substack{\mathtt{G}, \mathtt{G}' \in \mathbb{G}_n^* \\ \mathtt{G} \not\simeq \mathtt{G}'}} \frac{|\mathbb{E}[h_{\text{pert}}(\mathtt{G})] - \mathbb{E}[h_{\text{pert}}(\mathtt{G}')]|}{\mathrm{dist}(\mathtt{G}, \mathtt{G}'; \lambda)}. \tag{49}$$

Note that the function $\hat{f}$ is supported on $\mathcal{A}$, and thus we have

$$\mathbb{P}_{\tilde{\mathtt{G}}}(f(\mathtt{G}) \neq \mathbb{1}\{h_{\text{pert}}(\mathtt{G}) \geq \gamma\}) \leq \frac{1}{C_1}, \tag{50}$$

for all $\tilde{\mathtt{G}} \in \mathcal{A}$ (see (33) for the definition). Fix $\mathtt{G} \in \mathcal{A}$, and let us approximate the above probability from below. First let us consider the case that $\hat{f}(\mathtt{G}) = -1$. Using the definition, we have

$$\mathbb{P}_{\tilde{\mathtt{G}}}(h_{\text{pert}}(\tilde{\mathtt{G}}) \geq \gamma) \leq \frac{1}{C_1}. \tag{51}$$

But $h_{\text{pert}}(\tilde{\mathtt{G}})$ is a deterministic function of some jointly Gaussian random variables. In particular, let us recall that $h_{\text{pert}}(\tilde{\mathtt{G}})$ is the graph representation for $\tilde{\mathtt{G}}$, when we have perturbed initial attributes $\mathbf{x}_v \sim \mathcal{N}(\mu_v, \frac{\sigma_n^2}{d} I_d)$ and $\mathbf{e}_{(u,v)} \sim \mathcal{N}(\rho_{(u,v)}, \frac{\sigma_e^2}{d} I_d)$ where $v \in [n]$ and $(u,v) \in \mathcal{E}_{\tilde{\mathtt{G}}}$. Now we need the following lemma for the rest of analysis.

Lemma 1 (Lesley & Rotar (2003)). *Let X be a random variable with zero mean, unit variance, and $\mathbb{E}[X^4] = \zeta \geq 1$. Then, we have*

$$\mathbb{P}(X \geq t) \geq \frac{0.46}{\zeta} - \frac{1.4}{\sqrt{\zeta}} t + \frac{0.02}{\zeta^{1.5}} t, \tag{52}$$

for any $t \geq 0$.

The above lemma can be proved by carefully finding an upper bound for the indicator function $\mathbb{1}\{. \leq t\}$ in terms of a degree four polynomial, and do some calculation to optimize the bound. Consequently, under the conditions of the lemma,

$$\forall t \geq 0 : \mathbb{P}(X \geq t) \geq \frac{0.46}{\zeta} - \frac{1.4}{\sqrt{\zeta}} t \implies \mathbb{P}(X \geq \frac{0.17}{\sqrt{\zeta}}) \geq \frac{0.23}{\zeta}. \tag{53}$$

Let $X_{\tilde{\mathtt{G}}} := \frac{h_{\text{pert}}(\tilde{\mathtt{G}}) - \mathbb{E}[h_{\text{pert}}(\tilde{\mathtt{G}})]}{\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}}$. Clearly, $X_{\tilde{\mathtt{G}}}$ is a random variable with zero mean and unit variance, and that randomness comes merely from perturbations, since $\tilde{\mathtt{G}}$ is fixed. Thus, by the above lemma, we conclude that

$$\mathbb{P}_{\tilde{\mathtt{G}}}(h_{\text{pert}}(\tilde{\mathtt{G}}) \geq \gamma) = \mathbb{P}_{\tilde{\mathtt{G}}}\Big(X_{\tilde{\mathtt{G}}} \geq \frac{\gamma - \mathbb{E}[h_{\text{pert}}(\tilde{\mathtt{G}})]}{\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}}\Big) \leq \frac{1}{C_1}. \tag{54}$$

Now using (53), we conclude that

$$C_1 \geq \frac{\mathbb{E}[X_{\tilde{\mathtt{G}}}^4]}{0.23} \implies \frac{\gamma - \mathbb{E}[h_{\text{pert}}(\tilde{\mathtt{G}})]}{\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}} \geq \frac{0.17}{\sqrt{\mathbb{E}[X_{\tilde{\mathtt{G}}}^4]}}. \tag{55}$$

Now we need to upper bound the fourth moment of $X_{\tilde{\mathtt{G}}}$. To do that, we use the following result on Gaussian hypercontactivity.

Lemma 2 (Boucheron et al. (2013)). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a multivariate polynomial of degree at most k, and $(X_1, X_2, \ldots, X_n) \sim \mathcal{N}(0, I)$ be a sequence of i.i.d. Gaussian random variables. Then, we have*

$$(\mathbb{E}[|f(X_1, X_2, \ldots, X_n)|^q])^{1/q} \leq (q-1)^{k/2}(\mathbb{E}[|f(X_1, X_2, \ldots, X_n)|^2])^{1/2}. \tag{56}$$

Note that by assumption there is a polynomial $h_{\text{poly}}(\tilde{\mathtt{G}})$ of degree at most k_{poly} such that

$$|h_{\text{poly}}(\tilde{\mathtt{G}}) - h(\tilde{\mathtt{G}})| \leq C'_{\text{poly}} := C_{\text{poly}}\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}, \tag{57}$$

holds for any initial attributes, even if we consider the perturbed case $h_{\text{pert}}(\tilde{\mathtt{G}})$. The polynomial, however, may depend on $\mathtt{G}$. Thus, we obtain that

In particular, we can apply the above lemma to $X_{\tilde{\mathtt{G}}}$ since it is a real-valued random variable, which is a deterministic function of Gaussian random variables (i.e., perturbations). Thus, setting $q = 4$ in the lemma results in

$$(\mathbb{E}[X_{\tilde{\mathtt{G}}}^4])^{1/4} = \Big(\mathbb{E}[\big(\frac{h_{\text{pert}}(\tilde{\mathtt{G}}) - \mathbb{E}[h_{\text{pert}}(\tilde{\mathtt{G}})]}{\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}}\big)^4]\Big)^{1/4} \tag{58}$$

$$= \frac{1}{\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}}\Big(\mathbb{E}[\big(h_{\text{pert}}(\tilde{\mathtt{G}}) - \mathbb{E}[h_{\text{pert}}(\tilde{\mathtt{G}})]\big)^4]\Big)^{1/4} \tag{59}$$

$$= \frac{1}{\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}}\Big(\mathbb{E}[\big((h_{\text{pert}}(\tilde{\mathtt{G}}) - h_{\text{poly}}(\tilde{\mathtt{G}})) + (h_{\text{poly}}(\tilde{\mathtt{G}}) - \mathbb{E}[h_{\text{poly}}(\tilde{\mathtt{G}})]) \tag{60}$$

$$+ (\mathbb{E}[h_{\text{poly}}(\tilde{\mathtt{G}})] - \mathbb{E}[h_{\text{pert}}(\tilde{\mathtt{G}})])\big)^4]\Big)^{1/4} \tag{61}$$

$$\overset{(a)}{\leq} \frac{1}{\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}}\Big(2C'_{\text{poly}} + (\mathbb{E}[\big(h_{\text{poly}}(\tilde{\mathtt{G}}) - \mathbb{E}[h_{\text{poly}}(\tilde{\mathtt{G}})]\big)^4])^{1/4}\Big) \tag{62}$$

$$\overset{(b)}{\leq} \frac{1}{\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}}\Big(2C'_{\text{poly}} + 3^{\frac{1}{2}k_{\text{poly}}}(\mathbb{E}[(h_{\text{poly}}(\tilde{\mathtt{G}}) - \mathbb{E}[h_{\text{poly}}(\tilde{\mathtt{G}})])^2])^{1/2}\Big) \tag{63}$$

$$= \frac{1}{\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}}\Big(2C'_{\text{poly}} + 3^{\frac{1}{2}k_{\text{poly}}}(\mathbb{E}[((h_{\text{poly}}(\tilde{\mathtt{G}}) - h_{\text{pert}}(\tilde{\mathtt{G}})) \tag{64}$$

$$+ (h_{\text{pert}}(\tilde{\mathtt{G}}) - \mathbb{E}[h_{\text{pert}}(\tilde{\mathtt{G}})])) + (\mathbb{E}[h_{\text{pert}}(\tilde{\mathtt{G}})] - \mathbb{E}[h_{\text{poly}}(\tilde{\mathtt{G}})])^2)])^{1/2}\Big) \tag{65}$$

$$\overset{(c)}{\leq} \frac{1}{\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}}\Big(2C'_{\text{poly}} + 3^{\frac{1}{2}k_{\text{poly}}}(2C'_{\text{poly}} + (\mathbb{E}[(h_{\text{pert}}(\tilde{\mathtt{G}}) - \mathbb{E}[h_{\text{pert}}(\tilde{\mathtt{G}})])^2])^{1/2})\Big) \tag{66}$$

$$= 1 + \frac{2C'_{\text{poly}}}{\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}}\Big(1 + 3^{\frac{1}{2}k_{\text{poly}}}\Big) \tag{67}$$

$$= 1 + 2C_{\text{poly}}\Big(1 + 3^{\frac{1}{2}k_{\text{poly}}}\Big), \tag{68}$$

where (a) and (c) hold by Minkowski inequality, and (b) hold by Lemma 2. To proceed, we combine the above lemma with (55) and obtain the optimal constant C_1, thus if

$$C_1 = 4.34\Big(1 + 2C_{\text{poly}}(1 + 3^{\frac{1}{2}k_{\text{poly}}})\Big)^4, \tag{69}$$

then

$$\gamma \geq \mathbb{E}[h_{\text{pert}}(\tilde{\mathtt{G}})] + 0.17\frac{\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}}{\Big(1 + 2C_{\text{poly}}(1 + 3^{\frac{1}{2}k_{\text{poly}}})\Big)^2}, \tag{70}$$

holds for any $\tilde{\mathtt{G}} \in \mathcal{A}$ such that $\hat{f}(\mathtt{G}) = -1$. Thus, taking the expectation from both sides with respect to $\mathtt{G} \sim \eta^-_{\hat{f}}$ results in

$$\gamma \geq \mathbb{E}_{\mathtt{G}\sim\eta^-_{\hat{f}}}[\mathbb{E}[h_{\text{pert}}(\tilde{\mathtt{G}})]] + 0.17\frac{\mathbb{E}_{\eta^-_{\hat{f}}}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}]}{\Big(1 + 2C_{\text{poly}}(1 + 3^{\frac{1}{2}k_{\text{poly}}})\Big)^2}, \tag{71}$$

while again we emphasize that $h_{\text{pert}}(\tilde{\mathtt{G}})$ is a random variable which depends on on perturbations, given any graph $\mathtt{G} \in \mathcal{A}$.

In a similar way, for any $\mathtt{G} \in \mathcal{A}$ such that $f(\mathtt{G}) = 1$ it can be shown that

$$\gamma \leq \mathbb{E}_{\mathtt{G}\sim\eta^+_{\hat{f}}}[\mathbb{E}[h_{\text{pert}}(\tilde{\mathtt{G}})]] - 0.17\frac{\mathbb{E}_{\eta^-_{\hat{f}}}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}]}{\Big(1 + 2C_{\text{poly}}(1 + 3^{\frac{1}{2}k_{\text{poly}}})\Big)^2}, \tag{72}$$

with the same C_1 defined in (69). Combining (71) and (72) shows that

$$\mathbb{E}_{\mathtt{G}\sim\eta^+_{\hat{f}}}[\mathbb{E}[h_{\text{pert}}(\tilde{\mathtt{G}})]] - \mathbb{E}_{\mathtt{G}\sim\eta^-_{\hat{f}}}[\mathbb{E}[h_{\text{pert}}(\tilde{\mathtt{G}})]] \geq 0.17\frac{\mathbb{E}_{\eta^+_{\hat{f}}}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}] + \mathbb{E}_{\eta^-_{\hat{f}}}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}]}{\Big(1 + 2C_{\text{poly}}(1 + 3^{\frac{1}{2}k_{\text{poly}}})\Big)^2}. \tag{73}$$

Now using the duality principle (see 48) we conclude that

$$\text{disc}(\hat{f})\mathscr{L}(h) \geq 0.17\frac{\mathbb{E}_{\eta^+_{\hat{f}}}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}] + \mathbb{E}_{\eta^-_{\hat{f}}}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}]}{\Big(1 + 2C_{\text{poly}}(1 + 3^{\frac{1}{2}k_{\text{poly}}})\Big)^2}. \tag{74}$$

Note that

$$\mathbb{E}_{\eta^+_{\hat{f}}}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}] = \mathbb{E}_{\eta^+_{\hat{f}}}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}|\mathcal{A}]\mathbb{P}(\mathcal{A}) + \mathbb{E}_{\eta^+_{\hat{f}}}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}|\mathcal{A}^c]\mathbb{P}(\mathcal{A}^c) \tag{75}$$

$$\leq \mathbb{E}_{\eta^+_{\hat{f}}}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}|\mathcal{A}] + \mathbb{E}_{\eta^+_{\hat{f}}}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}|\mathcal{A}^c]\mathbb{P}(\mathcal{A}^c) \tag{76}$$

$$\leq \mathbb{E}_{\eta^+_{\hat{f}}}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}] \tag{77}$$

$$+ 4.34\delta\Big(1 + 2C_{\text{poly}}(1 + 3^{\frac{1}{2}k_{\text{poly}}})\Big)^4\mathbb{E}_{\eta^+_{\hat{f}}}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}|\mathcal{A}^c] \tag{78}$$

$$\leq \mathbb{E}_{\eta^+_{\hat{f}}}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}] \tag{79}$$

$$+ 4.34\delta\Big(1 + 2C_{\text{poly}}(1 + 3^{\frac{1}{2}k_{\text{poly}}})\Big)^4 \sup_{\mathtt{G}\in\mathbb{G}_n^*} \sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))} \tag{80}$$

$$\tag{81}$$

Similarly, one can obtain

$$\mathbb{E}_{\eta_f^-}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}] \leq \mathbb{E}_{\eta_{\hat{f}}^-}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}] \tag{82}$$

$$+4.34\delta\Big(1 + 2C_{\text{poly}}(1 + 3^{\frac{1}{2}k_{\text{poly}}})\Big)^4 \sup_{\mathtt{G}\in\mathbb{G}_n^*} \sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))} \tag{83}$$

$$\tag{84}$$

Hence[9],

$$\mathbb{E}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}] \leq \mathbb{E}_{\eta_f^+}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}] + \mathbb{E}_{\eta_f^-}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}] \tag{85}$$

$$\leq \mathbb{E}_{\eta_{\hat{f}}^+}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}] + \mathbb{E}_{\eta_{\hat{f}}^-}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}] \tag{86}$$

$$+ 8.68\delta\Big(1 + 2C_{\text{poly}}(1 + 3^{\frac{1}{2}k_{\text{poly}}})\Big)^4 \sup_{\mathtt{G}\in\mathbb{G}_n^*} \sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))} \tag{87}$$

Now combining with 74 results in

$$\text{disc}(\hat{f})\mathscr{L}(h) \geq \frac{0.17}{\Big(1 + 2C_{\text{poly}}(1 + 3^{\frac{1}{2}k_{\text{poly}}})\Big)^2} \max\Big\{ \inf_{\mathtt{G}\in\mathbb{G}_n^*} \sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}, \tag{88}$$

$$\mathbb{E}[\sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}] - 8.68\delta\Big(1 + 2C_{\text{poly}}(1 + 3^{\frac{1}{2}k_{\text{poly}}})\Big)^4 \sup_{\mathtt{G}\in\mathbb{G}_n^*} \sqrt{\text{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}\Big\}. \tag{89}$$

B.4 Analysis of the Lipschitz Constant

Now let us analyze $\mathscr{L}(h)$. By the definition

$$\mathscr{L}(h) = \sup_{\substack{\mathtt{G},\mathtt{G}'\in\mathbb{G}_n^* \\ \mathtt{G}\not\simeq\mathtt{G}'}} \frac{|\mathbb{E}[h_{\text{pert}}(\mathtt{G})] - \mathbb{E}[h_{\text{pert}}(\mathtt{G}')]|}{\text{dist}(\mathtt{G},\mathtt{G}';\lambda)} \tag{90}$$

$$\overset{(a)}{\leq} \sup_{\substack{\mathtt{G},\mathtt{G}'\in\mathbb{G}_n^* \\ \mathtt{G}\not\simeq\mathtt{G}'}} \frac{\mathbb{E}_\pi[|h_{\text{pert}}(\mathtt{G}) - h_{\text{pert}}(\mathtt{G}')|]}{\text{dist}(\mathtt{G},\mathtt{G}';\lambda)}, \tag{91}$$

where in (a) we couple the two perturbations into one according to the transportation map $\pi \in \mathcal{S}_n$ that achieves $\text{dist}(\mathtt{G},\mathtt{G}';\lambda)$. (see 3). Fix $\mathtt{G},\mathtt{G}' \in \mathbb{G}_n^*$ and the coupled perturbations, and without loss of generality, assume that $\pi = \text{id}$. Note that the initial attributes considered for $h_{\text{pert}}(\mathtt{G})$ are $\mathbf{x}_v = \mu_v + \mathbf{n}_v$ and $\mathbf{e}_{(u,v)} = \rho_{(u,v)} + \mathbf{n}_{(u,v)}$ for $v \in [n]$ and $(u,v) \in \mathcal{E}_{\mathtt{G}}$, and similarly for $h_{\text{pert}}(\mathtt{G}')$ they are $\mathbf{x}'_v = \mu'_v + \mathbf{n}_v$ and $\mathbf{e}'_{(u,v)} = \rho'_{(u,v)} + \mathbf{n}_{(u,v)}$ for $v \in [n]$ and $(u,v) \in \mathcal{E}_{\mathtt{G}'}$. Let us define a few auxiliary graphs as follows. Define $\mathtt{G}_1$ the same as $\mathtt{G}$, except with new initial node attributes $\mathbf{x}'_v = \mu'_v + \mathbf{n}_v$ for $v \in [n]$. Also, define $\mathtt{G}_2$ the same as $\mathtt{G}_1$, except with new edge attributes $\mathbf{e}_{2,(u,v)} = \rho_{(u,v)} + \mathbf{n}_{(u,v)}$ for all $(u,v) \in \mathcal{E}_{\mathtt{G}} \cap \mathcal{E}_{\mathtt{G}'}$, and $\mathbf{e}_{2,(u,v)} = \rho'_{(u,v)} + \mathbf{n}_{(u,v)}$ for all $(u,v) \in \mathcal{E}_{\mathtt{G}'} \setminus \mathcal{E}_{\mathtt{G}}$, and zero elsewhere.

One can write

$$|h_{\text{pert}}(\mathtt{G}) - h_{\text{pert}}(\mathtt{G}')| \leq |h_{\text{pert}}(\mathtt{G}) - h(\mathtt{G}_1)| + |h(\mathtt{G}_1) - h(\mathtt{G}_2)| + |h(\mathtt{G}_2) - h_{\text{pert}}(\mathtt{G}')|. \tag{92}$$

For the first term by the mean-value theorem,

$$|h_{\text{pert}}(\mathtt{G}) - h(\mathtt{G}_1)| \overset{(a)}{=} \Big| \sum_{v\in[n]} \langle \partial_v h_{\text{pert}}(\mathtt{G})(\xi_v), (\mu_v + \mathbf{n}_v) - (\mu'_v + \mathbf{n}_v)\rangle \Big| \tag{93}$$

[9]This expectation is with respect to the probability measure $\mathbb{P}$ which is the law of graph generation.

$$\overset{(b)}{\leq} \sum_{v\in[n]} \|\partial_v h_{\text{pert}}(\mathtt{G})(\xi_v)\|_2 \|\mu_v - \mu'_v\|_2 \tag{94}$$

$$\leq L_{\max} \sum_{v\in[n]} \|\mu_v - \mu'_v\|_2 \tag{95}$$

$$\leq L_{\max} \operatorname{dist}(\mathtt{G}, \mathtt{G}'; \lambda), \tag{96}$$

where (a) holds for some $\xi \in [\mu_v + \mathbf{n}_v, \mu'_v + \mathbf{n}_v]$, meaning that it belongs to the line segment between two endpoints, and also (b) holds by the Cauchy–Schwarz inequality.

For the second term, consider $\mathtt{G}_1$, and for instance, assume that $\mathtt{G}_2$ differs from $\mathtt{G}_1$ by only adding one edge. Observe that adding an edge such as (u, v) to $\mathtt{G}_1$, without changing node attributes, is equivalent to adding $\Psi_\ell\big(\mathbf{h}_u^{(\ell-1)}, \mathbf{e}_{(u,v)}\big)$ to node u in ℓth iteration (i.e., to $\mathbf{h}_u^{(\ell-1)}$) and the same for v. Note that each $\mathbf{h}_u^{(\ell-1)}$ in above is computed for $\mathtt{G}_1$, so the perturbations are also considered there. Thus, using the mean-value theorem,

$$|h(\mathtt{G}_1) - h(\mathtt{G}_2)| \overset{(a)}{=} \Big| \sum_{\ell\in[k]} \langle \Psi_\ell\big(\mathbf{h}_v^{(\ell-1)}, \mathbf{e}_{(u,v)}\big), \partial_{(u,\ell-1)} h(\mathtt{G}_1)(\xi_{u,\ell})\rangle \tag{97}$$

$$+ \sum_{\ell\in[k]} \langle \Psi_\ell\big(\mathbf{h}_u^{(\ell-1)}, \mathbf{e}_{(u,v)}\big), \partial_{(v,\ell-1)} h(\mathtt{G}_1)(\xi_{v,\ell})\rangle \Big| \tag{98}$$

$$\overset{(b)}{\leq} \sum_{\ell\in[k]} \|\Psi_\ell\big(\mathbf{h}_u^{(\ell-1)}, \mathbf{e}_{(u,v)}\big)\|_2 \|\partial_{(u,\ell-1)} h(\mathtt{G}_1)(\xi_{u,\ell})\|_2 \tag{99}$$

$$+ \sum_{\ell\in[k]} \|\Psi_\ell\big(\mathbf{h}_v^{(\ell-1)}, \mathbf{e}_{(u,v)}\big)\|_2 \|\partial_{(v,\ell-1)} h(\mathtt{G}_1)(\xi_{v,\ell})\|_2 \tag{100}$$

$$\leq L_{\max} \sum_{\ell\in[k]} \|\Psi_\ell\big(\mathbf{h}_u^{(\ell-1)}, \mathbf{e}_{(u,v)}\big)\|_2 + L_{\max} \sum_{\ell\in[k]} \|\Psi_\ell\big(\mathbf{h}_v^{(\ell-1)}, \mathbf{e}_{(u,v)}\big)\|_2 \tag{101}$$

$$\leq L_{\max} L_\Psi \sum_{\ell\in[k]} (\|\mathbf{h}_u^{(\ell-1)}\|_2 + \|\mathbf{h}_v^{(\ell-1)}\|_2) + 2kL_{\max} L_\Psi \|\mathbf{e}_{(u,v)}\|_2 \tag{102}$$

$$\leq 2kL_{\max} L_\Psi \Big(\frac{1}{n} \sum_{v\in[n]} L_h \|\mathbf{x}_v\|_2 + \frac{1}{|\mathcal{E}_{\mathtt{G}}|} \sum_{(u,v)\in\mathcal{E}_{\mathtt{G}}} L_h \|\mathbf{e}_{(u,v)}\|_2 + B_h\Big) \tag{103}$$

$$+ 2kL_{\max} L_\Psi \|\mathbf{e}_{(u,v)}\|_2 \tag{104}$$

$$\leq 2kL_{\max} L_\Psi \Big(\frac{1}{n} \sum_{v\in[n]} (L_h \|\mu_v\|_2 + L_h \|\mathbf{n}_v\|_2 + nB_h) \tag{105}$$

$$+ \frac{1}{|\mathcal{E}_{\mathtt{G}}|} \sum_{(u,v)\in\mathcal{E}_{\mathtt{G}}} L_h (\|\rho_{(u,v)}\|_2 + \|\mathbf{n}_{(u,v)}\|_2)\Big) \tag{106}$$

$$+ 2kL_{\max} L_\Psi (\|\rho_{(u,v)}\|_2 + \|\mathbf{n}_{(u,v)}\|_2), \tag{107}$$

where (a) holds for some $\xi_{v,\ell}, \xi_{u,\ell}$ for $\ell \in [k]$, and (b) follows from the Cauchy–Schwarz inequality. Thus, if we take expectation from both sides with respect to the perturbation, we have the bound

$$\mathbb{E}_\pi |h(\mathtt{G}_1) - h(\mathtt{G}_2)| \leq 2kL_{\max} L_\Psi \Big(\frac{1}{n} \sum_{\ell\in[k]} (L_h \|\mu_v\|_2 + L_h \sigma_n) \tag{108}$$

$$+ \frac{1}{|\mathcal{E}_{\mathtt{G}}|} \sum_{(u,v)\in\mathcal{E}_{\mathtt{G}}} (L_h \|\rho_{(u,v)}\|_2 + L_h \sigma_e) + nB_h\Big) \tag{109}$$

$$+ 2kL_{\max} L_\Psi (\|\rho_{(u,v)}\|_2 + \sigma_e) \tag{110}$$

$$\leq 2kL_{\max} L_\Psi (L_h B_\mu + L_h B_\rho + L_h \sigma_n + L_h \sigma_e + B_h) \tag{111}$$

$$+ 2kL_{\max} L_\Psi (\operatorname{dist}(\mathtt{G}_1, \mathtt{G}_2; \lambda)/\lambda + \sigma_e). \tag{112}$$

Note that in this case, $\operatorname{dist}(\mathtt{G}_1, \mathtt{G}_2; \lambda) \geq \lambda \|\rho_{(u,v)}\|_2 \geq \lambda$. Similarly, if we consider more than one edge difference between $\mathtt{G}_1, \mathtt{G}_2$, the same result holds as

$$\mathbb{E}_\pi |h(\mathtt{G}_1) - h(\mathtt{G}_2)| \tag{113}$$

$$\leq \Big(2kL_{\max}L_\Psi(L_hB_\mu + L_h\sigma_n + L_hB_\rho + L_h\sigma_e + B_h) + 2kL_{\max}L_\Psi(1+\sigma_e)\Big)\frac{\operatorname{dist}(\mathtt{G}_1,\mathtt{G}_2;\lambda)}{\lambda}. \tag{114}$$

Finally and similar to the previous arguments, we write

$$|h(\mathtt{G}_2) - h_{\text{pert}}(\mathtt{G}')| \leq \sum_{(u,v)\in\mathcal{E}_{\mathtt{G}_2}} \|\partial_{\mathbf{e}_{(u,v)}} h(\mathtt{G})(\xi_v)\|_2 \|(\rho_{(u,v)} + \mathbf{n}_{(u,v)}) - (\rho'_{(u,v)} + \mathbf{n}_{(u,v)})\|_2 \tag{115}$$

$$\leq L_{\max} \sum_{(u,v)\in\mathcal{E}_{\mathtt{G}_2}} \|\rho_{(u,v)} - \rho_{(u,v)}\|_2 \tag{116}$$

$$\leq L_{\max}\frac{\operatorname{dist}(\mathtt{G},\mathtt{G}';\lambda)}{\lambda}. \tag{117}$$

Combining (96), (114) and (117) shows that

$$\mathscr{L}(h) \leq \sup_{\substack{\mathtt{G},\mathtt{G}'\in\mathbb{G}_n^* \\ \mathtt{G}\not\simeq\mathtt{G}'}} \frac{\mathbb{E}_\pi[|h_{\text{pert}}(\mathtt{G}) - h_{\text{pert}}(\mathtt{G}')|]}{\operatorname{dist}(\mathtt{G},\mathtt{G}';\lambda)} \tag{118}$$

$$\leq L_{\max}(1 + \frac{1}{\lambda}(1 + 2kL_\Psi(L_hB_\mu + L_h\sigma_n + L_hB_\rho + L_h\sigma_e + B_h) + 2kL_\Psi(1+\sigma_e))). \tag{119}$$

B.5 Final Step of the Proof

Finally, we combine the above result and use (89) to achieve

$$\operatorname{disc}(\hat{f})\times L_{\max}(1 + \frac{1}{\lambda}(1 + 2kL_\Psi(L_hB_\mu + L_h\sigma_n + L_hB_\rho + L_h\sigma_e + B_h) + 2kL_\Psi(1+\sigma_e))) \tag{120}$$

$$\geq \frac{0.17}{\Big(1 + 2C_{\text{poly}}(1 + 3^{\frac{1}{2}k_{\text{poly}}})\Big)^2} \max\Big\{ \inf_{\mathtt{G}\in\mathbb{G}_n^*} \sqrt{\operatorname{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}, \tag{121}$$

$$\mathbb{E}[\sqrt{\operatorname{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}] - 8.68\delta\Big(1 + 2C_{\text{poly}}(1 + 3^{\frac{1}{2}k_{\text{poly}}})\Big)^4 \sup_{\mathtt{G}\in\mathbb{G}_n^*} \sqrt{\operatorname{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}\Big\}. \tag{122}$$

Now according to (45) we obtain

$$(\operatorname{disc}(f) + 8.68\Big(1 + 2C_{\text{poly}}(1 + 3^{\frac{1}{2}k_{\text{poly}}})\Big)^4\delta n(2B_\mu + \lambda B_\rho\Delta_{\max})) \tag{123}$$

$$\times L_{\max}(1 + \frac{1}{\lambda}(1 + 2kL_\Psi(L_hB_\mu + L_h\sigma_n + L_hB_\rho + L_h\sigma_e + B_h) + 2kL_\Psi(1+\sigma_e))) \tag{124}$$

$$\geq \frac{0.17}{\Big(1 + 2C_{\text{poly}}(1 + 3^{\frac{1}{2}k_{\text{poly}}})\Big)^2} \max\Big\{ \inf_{\mathtt{G}\in\mathbb{G}_n^*} \sqrt{\operatorname{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}, \tag{125}$$

$$\mathbb{E}[\sqrt{\operatorname{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}] - 8.68\delta\Big(1 + 2C_{\text{poly}}(1 + 3^{\frac{1}{2}k_{\text{poly}}})\Big)^4 \sup_{\mathtt{G}\in\mathbb{G}_n^*} \sqrt{\operatorname{var}(h_{\text{pert}}(\tilde{\mathtt{G}}))}\Big\}, \tag{126}$$

which completes the proof.

C Proof of Proposition 2

To prove Proposition 2, we introduce a natural class of simple graphs $\mathbb{G}_n$ that simply allows to couple graphs with different labels.

- *Number of edges.* In this case, simply consider the set of graphs $\mathbb{G}_n = \{G_0, G_1, \ldots, G_m\}$ where $m = \frac{n\Delta_{\max}}{2}$ constructed as follows. First let G_0 be the empty graph and for each G_i, add an arbitrary edge to G_{i-1} to construct it. For the number of edges function, denoted by $f : \mathbb{G}_n \to [0, \frac{n\Delta_{\max}}{2}]$, each interval in the definition of f_p has length $\frac{n\Delta_{\max}}{2p}$. To couple two intervals, we just shift the index: $i \to i + \frac{n\Delta_{\max}}{2p} \pm 1$ (since its an integer). This means that the discrepancy can be upper bounded by $\frac{n\Delta_{\max}}{2p} + O(1)$. The next part also follows from $\epsilon_{\text{additive}} = \frac{n\Delta_{\max}}{2p}$ which is the length of interval.
- *Number of triangles, number of subgraphs H.* Let us just prove the desired result for the latter. For any graph H, construct the set $\mathbb{G}_n = \{G_0, G_1, \ldots, G_m\}$ where $m = \frac{n}{n_H}$ constructed as follows. Start with empty graph G_0, and add independent copies of H (i.e., without overlapping nodes) sequentially to find G_i, $i \in [m]$. Now, for the subgraph count function $f : \mathbb{G}_n \to [0, \frac{n}{n_H}]$, we have the length of interval $\epsilon_{\text{additive}} = \frac{n}{pn_H}$. Similar to the previous part, to achieve a coupling, use index shift, and to increase the function by $\frac{n}{pn_H}$ we need to add $\frac{nm_H}{pn_H} \pm 1$ edges, which completes the proof.
- *Diameter.* Consider path graphs P_i for $i \in \{0, 1, \ldots, n-1\}$ as $\mathbb{G}_n$, and similar to the number of edges, construct the coupling. The proof is the same.
- *Max-cut, min-cut.* The two case are dual of each other if we compute the completion of the graphs (discrepancy bounds are invariant with respect to this operation). Thus, we only prove the max-cut case, and construct bipartite graph on $\frac{n}{2} \pm 1$ nodes (since the number is integer), and add edges to greedily construct $G_0, G_1, \ldots, G_m$ for $m = \frac{n\Delta_{\max}}{2} \pm 1$. Then, since adding edges will directly change the function value, we conclude the result similar to the number of edges case.
- *Clique number, independence number.* By duality we only prove the result for the clique number. Start from G_0 as empty set, and greedily enlarge a clique to achieve G_m which has a clique of size $\Delta_{\max}$. Then, $\epsilon_{\text{additive}} = \frac{\Delta_{\max}}{p}$, and to add this number we need to add at most $\Delta_{\max} \times \frac{\Delta_{\max}}{p}$ new edges. The proof is thus complete.
- *Number of connected components.* This case essentially the same as the analysis for the diameter of graphs.