Wasserstein Barycenter Matching for Graph Size Generalization of Message Passing Neural Networks

Xu Chu [*1] **Yujie Jin** [*2] **Xin Wang** [1 3] **Shanghang Zhang** [2] **Yasha Wang** [2] **Wenwu Zhu** [1 3] **Hong Mei** [2]

Abstract

Graph size generalization is hard for Message passing neural networks (MPNNs). The graph-level classification performance of MPNNs degrades across various graph sizes. Recently, theoretical studies reveal that a slow uncontrollable convergence rate w.r.t. graph size could adversely affect the size generalization. To address the uncontrollable convergence rate caused by correlations across nodes in the underlying dimensional signal-generating space, we propose to use Wasserstein barycenters as graph-level consensus to combat node-level correlations. Methodologically, we propose a Wasserstein barycenter matching (WBM) layer that represents an input graph by Wasserstein distances between its MPNN-filtered node embeddings versus some learned class-wise barycenters. Theoretically, we show that the convergence rate of an MPNN with a WBM layer is controllable and independent to the dimensionality of the signal-generating space. Thus MPNNs with WBM layers are less susceptible to slow uncontrollable convergence rate and size variations. Empirically, the WBM layer improves the size generalization over vanilla MPNNs with different backbones (e.g., GCN, GIN, and PNA) significantly on real-world graph datasets.

1. Introduction

In recent years, graph neural networks (GNNs) (Bruna et al., 2013; Defferrard et al., 2016; Kipf & Welling, 2017) have become the *de facto* choice for graph-level classification. Most GNNs used in practice can be reformulated into the common message passing neural network (MPNN) framework (Gilmer et al., 2017). An MPNN is size generalizable if it generalizes to testing graphs exhibiting a different average number of nodes from that of the training graphs. The size generalizability of an MPNN is desirable. Because the graph sizes can vary significantly, e.g., the size of traffic networks can be much larger in metropolitan areas than those in rural areas, and labeling large graphs can be costly, e.g., in combinatorial optimization (Bengio et al., 2021).

Empirical studies show that size generalization is hard for widely used MPNNs (Joshi et al., 2021; Gasteiger et al., 2022). Though efforts are made to promote the size generalization of MPNNs (Yehudai et al., 2021; Bevilacqua et al., 2021; Buffelli et al., 2022), there are still gaps between empirical success and reasonable theoretical understanding. Lately, based on a graphon (Lovász, 2012) random graph model and Monte Carlo theory, Maskey et al. (2022) developed a tight generalization bound decreasing with the graph size at a $-1/2D_{\mathcal{X}}$ rate, where $D_{\mathcal{X}}$ is the dimensionality of the underlying metric space generating the graph signals. The $-1/2D_{\mathcal{X}}$ convergence rate takes root in correlations across nodes entailed by the graph structure.

The rate at $-1/2D_{\mathcal{X}}$ is undesirable for the size generalization of MPNNs. The underlying metric space is not necessarily a low-dimensional manifold. When the sample size and the average graph size are limited, and $D_{\mathcal{X}}$ is large, the slow convergence rate $-1/2D_{\mathcal{X}}$ would inflate the generalization risk. In widely used MPNNs, there is a lack of mechanisms to tackle the uncontrollable $-1/2D_{\mathcal{X}}$ rate.

To address the uncontrollable $-1/2D_{\chi}$ rate that adversely affects the size generalization, we propose to use the Wasserstein barycenters as a graph-level consensus to combat the nodes-level consensus. Regarding the graphs as empirical measures in the Wasserstein metric measure space, the nodes across graphs of variant sizes can be registered (matched) by the Wasserstein metric. Methodologically, we propose the **W**asserstein **B**arycenter **M**atching (WBM) layer for MPNNs to improve size generalization with a controlled convergence rate. Specifically, the WBM layer approximates the class-wise empirical Wasserstein barycenters in end-to-end learning. When applying a WBM layer to an MPNN, an input graph is represented by the Wasserstein distances be-

[*]Equal contribution [1]Department of Computer Science and Technology, Tsinghua University, Beijing [2]Peking University, Beijing [3]BNRist, Tsinghua University, Beijing. Correspondence to: Xin Wang <xin_wang@tsinghua.edu.cn>, Yasha Wang <wangyasha@pku.edu.cn>, Wenwu Zhu <wwzhu@tsinghua.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

tween its MPNN-filtered node embeddings and each of the learned class-wise barycenters. We theoretically justify the proposed WBM layer with controllable convergence and generalization properties. Denote by F_L the output dimensionality of the node embeddings of the MPNNs, we show that the convergence rate of an MPNN with a WBM layer is independent of the uncontrollable D_χ: $-1/3$ in the low-dimensional regime and $-2/F_L$ in the high-dimensional regime. Therefore with an appropriate F_L, an MPNN with a WBM layer is theoretically guaranteed to be less susceptible to high underlying dimensionality and graph size variations, when only small training graphs are accessible.

In summary, we highlight the contributions of this paper: We address the adverse effect of uncontrollable convergence rate on size generalization. We propose a WBM layer for MPNNs, which employs Wasserstein barycenters as graph-level consensus to combat the correlation across nodes. We prove that an MPNN with a WBM layer enjoys a controllable and sharper convergence rate, in contrast to the uncontrollable rate of vanilla MPNNs. We demonstrate the effectiveness of the WBM layer with extensive experiments. With various MPNN backbones (GCN, GIN, and PNA), the size generalization of MPNNs with WBM layers significantly improves over the vanilla MPNNs and is competitive with the heuristic model in Buffelli et al. (2022).

2. Related Work

Size generalization of MPNNs. Empirical studies notice that the widely-used MPNNs are poor at size generalization, e.g., in combinatorial optimization (Joshi et al., 2021) and molecular biology (Gasteiger et al., 2022), the models trained on small graphs exhibit a large performance gap between testing sets of small graphs and sets of large graphs. To improve the size generalization of MPNNs. Yehudai et al. (2021) propose to minimize the discrepancy in the local structures between small and large graphs. Their framework is based on the assumption that either the testing graphs or the domain labels are accessible during training, which is often prohibited in practice. Bevilacqua et al. (2021) assume a complex causal model describing the generative process for graphs of different sizes and thereof design a size-invariant learning model. However, the size invariant model's performance decreases from synthetic graphs to real-world graphs, suggesting the model is susceptible to model misspecification. Aiming at a practical size-generalization method, Buffelli et al. (2022) develop a heuristic regularized model with impressive empirical performance. Their model simulates size shift by graph coarsening and penalizes the shift in the distribution of node embeddings. Howbeit the underlying invariance assumption and scope of application are unclear. In this paper, we aim at bridging empirical effectiveness and reasonable theoretical grounding.

There is also a relevant line of literature that aims at improving the general out-of-distribution (OOD) generalization on graphs. The graph augmentation methods combat the distributional shifts by increasing the data diversity (Zhao et al., 2021; Wang et al., 2021; Han et al., 2022). The invariant representation methods propose to learn representations invariant to distributional shifts (Sun & Saenko, 2016; Arjovsky et al., 2019; Wu et al., 2021). There are also methods modifying the training process to increase the models' robustness (Sagawa et al., 2019; Krueger et al., 2021; Wu et al., 2022a). Those methods usually impose strong assumptions on the graph data-generating process to defend various types of distributional shifts. While we assume in this paper, the graphs are generated from the graphon random graph model. Many graph models such as Erdős-Rényi model (Erdős et al., 1960), stochastic block model (Holland et al., 1983), and random geometric graphs (Penrose, 2003) are special cases of graphons (Lovász, 2012). Therefore our method is less susceptible to model misspecification.

Theoretical analysis of generalization for GNNs. From various perspectives of model complexity, generalization bounds are proposed for MPNNs, e.g., the bounds based on VC-dimension (Scarselli et al., 2018), the data-dependent bounds based on Rademacher complexity (Garg et al., 2020), and PAC-Bayesian bounds (Liao et al., 2020). However, the generalization bounds in those works increase with increasing average graph size N, implying a looseness. Lately, in the spirit of Monte Carlo theory, Maskey et al. (2022) develop a generalization bound for MPNNs with a pooling layer that decreases with increasing average graph size N, at the rate $-1/2(D_\chi + 1)$. Such a rate is also discovered in the earlier convergence analysis for spectral-based GNNs (Keriven et al., 2020). We argue that the dimensionality of the underlying graph-signal-generating space metric space D_χ is uncontrollable, as space χ is not necessarily a low-dimensional manifold. The $-1/2(D_\chi+1)$ rate is undesirable if only graphs of small sizes are accessible for training. We propose a Wasserstein barycenter matching (WBM) layer for MPNNs. We demonstrate that the convergence rate of an MPNN with a WBM layer is controllable: a constant $-1/3$ rate in the low-dimensional regime, and a controllable $-2/F_L$ in the high-dimensional regime, where F_L is the dimensionality of the last-layer node embeddings of MPNN. There is also a notion of transferability quantifying bounds between the GNN output of a finite graph versus its graphon limit (Ruiz et al., 2020; Levie et al., 2021), which is similar to our analysis conceptually. The transferability analysis focuses on a sequence of a deterministic sequence of graphs, while the generalization analysis focuses on random graphs.

Miscellaneous We mention two methodologically similar methods for completeness. The OT-GNN model (Chen et al., 2020) represents an input graph by the Wasserstein distances between node embeddings versus some learned

templates. The TFGW model (Vincent-Cuaz et al., 2022) extends OT-GNN by considering a trade-off between Wasserstein distance across nodes and Gromov-Wasserstein distance (Mémoli, 2011) across adjacency matrices. We highlight three differences: (1) The learned graph templates in OT-GNN and TFGW are not required to be the mean of the data cluster. While our Wasserstein barycenters are related to particular classes. (2) The MPNNs in OT-GNN and TFGW are seen as optional pre-processing units for graph signals. While we regard the proposed MPNN layer as a substitution for the pooling layer. (3) The OT-GNN and TFGW are designed for general graph classification. While our WBM layer is designed specifically for size generalization.

Wasserstein barycenters (Agueh & Carlier, 2011) are appealing, as empirical barycenters enjoy a guaranteed convergence in the Wasserstein space (Le Gouic et al., 2022). Besides, the Wasserstein barycenter is able to take into account the underlying geometry of the measures that a Euclidean barycenter cannot (Backhoff-Veraguas et al., 2022).

3. Preliminaries

As a starting point for theoretical analysis, we follow Keriven et al. (2020) and consider graphs and graph MPNNs as discretizations of continuous graphon random graph models and continuous graphon MPNNs, respectively. In this section, we introduce the relative concepts.

An N-node *weighted feature graph* (*graph* for short) is a tuple $G = (V, E, A, \mathbf{f})$, where $V = \{1, \dots, N\}$ is the node set and $E = \{(i, j)\} \subset V \times V$ is the edge set. The matrix $A = \{a_{ij}\}_{i,j}$ is the weight matrix of G, with $a_{ij} \in (0, 1]$ if the edge $(i, j) \in E$ and $a_{ij} = 0$ if $(i, j) \notin E$. A graph signal is defined as the function[1] $\mathbf{f} : V \to \mathbb{R}^F$ mapping each node to its $F \in \mathbb{N}$ dimensional signal in $\mathbb{R}^F$. We abuse notations that $\mathbf{f}$ also denotes the *graph signal* matrix of an N-node graph, i.e., $\mathbf{f} = (\mathbf{f}_1, \dots, \mathbf{f}_N)^\top \in \mathbb{R}^{N \times F}$, where $\mathbf{f}_i \in \mathbb{R}^F$ is the graph signal evaluated at node i.

The graphs can be viewed as discretizations of a continuous *graphon RGM* (cf. Def. 3.2). Firstly, we define graphons.

Definition 3.1 (graphon Lovász (2012))**.** Given a metric measure space $\Xi = (\chi, d, \mu)$, a *graphon* is a bivariate measurable mapping $\mathcal{A} : \chi \times \chi \to [0, 1]$. The sets of points in the metric space $V \subset \chi$ are sets of graph nodes and the corresponding images of the mapping $\mathcal{A}$ are the graph weight matrices, i.e., $A = \mathcal{A}|_V : V \times V \to [0, 1]$.

We then generalize the notion of graph signals $\mathbf{f}$ of a graph G by formally introducing the *graphon random graph model*.

Definition 3.2 (graphon RGM Keriven et al. (2020))**.** Given a space $\Xi = (\chi, d, \mu)$, a *graphon random graph model* (graphon RGM) is a pair of measurable functions $(\mathcal{A}, f)$, where $\mathcal{A}$ is the graphon in Def. 3.1 and $f : \chi \to \mathbb{R}^F$ is a metric-space signal. An N-nodes random weighted feature graph $(V, A, \mathbf{f})$ is defined by sampling N i.i.d. random points $\{X_1, \dots, X_N\} = V$ from χ according to measure μ. The weight matrix $A = \{a_{ij}\}_{i,j}$ is given by $a_{ij} := \mathcal{A}(X_i, X_j)$ for $i, j \in \{1, \dots, N\}$. The graph signal at node i is defined by $\mathbf{f}_i := f(X_i)$. We say that the random graph $(V, A, \mathbf{f})$ is sampled from the graphon $\mathcal{A}$, and denote $(A, \mathbf{f}) \sim (\mathcal{A}, f)$, where $\mathbf{f} = (\mathbf{f}_1, \dots, \mathbf{f}_N)^\top$.

With a graphon RGM, we may extend concepts of a graph to their continuous counterparts. Given a graph with weight matrix $A = \{a_{ij}\}_{i,j}$, the *degree* of the node i is defined by $d_i := \sum_{j=1}^N a_{ij}$. Given a graphon $\mathcal{A}$ on space (χ, d, μ), the *kernel degree* of $\mathcal{A}$ at $s \in \chi$ is $d_{\mathcal{A}}(s) := \int_\chi \mathcal{A}(s, t) d\mu(t)$.

Similar to the relationship between discrete random graphs and continuous graphon RGMs, we may also extend *graph MPNNs* (cf. Def. 3.4) to *graphon MPNNs* (cf. Def. 3.5), by applying *MPNNs* to random graphs and graphon RGMs, respectively. We formalize an MPNN as follows.

Definition 3.3 (message passing neural networks)**.** We define an L-layer *MPNN* Θ as a sequence of functions,

$$\Theta \overset{def.}{=} (\{\Phi^{(l)}, \Psi^{(l)}\}_{l=1}^L), \tag{1}$$

where $\Phi^{(l)} : \mathbb{R}^{2F_{l-1}} \to \mathbb{R}^{H_{l-1}}$ and $\Psi^{(l)} : \mathbb{R}^{F_{l-1}+H_{l-1}} \to \mathbb{R}^{F_l}$ are called message and update functions, respectively, with F_l being the feature dimension of layer l and $F_0 = F$ in convention. The functions $\{\Phi^{(l)}\}_{l=1}^L$ and $\{\Psi^{(l)}\}_{l=1}^L$ are usually parameterized by multi-layer perceptrons (MLPs).

Taking instantiations of random graphs as inputs, the graph MPNN is the mapping that maps the graph signals of graph nodes to the corresponding node embeddings.

Definition 3.4 (graph message passing neural networks)**.** Given an MPNN Θ, a space Ξ and a random graph $(A, \mathbf{f})$, a *graph MPNN* $\Theta_A(\mathbf{f})$ is defined as the mapping $\Theta_A(\mathbf{f}) : \mathbb{R}^{N \times F} \to \mathbb{R}^{N \times F_L}, \mathbf{f} \mapsto \mathbf{f}^{(L)} = (\mathbf{f}_1^{(L)}, \dots, \mathbf{f}_N^{(L)})^\top$. Let $\mathbf{f}^{(0)} = \mathbf{f}$ be the initial signal, the node embeddings $\mathbf{f}^{(l)} \in \mathbb{R}^{N \times F_l}$ at layer $l = 1, \dots, L$ are iteratively defined

$$\begin{cases} \mathbf{m}_i^{(l)} \overset{def.}{=} \sum_{j=1}^N \frac{a_{ij}}{d_i} \Phi^{(l)}(\mathbf{f}_i^{(l-1)}, \mathbf{f}_j^{(l-1)}), & \{\textit{message passing}\} \quad (2) \\ \mathbf{f}_i^{(l)} \overset{def.}{=} \Psi^{(l)}(\mathbf{f}_i^{(l-1)}, \mathbf{m}_i^{(l)}). & \{\textit{message updating}\} \quad (3) \end{cases}$$

The *graphon MPNN* extends graph MPNN by replacing $(A, \mathbf{f})$ by its continuous counterpart $(\mathcal{A}, f)$.

Definition 3.5 (graphon message passing neural networks)**.** Given an L-layer MPNN Θ, a space Ξ and a graphon RGM $(\mathcal{A}, f)$, a *graphon MPNN* $\Theta_{\mathcal{A}}(f)$ is defined as the mapping $\Theta_{\mathcal{A}}(f) : L^2(\chi) \to L^2(\chi), f \mapsto f^{(L)}$. A graphon MPNN maps a metric-space signal to another signal. With initial

[1] Conventionally, a graph signal is a scalar-valued function (Ortega et al., 2018), with F signals in the multi-dimensional setting.

input $f^{(0)} = f : \chi \to \mathbb{R}^F$, the metric-space signal $f^{(l)} : \chi \to \mathbb{R}^{F_l}$ at layer $l = 1, \ldots, L$ is iteratively defined as

$$\begin{cases} m^{(l)}(s) \overset{def.}{=} \int_{\chi} \frac{A(s,t)}{d_{\mathcal{A}}(s)} \Phi^{(l)}(f^{(l-1)}(s), f^{(l-1)}(t)) d\mu(t), & (4) \\ f^{(l)}(s) \overset{def.}{=} \Psi^{(l)}(f^{(l-1)}(s), m^{(l)}(s)). & (5) \end{cases}$$

The $m^{(l)}(s)$ in Eq.(4) and $f^{(l)}(s)$ in Eq.(5) can be viewed as the continuous version of message passing and updating in Eq.(2) and Eq.(3), respectively. The graphon MPNN $\Theta_{\mathcal{A}}(f) : \chi \to \mathbb{R}^{F_L}$ in Def. 3.5 can be viewed as the continuous version of the graph MPNN $\Theta_A(\mathbf{f}) \in \mathbb{R}^{N \times F_L}$ in Def. 3.4 for a random graph $(A, \mathbf{f}) \sim (\mathcal{A}, f)$[2].

4. Method

In this section, we first introduce classification task and size generalization in Sect. 4.1. We emphasize the impact of graph size on the generalization for MPNNs with pooling layers (cf. Def. 4.1) by quoting the result from Maskey et al. (2022) (cf. Thm. 4.2). Then we formalize the *Wasserstein Barycenter Matching* layer and MPNNs with WBM layers (cf. Def. 4.6) in Sect. 4.2. Finally, we discuss the convergence (cf. Thm. 4.9) and generalization (cf. Thm. 4.10) of an MPNN with a WBM layer in Sect 4.3.

4.1. Problem Formulation and Analysis

Data generation. In a C-class graph-level classification task, we are provided with a training dataset $\mathcal{S} = \{\mathbf{x}_k = (A_k, \mathbf{f}_k), \mathbf{y}_k\}_{k=1}^n$ consisting of n graph instances. Each instance in $\mathcal{S}$ is from a unique class from $\{1, \ldots, C\}$. We assume that graph class j is associated with a metric measure space (χ^j, d^j, μ^j) and a graphon RGM $(\mathcal{A}^j, f^j)$ for $j \in \{1, \ldots, C\}$. The instances in set $\mathcal{S}$ are assumed to be i.i.d. drawn from a probabilistic measure $\mu_{\mathcal{G}} := \sum_{j=1}^C h^j \mu_{\mathcal{G}_j}$, with $h^j := P(\mathbf{y} = j)$ for $j = 1, \ldots, C$ denoting the probability of an instance sampled from the class j. For simplicity of exposition, we assume that all the graphs in the training dataset are $\boldsymbol{N}$**-nodes graphs**. The measure of a measurable N-element set $V = \{X_1, \ldots, X_N\} \subset (\chi^j)^N$ is defined as $\mu_{\mathcal{G}_j}(V) := \prod_{i=1}^N \mu^j(X_i)$. Therefore sampling one N-node graph instance $(\mathbf{x}, \mathbf{y})$ w.r.t. measure $\mu_{\mathcal{G}}$ can be viewed as choosing a class $\mathbf{y} \in \{1, \ldots, C\}$ w.r.t. the simplex $(h^1, \ldots, h^C)$ first, then a random graph $(A, \mathbf{f}) \sim (\mathcal{A}^{\mathbf{y}}, f^{\mathbf{y}})$ is drawn from the space $\chi^{\mathbf{y}}$ w.r.t. the measure $(\mu^{\mathbf{y}})^N$.

The **goal** of classification tasks is to minimize the *generalization risk* (*risk* for short) $R_{exp}(\Theta) := \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mu_{\mathcal{G}}}[\ell(\Theta(\mathbf{x}), \mathbf{y})]$ for a *loss function*[3] ℓ. Explicit computation of $R_{exp}(\Theta)$ is often intractable. In practice, in an *empirical risk minimization* (ERM) (Vapnik, 1999) framework, one computes the *empirical risk* $R^S_{emp}(\Theta) := \frac{1}{n}\sum_{k=1}^n \ell(\Theta(\mathbf{x}_k), \mathbf{y}_k)$, based on the training set $\mathcal{S} = \{\mathbf{x}_k = (A_k, \mathbf{f}_k), \mathbf{y}_k\}_{k=1}^n$. Ideally, the gap between the empirical risks and the expected risk $|\hat{R}^S_{emp}(\Theta) - R_{exp}(\Theta)|$ should be small with a high probability for moderate sample size n. Meanwhile, if the instances in sets $\mathcal{S}$ and $\mathcal{T} = \{(A_k, \mathbf{f}'_k), \mathbf{y}'_k\}_{k=1}^{n'}$ are i.i.d. and n, n' satisfying the sample complexity required by ERM, the empirical risk gap $|\hat{R}^S_{emp}(\Theta) - \hat{R}^T_{emp}(\Theta)|$ should be small to guarantee testing performance on the testing set $\mathcal{T}$.

Size generalizability across graph sets $\mathcal{S}$ and $\mathcal{T}$ is not guaranteed for n, n' satisfying ERM sample complexity when the graph sizes vary from $\mathcal{S}$ to $\mathcal{T}$, denoted by $N_S \neq N_T$. In other words, the empirical risk gap $|\hat{R}^S_{emp}(\Theta) - \hat{R}^T_{emp}(\Theta)|$ could be large. The N_S-nodes instances in $\mathcal{S}$ and N_T-nodes instances in $\mathcal{T}$ are non-i.i.d. Recalling the generating process, the N_S-nodes (N_T-nodes resp.) graphs are sampled w.r.t. the product measure $(\mu^j)^{N_S}$ ($(\mu^j)^{N_T}$ resp.). More importantly, the stochastic sampling procedure for N nodes in each graph inflates the usual ERM sample complexity that depends only on sample size n to stabilize the learning.

Taking the stochasticity of graph nodes into account, Maskey et al. (2022) quantify the impact of graph size on the generalization of MPNNs with an average pooling layer (cf. Thm. 4.2). The function of the pooling layer converts the matrix of node embeddings $\Theta_A(\mathbf{f}) \in \mathbb{R}^{N \times F_L}$ to a *vectorized graph representation* for graph-level classification.

Definition 4.1 (pooling layer of MPNNs)**.** Given an MPNN Θ, the *average pooling layer* over nodes of a random graph $(A, \mathbf{f}) \sim (\mathcal{A}, f)$ for a graph MPNN in Def. 3.4 and the average pooling layer for a graphon MPNN in Def. 3.5 are

$$\begin{cases} \Theta^P_A(\mathbf{f}) \overset{def.}{=} 1/N(\Theta_A(\mathbf{f}))^\top \mathbf{1}_N, \ \{\textit{graph MPNN pooling}\} & (6) \\ \Theta^P_{\mathcal{A}}(f) \overset{def.}{=} \int_{\chi} \Theta_{\mathcal{A}}(f)(s) d\mu. \ \{\textit{graphon MPNN pooling}\} & (7) \end{cases}$$

We rehearse a simplified Thm.3.3 in Maskey et al. (2022).

Theorem 4.2 (generalization of MPNNs with a pooling layer, Thm3.3 Maskey et al. (2022))**.** *Given an MPNN Θ, a loss function ℓ and a set of N-nodes graphs $\mathcal{S} \sim \mu^n_{\mathcal{G}}$. Under regularity assumptions, there exists a constant B such that*

$$\mathop{\mathbb{E}}_{\mathcal{S}\sim\mu^n_{\mathcal{G}}} \left[\left(\hat{R}_{emp}(\Theta^P_A) - R_{exp}(\Theta^P_{\mathcal{A}}) \right)^2 \right] \leq \frac{8\|\ell\|^2_\infty \pi 2^C}{n} + \frac{BL^2_\ell 2^C C}{n} \sum_j h_j (\|f^j\|_\infty + L^2_{f^j}) \alpha(N, D_{\chi^j}), \textit{where} \quad (8)$$

$$\alpha(N, D_{\chi^j}) = \frac{1}{N} + \frac{1 + \log(N)}{N^{1/(D_{\chi^j}+1)}} + \mathcal{O}(N^{\frac{3(L-1)}{2}}/e^N). \quad (9)$$

In Eq.(8), L_ℓ and L_{f^j} are Lipschitz constants of loss ℓ and metric-space signal f^j, respectively. The D_{χ^j} is the dimension of the underlying space χ^j. The regularity assumptions are specified in the Appx. A.

[2] We use $\Theta_A(\mathbf{f})$ to denote the mapping itself or thereof image interchangeably, whenever it is clear from the context.

[3] Conventionally the loss function ℓ is taken as a fixed function, whereas we integrate some end-to-end learnable parameters (e.g., the cross-entropy loss composed on softmax, making ℓ Lipschitz continuous) in ℓ to isolate the MPNN Θ for ease of exposition.

The generalization upper bound for MPNNs with a pooling layer in Thm. 4.2 consists of two terms. The first term $8\|\ell\|_\infty^2 \pi 2^C/n$ is much smaller than the second term for typical neural networks. For a fixed sample size n and model complexity, the upper bound in Eq.(8) is dominated by the $\mathcal{O}(log(N)N^{-1/(D_\chi+1)})$ in the second term.

Given an MPNN Θ with a pooling layer, suppose that instances of an N_S-nodes graph set $\mathcal{S}$ and an N_T-nodes graph set $\mathcal{T}$ are drawn from the same graphon RGM $\mu_\mathcal{G}$, then we have a quantitative assessment on the size generalizability.

Proposition 4.3 (size generalizability of a graph MPNN with a pooling layer)**.** *Suppose the conditions of Thm. 4.2 are satisfied, then for a set of N_S-nodes graphs $\mathcal{S} \sim \mu_\mathcal{G}^n$ and a set of N_T-nodes graphs $\mathcal{T} \sim \mu_\mathcal{G}^{n'}$, fixing n, n' we have*

$$\left| \mathop{\mathbb{E}}_{\mathcal{S}\sim\mu_\mathcal{G}^n}[\hat{R}_{emp}(\Theta_A^P)] - \mathop{\mathbb{E}}_{\mathcal{T}\sim\mu_\mathcal{G}^{n'}}[\hat{R}_{emp}(\Theta_A^P)] \right| \leq \sum_j \mathcal{O}\left(\frac{log(N_S)^{1/2}}{N_S^{-1/2(D_{\mathcal{X}^j}+1)}} \right) + \mathcal{O}\left(\frac{log(N_T)^{1/2}}{N_T^{-1/2(D_{\mathcal{X}^j}+1)}} \right) \tag{10}$$

Proof. The proof of Prop. 4.3 is straightforward according to Thm. 4.2 and the triangle inequality. □

The $-1/2D_{\mathcal{X}^j}$ rate roots in the correlations across nodes entailed by the graph structure. Because not every metric spaces χ^j is necessarily a low-dimensional manifold, we are urged to develop a mechanism for MPNNs to alleviate the slow rate caused by a possibly large uncontrollable $D_{\mathcal{X}^j}$.

4.2. The Wasserstein Barycenter Matching Layer

We propose a *Wasserstein Barycenter Matching* (WBM) layer that exploits Wasserstein barycenters in the Wasserstein space as graph-level consensus to combat the adverse effect caused by the nodes-level correlation. We introduce the formalization of the WBM layer in this subsection. The controllable convergence rate will be discussed in Sect. 4.3. We start with introducing the *p-Wasserstein space*.

Definition 4.4 (p-Wasserstein space Ambrosio et al. (2005))**.** Given $p \in [1, +\infty)$ and a closed convex set $\Omega \in \mathbb{R}^D$. Let $\mathbb{P}_p(\Omega)$ be the set of probability measures over Ω with finite p-order moments. The metric space $\mathbb{W}_p(\Omega) = (\mathbb{P}_p(\Omega), \mathcal{W}_p)$ is called the *p-Wasserstein space*, with *p-Wasserstein distance* between measures $\rho, \nu \in \mathbb{P}_p(\Omega)$ defined as

$$\mathcal{W}_p(\rho,\nu) \overset{def.}{=} \left(\inf_{\pi\in\Pi(\rho,\nu)} \int_{\Omega^2} \|s-t\|^2 d\pi(s,t) \right)^{1/p}, \tag{11}$$

where $\Pi(\rho,\nu)$ is the set of couplings on $\mathbb{R}^D \times \mathbb{R}^D$ with ρ and ν as marginals, and $\|\cdot\|$ is the Euclidean norm.

The *Wasserstein barycenter* is a natural extension of the mean of probability distributions on the Wasserstein space.

Definition 4.5 (Wasserstein barycenter Agueh & Carlier (2011))**.** For $p \in [1, +\infty))$, the *p-Wasserstein barycenter* $\boldsymbol{b}_p(\mathcal{P})$ of $\mathcal{P} \in \mathbb{P}_p(\mathbb{W}_p(\Omega))$ is defined as follows,

$$\boldsymbol{b}_p(\mathcal{P}) \overset{def.}{=} \mathop{\arg\min}_{\rho\in\mathbb{P}_p(\mathbb{R}^D)} \mathop{\mathbb{E}}_{\nu\sim\mathcal{P}}[\mathcal{W}_p^p(\rho,\nu)]. \tag{12}$$

The *(uniform-weighted) empirical p-Wasserstein barycenter* of the empirical distribution $\hat{\mathcal{P}}_n = 1/n \sum_{k=1}^n \delta_{\nu_k}$ is

$$\hat{\boldsymbol{b}}_p(\{\nu_1,\ldots,\nu_n\}) \overset{def.}{=} \mathop{\arg\min}_{\rho\in\mathbb{P}_p(\mathbb{R}^D)} \frac{1}{n}\sum_{k=1}^n \mathcal{W}_p^p(\rho,\nu_k). \tag{13}$$

Throughout the paper, we assume the existence of at least one Wasserstein barycenter per class, which is shown to hold in reasonable scenarios (Afsari, 2011).

Similar to the pooling layer in Def. 4.1, a WBM layer attaches graph MPNNs and vectorizes the matrix of node embeddings $\Theta_A(\mathbf{f}) \in \mathbb{R}^{N\times F_L}$. Given an MPNN Θ, the WBM layer collects node embeddings of an input graph and represents the graph by distances between the graph-wise measure and class-wise empirical Wasserstein barycenters.

Concretely, let $\{(\chi^j, d^j, \mu^j)\}_{j=1}^C$ be the metric measure spaces of different graph classes. Suppose that $\mathcal{S} = \{\mathbf{x}_k = (A_k, \mathbf{f}_k), \mathbf{y}_k\}_{k=1}^n$ is a dataset of N-nodes graphs drawn w.r.t. measure $\mu_\mathcal{G} = \sum_{j=1}^C h^j \mu_{\mathcal{G}_j}$. Let R be the same-class *equivalence relation* on $\mathcal{S}$, suppose the *quotient set* of R on $\mathcal{S}$ denoted as $\mathcal{S}/R = \{\mathcal{S}^1,\ldots,\mathcal{S}^C\}$. In other words, $\mathcal{S}$ can be partitioned into the disjoint union: $\mathcal{S} = \bigsqcup_{j=1}^C \mathcal{S}^j$. Denote by n_j the cardinality of the j class set $\mathcal{S}^j$ with $\sum_{j=1}^C n_j = n$. Denote by $\Theta_\# \mu$ the push-forward measure of a measure μ by a measurable MPNN Θ. For a random graph $(A_k^j, \mathbf{f}_k^j) \sim (\mathcal{A}^j, f^j)$ from class j, we consider the push-forward of the measure $\hat{\mu}_k^j = 1/N \sum_{i=1}^N \delta_{X_{k,i}^j}$ by the composite mapping $\Theta_{A_k^j}(\mathbf{f}_k^j) \circ f_k^j$, denoted by $\hat{\nu}_k^j = (\Theta_{A_k^j}(\mathbf{f}_k^j) \circ f_k^j)_\# \hat{\mu}_k^j$.

In contrast to the parameter-free average pooling layer, for each class $j \in \{1,\ldots,C\}$, the WBM layer estimates the empirical 2-Wasserstein barycenter of the empirical distribution $\hat{\mathcal{P}}_{n_j}^j = 1/n_j \sum_{k=1}^{n_j} \delta_{\hat{\nu}_k^j}$, i.e.,

$$\hat{\boldsymbol{b}}_2(\mathcal{S}^j) \overset{def.}{=} \mathop{\arg\min}_{\rho\in\mathbb{P}_2(\mathbb{R}^{F_L})} \frac{1}{n_j}\sum_{k=1}^{n_j} \mathcal{W}_2^2(\rho,\hat{\nu}_k^j). \tag{14}$$

In practice, we estimate $\hat{\boldsymbol{b}}_2(\mathcal{S}^j)$ for class $j \in \{1,\ldots,C\}$ with other neural network parameters in the end-to-end learning process with the following optimization objective,

$$\ell_{WBM} \overset{def.}{=} \sum_{j=1}^C \frac{1}{n_j}\sum_{k=1}^{n_j} \mathcal{W}_2^2(\hat{\boldsymbol{b}}_2(\mathcal{S}^j), \hat{\nu}_k^j). \tag{15}$$

Finally, we formalize the MPNNs with a WBM layer, which represents each input graph by 2-Wasserstein distances be-

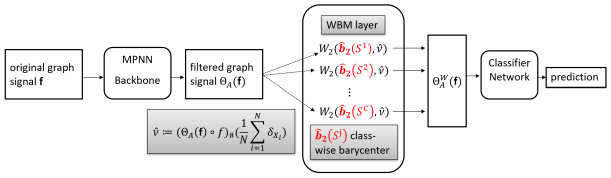

Figure 1. The framework of an MPNN with a WBM layer.

tween the graph-wise (empirical) measure and class-wise empirical Wasserstein barycenters $\{\hat{\boldsymbol{b}}_2(\mathcal{S}^j)\}_{j=1}^C$.

Definition 4.6 (MPNNs with a WBM layer)**.** Given an MPNN Θ and a set of graphs $\mathcal{S} = \{(A_k, \mathbf{f}_k), \mathbf{y}_k\}_{k=1}^n$, the graph MPNN in Def. 3.4 with a WBM layer $\Theta_A^{\mathcal{W}}(\mathbf{f})$ and graphon MPNN in Def. 3.5 with a WBM layer $\Theta_{\mathcal{A}}^{\mathcal{W}}(f)$ for an N-nodes random graph $(A, \mathbf{f}) \sim (\mathcal{A}, f)$ are

$$\begin{cases} \Theta_A^{\mathcal{W}}(\mathbf{f}) \overset{def.}{=} \left(\mathcal{W}_2(\hat{\boldsymbol{b}}_2(\mathcal{S}^1), \hat{\nu}), \ldots, \mathcal{W}_2(\hat{\boldsymbol{b}}_2(\mathcal{S}^C), \hat{\nu})\right), & (16) \\ \Theta_{\mathcal{A}}^{\mathcal{W}}(f) \overset{def.}{=} \left(\mathcal{W}_2(\boldsymbol{b}_2^1, \nu), \ldots, \mathcal{W}_2(\boldsymbol{b}_2^j, \nu)\right), & (17) \end{cases}$$

where $\hat{\nu}$ in Eq.(16) is the push-forward measure of $\hat{\mu} = 1/N \sum_{i=1}^N \delta_{X_i^j}$ by the composite mapping $\Theta_A(\mathbf{f}) \circ f$, i.e., $\hat{\nu} = (\Theta_A(\mathbf{f}) \circ f)_{\#}\hat{\mu}$, and ν in Eq.(17) is the push-forward of μ by the composition $\Theta_{\mathcal{A}}(f) \circ f$, i.e., $\nu = (\Theta_{\mathcal{A}}(f) \circ f)_{\#}\mu$. In particular, denoting $\nu^j = (\Theta_{\mathcal{A}}(f) \circ f)_{\#}\mu^j$, $\{\boldsymbol{b}_2^j\}_{j=1}^C$ in Eq.(17) are the graphon extensions of empirical Wasserstein barycenters $\{\hat{\boldsymbol{b}}_2(\mathcal{S}^j)\}_{j=1}^C$, we have the following equality

$$\boldsymbol{b}_2^j \overset{def.}{=} \underset{\rho \in \mathbb{P}_q(\mathbb{R}^{F_L})}{\arg\min} \frac{1}{n_j} \sum_{k=1}^{n_j} \mathcal{W}_2^2(\rho, \nu_k^j) = \nu^j. \quad (18)$$

The graph classifier of an MPNN with a WBM layer is learned by optimizing the cross-entropy loss with the WBM loss in Eq.(15) in a common end-to-end learning fashion.

We illustrate the framework of an MPNN with a WBM layer in Figure 1.The detailed learning algorithm of an MPNN with a WBM layer is in Appx. H.

Remark 4.7. For ease of exposition, we assume that each graph class is related to a single graphon. It is reasonable that graphs sampled from the same graphon RGM are from the same class, but not necessarily vice versa. In practice, we may assume a hyperparameter M accommodating the number of graphons corresponding to each class. When $M > 1$, we actually construct an extended probabilistic measure reads $\mu_G = \sum_{j=1}^C P(y = j)/M \sum_{m=1}^M \mu_{G_{jm}}$, where $\{\mu_{G_{jm}}\}_{m=1}^M$ corresponds to the M different graphon RGMs of class j. We may think of the measure corresponding to each class as a mixture of M components with the same weight. We then feed the vector of Wasserstein distances to a non-linear MLP.

Remark 4.8. The graph size in real-world datasets possibly varies. The N-nodes graph assumption is placed for ease of illustration and theoretical analysis. In practice, the proposed WBM layer can process graphs of finite size, by considering the push-forward measures of measures $\{\hat{\mu}_k = 1/N_k \sum_{i=1}^{N_k} \delta_{X_{k,i}}\}_{k=1}^n$ for varying N_ks.

4.3. Theoretical Analysis on the MPNNs with WBM

The graph MPNN with a WBM layer has a controllable convergence rate that is independent of the dimension $D_{\mathcal{X}^j}$. We provide our theoretical results on convergence (Thm. 4.9), generalization (Thm. 4.10), and size generalization (Prop. 4.11) of a graph MPNN with a WBM layer.

Following high-dimensional statistics practice (Vershynin, 2018), we restrict the tail behaviour of the measures $\{\mu^j\}_{j=1}^C$ by assuming they are K-sub-Gaussians on $\mathcal{X}^j$, i.e.,

$$\int_{\mathcal{X}^j} e^{\|s\|^2/(2D_{\mathcal{X}^j}K^2)} d\mu^j(s) \leq 2, \quad j = 1, \ldots, C. \quad (19)$$

Theorem 4.9 (convergence of an MPNN with a WBM layer)**.** *Given an MPNN Θ, a loss function ℓ and a set of N-nodes graphs $\mathcal{S} = \{(A_k, \boldsymbol{f}_k), \boldsymbol{y}_k\}_{k=1}^n \sim \mu_{\mathcal{G}}^n$, with $\mu_{\mathcal{G}} = \sum_{j=1}^C h^j \mu_{\mathcal{G}_j}$, $\{\mu^j\}_{j=1}^C$ are K-sub-Gaussians and $h^j = P(\boldsymbol{y} = j)$. Denote by $\mathcal{S} = \bigsqcup_{j=1}^C \mathcal{S}^j$ a same-class-partition of $\mathcal{S}$, with $n_j = |S^j| \geq 1$. Let $\Theta_A^{\mathcal{W}}(\boldsymbol{f})$ be a graphon MPNN with a WBM layer and $\Theta_{\mathcal{A}}^{\mathcal{W}}(f)$ be a graphon MPNN with a WBM layer. Let $\eta \in (0, 1)$. Under the regularity assumptions in Appx. A (same as in Thm. 4.2), we have the following with probability (w.p.) $\geq 1 - \eta - e^{-C_2 N} - e^{-C_2' n^*}$*

$$\|\Theta_A^{\mathcal{W}}(\boldsymbol{f}) - \Theta_{\mathcal{A}}^{\mathcal{W}}(f)\|^2 \leq C[\alpha_{C_1}(N, n^*, \eta) + \beta_{K', C_3}(N, \eta)]^2, \quad (20)$$

$$\text{with } \alpha_{C_1}(N, n^*, \eta) = \sqrt{\frac{C_1}{N}\log(\frac{8}{\eta}) + \frac{C_1}{n^*}\log(\frac{8}{\eta})}, \quad (21)$$

$$\beta_{K', C_3}(N, \eta) = 2N^{-\frac{1}{2}}\log(\frac{4}{\eta}) + \sqrt{16K'^2\log(\frac{4}{\eta}) + \frac{8}{N}\log(\frac{4}{\eta})}$$
$$+ C_3 K'^2 \times \begin{cases} \mathcal{O}(N^{-1/3}) & \text{if } F_L \in \{1, 2, 3, 4\} \\ N^{-2/F_L} + N^{-1/3} & \text{if } F_L > 4 \end{cases}, \quad (22)$$

where $n^* = min(n_1, \ldots, n_C)$, C_1, C_2, C_2', C_3 and K' are constants. The proof and details of the constants are specified in the Appx. B.1.

We further derive the generalization upper bound of a graph MPNN when the loss function ℓ is Lipschitz-continuous.

Theorem 4.10 (generalization of an MPNN with a WBM layer)**.** *Under the same conditions as in Thm. 4.9, assuming that there is at least one instance per class in $\mathcal{S}$, the following inequality holds w.p. $\geq 1 - \eta - e^{-C_2 N} - e^{-C_2'}$,*

$$\mathop{\mathbb{E}}_{\mathcal{S} \sim \mu_{\mathcal{G}}^n}\left[\left(\hat{R}_{emp}(\Theta_A^{\mathcal{W}}) - R_{exp}(\Theta_{\mathcal{A}}^{\mathcal{W}})\right)^2\right] \leq \frac{8\|\ell\|_\infty^2 \pi 2^C}{n} +$$
$$\frac{\pi^{\frac{1}{2}} L_\ell^2 2^C C^2}{n} \sum_{j=1}^C h_j [\alpha_{C_1}(N, 1, \eta) + \beta_{K', C_3}(N, \eta)]^2, \quad (23)$$

where $\alpha_{C_1}(N,1,\eta)$ is defined in Eq.(21) and $\beta_{K',C_3}(N,\eta)$ is defined Eq.(22). The constants C_1, C_2, C_2', C_3 and K' are the same as in Thm. 4.9. More details and the proof are specified in Appx. B.2.

Finally, we may also quantify the size generalizability of a graph MPNN with a WBM layer across graph sets $\mathcal{S}$ and $\mathcal{T}$ with different graph sizes $N_S \neq N_T$, assuming the instances in both sets are drawn from the same graphon RGM $\mu_{\mathcal{G}}$. Based on Thm. 4.10, the triangle inequality, and the union bound, we have the following proposition.

Proposition 4.11 (size generalizability of a graph MPNN with a WBM layer)**.** *Given a set of N_S-nodes graphs $\mathcal{S} \sim \mu_{\mathcal{G}}^n$ and a set of N_T-nodes graphs $\mathcal{T} \sim \mu_{\mathcal{G}}^{n'}$. Let Θ_A be a graph MPNN, under conditions of Thm. 4.10 and fixed same size n, n', w.p. $\geq 1 - 2\eta - e^{-C_2 min(N_S, N_T)} - e^{-C_2'}$*

$$\left| \mathop{\mathbb{E}}_{\mathcal{S}\sim\mu_{\mathcal{G}}^n}[\hat{R}_{emp}(\Theta_A^{\mathcal{W}})] - \mathop{\mathbb{E}}_{\mathcal{T}\sim\mu_{\mathcal{G}}^{n'}}[\hat{R}_{emp}(\Theta_A^{\mathcal{W}})] \right| \leq \begin{cases} \sum_j \mathcal{O}(N_S^{-1/3}) + \mathcal{O}(N_T^{-1/3}) & \text{if } 1 \leq F_L \leq 6 \\ \sum_j \mathcal{O}(N_S^{-2/F_L}) + \mathcal{O}(N_T^{-2/F_L}) & \text{if } F_L > 6 \end{cases} . \tag{24}$$

Remark 4.12. In the low-dimension regime, an MPNN with a WBM layer has a convergence rate at $-1/3$, which is provably faster than $-1/2(D_\chi+1)$ when $D_\chi \in [1,6]$. In the high-dimensional regime, an MPNN with a WBM layer has a convergence rate at $-2/F_L$, which is tuneable compared to the uncontrollable $-1/2(D_\chi+1)$. Our non-asymptotic results in Thm. 4.9, Thm. 4.10 and Prop. 4.11 show that the proposed WBM provably improves the vanilla MPNN in convergence rate, generalization, and size generalizability for limited sample size and average graph size.

4.4. Complexity Analysis

The main source of additional time complexity in our method arises from the computation of the Wasserstein distance between every example and every Barycenter. This complexity is determined by the mean graph size, the number of barycenters, and the Wasserstein solver we used. Computing Wasserstein distance involves solving a linear programming optimization problem under linear constraints. It can be performed by using the network simplex algorithm as done by Pele & Werman (2009); Bonneel et al. (2011) in $\mathcal{O}(n^3)$ times, or approximately up to ϵ via the Sinkhorn algorithm (Cuturi, 2013) in $\mathcal{O}(n^2/\epsilon^3)$ time. In our implementation, we used the EMD solver from the POT library (Flamary et al., 2021), whose complexity is up to $\mathcal{O}(n^3)$. Assume that the training set consists of N graphs with mean size N_s, and there are M barycenters with size N_g per class (totally C classes). For simplicity, we assume $N_s = N_g$. In each epoch, the time for computing the cost matrix is $\mathcal{O}(NMC)$, and the time for computing the optimal transport plan by the EMD solver is up to $\mathcal{O}(NMCN_s^3)$. Therefore the overall overhead is $\mathcal{O}(NMCN_s^3)$.

The main space complexity overhead is the space needed to store the cost matrix and transport plan between each sample and each barycenter. For a graph with N_s nodes and a barycenter with N_g nodes, it occupies $O(N_sN_g)$ space. Assume there are totally N samples and MC barycenters, the overall space overhead is $O(NMC \times N_sN_g)$. The space required to store additional parameters of WBM layer is $O(MCN_gF_L)$ where F_L is the hidden layers dimension, and is negligible when $NN_s >> F_L$.

5. Experiments

To validate the effectiveness of the WBM layer for size generalization, we conduct experiments on two sets of datasets[4]. The first set of four datasets (NCI1, NCI109, PROTEINS, and DD) from the TUDataset (Morris et al., 2020) is the standard protocol in prior works for evaluating size generalization (Bevilacqua et al., 2021; Buffelli et al., 2022). We report the corresponding results in the main text. The second set contains two **larger** datasets (GOOD-Motif and GOOD-HIV) for graph-level classification with covariate shifts in graph sizes from the Graph OOD (Gui et al., 2022) benchmark. We delay the corresponding results in Appx. F. More details on the datasets such as data splits are in Appx. C.

5.1. Experimental Settings

Following Buffelli et al. (2022), we employ the proposed WBM layer on three different MPNN backbones: GCN (Kipf & Welling, 2017), GIN (Xu et al., 2019), and PNA (Corso et al., 2020). We compare our method with the following baselines: (1) Two graph kernels, the Graphlet Counting kernel (GC kernel) (Shervashidze et al., 2009) and Weisfeiler-Lehman kernel (WL kernel) (Shervashidze et al., 2011). (2) Invariant Risk Minimization (IRM) (Arjovsky et al., 2019). (3) E-invariant models ($\Gamma_{1-hot}, \Gamma_{GIN}, \Gamma_{RPGIN}$) introduced in Bevilacqua et al. (2021). (4) Central Moment Discrepancy regularization (CMDr) introduced in Buffelli et al. (2022). More details of baselines are in Appx. D. Following Buffelli et al. (2022), we use Matthews correlation coefficient (MCC) as the evaluation metric for its reliability in imbalanced classification (Chicco & Jurman, 2020). MCC ranges from -1 to 1, with 1 indicating perfect agreement of predictions with ground truth. We report the mean MCC and standard deviation of 10 independent trials. We use a 3-layer MPNN (with different backbones) before a WBM layer. We select the size of the Wasserstein barycenters from the {max, median} size of the observed graphs based on the validation set. We initialize barycenters by sampling from the training set. We fix the number of graphons (barycenters) per class $M = 3$. Our method and baselines adopt the JKNet architecture (Xu et al., 2018). More implementation details are in Appx. E.

[4]Codes are available at https://github.com/JinYujie99/WBM

Table 1. MCC (mean ± std) on the size generalization test set. The models are original MPNNs without (×) and with (✓) the WBM layer. The right-most column shows the average improvement brought by the WBM layer.

Backbone	GIN		GCN		PNA		Avg Impr
WBM layer	×	✓	×	✓	×	✓	
NCI109	0.18 ± 0.05	0.24 ± 0.05	0.15 ± 0.06	0.22 ± 0.04	0.23 ± 0.07	0.25 ± 0.04	$\uparrow 29.6\%$
NCI1	0.19 ± 0.06	0.24 ± 0.04	0.17 ± 0.06	0.19 ± 0.05	0.19 ± 0.08	0.21 ± 0.08	$\uparrow 16.2\%$
PROTEINS	0.25 ± 0.07	0.37 ± 0.08	0.21 ± 0.10	0.35 ± 0.09	0.22 ± 0.12	0.25 ± 0.09	$\uparrow 42.8\%$
DD	0.23 ± 0.09	0.27 ± 0.06	0.24 ± 0.07	0.28 ± 0.10	0.23 ± 0.09	0.26 ± 0.09	$\uparrow 15.7\%$

5.2. Main Results

Table 1 shows the performances of different MPNN backbones on the size generalization test set, with and without the proposed WBM layer. The MPNNs with WBM layers consistently outperform the vanilla MPNNs by a large margin, with average improvement brought by the WBM layer up to 42.8%, which manifests the effectiveness of the WBM layer for improving the size generalization of MPNNs. In Table 2, we compare our MPNN with a WBM layer against baselines. It shows that on three of the four datasets, an original MPNN with the WBM layer achieves the best mean performance on the test set. Specifically, our method outperforms the IRM and E-invariant models by a large margin and is competitive with the pure heuristic model CMDr. Furthermore, we observe that always one of the MPNN with WBM layer is among the top 4 best models on all datasets. These results validate the effectiveness of our method.

Table 2. Performance comparisons in MCC between the MPNN with a WBM layer and baselines. Bold emphasizes the top-4 models (in average MCC) for each dataset.

Dataset	NCI109	NCI1	PROTEINS	DD
PNA + IRM	0.20 ± 0.07	0.17 ± 0.07	0.21 ± 0.12	0.24 ± 0.08
GCN + IRM	0.20 ± 0.06	0.22 ± 0.06	0.23 ± 0.16	0.23 ± 0.08
GIN + IRM	0.15 ± 0.04	0.18 ± 0.06	0.24 ± 0.08	0.21 ± 0.10
WL kernel	0.21 ± 0.00	$\mathbf{0.39} \pm 0.00$	0.00 ± 0.00	0.00 ± 0.00
GC kernel	0.01 ± 0.00	0.02 ± 0.00	0.29 ± 0.00	0.00 ± 0.00
Γ_{1-hot}	$\mathbf{0.22} \pm 0.06$	0.15 ± 0.05	0.18 ± 0.08	0.22 ± 0.09
Γ_{GIN}	0.16 ± 0.07	$\mathbf{0.24} \pm 0.05$	0.28 ± 0.10	$\mathbf{0.27} \pm 0.05$
Γ_{RPGIN}	0.19 ± 0.06	$\mathbf{0.26} \pm 0.05$	0.26 ± 0.07	0.20 ± 0.05
PNA + CMDr	$\mathbf{0.24} \pm 0.07$	0.22 ± 0.07	$\mathbf{0.33} \pm 0.09$	$\mathbf{0.27} \pm 0.08$
GCN + CMDr	0.19 ± 0.06	$\mathbf{0.25} \pm 0.06$	0.29 ± 0.13	0.26 ± 0.07
GIN + CMDr	0.20 ± 0.05	0.23 ± 0.08	$\mathbf{0.36} \pm 0.11$	0.25 ± 0.09
PNA + WBM	$\mathbf{0.25} \pm 0.04$	0.21 ± 0.08	0.25 ± 0.09	0.26 ± 0.09
GCN + WBM	0.22 ± 0.04	0.19 ± 0.05	$\mathbf{0.35} \pm 0.09$	$\mathbf{0.28} \pm 0.10$
GIN + WBM	$\mathbf{0.24} \pm 0.05$	$\mathbf{0.24} \pm 0.04$	$\mathbf{0.37} \pm 0.08$	$\mathbf{0.27} \pm 0.06$

Table 3. Ablation studies. Table shows mean MCC over the test data on four datasets with GIN as the backbone.

Dataset	NCI109	NCI1	PROTEINS	DD	Avg
EBM	0.18 ± 0.08	0.22 ± 0.06	0.35 ± 0.09	0.21 ± 0.06	0.24
WBM^0	0.19 ± 0.07	0.21 ± 0.06	0.33 ± 0.11	0.24 ± 0.11	0.24
$WBM^{1/N}$	0.20 ± 0.05	0.16 ± 0.03	0.36 ± 0.10	0.21 ± 0.07	0.23
WBM	$\mathbf{0.24} \pm 0.05$	$\mathbf{0.24} \pm 0.04$	$\mathbf{0.37} \pm 0.08$	$\mathbf{0.27} \pm 0.06$	**0.28**

We conduct PCA visualizations of the WBM embeddings for the PROTEINS dataset. The figures are shown in Figure 2. We observe examples from different categories form well-separated cluster structures as desired. We can also see that the learned Wasserstein barycenters are extreme points in the embedding space of the PCA, which is similar to the phenomenon of TFGW embeddings in Vincent-Cuaz et al. (2022), as the result of using distances as representations.

5.3. The Sensitivity on F_L

In this subsection, we study the empirical impact of F_L on size generalization performance over the test data. We use a GIN model with the WBM layer and vary F_L in $\{8, 16, 32, 64\}$. Figure 3 shows the mean test MCC on the four datasets, respectively. We can observe that the results approximately present bell-shaped curves, which implies a trade-off between discrimination and convergence with regard to hidden layers dimension F_L. On the one hand, a low dimensionality constrains the expressiveness of MPNN, thus impairing the expressiveness of the model and degrading the discrimination. On the other hand, a high dimensionality leads to slow convergence of size generalization, as shown in Thm. 4.9, Thm. 4.10, and Prop. 4.11.

5.4. Ablation Studies

In this subsection, we use a GIN model to do ablation studies to verify the effectiveness of all components of WBM, including using Wasserstein metric space against Euclidean space, imposing the WBM loss ℓ_{WBM} explicitly, and learning Wasserstein barycenters nodes weights. The results are shown in Table 3 and are also obtained from 10 independent trials with different random seeds.

Effect of Wasserstein metric space. To validate the effectiveness of using Wasserstein barycenters, rather than Euclidean barycenters, we compare our model with Euclidean barycenter matching (**EBM**). Specifically, EBM first pools the node embeddings to obtain the graph embeddings for each input graph, and then the EBM layer represents each graph by Euclidean distances between its graph embedding and the class-wise Euclidean barycenters. The class-wise Euclidean barycenters are learned end-to-end by optimiz-

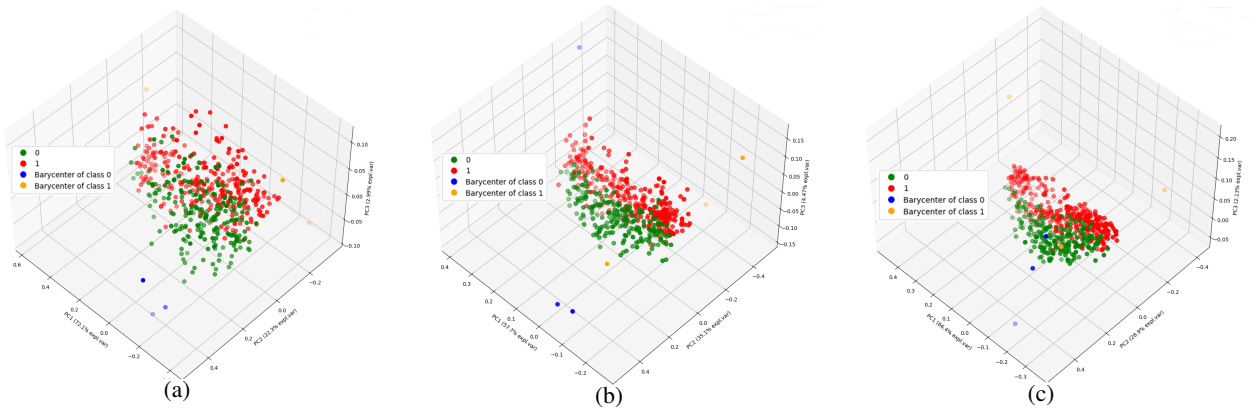

Figure 2. **Visualization of the WBM embeddings using PCA.** Three different runs with three random seeds of the PROTEINS.

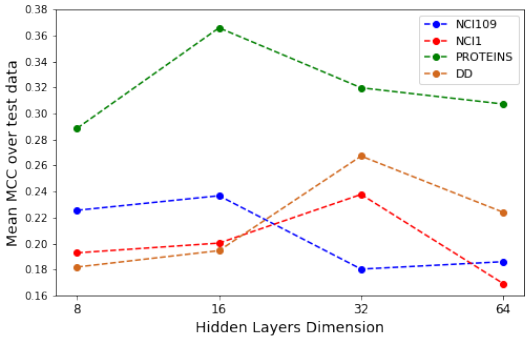

Figure 3. Mean MCC of a GIN model with the WBM layer over the test data, with hidden layers dimension varying in $\{8, 16, 32, 64\}$.

ing the classification loss and a matching loss similar to ℓ_{WBM}. It is shown in Table 3 that WBM outperforms EBM, which validates the superiority of the Wasserstein barycenter. Since it is able to take into account the underlying geometry of the measures while a Euclidean barycenter cannot (Backhoff-Veraguas et al., 2022).

Effect of explicitly imposing ℓ_{WBM}. To show the effectiveness of imposing ℓ_{WBM} explicitly, we compare with the models which only optimize the classification loss (dubbed WBM^0 in Table 3). Imposing ℓ_{WBM} explicitly endows the learned Wasserstein barycenters with semantic meanings related to a particular class. The results show that the performances degrade without ℓ_{WBM}, and empirically justify the necessity of considering class-wise Wasserstein barycenters for graph size generalization.

Learning Wasserstein barycenters nodes weights. In a WBM layer, both the node embeddings and the node weights of the Wasserstein barycenters are learned end-to-end. To analyze the effect of learning node weights, we compare with the models without learning node weights (dubbed $WBM^{1/N}$ in Table 3). These models fix the node weights to uniform $1/N$, where N is the number of nodes of the barycenter. From the results, we can find that the performances degrade when fixing node weights. We perceive that learning node weights weakens the effect of incorrect size settings of barycenters, thus leading to better performance. *Additional ablations studies can be found in the Appendix* G, validating the effectiveness of the components of the proposed WBM method.

6. Conclusion and Discussion

In this paper, we propose a WBM layer, aiming at bridging theoretical understanding and empirical success of size generalization for MPNNs. We give non-asymptotic bounds in convergence, generalization, and size generalizability for an MPNN with a WBM layer. We validate the effectiveness of the WBM layer for size generalization on real-world datasets. There are many future directions on the WBM layer: e.g., exploiting its stability against various types of perturbations, discussing its induced discriminability based on the theory in Balcan et al. (2008), etc. The additional overhead of time complexity is a limitation, potential speedup can be obtained by adapting advanced optimal transport algorithms, such as the primal-dual method in Dvurechensky et al. (2018).

Acknowledgements

This work was supported by the National Key Research and Development Program of China No. 2022ZD0115903, National Natural Science Foundation of China (No. 62250008,62222209), Beijing National Research Center for Information Science and Technology (BNRist) under Grant No.BNR2023RC01003, BNR2023TD03006 and Beijing Key Lab of Networked Multimedia. We appreciate the discussion with Ziwei Zhang, Haoyang Li, and reviewers.

References

Afsari, B. Riemannian L^p center of mass: existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 139(2):655–673, 2011.

Agueh, M. and Carlier, G. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2): 904–924, 2011.

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Backhoff-Veraguas, J., Fontbona, J., Rios, G., and Tobar, F. Bayesian learning with wasserstein barycenters. *ESAIM: Probability and Statistics*, 26:436–472, 2022.

Balcan, M.-F., Blum, A., and Srebro, N. A theory of learning with similarity functions. *Machine Learning*, 72(1-2):89–112, 2008.

Bengio, Y., Lodi, A., and Prouvost, A. Machine learning for combinatorial optimization: a methodological tour d'horizon. *European Journal of Operational Research*, 290(2):405–421, 2021.

Bevilacqua, B., Zhou, Y., and Ribeiro, B. Size-invariant graph representations for graph classification extrapolations. In *International Conference on Machine Learning*, pp. 837–851. PMLR, 2021.

Bonneel, N., Van De Panne, M., Paris, S., and Heidrich, W. Displacement interpolation using lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia conference*, pp. 1–12, 2011.

Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. Spectral networks and locally connected networks on graphs. In *International Conference on Learning Representations*, 2013.

Buffelli, D., Lio, P., and Vandin, F. Sizeshiftreg: a regularization method for improving size-generalization in graph neural networks. In *Advances in Neural Information Processing Systems*, 2022.

Chen, B., Bécigneul, G., Ganea, O.-E., Barzilay, R., and Jaakkola, T. Optimal transport graph neural networks. *arXiv preprint arXiv:2006.04804*, 2020.

Chicco, D. and Jurman, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1): 1–13, 2020.

Corso, G., Cavalleri, L., Beaini, D., Liò, P., and Veličković, P. Principal neighbourhood aggregation for graph nets. *Advances in Neural Information Processing Systems*, 33: 13260–13271, 2020.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Defferrard, M., Bresson, X., and Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Dvurechensky, P., Gasnikov, A., and Kroshnin, A. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In *International Conference on Machine Learning*, pp. 1367–1376. PMLR, 2018.

Erdős, P., Rényi, A., et al. On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, 5(1):17–60, 1960.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., et al. Pot: Python optimal transport. *The Journal of Machine Learning Research*, 22(1):3571–3578, 2021.

Fournier, N. and Guillin, A. On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3-4):707–738, 2015.

Garg, V., Jegelka, S., and Jaakkola, T. Generalization and representational limits of graph neural networks. In *International Conference on Machine Learning*, pp. 3419–3430. PMLR, 2020.

Gasteiger, J., Shuaibi, M., Sriram, A., Günnemann, S., Ulissi, Z., Zitnick, C. L., and Das, A. How do graph networks generalize to large and diverse molecular systems? *arXiv preprint arXiv:2204.02782*, 2022.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pp. 1263–1272. PMLR, 2017.

Gui, S., Li, X., Wang, L., and Ji, S. Good: A graph out-of-distribution benchmark. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Han, X., Jiang, Z., Liu, N., and Hu, X. G-mixup: Graph data augmentation for graph classification. In *International Conference on Machine Learning*. PMLR, 2022.

Holland, P. W., Laskey, K. B., and Leinhardt, S. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

Joshi, C. K., Cappart, Q., Rousseau, L.-M., and Laurent, T. Learning tsp requires rethinking generalization. In *27th International Conference on Principles and Practice of Constraint Programming (CP 2021)*, volume 210, 2021.

Keriven, N., Bietti, A., and Vaiter, S. Convergence and stability of graph convolutional networks on large random graphs. *Advances in Neural Information Processing Systems*, 33:21512–21523, 2020.

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

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Le Gouic, T., Paris, Q., Rigollet, P., and Stromme, A. J. Fast convergence of empirical barycenters in alexandrov spaces and the wasserstein space. *Journal of the European Mathematical Society*, 2022.

Lei, J. Convergence and concentration of empirical measures under wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767–798, 2020.

Levie, R., Huang, W., Bucci, L., Bronstein, M. M., and Kutyniok, G. Transferability of spectral graph convolutional neural networks. *Journal of Machine Learning Research*, 22:272–1, 2021.

Liao, R., Urtasun, R., and Zemel, R. A pac-bayesian approach to generalization bounds for graph neural networks. In *International Conference on Learning Representations*, 2020.

Lovász, L. *Large networks and graph limits*, volume 60. American Mathematical Society, 2012.

Maskey, S., Levie, R., Lee, Y., and Kutyniok, G. Generalization analysis of message passing neural networks on large random graphs. In *Advances in Neural Information Processing Systems*, 2022.

Mémoli, F. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

Morris, C., Kriege, N. M., Bause, F., Kersting, K., Mutzel, P., and Neumann, M. Tudataset: A collection of benchmark datasets for learning with graphs. In *ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+ 2020)*, 2020. URL www.graphlearning.io.

Ortega, A., Frossard, P., Kovačević, J., Moura, J. M., and Vandergheynst, P. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5):808–828, 2018.

Pele, O. and Werman, M. Fast and robust earth mover's distances. In *IEEE 12th International Conference on Computer Vision*, pp. 460–467, 2009.

Penrose, M. *Random geometric graphs*, volume 5. Oxford University Press, 2003.

Ruiz, L., Chamon, L., and Ribeiro, A. Graphon neural networks and the transferability of graph neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1702–1712, 2020.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019.

Scarselli, F., Tsoi, A. C., and Hagenbuchner, M. The vapnik–chervonenkis dimension of graph and recursive neural networks. *Neural Networks*, 108:248–259, 2018.

Shervashidze, N., Vishwanathan, S., Petri, T., Mehlhorn, K., and Borgwardt, K. Efficient graphlet kernels for large graph comparison. In *Artificial intelligence and statistics*, pp. 488–495. PMLR, 2009.

Shervashidze, N., Schweitzer, P., Van Leeuwen, E. J., Mehlhorn, K., and Borgwardt, K. M. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(9), 2011.

Sun, B. and Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer, 2016.

Van Der Vaart, A. W., Wellner, J. A., van der Vaart, A. W., and Wellner, J. A. *Weak convergence*. Springer, 1996.

Vapnik, V. N. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5), 1999.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Vincent-Cuaz, C., Flamary, R., Corneli, M., Vayer, T., and Courty, N. Template based graph neural network with optimal transport distances. In *Advances in Neural Information Processing Systems*, 2022.

Wang, Y., Wang, W., Liang, Y., Cai, Y., and Hooi, B. Mixup for node and graph classification. In *Proceedings of the Web Conference 2021*, pp. 3663–3674, 2021.

Wu, Y., Wang, X., Zhang, A., He, X., and Chua, T.-S. Discovering invariant rationales for graph neural networks. In *International Conference on Learning Representations*, 2021.

Wu, Y., Bojchevski, A., and Huang, H. Adversarial weight perturbation improves generalization in graph neural network. *arXiv preprint arXiv:2212.04983*, 2022a.

Wu, Y., Wang, X., Zhang, A., He, X., and Chua, T.-S. Discovering invariant rationales for graph neural networks. In *International Conference on Learning Representations*, 2022b.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K.-i., and Jegelka, S. Representation learning on graphs with jumping knowledge networks. In *International Conference on Machine Learning*, pp. 5453–5462. PMLR, 2018.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Yehudai, G., Fetaya, E., Meirom, E., Chechik, G., and Maron, H. From local structures to size generalization in graph neural networks. In *International Conference on Machine Learning*, pp. 11975–11986. PMLR, 2021.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zhao, T., Liu, Y., Neves, L., Woodford, O., Jiang, M., and Shah, N. Data augmentation for graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11015–11023, 2021.

Organization of the appendix.
In Section A, we introduce the widely adopted regularity assumptions for theoretical analysis of MPNNs.
In Section B, we detail the convergence and generalization results along with the proofs.
In Section C, we introduce more details on the datasets used for numerical experiments.
In Section D, we give more information about the baselines being compared.
In Section E, we give more explanation on the implementation details of the empirical realization.
In Section F, we provide empirical proofs of the effectiveness of the MPNN with a WBM layer in size generalization in larger datasets.
In Section G, we provide additional experiments validating the effectiveness of the components of the WBM method. The Section F and Section G act as a complementary for Section 5.
In Section H, we give an algorithmic framework description for our method.
Lastly, in Section I, we rehearse some useful results from third-party theoretical studies.

A. Regularity Assumptions

Assumption A.1 (Regularity assumptions, Assumption A.10 in Maskey et al. (2022))**.** Let (χ, d) be a metric space and $\mathcal{A} : \chi \times \chi \to [0, +\infty)$ be a graphon. Let $\Theta = (\{\Phi^{(l)}, \Psi^{(l)}\}_{l=1}^{L})$ be an L-layer MPNN with message functions $\Phi^{(l)} : \mathbb{R}^{2F_{l-1}} \to \mathbb{R}^{H_{l-1}}$ and update functions $\Psi^{(l)} : \mathbb{R}^{F_{l-1}+H_{l-1}} \to \mathbb{R}^{F_l}$, for $l = 1, \ldots, L$. The regularity assumptions are as follows,

1. The space χ is compact, and there exist $D_\chi, C_\chi > 0$ such that $\mathcal{C}(\chi, \varepsilon, d) \leq C_\chi \varepsilon^{-D_\chi}$ for every $\epsilon > 0$, where $\mathcal{C}(\chi, \varepsilon, d)$ is the ϵ-covering numbers of the space χ.
2. The diameter of space χ is bounded by 1, i.e., $\mathrm{diam}(\chi) := \sup_{x,y\in\chi} d(x, y) \leq 1$.
3. The graphon satisfies $\|\mathcal{A}\|_\infty < \infty$.
4. The graphon function $\mathcal{A}(\cdot, \cdot)$ is $L_\mathcal{A}$-Lipschitz continuous with respect to both of its variables if fixing the other variable, i.e., $\forall t \in \chi$, $\mathcal{A}(\cdot, t)$ is $L_\mathcal{A}$-Lipschitz continuous. And $\forall s \in \chi$, $\mathcal{A}(s, \cdot)$ is $L_\mathcal{A}$-Lipschitz continuous.
5. There exists a constant $d_{min} > 0$ such that the graphon degree $d_\mathcal{A}(\cdot)$ is bounded from below by $d_{min} > 0$, i.e., $\forall s \in \chi$, $d_\mathcal{A}(s) \geq d_{min}$.
6. For every $l = 1, \ldots, L$, the message function $\Phi^{(l)}$ is $L_{\Phi^{(l)}}$-Lipschitz continuous with $\Phi^{(l)}(0, 0) = 0$. And the update function $\Psi^{(l)}$ is $L_{\Psi^{(l)}}$-Lipschitz continuous with $\Psi^{(l)}(0, 0) = 0$.
7. There exists a constant $A_{diag} > 0$ such that for every $s \in \chi$, we have $\mathcal{A}(s, s) \geq A_{diag} > 0$.

Remark A.2 (non-necessity of the zero condition). The key for Assumption A.1.6 is the Lipschitz continuity. We introduce the zero condition for ease of notation. The zero condition is not necessary to arrive at the proposed upper bounds (up to a rescaling of the constant terms). In fact, Lemma I.1 used in the proofs is a simplified version adopting the zero condition of Lemma B.9 in Maskey et al. (2022). In their Lemma B.9, there are more involved terms related to $\|\Phi^{(l)}(0, 0)\|$ and $\|\Psi^{(l)}(0, 0)\|$ that only affect the constant terms in our bounds. Additionally, the conclusion of our Lemma B.1 still holds without the zero condition, by rescaling the K^* again with the non-null bias term.

B. Convergence and Generalization of an MPNN with a WBM layer

In this section, we give the details and the proofs for Theorem 4.9 and Theorem 4.10.

B.1. Convergence of an MPNN with a WBM layer

Give an L-layer MPNN Θ, we will use $f^{(L)}$ and $(\Theta_\mathcal{A}(f) \circ f)$ to denote the same mapping from χ to $\mathbb{R}^{F_L}$ interchangeably. We also use $f^{(L)}$ to denote the mapped image or the variable whenever it is clear from the context.

We first give a Lemma restricting the tail behaviour of the push-forward measure $\nu^j = (\Theta_\mathcal{A}(f) \circ f)_{\#}\mu^j$ for every $j = 1, \ldots, C$.

Lemma B.1 (the push-forward measure ν of a sub-Gaussian μ)**.** *Let (χ, d, μ) be a metric measure space and $\mathcal{A} : \chi \times \chi \to [0, +\infty)$ be a graphon. Let $\Theta = (\{\Phi^{(l)}, \Psi^{(l)}\}_{l=1}^{L})$ be an L-layer MPNN s.t. the regularity assumptions in Appendix A are satisfied. Consider an L_f-Lipschitz continuous metric-space signal $f : \chi \to \mathbb{R}^F$ with finite infinity norm $\|f\| < \infty$. If μ is K-sub-Gaussians on χ, i.e.,*

$$\int_{\chi} e^{\|s\|^2/(2D_\chi K^2)} d\mu(s) \leq 2, \tag{B.1}$$

then ν is K^*-sub-Gaussian, with $K^* = L_{f^{(L)}} K \sqrt{\frac{D_\chi}{F_L}}$.

Moreover, for $q \in \mathbb{N}$, there is

$$M_q(\nu) \overset{def.}{=} \int_{\mathbb{R}^{F_L}} \|t\|^q d\nu(t) \leq 2K^{*q}\Gamma(\frac{q}{2}+1), \tag{B.2}$$

where $\Gamma(x)$ is the Gamma function for $x > 0$.

Proof. We check ν is K^*-sub-Gaussian by definition Eq.(B.1),

$$\begin{aligned}
&\int_{\mathbb{R}^{F_L}} e^{\|t\|^2/(2F_L K^{*2})} d\nu(t) = \int_{\chi} e^{\|\Theta_{\mathcal{A}}(f)\circ f)(s)\|^2/(2F_L K^{*2})} d\mu(s) \\
\overset{(a)}{\leq} &\int_{\chi} e^{L_{f^{(L)}}\|s\|^2/(2F_L K^{*2})} d\mu(s) \overset{(b)}{=} \int_{\chi} e^{\|s\|^2/(2D_\chi K^2)} d\mu(s) \overset{(c)}{\leq} 2,
\end{aligned}$$

where (a) invokes the Lipschitz continuity of $f^{(L)}$ and regularity assumption A.1.6, with Lemma B.9 in (Maskey et al., 2022) (cf. Lemma I.1) assuring the Lipschitz-continuity of $f^{(L)}$ and the existence of Lipschitz constant $L_{f^{(L)}}$. The equality (b) uses $K^* = L_{f^{(L)}} K \sqrt{\frac{D_\chi}{F_L}}$, and (c) invokes the assumption that μ is K-sub-Gaussian on χ.

Then by Markov's inequality,

$$\begin{aligned}
&Prob\left(\|f^{(L)}(t)\| > s\right) = Prob\left(e^{\frac{\|f^{(L)}(t)\|^2}{K^{*2}}} \leq e^{\frac{s^2}{K^{*2}}}\right) \\
\leq &\frac{\mathbb{E}[e^{\|f^{(L)}\|^2/K^{*2}}]}{e^{\frac{s^2}{K^{*2}}}} \leq 2e^{-\frac{s^2}{K^{*2}}}.
\end{aligned} \tag{B.3}$$

By the layer cake representation

$$\begin{aligned}
M_q(\nu) &= \int_{\mathbb{R}^{F_L}} \|f^{(L)}\|^q d\nu = \int_0^\infty Prob(\|f^{(L)}\|^q \geq s) ds \\
&\overset{(d)}{\leq} 2\int_0^\infty q s^{q-1} e^{-\frac{s^2}{K^{*2}}} ds \\
&\overset{(e)}{\leq} 2K^{*q}\frac{q}{2}\int_0^\infty t^{q/2-1} e^{-t} dt = 2K^{*q}\Gamma(\frac{q}{2}+1),
\end{aligned}$$

where (d) invokes Eq.(B.3) and (e) changes the variable with $t = s^2/K^{*2}$. □

We rewrite Theorem 4.9. Then we prove it with details on the constants.

Theorem B.2 (convergence of an MPNN with a WBM layer)**.** *Given an MPNN Θ, a loss function ℓ and a set of N-nodes graphs $\mathcal{S} = \{(A_k, \boldsymbol{f}_k), \boldsymbol{y}_k\}_{k=1}^n \sim \mu_{\mathcal{G}}^n$, with $\mu_{\mathcal{G}} = \sum_{j=1}^C h^j \mu_{\mathcal{G}_j}$, $\{\mu^j\}_{j=1}^C$ are K-sub-Gaussians and $h^j = P(\boldsymbol{y} = j)$. Denote by $\mathcal{S} = \bigsqcup_{j=1}^C \mathcal{S}^j$ a same-class-partition of $\mathcal{S}$, with $n_j = |S^j| \geq 1$. Let $\Theta_A^{\mathcal{W}}(\boldsymbol{f})$ and $\Theta_{\mathcal{A}}^{\mathcal{W}}(f)$ be the MPNNs with WBM layers, i.e.,*

$$\begin{cases}
\Theta_A^{\mathcal{W}}(\boldsymbol{f}) = \left(\mathcal{W}_2(\hat{\boldsymbol{b}}_2(\mathcal{S}^1), \hat{\nu}), \ldots, \mathcal{W}_2(\hat{\boldsymbol{b}}_2(\mathcal{S}^C), \hat{\nu})\right), & \text{(B.4)} \\
\Theta_{\mathcal{A}}^{\mathcal{W}}(f) = \left(\mathcal{W}_2(\boldsymbol{b}_2^1, \nu), \ldots, \mathcal{W}_2(\boldsymbol{b}_2^j, \nu)\right), & \text{(B.5)}
\end{cases}$$

where $\hat{\nu} = (\Theta_A(\boldsymbol{f}) \circ f)_{\#}\hat{\mu}$ *with* $\hat{\mu} = 1/N \sum_{i=1}^{N} \delta_{X_i^j}$ *for a given random graph. Let* $\eta \in (0,1)$*. Under the regularity assumptions in Appendix A, with probability no less than* $1 - \eta - e^{-C_2 N} - e^{-C_2' n^*}$ *we have the following*

$$\|\Theta_A^{\mathcal{W}}(\boldsymbol{f}) - \Theta_{\mathcal{A}}^{\mathcal{W}}(f)\|^2 \leq C[\alpha_{C_1}(N, n^*, \eta) + \beta_{K',C_3}(N,\eta)]^2, \tag{B.6}$$

where

$$\alpha_{C_1}(N, n^*, \eta) = \sqrt{\frac{C_1}{N}\log(\frac{8}{\eta}) + \frac{C_1}{n^*}\log(\frac{8}{\eta})}, \tag{B.7}$$

$$\beta_{K',C_3}(N,\eta) = 2N^{-\frac{1}{2}}\log(\frac{4}{\eta}) + \sqrt{16K'^2\log(\frac{4}{\eta}) + \frac{8}{N}\log(\frac{4}{\eta})} + C_3 K'^2 \times \begin{cases} \mathcal{O}(N^{-1/3}) & \text{if } F_L \in \{1,2,3,4\} \\ N^{-2/F_L} + N^{-1/3} & \text{if } F_L > 4 \end{cases}, \tag{B.8}$$

where $n^* = min(n_1, \ldots, n_C)$*.* C_1, C_2, C_2', C_3 *and* K' *are constants to be specified in the proof (cf. Eq.*(B.20)*).*

Proof. For simplicity of notation, denote $\hat{\boldsymbol{b}}_2^j := \hat{\boldsymbol{b}}_2(\mathcal{S}^j)$. For every $j = 1, \ldots, C$ we can conclude the following inequality from the fact 2-Wasserstein distance is a metric in the Wasserstein space,

$$\mathcal{W}_2(\hat{\nu}, \hat{\boldsymbol{b}}_2^j) \leq \mathcal{W}_2(\hat{\nu}, \nu) + \mathcal{W}_2(\nu, \hat{\boldsymbol{b}}_2^j) \leq \mathcal{W}_2(\nu, \hat{\nu}) + \mathcal{W}_2(\boldsymbol{b}_2^j, \hat{\boldsymbol{b}}_2^j) + \mathcal{W}_2(\nu, \boldsymbol{b}_2^j).$$

Rearranging both sides, for every $j = 1, \ldots, C$ we have

$$|\mathcal{W}_2(\hat{\nu}, \hat{\boldsymbol{b}}_2^j) - \mathcal{W}_2(\nu, \boldsymbol{b}_2^j)| \leq \mathcal{W}_2(\nu, \hat{\nu}) + \mathcal{W}_2(\boldsymbol{b}_2^j, \hat{\boldsymbol{b}}_2^j). \tag{B.9}$$

Next, we investigate $\|\Theta_A^{\mathcal{W}}(\mathbf{f}) - \Theta_{\mathcal{A}}^{\mathcal{W}}(f)\|^2$,

$$\|\Theta_A^{\mathcal{W}}(\mathbf{f}) - \Theta_{\mathcal{A}}^{\mathcal{W}}(f)\|^2 = \sum_{j=1}^{C}\left(\mathcal{W}_2(\hat{\nu}, \hat{\boldsymbol{b}}_2^j) - \mathcal{W}_2(\nu, \boldsymbol{b}_2^j)\right)^2 \overset{Eq.(B.9)}{\leq} \sum_{j=1}^{C}\left(\mathcal{W}_2(\nu, \hat{\nu}) + \mathcal{W}_2(\boldsymbol{b}_2^j, \hat{\boldsymbol{b}}_2^j)\right)^2 \tag{B.10}$$

We investigate the bound on $\mathcal{W}_2(\nu, \hat{\nu})$ and the bound on $\mathcal{W}_2(\boldsymbol{b}_2^j, \hat{\boldsymbol{b}}_2^j)$ separately.

The bound on $\mathcal{W}_2(\nu, \hat{\nu})$. We first bound the expectation $\mathbb{E}\left(\mathcal{W}_p\left(\hat{\nu}, \nu\right)\right)$ with the acknowledged Theorem 1 in Fournier & Guillin (2015) (cf. Lemma I.2). Then we bound the difference between $\mathcal{W}_2(\nu, \hat{\nu})$ and $\mathbb{E}\left(\mathcal{W}_p\left(\hat{\nu}, \nu\right)\right)$ with a mean-concentration inequality for $\mathcal{W}_2(\nu, \hat{\nu})$, which is the Corollary 5.5 from Lei (2020) (cf. Lemma I.3.) The Theorem 1 in Fournier & Guillin (2015) (cf. Lemma I.2) says that if $M_q(\nu) < \infty$ fome some $q > p$, then there exists a constant D depending only on p, d, q such that for all $N \geq 1$,

$$\mathbb{E}\left(\mathcal{W}_p\left(\hat{\nu}, \nu\right)\right) \leq D M_q^{p/q}(\nu) \times \begin{cases} N^{-1/2} + N^{-(q-p)/q} & \text{if } p > d/2 \text{ and } q \neq 2p, \\ N^{-1/2}\log(1+N) + N^{-(q-p)/q} & \text{if } p = d/2 \text{ and } q \neq 2p, \\ N^{-p/d} + N^{-(q-p)/q} & \text{if } p \in (0, d/2) \text{ and } q \neq d/(d-p). \end{cases}$$

Then, guaranteed by Eq. (B.2) in Lemma B.1, taking $d = F_L$ $p = 2$ and $q = 3$ we get

$$\begin{aligned} \mathbb{E}\left(\mathcal{W}_2\left(\hat{\nu}, \nu\right)\right) &\leq D M_3^{2/3}(\nu) \times \begin{cases} \mathcal{O}(N^{-1/3}) & \text{if } F_L \in \{1,2,3,4\} \\ N^{-2/F_L} + N^{-1/3} & \text{if } F_L > 4 \end{cases} \\ &= D 2^{2/3} K^{*2} \Gamma(\frac{5}{2}) \times \begin{cases} \mathcal{O}(N^{-1/3}) & \text{if } F_L \in \{1,2,3,4\} \\ N^{-2/F_L} + N^{-1/3} & \text{if } F_L > 4 \end{cases}. \end{aligned} \tag{B.11}$$

According to the Corollary 5.5 from Lei (2020) (cf. Lemma I.3.), if $M_q(\nu) \leq \frac{1}{2}s^2 q! V^{q-2}$ for all integer $q \geq 2$ and some constants s, V, then for all $t > 0$

$$Prob\left[|\mathcal{W}_2(\hat{\nu}, \nu) - \mathbb{E}(\mathcal{W}_2(\hat{\nu}, \nu))| \geq t\right] \leq 2\exp\left(-\frac{t^2}{8s^2 + 4Vt N^{-1/2}}\right).$$

Based on $M_q(\nu) \leq 2K^{*q}\Gamma(\frac{q}{2}+1)$ from Lemma B.1. It is easy to check that $s = \sqrt{2}K^*$ and $V = K^*$ satisfying the condition $M_q(\nu) \leq \frac{1}{2}s^2 q! V^{q-2}$ for all integer $q \geq 2$. Therefore we have the following

$$Prob\left[|\mathcal{W}_2(\hat{\nu},\nu) - \mathbb{E}(\mathcal{W}_2(\hat{\nu},\nu))| \geq t\right] \leq 2\exp\left(-\frac{t^2}{16K^{*2}+4K^*tN^{-1/2}}\right). \tag{B.12}$$

For $\eta \in (0,1)$, let $\frac{\eta}{2} = 2\exp\left(-\frac{t^2}{16K^{*2}+4K^*tN^{-1/2}}\right)$, we solve this the quadratic equation and get

$$t = 2N^{-\frac{1}{2}}\log(\frac{4}{\eta}) \pm \sqrt{16K^{*2}\log(\frac{4}{\eta}) + \frac{8}{N}\log(\frac{4}{\eta})} \tag{B.13}$$

Therefore, we have

$$\text{w.p.} \geq 1-\eta/2, \quad |\mathcal{W}_2(\hat{\nu},\nu) - \mathbb{E}(\mathcal{W}_2(\hat{\nu},\nu))| \leq 2N^{-\frac{1}{2}}\log(\frac{4}{\eta}) + \sqrt{16K^{*2}\log(\frac{4}{\eta}) + \frac{8}{N}\log(\frac{4}{\eta})}. \tag{B.14}$$

Combining Eq.(B.11) and Eq.(B.14), we get a bound for $\mathcal{W}_2(\nu,\hat{\nu})$,

$$\begin{aligned}&\text{w.p.} \geq 1-\eta/2,\\ &\mathcal{W}_2(\hat{\nu},\nu) \leq 2N^{-\frac{1}{2}}\log(\frac{4}{\eta}) + \sqrt{16K^{*2}\log(\frac{4}{\eta}) + \frac{8}{N}\log(\frac{4}{\eta})} + D2^{2/3}K^{*2}\Gamma(\frac{5}{2}) \times \begin{cases}\mathcal{O}(N^{-1/3}) & \text{if } F_L \in \{1,2,3,4\}\\ N^{-2/F_L}+N^{-1/3} & \text{if } F_L > 4\end{cases}.\end{aligned} \tag{B.15}$$

The bound on $\mathcal{W}_2(\boldsymbol{b}_2^j, \hat{\boldsymbol{b}}_2^j)$. We apply the concentration results for Wasserstein barycenters, i.e., Theorem 12 in (Le Gouic et al., 2022) (cf. Lemma I.4) to derive the uniform bound on $\mathcal{W}_2(\boldsymbol{b}_2^j, \hat{\boldsymbol{b}}_2^j)$ for $j = 1,\ldots,C$.
Fixing j, consider the class j empirical barycenter of empirical measure $\hat{\mathcal{P}}^j_{n_j} = 1/n_j \sum_{k=1}^{n_j} \delta_{\hat{\nu}_k^j}$.
Notice that $\hat{\nu}_k^j = 1/N \sum_{i=1}^N \delta_{\mathbf{f}_{k,i}^{(L)j}}$. We can rewrite the empirical measure $\hat{\mathcal{P}}^j_{n_j}$ as

$$\hat{\mathcal{P}}^j_{n_j} = 1/n_j \sum_{k=1}^{n_j} 1/N \sum_{i=1}^{N} \delta_{\mathbf{f}_{k,i}^{(L)j}}. \tag{B.16}$$

Considering the relationship between $\hat{\boldsymbol{b}}_2^j$ and $\boldsymbol{b}_2^j = \nu^j$, and considering the limiting process w.r.t. the index i and the index k sequentially, we may apply the concentration inequality in Lemma I.4 twice to bound the $\mathcal{W}_2(\boldsymbol{b}_2^j, \hat{\boldsymbol{b}}_2^j)$. For $\eta \in (0,1)$, based on the union bound and Lemma I.4, there is

$$w.p. \geq 1-\eta/2-e^{-C_2N}-e^{-C_2'n^j}, \quad \mathcal{W}_2^2(\boldsymbol{b}_2^j, \hat{\boldsymbol{b}}_2^j) \leq \frac{E}{N}\log(\frac{8}{\eta}) + \frac{E'}{n^j}\log(\frac{8}{\eta}), \tag{B.17}$$

where C_2, C_2, E, and E' are constants from the Wasserstein barycenter concentration inequality and are independent of N and n_j.
Taking $n^* = min(n_1,\ldots,n_C)$ and $C_1 = max(E,E')$, for every class $j = 1,\ldots,C$ we have a uniform bound

$$w.p. \geq 1-\eta/2-e^{-C_2N}-e^{-C_2'n^*}, \quad \mathcal{W}_2^2(\boldsymbol{b}_2^j, \hat{\boldsymbol{b}}_2^j) \leq \frac{C_1}{N}\log(\frac{8}{\eta}) + \frac{C_1}{n^*}\log(\frac{8}{\eta}), \tag{B.18}$$

Finally, combining the Eq.(B.10), and uniform bounds in Eq.(B.15) and Eq.(B.18), with union bound, we have with probability no less than $1-\eta-e^{-C_2N}-e^{-C_2'}$

$$\begin{aligned}&\|\Theta_A^{\mathcal{W}}(\mathbf{f}) - \Theta_{\mathcal{A}}^{\mathcal{W}}(f)\|^2 \leq C \times [\sqrt{\frac{C_1}{N}\log(\frac{8}{\eta}) + \frac{C_1}{n^*}\log(\frac{8}{\eta})} +\\ &2N^{-\frac{1}{2}}\log(\frac{4}{\eta}) + \sqrt{16K^{*2}\log(\frac{4}{\eta}) + \frac{8}{N}\log(\frac{4}{\eta})} + D2^{2/3}K^{*2}\Gamma(\frac{5}{2}) \times \begin{cases}\mathcal{O}(N^{-1/3}) & \text{if } F_L \in \{1,2,3,4\}\\ N^{-2/F_L}+N^{-1/3} & \text{if } F_L > 4\end{cases}]^2.\end{aligned} \tag{B.19}$$

Let $K' = K\sqrt{\frac{D_X}{F_L}} \times (Z_1^{(L)}\|f\|_\infty + Z_2^{(L)} L_f)$, where $Z_1^{(L)}, Z_2^{(L)}$ are specified in Eq. I.2 in Lemma. I.1.Recalling $K^* = L_{f^{(L)}} K\sqrt{\frac{D_X}{F_L}}$ in Lemma. B.1 and $L_{f^{(l)}} \leq Z_1^{(l)}\|f\|_\infty + Z_2^{(l)} L_f$ in Lemma. I.1. We have $K' \geq K*$.
The theorem is concluded by taking

$$\begin{cases} C_1 = max(E, E') \\ C_3 = 2^{2/3} D\Gamma(\frac{5}{2}) \\ K' = K\sqrt{\frac{D_X}{F_L}}(Z_1^{(L)}\|f\|_\infty + Z_2^{(L)} L_f) \\ \alpha_{C_1}(N, n^*, \eta) = \sqrt{\frac{C_1}{N}\log(\frac{8}{\eta}) + \frac{C_1}{n^*}\log(\frac{8}{\eta})} \\ \beta_{K',C_3}(N,\eta) = 2N^{-\frac{1}{2}}\log(\frac{4}{\eta}) + \sqrt{16K'^2\log(\frac{4}{\eta}) + \frac{8}{N}\log(\frac{4}{\eta})} + C_3 K'^2 \times \begin{cases} \mathcal{O}(N^{-1/3}) & \text{if } F_L \in \{1,2,3,4\} \\ N^{-2/F_L} + N^{-1/3} & \text{if } F_L > 4 \end{cases} \end{cases} \tag{B.20}$$

□

Remark B.3. For ease of exposition, we make a fixed graph size assumption. However, the proven upper bounds would be maintained even when the distribution of the number of nodes is considered. Denote by N the random variable for the number of nodes. Suppose that N follows the distribution Q, i.e., $N \sim Q$. Whereof we need an extended version of the graph data generation process considering Q, in contrast to the process described in Section 4.1 of the main context. Specifically, the novel measure on class j graphs is

$$\mu_{\mathcal{G}_j} = \sum_{N=1}^{\infty} Q(N)(\mu^j)^N.$$

The upper bounds in the main paper can be thought of as the result of conditioning on $N = N_s$. Combining the extended measure and Theorem 4.9, it is straightforward to arrive at the following result,

$$\|\Theta_A^{\mathcal{W}}(\mathbf{f}) - \Theta_{\mathcal{A}}^{\mathcal{W}}(f)\|^2 \leq C\mathbb{E}_{N\sim Q}\left[\alpha_{C_1}(N, n^*, \eta) + \beta_{K',C_3}(N,\eta)\right]^2.$$

Similar corollaries are also straightforward for Theorem 4.10 and Proposition 4.11.

B.2. Generalization of an MPNN with a WBM layer

We rewrite Theorem 4.10 and prove it. We follow the proof technique of Theorem 3.3 in Maskey et al. (2022) that uses a concentration inequality for multinomial measures proposed in Van Der Vaart et al. (1996) (cf. Lemma I.5) and "ghost samples".

Theorem B.4 (generalization of an MPNN with a WBM layer)**.** *Given an MPNN Θ, a loss function ℓ. Consider a set of N-nodes random graphs $\mathcal{S} = \{(A_k, \boldsymbol{f}_k), \boldsymbol{y}_k\}_{k=1}^n \sim \mu_{\mathcal{G}}^n$, with $\mu_{\mathcal{G}} = \sum_{j=1}^C h^j \mu_{\mathcal{G}_j}$, $\{\mu^j\}_{j=1}^C$ are K-sub-Gaussians and $h^j = P(\boldsymbol{y} = j)$. Denote by $\mathcal{S} = \bigsqcup_{j=1}^C \mathcal{S}^j$ a same-class-partition of $\mathcal{S}$, with $n_j = |S^j| \geq 1$. Let $\Theta_A^{\mathcal{W}}(\boldsymbol{f})$ and $\Theta_{\mathcal{A}}^{\mathcal{W}}(f)$ be the MPNNs with WBM layers. We also assume that there is at least one instance per class[5] in $\mathcal{S}$. We have the following inequality w.p. $\geq 1 - \eta - e^{-C_2 N} - e^{-C_2'}$,*

$$\mathop{\mathbb{E}}_{\mathcal{S}\sim\mu_{\mathcal{G}}^n}\left[\left(\hat{R}_{emp}(\Theta_A^{\mathcal{W}}) - R_{exp}(\Theta_{\mathcal{A}}^{\mathcal{W}})\right)^2\right] \leq \frac{8\|\ell\|_\infty^2 \pi 2^C}{n} + \frac{\pi^{\frac{1}{2}} L_\ell^2 2^C C^2}{n} \sum_{j=1}^C h_j[\alpha_{C_1}(N,1,\eta) + \beta_{K',C_3}(N,\eta)]^2, \tag{B.21}$$

where $\alpha_{C_1}(N,1,\eta)$ and $\beta_{K',C_3}(N,\eta)$ are defined in Eq.(B.20), and C_1, C_2, C_2', C_3 and K' are the same constants as in Theorem B.2.

Proof. Denote by $\boldsymbol{n} = (n_1, \ldots, n_C)$ the random vector that is multinomially distributed with parameters n and $h^1, \ldots, h^C$. For ease of notation, we write $\mathcal{S}^j = \{(A_k^j, \mathbf{f}_k^j), y_j\}_{k=1}^{n_j}$ to denote the set of random graphs of the j-th class, for $j = 1, \ldots, C$. Denote by $\mathcal{E}$ to be the event that at least one instance per class is observed. Denote by $\mathcal{D}_d$ the event defined by

$$\mathcal{D}_z \overset{def.}{=} \{\boldsymbol{n} = (n_1, \ldots, n_j) | \sum_{j=1}^C n_j = n,\ 2\sqrt{n}z \leq |n_j - nh^j| \leq 2\sqrt{n}(z+1)\}. \tag{B.22}$$

[5]We focus on closed set classification in this paper.

We decompose the generalization risk as follows.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{S}\sim\mu_{\mathcal{G}}^n}\left[\left(\frac{1}{n}\sum_{k=1}^{n}\ell\left(\Theta_{A_k}^{\mathcal{W}}\left(\mathbf{f}_k\right),y_j\right)-\mathbb{E}_{\mu_{\mathcal{G}}}\left[\ell\left(\Theta_{A}^{\mathcal{W}}(\mathbf{f}),y\right)\right]\right)^2\right]\\
&=\mathop{\mathbb{E}}_{\mathcal{S}\sim\mu_{\mathcal{G}}^n}\left[\left(\frac{1}{n}\sum_{j=1}^{C}\sum_{k=1}^{n_j}\ell\left(\Theta_{A_k^j}^{\mathcal{W}}\left(\mathbf{f}_k^j\right),y_j\right)-\mathbb{E}_{\mu_{\mathcal{G}}}\left[\ell\left(\Theta_{A}^{\mathcal{W}}(\mathbf{f}),y\right)\right]\right)^2\right]\\
&=\mathop{\mathbb{E}}_{\mathcal{S}\sim\mu_{\mathcal{G}}^n}\left[\left(\sum_{j=1}^{C}\left(\frac{1}{n}\sum_{i=1}^{n_j}\ell\left(\Theta_{A_k^j}^{\mathcal{W}}\left(\mathbf{f}_k^j\right),y_j\right)-h^j\mathbb{E}_{\mu_{\mathcal{G}_j}}\left[\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)\right]\right)\right)^2\right]
\end{aligned}\tag{B.23}$$

Regarding $\mathbb{E}_{\mu_{\mathcal{G}_j}}\left[\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)\right]$ as a scalar and invoking the law of total probability, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{S}\sim\mu_{\mathcal{G}}^n}\left[\left(\sum_{j=1}^{C}\left(\frac{1}{n}\sum_{i=1}^{n_j}\ell\left(\Theta_{A_k^j}^{\mathcal{W}}\left(\mathbf{f}_k^j\right),y_j\right)-h^j\mathbb{E}_{\mu_{\mathcal{G}_j}}\left[\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)\right]\right)\right)^2\right]\\
&=\mathop{\mathbb{E}}_{\mathcal{S}\sim\mu_{\mathcal{G}}^n}\left[\left(\sum_{j=1}^{C}\left(\frac{1}{n}\sum_{i=1}^{n_j}\ell\left(\Theta_{A_k^j}^{\mathcal{W}}\left(\mathbf{f}_k^j\right),y_j\right)-\frac{1}{n}\times nh^j\times\mathbb{E}_{\mu_{\mathcal{G}_j}}\left[\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)\right]\right)\right)^2\right]\\
&=\mathop{\mathbb{E}}_{\mathcal{S}\sim\mu_{\mathcal{G}}^n}\left[\left(\sum_{j=1}^{C}\left(\frac{1}{n}\sum_{k=1}^{n_j}\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)-\frac{1}{n}\sum_{j=1}^{nh^j}\mathbb{E}_{\mu_{\mathcal{G}_j}}\left[\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)\right]\right)\right)^2\right]\\
&=\sum_z Prob\left(\boldsymbol{n}\in\mathcal{D}_z\right)\times\mathop{\mathbb{E}}_{\mathcal{S}\sim\mu_{\mathcal{G}}^n\text{ given }\boldsymbol{n}\in\mathcal{D}_z}\left[\left(\sum_{j=1}^{C}\left(\frac{1}{n}\sum_{k=1}^{n_j}\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)-\frac{1}{n}\sum_{j=1}^{nh^j}\mathbb{E}_{\mu_{\mathcal{G}_j}}\left[\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)\right]\right)\right)^2\right]\\
&\le\sum_z Prob\left(\boldsymbol{n}\in\mathcal{D}_z\right)\times\sup_{\boldsymbol{n}\in\mathcal{D}_z}\mathop{\mathbb{E}}_{\mathcal{S}_n\sim\mu_{\mathcal{G}}^n}\left[\left(\sum_{j=1}^{C}\left(\frac{1}{n}\sum_{k=1}^{n_j}\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)-\frac{1}{n}\sum_{j=1}^{nh^j}\mathbb{E}_{\mu_{\mathcal{G}_j}}\left[\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)\right]\right)\right)^2\right]
\end{aligned}\tag{B.24}$$

where we use the notation of $\mathop{\mathbb{E}}_{\mathcal{S}_n\sim\mu_{\mathcal{G}}^n}[\cdot]$ indicating the conditional choice of the dataset on the choice of $\boldsymbol{n}=\{n_1,\ldots,n_C\}$ (with $\sum_{j=1}^{C}n_j=n$) by $S_n:=\{\{(A_k^j,\boldsymbol{f}_k^j),\boldsymbol{y}_k^j\}_{k=1}^{n_j}\}_{j=1}^{C}$.

For $j=1,\ldots,C$, if $n_j<nh^j$, add additional i.i.d. random graphs $\{(A_k^j,\mathbf{f}_k^j)\}_{k=n_j}^{nh^j}$ sampled from $(\mathcal{A}^j,f^j)$. We use the notation $\sum_{k=m}^{n}a_k:=-\sum_{k=m}^{n}a_k$ for real sequence $\{a_k\}_{k=n}^{m}$ for $n<m$. Then we manipulate

$\mathop{\mathbb{E}}_{\mathcal{S}_n\sim\mu_{\mathcal{G}}^n}\left[\left(\sum_{j=1}^{C}\left(\frac{1}{n}\sum_{k=1}^{n_j}\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)-\frac{1}{n}\sum_{j=1}^{nh^j}\mathbb{E}_{\mu_{\mathcal{G}_j}}\left[\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)\right]\right)\right)^2\right]$ with "ghost samples" as

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{S}_n\sim\mu_{\mathcal{G}}^n}\left[\left(\sum_{j=1}^{C}\left(\frac{1}{n}\sum_{k=1}^{n_j}\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)-\frac{1}{n}\sum_{j=1}^{nh^j}\mathbb{E}_{\mu_{\mathcal{G}_j}}\left[\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)\right]\right)\right)^2\right]\\
=&\mathop{\mathbb{E}}_{\mathcal{S}_n\sim\mu_{\mathcal{G}}^n}\left[\left(\sum_{j=1}^{C}\left(\frac{1}{n}\sum_{k=1}^{nh^j}\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)+\frac{1}{n}\sum_{k=nh^j}^{n_j}\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)-\frac{1}{n}\sum_{j=1}^{nh^j}\mathbb{E}_{\mu_{\mathcal{G}_j}}\left[\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)\right]\right)\right)^2\right]\\
\le&\mathop{\mathbb{E}}_{\mathcal{S}_n\sim\mu_{\mathcal{G}}^n}\left[2\left(\sum_{j=1}^{C}\left(\frac{1}{n}\sum_{k=1}^{nh^j}\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)-\frac{1}{n}\sum_{j=1}^{nh^j}\mathbb{E}_{\mu_{\mathcal{G}_j}}\left[\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)\right]\right)\right)^2\right]\\
&+\mathop{\mathbb{E}}_{\mathcal{S}\sim\mu_{\mathcal{G}}^n}\left[2\left(\sum_{j=1}^{C}\left(\frac{1}{n}\sum_{k=nh^j}^{n_j}\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)\right)\right)^2\right]\\
\le&\mathop{\mathbb{E}}_{\mathcal{S}_n\sim\mu_{\mathcal{G}}^n}\left[2\left(\sum_{j=1}^{C}\left(\frac{1}{n}\sum_{k=1}^{nh^j}\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)-\frac{1}{n}\sum_{j=1}^{nh^j}\mathbb{E}_{\mu_{\mathcal{G}_j}}\left[\ell\left(\Theta_{A^j}^{\mathcal{W}}\left(\mathbf{f}^j\right),y_j\right)\right]\right)\right)^2\right]\\
&+\mathop{\mathbb{E}}_{\mathcal{S}\sim\mu_{\mathcal{G}}^n}\left[2\left(\sum_{j=1}^{C}\left(\frac{1}{n}|n_j-nh^j|\|\ell\|_\infty\right)\right)^2\right]
\end{aligned}
\tag{B.25}
$$

We first investigate the second term in Eq. (B.25). Conditioned on event $\mathcal{D}_z$, we have $\sum_{j=1}^{C}|n_j-nh^j|<2\sqrt{n}(z+1)$, therefore

$$
\mathop{\mathbb{E}}_{\mathcal{S}_n\sim\mu_{\mathcal{G}}^n}\left[2\left(\sum_{j=1}^{C}\left(\frac{1}{n}|n_j-nh^j|\|\ell\|_\infty\right)\right)^2\right]\le\frac{2}{n^2}\|\ell\|^2(\sum_{j=1}^{c}|n_j-nh^j|)^2\le\frac{2}{n^2}\|\ell\|^2 4n(z+1)^2=\frac{8\|\ell\|_\infty^2}{n}(z+1)^2. \tag{B.26}
$$

Then

$$
\begin{aligned}
&\sum_d Prob\left(\boldsymbol{n}\in\mathcal{D}_z\right)\times\sup_{\boldsymbol{n}\in\mathcal{D}_z}\mathop{\mathbb{E}}_{\mathcal{S}_n\sim\mu_{\mathcal{G}}^n}\left[2\left(\sum_{j=1}^{C}\left(\frac{1}{n}|n_j-nh^j|\|\ell\|_\infty\right)\right)^2\right]\\
\le&\sum_z Prob\left(\boldsymbol{n}\in\mathcal{D}_z\right)\times\frac{8\|\ell\|_\infty^2}{n}(z+1)^2\\
\overset{(a)}{\le}&\sum_z 2^C\exp\left(-2z^2\right)\frac{8\|\ell\|_\infty^2}{n}(z+1)^2\\
\le&\int_0^\infty 2^C\exp\left(-2z^2\right)\frac{8\|\ell\|_\infty^2}{n}(z+1)^2dz\\
=&2^C\frac{8\|\ell\|_\infty^2}{n}\int_0^\infty\exp\left(-2z^2\right)(z+1)^2dz\le 2^C\frac{8\|\ell\|_\infty^2}{m}\pi,
\end{aligned}
\tag{B.27}
$$

where the inequality (a) invokes Lemma I.5.

We next investigate the second term in Eq. (B.25). Using the fact that $\sum_{j=1}^{C} a_j^2 \leq C \sum_{j=1}^{C} a_j^2$, we have the following

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{S}_n \sim \mu_{\mathcal{G}}^n} \left[2 \left(\sum_{j=1}^{C} \left(\frac{1}{n} \sum_{k=1}^{nh^j} \ell \left(\Theta_{A^j}^{\mathcal{W}} \left(\mathbf{f}^j \right), y_j \right) - \frac{1}{n} \sum_{j=1}^{nh^j} \mathbb{E}_{\mu_{\mathcal{G}_j}} \left[\ell \left(\Theta_{A^j}^{\mathcal{W}} \left(\mathbf{f}^j \right), y_j \right) \right] \right) \right)^2 \right] \\
\leq & 2C \sum_{j=1}^{C} \mathop{\mathbb{E}}_{\mathcal{S}_n \sim \mu_{\mathcal{G}}^n} \left[\left(\frac{1}{n} \sum_{k=1}^{nh^j} \ell \left(\Theta_{A^j}^{\mathcal{W}} \left(\mathbf{f}^j \right), y_j \right) - \frac{1}{n} \sum_{j=1}^{nh^j} \mathbb{E}_{\mu_{\mathcal{G}_j}} \left[\ell \left(\Theta_{A^j}^{\mathcal{W}} \left(\mathbf{f}^j \right), y_j \right) \right] \right)^2 \right] \\
= & 2C \sum_{j=1}^{C} Var_{\mu_{\mathcal{G}_j}} \left[\frac{1}{n} \sum_{k=1}^{nh^j} \ell \left(\Theta_{A^j}^{\mathcal{W}} \left(\mathbf{f}^j \right), y_j \right) \right] = 2C \sum_{j=1}^{C} \frac{h^j}{n} Var_{\mu_{\mathcal{G}_j}} \left[\ell \left(\Theta_{A^j}^{\mathcal{W}} \left(\mathbf{f}^j \right), y_j \right) \right] \\
\leq & 2C \sum_{j=1}^{C} \frac{h^j}{n} \mathbb{E}_{\mu_{\mathcal{G}_j}} \left[\left(\ell \left(\Theta_{A^j}^{\mathcal{W}} \left(\mathbf{f}^j \right), y_j \right) - \ell \left(\Theta_{\mathcal{A}^j}^{\mathcal{W}} \left(f^j \right), y_j \right) \right)^2 \right] \\
\leq & 2C \sum_{j=1}^{C} \frac{h^j}{n} \left[L_\ell^2 \| \Theta_{A^j}^{\mathcal{W}} \left(\mathbf{f}^j \right) - \Theta_{\mathcal{A}^j}^{\mathcal{W}} \left(f^j \right) \|^2 \right] \\
\overset{(b)}{\leq} & 2C^2 \sum_{j=1}^{C} \frac{h^j}{n} L_\ell^2 [\alpha_{C_1}(N, \min(n_1, \ldots, n_C), \eta) + \beta_{K', C_3}(N, \eta)]^2 \quad \text{w.p.1} - \eta - e^{-C_2 N} - e^{-C_2' \min(n_1, \ldots, n_C)}
\end{aligned}
\tag{B.28}
$$

where (b) invokes Theorem B.2.

Because we assume that at least one graph per class should be observed, thus the probability of any events being discussed is conditioned on event $\mathcal{E}$. Thereof $\alpha_{C_1}(N, 1, \eta)$ is an uniform bound on $\alpha_{C_1}(N, \boldsymbol{n}^*, \eta)$ for every z and any $\boldsymbol{n} \in \mathcal{D}_z \cap \mathcal{E}$ with $\boldsymbol{n}^* := min(n_1, \ldots, n_C)$. Meanwhile, $\beta_{K', C_3}(N, \eta)$ is independent of $\boldsymbol{n}$. Therefore, by Lemma I.5

$$
\begin{aligned}
&\sum_{d} Prob\left(\boldsymbol{n} \in \mathcal{D}_z \right) \times \sup_{\boldsymbol{n} \in \mathcal{D}_z} \mathop{\mathbb{E}}_{\mathcal{S}_n \sim \mu_{\mathcal{G}}^n} \left[2 \left(\sum_{j=1}^{C} \left(\frac{1}{n} \sum_{k=1}^{nh^j} \ell \left(\Theta_{A^j}^{\mathcal{W}} \left(\mathbf{f}^j \right), y_j \right) - \frac{1}{n} \sum_{j=1}^{nh^j} \mathbb{E}_{\mu_{\mathcal{G}_j}} \left[\ell \left(\Theta_{A^j}^{\mathcal{W}} \left(\mathbf{f}^j \right), y_j \right) \right] \right) \right)^2 \right] \\
\leq & \frac{\sqrt{\pi}}{2} 2^C 2C^2 \sum_{j=1}^{C} \frac{h^j}{n} L_\ell^2 [\alpha_{C_1}(N, 1, \eta) + \beta_{K', C_3}(N, \eta)]^2 \quad \text{w.p.1} - \eta - e^{-C_2 N} - e^{-C_2'} \\
= & \frac{\pi^{\frac{1}{2}} L_\ell^2 2^C C^2}{n} \sum_{j=1}^{C} h_j [\alpha_{C_1}(N, 1, \eta) + \beta_{K', C_3}(N, \eta)]^2 \quad \text{w.p.1} - \eta - e^{-C_2 N} - e^{-C_2'}.
\end{aligned}
\tag{B.29}
$$

Finally, combining Eq.(B.27), Eq.(B.29), and Eq.(B.23) conclude the proof of the theorem. □

C. Dataset Information

C.1. Small-scale datasets

NCI109, NCI1, PROTEINS, and DD are four vertex-attributed graph datasets collected from real-world (Morris et al., 2020), and are commonly used in graph size generalization literature (Yehudai et al., 2021; Bevilacqua et al., 2021; Buffelli et al., 2022). The prediction tasks for these datasets are binary classification. Following previous work, we explicitly split the dataset to create a domain shift on size: graphs with sizes smaller than 50-percentile are assigned to the training set, while graphs with sizes larger than 90-percentile are assigned to the test set. 10% of the training examples are split out as a validation set for model selection and hyperparameter tuning. With this split, the average size of the test graphs is 3 to 9 times larger than the average size of the training graphs (in more detail, it is 3 for NCI109 and NCI1, 9 for PROTEINS, and 5 for DD). This split leads to an imbalanced training set, as shown in Table 4.

Table 4. Dataset statistics, taken from (Buffelli et al., 2022).

Dataset	NCI1			NCI109		
	ALL	SMALLEST 50%	LARGEST 10%	ALL	SMALLEST 50%	LARGEST 10%
CLASS A	49.95%	62.30%	19.17%	49.62%	62.04%	21.37%
CLASS B	50.04%	37.69%	80.82%	50.37%	37.95%	78.62%
NUM OF GRAPHS	4110	2157	412	4127	2079	421
AVG GRAPH SIZE	29	20	61	29	20	61
Dataset	PROTEINS			DD		
	ALL	SMALLEST 50%	LARGEST 10%	ALL	SMALLEST 50%	LARGEST 10%
CLASS A	59.56%	41.97%	90.17%	58.65%	35.57%	79.66%
CLASS B	40.43%	58.02%	9.82%	41.34%	64.52%	20.33%
NUM OF GRAPHS	1113	567	112	1178	592	118
AVG GRAPH SIZE	39	15	138	284	144	746

To combat the class imbalance, we follow (Bevilacqua et al., 2021; Buffelli et al., 2022) to weight different classes in the classification loss according to the frequency of a class in the training set.

C.2. GOOD datasets

GOOD (Gui et al., 2022) is a recently established benchmark for testing graph out-of-distribution algorithms. Its designed data splitting creates various covariate shifts and concept shifts between the training set and test set, including base, color, size, scaffold, degree, and so on. In our experiment, we consider GOOD-Motif and GOOD-HIV using the default data splitting which creates covariate shifts in graph sizes (i.e., size splitting). We choose them because the prediction tasks of the two datasets are multi-class graph classification. Note that each instance in GOOD datasets has an accessible domain label (indicates the level of its size), but we do not use them in our method and treat the problem as a single-source generalization. We give a brief introduction to the two datasets as below:

- **GOOD-Motif** is synthetic base-motif dataset motivated by Spurious-Motif (Wu et al., 2022b). Each graph in the dataset is generated by connecting a base graph and a motif, and the label is determined by the motif. The task is to predict the label (3-way classification). For size covariate shift, the training set contains small size graphs, while the validation and the test sets include the middle and the largest size ranges, respectively.

- **GOOD-HIV** is a real-world molecular dataset adapted from MoleculeNet (Wu et al., 2018). The inputs are molecular graphs in which nodes are atoms, and edges are chemical bonds. The task is to predict whether the molecule can inhibit HIV replication (binary classification). For size covariate shift ,the larger-size graphs are used for training and the smaller ones are used for validation and testing.

The statistics of the two datasets (with size covariate shift splitting) are shown in Table 5.

Table 5. Statistics of GOOD datasets (with size covariate shift splitting).

Dataset		GOOD-Motif	GOOD-HIV
Train	NUM OF GRAPHS	18000	26169
	AVG GRAPH SIZE	16.9	27.9
Val	NUM OF GRAPHS	3000	2773
	AVG GRAPH SIZE	39.2	15.5
Test	NUM OF GRAPHS	3000	3961
	AVG GRAPH SIZE	87.2	12.1

D. Baseline Details

This section provides a detailed description of the baseline methods used for benchmark comparisons.

- Invariant Risk Minimization (**IRM**, (Arjovsky et al., 2019)) searches for graph representations that perform well across all environments by penalizing feature distributions that have different optimal linear classifiers for each environment.
- Variance Risk Extrapolation (**VREx**, (Krueger et al., 2021)) is a form of robust optimization over a perturbation set of extrapolated domains and minimizes the variance of training risks across domains.
- Group Distributionally Robust Optimization (**GroupDRO**, (Sagawa et al., 2019)) is a fair optimization method that tackles the problem that the distribution minority lacks sufficient training. It explicitly minimizes the loss in the worst training environment.
- Deep Correlation Alignment (**Deep Coral**, (Sun & Saenko, 2016)) encourages similar features in different domains and minimizes the deviation of covariant matrices from different training domains.
- **Mixup** (Zhang et al., 2017) is a data augmentation method to improve generalization, which interpolates both features and labels of a pair of instances to produce synthetic samples. The GOOD implementation uses the Mixup technique designed for graph classification (Wang et al., 2021), which interpolates complex and diverse graphs in the semantic space rather than in input space.
- E-invariant models ($\mathbf{\Gamma_{1-hot}}, \mathbf{\Gamma_{GIN}}, \mathbf{\Gamma_{RPGIN}}$, (Bevilacqua et al., 2021)) assume a causal model describing the generating process for graphs of different sizes, and are invariant to the train/test size shifts of the causal model.
- Central Moment Discrepancy regularization (**CMDr**, (Buffelli et al., 2022)) is a heuristic method that simulates size shift by graph coarsening and penalizes the shift in the distribution of node embeddings of different coarsened versions.
- Graphlet Counting kernel (**GC kernel**, (Shervashidze et al., 2009)) is a kernel method that compares graphs by counting graphlets, i.e., subgraphs with k nodes where k is some specified value.
- Weisfeiler-Lehman kernel (**WL kernel**, (Shervashidze et al., 2011)) is a family of graph kernels that extract features based on Weisfeiler-Lehman test of isomorphism on graphs.

E. More Implementation Details

E.1. Evaluation protocol

For the four real-world datasets (NCI109, NCI1, PROTEINS and DD), as the data splitting process leads to class imbalance, we follow (Buffelli et al., 2022) to use Matthews correlation coefficient (MCC) as the evaluation metric, which has been shown to be more reliable than other metrics in imbalanced classification settings (Chicco & Jurman, 2020). MCC gives a value between -1 and 1, where -1 indicates perfect disagreement and 1 indicates perfect agreement between the predictions and the ground-truth labels. For the GOOD benchmark, we use accuracy for GOOD-Motif and ROC-AUC for GOOD-HIV as the evaluation metrics.

For all methods, the model with the highest metric on the validation set is evaluated on the test set, and we report the mean test metric and standard deviation of ten independent trials with different random seeds. For all the baselines we use the hyperparameters or the original results introduced in their respective papers.

E.2. Hyperparameters

For the four real-world datasets (NCI109, NCI1, PROTEINS and DD), we use a 3-layer MPNN (GIN, GCN and PNA) to filter the original graph signals to conduct barycenter matching in the Wasserstein metric space. For the WBM layer, we set the size of the Wasserstein barycenters to the max/median size of the observed graphs. The Wasserstein barycenters are initialized by randomly sampling from the observed graphs. The number of barycenters per class M is chosen from $\{1, 2, 3, 4\}$ on NCI109. We find that $M = 3$ is good to balance performance and efficiency and fix it for the other datasets. The trade-off hyperparameter λ is tuned in $\{0.005, 0.01, 0.05, 0.1, 0.2\}$. **The WBM embedding is normalized using L_2 normalization**. We find empirically that normalizing the WBM embeddings could make the training more stable. The network is trained for 500 epochs, using the Adam optimizer with a weight decay of 1e-5. For fairness, we validate the batch size, learning rate, and MPNN hidden layers dimension, similar to (Bevilacqua et al., 2021; Buffelli et al., 2022). The batch size is selected from $\{64, 128\}$ and the learning rate is selected from $\{1e\text{-}3, 5e\text{-}3, 1e\text{-}2\}$. The hidden layer's dimension is chosen from $\{8, 16, 32, 64\}$.

For GOOD datasets, we use GIN for GOOD-Motif and GIN-Virtual Node (Gilmer et al., 2017; Xu et al., 2019)(vGIN) for GOOD-HIV as the backbones. we use the default backbone architecture, batch size, learning rate, and training epochs in GOOD official implementation. We only reduce the MPNN hidden layers dimension from 300 to 32, since too large dimensionality increases the computational overhead in computing Wasserstein distance. The number of total Wasserstein barycenters is fixed as 6 (so $M = 2$ for GOOD-Motif and $M = 3$ for GOOD-HIV). The trade-off hyperparameter λ is tuned in $\{0.005, 0.01, 0.05, 0.1, 0.2\}$.

E.3. Experimental environments

Hardware environments. We perform our experiments on three machines: one with 8 Nvidia RTX3090s and Xeon E5-2680, one with 8 Nvidia RTX3090s and Xeon Platinum 8358P, and one with 2 Nvidia RTX8000s and Xeon Gold 6230.

Software environments. Our experiments are conducted using Python 3.8, Pytorch 1.11.0, Pytorch-Geometric (PyG) 2.1.0 and Python Optimal Transport (POT) 0.8.2. Our GIN, GCN, and PNA implementations are based on their PyG implementations. For GOOD-HIV, we use the official implementation of GIN-Virtual Node (vGIN) in GOOD benchmark. We use POT package to compute the Wasserstein distance.

F. Quantitive Results on Good-Motif and Good-HIV

F.1. Baselines

For GOOD benchmark, We consider the heuristic CMDr (Buffelli et al., 2022) along with five mainstream out-of-distribution algorithms including IRM, VREx (Krueger et al., 2021), GroupDRO (Sagawa et al., 2019), Deep Coral (Sun & Saenko, 2016) and Mixup (Wang et al., 2021), as the baselines. More details of each baseline can be found in Appx. D. More implementation details are included in Appx. E.

F.2. Main results

Table 6 shows the performances of GIN (vGIN) backbones on the size generalization test set, with and without the proposed WBM layer. The GIN (vGIN) with a WBM layer outperforms the vanilla version by a large margin, with average improvement brought by the WBM layer up to 14.7%, which manifests the effectiveness of the WBM layer for improving the size generalization of MPNNs.

Table 6. Accuracy on GOOD-Motif and ROC-AUC on GOOD-HIV (mean $\pm$ std). The models are original GIN without ($\times$) and with ($\checkmark$) the WBM layer. The right-most column shows the improvement brought by the WBM layer.

Backbone	GIN(vGIN)		Improvement
WBM layer	$\times$	$\checkmark$	
GOOD-Motif	$51.75 \pm 2.88\%$	$59.36 \pm 5.02\%$	$\uparrow 14.7\%$
GOOD-HIV	$59.94 \pm 2.86\%$	$62.46 \pm 5.59\%$	$\uparrow 4.2\%$

In Table 7, we compare our method with baselines on GOOD benchmark. As can be seen, most mainstream out-of-distribution generalization methods fail under size shifts in the graph domain, while our method significantly outperforms all baseline methods. Specifically, for GOOD-Motif, WBM achieves a performance improvement of 2.61% compared to the best baseline CMDr. For GOOD-HIV, the performance is improved by 2.11%.

Table 7. Performance comparisons on GOOD benchmark between WBM and the baselines. Table shows the accuracy for GOOD-Motif and ROC-AUC for GOOD-HIV of the classifiers over the test data. The best (in mean) results are highlighted in bold.

Dataset	GOOD-Motif	GOOD-HIV
IRM	51.41 ± 3.30	59.00 ± 2.74
VREx	52.67 ± 2.87	58.53 ± 2.22
GroupDRO	51.95 ± 2.80	58.98 ± 1.84
Deep Coral	50.97 ± 1.76	60.11 ± 3.53
Mixup	51.48 ± 3.35	59.03 ± 3.07
CMDr	56.75 ± 7.14	60.35 ± 1.99
WBM(ours)	$\mathbf{59.36} \pm 5.02$	$\mathbf{62.46} \pm 5.59$

G. Additional Ablation Study

To inspect the **generalization convergence rate** by MPNN with and without WBM, we choose the GOOD-Motif dataset. The GOOD-Motif is a relatively large-scale dataset which allows us to construct different training subsets based on the graph size. We construct three splits for training, with different average graph sizes. We provide the average error rate (the lower the better) of the validation and test sets (with larger graph size, i.e., $N_s \leq N_t$) over multiple runs with GIN as the backbone in Table 8 below. The results show that the larger the average graph size for training, the better performance an MPNN with a WBM achieves, which implies a faster convergence rate with the presence of a WBM layer.

Table 8. The average error rate on GOOD-Motif of GIN backbone with a WBM layer and without a WBM layer.

Avg training graph size	**8.21**	**18.33**	**29.81**
with WBM	0.57	0.39	0.24
without WBM	0.58	0.45	0.31

We inspect using **a simple softmax as a classifier of the vanilla WBM model** studied in theory. We set $M = 1$ and we remove the MLP layer between the WBM layer and the softmax layer) with GIN as the backbone. We provide in Table 9 the average MCC of GIN without the vanilla WBM layer. The results are the average MCC over multiple runs. We can see that, though the comparison is not fair for the vanilla WBM (GIN without vanilla WBM uses an MLP layer between the average pooling layer and the softmax layer. In our analysis, we assume Lipschitz continuity of the classifier on top of the WBM embedding), applying the vanilla WBM achieves competitive performance on four datasets.

Table 9. The average MCC of GIN with (vanilla WBM) and without (vanilla GIN) the vanilla WBM layer.

Datasets	**NC1**	**NC109**	**PROTEINS**	**DD**
Vanilla GIN	0.19	0.28	0.25	0.23
Vanilla WBM	0.22	0.20	0.35	0.23

We inspect the importance of **enforcing a class-related semantic structure on the data clusters** in the Wasserstein space. Specifically, we model the learned template graphs explicitly as barycenters associated with various classes. We seek to minimize the Wasserstein distance between the learned graph and the graphs of the same class, whereas OT-based methods such as OT-GNN(Chen et al., 2020) seek to minimize the distance between the learned graph and any input graph. The semantic constraint arises from the basic assumption that graphs sampled from the same graphon RGM belong to the same class, and is critical for the convergence analysis from the graphon RGM perspective. To inspect the empirical effect of imposing semantic structure on the data clusters, we conduct a complementary experiment by minimizing the W-distance of the learned graphs against any input graph. The number of the learned graph equals $C \times M$. Table 10 below shows the average MCC over multiple runs with GIN as the backbone. The performance of minimizing W-distance against all input graphs is not satisfactory on the graph size generalization benchmarks, implying that imposing the semantic structure on the data clusters indeed makes a large difference.

Table 10. The average MCC of a GIN of WBM and a GIN minimizing W-distance against all graphs

Datasets	**NC1**	**NC109**	**PROTEINS**	**DD**
WBM	0.24	0.24	0.37	0.27
min W-distance against all graphs	0.19	0.15	0.34	0.17

H. The Algorithm of WBM

Algorithm 1 WBM: an MPNN with a Wasserstein Barycenter Matching layer

Input: Training dataset of graphs $\mathcal{S} = \{\mathbf{x}_k = (A_k, \mathbf{f}_k), \mathbf{y}_k\}_{k=1}^n$, MPNN Θ, classifier head ϕ, batch size B, total iterations for training I_{max}, trade-off hyperparameter λ, number of Wasserstein barycenters corresponding to each class M, graphon size N_g.
Output: learned models Θ, ϕ, and Wasserstein barycenters $\{\hat{\boldsymbol{b}}_2(\mathcal{S}_m^j) | j = 1 \cdots C, m = 1 \cdots M\}$.
Initialize: $\{\hat{\boldsymbol{b}}_2(\mathcal{S}_m^j)\}_{m=1}^M \leftarrow$ randomly sample M graphs of size N_g in each class j.
for $I = 1$ to I_{max} **do**
  randomly fetch a mini-batch B from $\mathcal{S}$
  $\ell \leftarrow 0$
  for each graph $(\mathbf{x} = (A, \mathbf{f}), \mathbf{y})$ instance in B **do**
    $\nu = \Theta_A(\mathbf{f}) \leftarrow$ calculate the node embeddings through the MPNN
    $\Theta_A^{\mathcal{W}}(\mathbf{f}) = (\mathcal{W}_2^2(\hat{\boldsymbol{b}}_2(\mathcal{S}_1^1), \nu), \cdots, \mathcal{W}_2^2(\hat{\boldsymbol{b}}_2(\mathcal{S}_M^1), \nu), \cdots, \mathcal{W}_2^2(\hat{\boldsymbol{b}}_2(\mathcal{S}_M^C), \nu)) \leftarrow$ calculate the WBM embedding
    $\hat{\mathbf{y}} = \phi(\Theta_A^{\mathcal{W}}(\mathbf{f})) \leftarrow$ classifier prediction
    $\ell_{CLS} = cross - entropy(\hat{\mathbf{y}}, \mathbf{y}) \leftarrow$ calculate the supervised classification loss
    $\ell_{WBM} = \frac{1}{n} \sum_{m=1}^M \mathcal{W}_2^2(\hat{\boldsymbol{b}}_2(\mathcal{S}_m^y), \nu) \leftarrow$ calculate the WBM loss
    $\ell = \ell + \ell_{CLS} + \lambda \ell_{WBM}$
  end for
  update parameters of Θ, ϕ and $\{\hat{\boldsymbol{b}}_2(\mathcal{S}_m^j)\}$ using $\nabla \ell$
end for

I. Third-party Lemmas

Lemma I.1 (Lipschitz-continuity of MPNN $\Theta_{\mathcal{A}}(f)$, a simplified version of Lemma B.9 in Maskey et al. (2022))**.** *Let (χ, d, μ) be a metric measure space and $\mathcal{A} : \chi \times \chi \to [0, +\infty)$ be a graphon. Let $\Theta = (\{\Phi^{(l)}, \Psi^{(l)}\}_{l=1}^L)$ be an L-layer MPNN s.t. the regularity assumptions in Appendix A are satisfied. Consider an L_f-Lipschitz continuous metric-space signal $f : \chi \to \mathbb{R}^F$ with finite infinity norm $\|f\| < \infty$. Then for $\ell = 1, \ldots, L$, the graphon MPNN output f^l is L_{f^l}-Lipschitz continuous and satisfying*

$$L_{f^{(l)}} \leq Z_1^{(l)} \|f\|_\infty + Z_2^{(l)} L_f, \tag{I.1}$$

where Z_1^l and Z_2^l are independent of f and defined as

$$\begin{aligned} Z_1^{(l)} &= \sum_{k=1}^l B^{(k-1)} \left(L_{\Psi^{(k)}} \frac{L_{\mathcal{A}}}{\mathrm{d}_{\min}} L_{\Phi^{(k)}} + L_{\Psi^{(k)}} \|\mathcal{A}\|_\infty L_{\Phi^{(k)}} \frac{L_{\mathcal{A}}}{\mathrm{d}_{\min}^2} \right) \prod_{l'=k+1}^l L_{\Psi^{(l')}} \left(1 + \frac{\|\mathcal{A}\|_\infty}{\mathrm{d}_{\min}} L_{\Phi^{(l')}} \right), \\ Z_2^{(l)} &= \prod_{k=1}^l L_{\Psi^{(k)}} \left(1 + \frac{\|\mathcal{A}\|_\infty}{\mathrm{d}_{\min}} L_{\Phi^{(k)}} \right), \end{aligned} \tag{I.2}$$

where $B^{(k)}$ is given by

$$B^{(k)} = \prod_{i=1}^k L_{\Psi^{(i)}} \left(1 + L_{\Phi^{(i)}} \right).$$

Lemma I.2 (bound on the mean $\mathbb{E}(\mathcal{W}_p(\hat{\nu}, \nu))$, Theorem 1 in Fournier & Guillin (2015))**.** *Let $\nu \in \mathbb{P}(\mathbb{R}^d)$ and let $p > 0$. Assume that $M_q(\nu) < \infty$ fome some $q > p$. There exists a constant D depending only on p, d, q such that for all $N \geq 1$,*

$$\mathbb{E}\left(\mathcal{W}_p\left(\hat{\nu}, \nu\right)\right) \leq D M_q^{p/q}(\nu) \times \begin{cases} N^{-1/2} + N^{-(q-p)/q} & \text{if } p > d/2 \text{ and } q \neq 2p, \\ N^{-1/2} \log(1+N) + N^{-(q-p)/q} & \text{if } p = d/2 \text{ and } q \neq 2p, \\ N^{-p/d} + N^{-(q-p)/q} & \text{if } p \in (0, d/2) \text{ and } q \neq d/(d-p). \end{cases} \tag{I.3}$$

Lemma I.3 (mean-concentration of $\mathcal{W}_p(\hat{\nu}, \nu)$, Corollary 5.5 in Lei (2020))**.** *For $X \sim \nu$, if $\mathbb{E}\left(\|X\|^k\right) \leq \frac{1}{2} s^2 k! V^{k-2}$, for all integer $k \geq 2$ and some constants s, V, then for all $t > 0$*

$$Prob\left[|\mathcal{W}_p(\hat{\nu}, \nu) - \mathbb{E}(\mathcal{W}_p(\hat{\nu}, \nu))| \geq t\right] \leq 2 \exp\left(-\frac{t^2}{8 s^2 n^{1-2/p} + 4 V t n^{-1/p}} \right). \tag{I.4}$$

Lemma I.4 (concentration of Wasserstein barycenters, simplified version of Theorem 12 in Le Gouic et al. (2022))**.** *Suppose the curvature $curv(\chi)$ is bounded from below. Fix a sub-Gaussian probability measure μ on χ and barycenter b^*. Let b_n be an empirical barycenter with n observations. For $\eta \in (0,1)$, then*

$$w.p. \geq 1-\eta-e^{-C_2 n} \quad \mathcal{W}_2(b_n, b^*) \leq \frac{E}{n}\log(\frac{2}{\eta}), \tag{I.5}$$

where $E, C_2 > 0$ are constants independent of n.

Lemma I.5 (concentration of multinomial measures, Proposition A.6 in Van Der Vaart et al. (1996))**.** *If the random vector $\boldsymbol{n} = (n_1, \ldots, n_C)$ is multinomially distributed with parameters n and $h^1, \ldots, h^C$, then*

$$Prob(\sum_{j=1}^{C} |n_j - nh^j| \geq 2\sqrt{n}t) \leq 2^C exp(-2t^2). \tag{I.6}$$

Lemma I.6 (Lipschitz-continuity of cross-entropy composed on softmax, Lemma D.1 in Maskey et al. (2022))**.** *Denoted by ℓ_{CE} the cross-entropy loss composed on softmax. Considering the simple binary case, ℓ_{CE} is defined by*

$$\ell_{\mathrm{CE}}(\mathbf{x}; \mathbf{y}) = -y_1 \log\left(\frac{e^{x_1}}{e^{x_1}+e^{x_2}}\right) - y_2 \log\left(\frac{e^{x_2}}{e^{x_1}+e^{x_2}}\right).$$

The loss ℓ_{CE} is 1=Lipschitz continuous. Additionally, ℓ_{CE} is locally bounded in the following sense:

$$\|\mathcal{L}_{\mathrm{CE}}\|_{L^\infty([-K,K]^2)} \leq \log\left(1+e^{2K}\right), \tag{I.7}$$

where $\|\mathcal{L}_{\mathrm{CE}}\|_{L^\infty([-K,K]^2)} = \max_{\mathbf{x}\in[-K,K]^2} \|\mathcal{L}_{\mathrm{CE}}(\mathbf{x})\|$.