Graph Neural Differential Equations in the Infinite-Node Limit: Convergence and Rates via Graphon Theory

Mingsong Yan
Department of Mathematics
University of California, Santa Barbara
Santa Barbara, CA 93106
`mingsongyan@ucsb.edu`

Charles Kulick
Department of Mathematics
University of California, Santa Barbara
Santa Barbara, CA 93106
`charleskulick@ucsb.edu`

Sui Tang
Department of Mathematics
University of California, Santa Barbara
Santa Barbara, CA 93106
`suitang@ucsb.edu`

Abstract

Graph Neural Differential Equations (GNDEs) combine the structural inductive bias of Graph Neural Networks (GNNs) with the continuous-depth architecture of Neural ODEs, offering an effective framework for modeling dynamics on graphs. In this paper, we present the first rigorous convergence analysis of GNDEs with time-varying parameters in the infinite-node limit, providing theoretical insights into their *size transferability*. We introduce Graphon Neural Differential Equations (Graphon-NDEs) as the infinite-node limit of GNDEs and establish their well-posedness. Leveraging tools from graphon theory and dynamical systems, we prove the *trajectory-wise* convergence of GNDE solutions to Graphon-NDE solutions. Moreover, we derive explicit convergence rates for GNDEs over weighted graphs sampled from Lipschitz-continuous graphons and unweighted graphs sampled from $\{0, 1\}$-valued (discontinuous) graphons. We further obtain size transferability bounds, providing theoretical justification for the practical strategy of transferring GNDE models trained on moderate-sized graphs to larger, structurally similar graphs without retraining. Numerical experiments support our theoretical findings.

1 Introduction

Graph Neural Networks (GNNs) [Scarselli et al., 2008] have achieved remarkable success across diverse graph-based learning tasks [Duvenaud et al., 2015, Battaglia et al., 2016, Hamilton et al., 2017, Sanchez-Gonzalez et al., 2020, Derrow-Pinion et al., 2021], in part due to their potential for *size transferability*: a model trained on smaller graphs can often be deployed on larger, structurally similar graphs without retraining [Ruiz et al., 2020, Levie et al., 2021]. This property is typically justified by convergence analyses under assumptions on graph sequences, message-passing operators, and activation functions, often modeled via graphons [Lovász, 2012]. Under such assumptions, GNN outputs converge to a continuous limit as graph size grows, and convergence rates provide explicit bounds on transferability errors. These results have been established for a wide range of architectures, including spectral, message-passing, invariant, and higher-order GNNs [Ruiz et al.,

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

Attribute	GNNs	GNDEs (ours)
Layer Type	Discrete	Continuum
Coefficient Type	Static	Temporally Continuous
Convergence Notion	Layer-wise	Trajectory-wise
Graphon Type	**Convergence Rates**	
Lipschitz Continuous	$\mathcal{O}(1/\sqrt{n})$ [Ruiz et al., 2020]; $\mathcal{O}(1/n)$ [Maskey et al., 2023]	$\mathcal{O}(1/n)$
$\{0,1\}$-valued	Inexplicit[1] [Morency and Leus, 2021, Kenlay et al., 2021a,b]	$\mathcal{O}(1/n^c)$, $c \in (0,1)$

Table 1: Comparison of Infinite-Node Limit Results for Spectral GNNs and GNDEs

2020, Keriven et al., 2020, Kenlay et al., 2021a, Levie et al., 2021, Cai and Wang, 2022, Maskey et al., 2023, Cordonnier et al., 2023, Le and Jegelka, 2024, Herbst and Jegelka, 2025].

Continuous-depth GNNs, often referred to as Graph Neural Differential Equations (GNDEs) [Poli et al., 2019, Liu et al., 2025], extend this paradigm by modeling node features as solutions of ODEs parameterized by GNNs, combining the expressivity of Neural ODEs [Chen et al., 2018] with graph inductive biases. Distinct from discrete-layer GNNs, GNDEs generate infinitely many intermediate states over time, requiring a stronger notion of convergence: the entire feature trajectory should converge *uniformly* in the infinite-node limit. Discretizing GNDEs into residual GNNs cannot guarantee this, since step sizes may not scale with graph size and discrete evaluations neglect error accumulation between time points. To overcome these limitations, we analyze GNDEs directly in continuous time, obtaining simultaneous, *uniform-in-time* convergence of trajectories via dynamical systems tools such as Grönwall-type stability estimates, thereby laying theoretical foundations for size transferability in continuous-depth models.

Contributions. We summarize our contributions as follows:

- **Infinite-node limits of GNDEs.** We introduce an infinite-node limit of GNDEs, termed *Graphon Neural Differential Equations* (Graphon-NDEs), which are a class of partial differential equations (PDEs) defined on graphon spaces. We establish sufficient conditions for their well-posedness, ensuring the existence and uniqueness of solutions. To the best of our knowledge, this is the first work considering infinite-node limits of GNDEs.
- **Trajectory-wise convergence.** We prove that solution trajectories of GNDEs (a sequence of ODEs) uniformly converge to a Graphon-NDE (a PDE) whenever the underlying graph sequences and initial features converge. Our analysis relies on Grönwall-type inequalities from dynamical systems and accommodates time-varying (temporally continuous) parameters.
- **Convergence rates.** We derive explicit convergence rates for both weighted and unweighted graphs which are generated deterministically from graphons. For weighted graphs sampled from Lipschitz graphons, we present a convergence rate of $\mathcal{O}(1/n)$; for unweighted graphs sampled from $\{0,1\}$-valued (hence non-continuous) graphons, we show a convergence rate of $\mathcal{O}(1/n^c)$, with $c \in (0,1)$ depending on the box-counting dimension of the boundary of the graphon's support.
- **Size transferability bounds.** Leveraging our derived convergence rates, we establish upper bounds on the solution discrepancy of GNDEs over graphs of different sizes. This provides theoretical justification for the size transferability of GNDEs – models trained on smaller graphs can reliably generalize to larger, structurally similar graphs without retraining.

2 Preliminaries

Graph Neural Differential Equations (GNDEs) [Poli et al., 2019] extend Neural ODEs to the graph domain by modeling the continuous-time dynamics with a Graph Neural Network (GNN). Formally,

[1] "Inexplicit" means that convergence is established, but no explicit rate is provided.

a GNDE is defined as

$$\frac{d}{dt}\boldsymbol{X}(t) = \Phi\big(\boldsymbol{S}; \boldsymbol{X}(t); \mathbf{H}(t)\big), \\ \boldsymbol{X}(0) = \boldsymbol{Z} \in \mathbb{R}^{|V(\mathcal{G})|\times F}, \tag{1}$$

in which $\boldsymbol{X}(t) \in \mathbb{R}^{|V(\mathcal{G})|\times F}$ denotes the node feature matrix at time t and is initialized by the input node feature matrix $\boldsymbol{Z}$ at $t = 0$; and Φ is an L-layer GNN parameterized by a graph shift operator $\boldsymbol{S}$ and a collection of trainable, time-varying K-hop filter coefficients $\mathbf{H}(t) = \{\boldsymbol{h}_{fgk}^{(\ell,t)} : f, g \in [F], k \in \mathbb{Z}_K, \ell \in [L]\}$, $t \in [0, T]$.

3 Main Results

3.1 Infinite-Node Limits: Graphon Neural Differential Equations and Well-Posedness

To explore the infinite-node limiting structure of GNDEs, we introduce *Graphon Neural Differential Equations* (Graphon-NDEs). Recalling that a graphon, as the limiting object of finite graphs, can be viewed as a graph with a continuum of nodes over the unit interval, we define Graphon-NDEs in a form similar to GNDEs (1), but tailored to operate on graphons rather than finite graphs. Specifically, we formulate Graphon-NDEs as

$$\frac{\partial}{\partial t}\mathbf{X}(u,t) = \Phi(\mathbf{W}; \mathbf{X}(u,t); \mathbf{H}(t)), \\ \mathbf{X}(u,0) = \mathbf{Z}(u), \tag{2}$$

where $I := [0,1]$, $\mathbf{X}(\cdot,t) : I \to \mathbb{R}^{1\times F}$ is the graphon node feature function at time t and initialized by an input node feature function $\mathbf{Z}$ at $t = 0$; and Φ is a Graphon-NN applying on $\mathbf{X}(\cdot,t)$ through graphon $\mathbf{W}$ and time-varying parameters $\mathbf{H}(t)$.

The continuum nature of both the node and time variables in Graphon-NDEs necessitates careful technical treatment to establish their *well-posedness* (i.e., the existence and uniqueness of solutions). We prove that the temporal continuity of the filter evolution and the non-amplifying Lipschitz property of the activation function (see Assumptions AS0 and AS1 below) suffice to guarantee well-posedness.

- **AS0.** The convolutional filters evolves continuously in time, i.e., $\boldsymbol{h}_{fgk}^{(\ell,t)}$ is a continuous function about $t \in [0,T]$, for each $f, g \in [F]$, $\ell \in [L]$, $k \in \mathbb{Z}_K$.
- **AS1.** The activation function σ is normalized Lipschitz, i.e., $|\sigma(x) - \sigma(y)| \leq |x - y|$, for all $x, y \in \mathbb{R}$; and $\sigma(0) = 0$.

Theorem 3.1 (Well-posedness, proof in Appendix A.3)**.** *Suppose that AS0 and AS1 hold. If* $\mathbf{W} \in L^\infty(I^2)$ *and* $\mathbf{Z} \in L^\infty(I; \mathbb{R}^{1\times F})$*, then for any* $T > 0$*, there exists a unique solution* $\mathbf{X} \in C^1\left([0,T]; L^\infty(I; \mathbb{R}^{1\times F})\right)$ *to the Graphon-NDE* (2)*.*

The well-posedness result established in Theorem 3.1 paves the way for the subsequent convergence analysis of GNDE solutions to the Graphon-NDE solution as the sequence of structurally similar graphs converges to a graphon. Theorem 3.1 presents that the unique solution $\mathbf{X}$ of the Graphon-NDE is uniformly bounded, which immediately implies that $\mathbf{X}$ is square integrable, i.e., $\mathbf{X} \in C\left([0,T]; L^2(I; \mathbb{R}^{1\times F})\right)$. Our forthcoming convergence results and rate estimates for GNDE solutions will be formulated in this L^2-based function space.

3.2 Trajectory-Wise Convergence

We proceed to study the convergence of GNDEs to Graphon-NDEs in terms of their solution trajectories. Let $\{\mathcal{G}_n\}$ be a sequence of graphs with adjacency matrices $\{\boldsymbol{W}_{\mathcal{G}_n}\}$. Let the GSO $\boldsymbol{S}_{\mathcal{G}_n}$ be defined as the adjacency matrix $\boldsymbol{W}_{\mathcal{G}_n}$ normalized by $1/|V(\mathcal{G}_n)|$, i.e., $\boldsymbol{S}_{\mathcal{G}_n} := \boldsymbol{W}_{\mathcal{G}_n}/|V(\mathcal{G}_n)|$. Recalling (1), we formulate a sequence of GNDEs as

$$\frac{d}{dt}\boldsymbol{X}_{\mathcal{G}_n}(t) = \Phi(\boldsymbol{S}_{\mathcal{G}_n}; \boldsymbol{X}_{\mathcal{G}_n}(t); \mathbf{H}(t)), \\ \boldsymbol{X}_{\mathcal{G}_n}(0) = \boldsymbol{Z}_{\mathcal{G}_n} \in \mathbb{R}^{|V(\mathcal{G}_n)|\times F}, \tag{3}$$

where $\boldsymbol{Z}_{\mathcal{G}_n}$ is the initial node feature matrix for graph $\mathcal{G}_n$. Below we establish the *trajectory-wise* convergence of GNDE solutions to Graphon-NDE solutions.

Theorem 3.2 (Trajectory-wise convergence, proof in Appendix A.5)**.** *Suppose that AS0 and AS1 hold, and let $\mathbf{W} \in L^\infty(I^2)$ and $\mathbf{Z} \in L^\infty(I; \mathbb{R}^{1\times F})$. Let $\mathbf{X}$ and $\boldsymbol{X}_{\mathcal{G}_n}$ denote the solutions of Graphon-NDE* (2) *and GNDE* (3)*, respectively. If $\{(\mathcal{G}_n, \boldsymbol{Z}_{\mathcal{G}_n})\}$ converges to $(\mathbf{W}, \mathbf{Z})$ (cf. Definition 1), then for any $T > 0$, there exists a sequence $\{\pi_n\}$ of permutations such that*

$$\lim_{n\to\infty} \|\mathbf{X} - \mathbf{X}_{\pi_n(\mathcal{G}_n)}\|_{C([0,T];L^2(I;\mathbb{R}^{1\times F}))} = 0,$$

where $\mathbf{X}_{\pi_n(\mathcal{G}_n)}$ denotes the induced graphon feature function of $\boldsymbol{X}_{\pi_n(\mathcal{G}_n)}$.

Discussion. The norm in the function space $C([0,T]; L^2(I; \mathbb{R}^{1\times F}))$ (cf. Appendix A.2) involves taking the supremum over $t \in [0,T]$. Consequently, the convergence we establish is uniform in time; that is, as $n \to \infty$, the approximation error diminishes uniformly along the entire trajectory, which consists of infinitely many intermediate states. In contrast, the convergence results in the literature for GNNs with finitely many layers [Ruiz et al., 2020, Keriven et al., 2020, Maskey et al., 2023] establish convergence only at the discrete set of layer outputs as the graph size grows. The trajectory-wise convergence we prove for GNDEs is therefore fundamentally stronger. Moreover, we remark that the established trajectory-wise convergence relies on Grönwall-type inequalities from dynamical systems and stability theory, which are tools not required in the existing GNN literatures.

The convergence property established in Theorem 3.2 suggests that GNDEs exhibit stability on large-scale, structurally similar graphs and are robust to perturbations in the graph structure or node features. It hinges on the temporal continuity of convolutional filters and Lipschitz continuity for the activation function. The latter assumption aligns with recent empirical studies of GNNs [Dasoulas et al., 2021, Arghal et al., 2022], which demonstrate that enhanced Lipschitz continuity in GNNs improves robustness, generalization, and performance on large-scale tasks. Moreover, Theorem 3.2 rigorously characterizes the function space $C([0,T]; L^2(I; \mathbb{R}^{1\times F}))$ in which GNDEs can approximate in the continuum regime. This complements recent advancements in the study of GNN limits and their expressive capabilities [Keriven et al., 2021, Keriven and Vaiter, 2023].

3.3 Convergence Rates

In this section, we use graphons as generative models to construct convergent graph sequences: weighted graphs sampled from Lipschitz-continuous graphons and unweighted graphs sampled from $\{0,1\}$-valued (discontinuous) graphons. We further refine our convergence theorem by deriving explicit convergence rates for each case.

3.3.1 Weighted Graphs

Let $\mathbf{W} : I^2 \to I$ be a graphon and $\mathbf{Z} \in L^\infty(I; \mathbb{R}^{1\times F})$ be a graphon feature function. For each $n \in \mathbb{N}$, we partition the unit interval I into n sub-intervals by defining $u_i := (i-1)/n$ and $I_i := [u_i, u_{i+1})$ for $i \in [n]$. We define a graph $\mathcal{G}_n$ of n nodes as $\mathcal{G}_n := \langle [n], [n] \times [n], \boldsymbol{W}_{\mathcal{G}_n} \rangle$, where we generate the weighted adjacency matrix $\boldsymbol{W}_{\mathcal{G}_n} \in \mathbb{R}^{n\times n}$ by direct sampling on the graphon $\mathbf{W}$ over the mesh grid as

$$[\boldsymbol{W}_{\mathcal{G}_n}]_{ij} := \mathbf{W}(u_i, u_j), \quad i,j \in [n]. \tag{4}$$

The corresponding node feature matrix $\boldsymbol{Z}_{\mathcal{G}_n} \in \mathbb{R}^{n\times F}$ of graph $\mathcal{G}_n$ is generated by sampling on the graphon feature function $\mathbf{Z}$ as

$$[\boldsymbol{Z}_{\mathcal{G}_n}]_{i,:} := \mathbf{Z}(u_i), \quad i \in [n]. \tag{5}$$

This weighted graph model is particularly well-suited for applications requiring fully connected network structures, such as dense communication networks and recommendation systems [Barrat et al., 2004, Newman, 2004, Aggarwal, 2016]. In these settings, the graphons are typically assumed to be Lipschitz continuous, reflecting the fact that interactions between entities (e.g., users, devices, or items) evolve gradually and predictably. We summarize the assumptions below.

- **AS2.** The graphon $\mathbf{W}$ is A_1-Lipschitz, that is, $|\mathbf{W}(u_2, v_2) - \mathbf{W}(u_1, v_1)| \le A_1(|u_2 - u_1| + |v_2 - v_1|)$, for all $v_1, v_2, u_1, u_2 \in I$.
- **AS3.** The initial graphon feature function $\mathbf{Z} = [Z_f : f \in [F]] \in L^\infty(I; \mathbb{R}^{1\times F})$ is A_2-Lipschitz, that is, for each $f \in [F]$, $|Z_f(u_2) - Z_f(u_1)| \le A_2|u_2 - u_1|$, for all $u_1, u_2 \in I$.

Theorem 3.3 (Rates for weighted graphs, proof in Appendix A.6)**.** *Suppose that AS0-AS3 hold. Let the adjacency matrices and node feature matrices of graphs $\{\mathcal{G}_n\}$ be generated according to* (4) *and* (5)*, respectively. Let $T \in \mathbb{R}^+$. Let $\mathbf{X}$ be the solution of Graphon-NDE* (2) *and $\mathbf{X}_{\mathcal{G}_n}$ be the induced graphon function of the solution $\boldsymbol{X}_{\mathcal{G}_n}$ of GNDE* (3)*. Then it holds that*

$$\|\mathbf{X} - \mathbf{X}_{\mathcal{G}_n}\|_{C([0,T];L^2(I;\mathbb{R}^{1\times F}))} \leq \frac{C}{n}, \tag{6}$$

where C is constant independent of n with explicit formula provided in equation (32)*. As a result, for any $n_1, n_2 \in \mathbb{N}$, it holds that*

$$\|\mathbf{X}_{\mathcal{G}_{n_1}} - \mathbf{X}_{\mathcal{G}_{n_2}}\|_{C([0,T];L^2(I;\mathbb{R}^{1\times F}))} \leq C\left(\frac{1}{n_1} + \frac{1}{n_2}\right). \tag{7}$$

Discussion. We remark that Theorem 3.3 establishes an $\mathcal{O}(1/n)$ convergence rate for weighted graphs sampled from Lipschitz-continuous graphons. This rate is known to be optimal for approximating Lipschitz-continuous functions [Schumaker, 2007]. Furthermore, the rate for GNDEs we obtain is trajectory-wise (i.e., uniform-in-time), which is strictly stronger than the linear convergence rates established for discrete-layer GNNs [Maskey et al., 2023, Krishnagopal and Ruiz, 2023].

3.3.2 Unweighted Graphs

Let $\mathbf{W} : I^2 \to \{0,1\}$ be a binary-valued graphon and $\mathbf{Z} \in L^\infty(I;\mathbb{R}^{1\times F})$ be a graphon feature function. We denote by $\mathbf{W}^+$ the support set of function $\mathbf{W}$, that is $\mathbf{W}^+ := \{(u,v) : \mathbf{W}(u,v) = 1\}$. For each $n \in \mathbb{N}$, we construct an unweighted graph $\mathcal{G}_n$ as $\mathcal{G}_n := \langle [n], E(\mathcal{G}_n), \boldsymbol{W}_{\mathcal{G}_n}\rangle$, where the edge set $E(\mathcal{G}_n)$ is defined by $E(\mathcal{G}_n) := \{(i,j) \in [n]\times[n] : (I_i \times I_j) \cap \mathbf{W}^+ \neq \emptyset\}$, and the adjacency matrix $\boldsymbol{W}_{\mathcal{G}_n}$ is defined as

$$[\boldsymbol{W}_{\mathcal{G}_n}]_{ij} := \begin{cases} 1, & \text{if } (i,j) \in E(\mathcal{G}_n), \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where $[\boldsymbol{W}_{\mathcal{G}_n}]_{ij}$ represents the binary connectivity between nodes i and j of the graph $\mathcal{G}_n$. The corresponding node feature matrix $\boldsymbol{Z}_{\mathcal{G}_n}$ for graph $\mathcal{G}_n$ is generated, from a Lipschitz continuous graphon feature function $\mathbf{Z}$, as

$$[\boldsymbol{Z}_{\mathcal{G}_n}]_{i,:} := \frac{1}{|I_i|}\int_{I_i} \mathbf{Z}(u)\,du, \quad i \in [n]. \tag{9}$$

This model is for generating network structures with binary relations, which are prevalent in social networks, citation graphs, and biological networks [Jeong et al., 2000, Milo et al., 2002, Girvan and Newman, 2002, Leskovec et al., 2009, Easley and Kleinberg, 2010].

The discontinuity of graphons prevents AS2 from being satisfied. To tackle this issue, we introduce a new metric—the upper box-counting dimension [Falconer, 2014] for the boundary $\partial\mathbf{W}^+$, where $\mathbf{W}^+$ is the support of the graphon $\mathbf{W}$. We review the definition of upper box-counting dimension as follows. Let Ω be any non-empty bounded subset of $\mathbb{R}^2$ and let $\mathcal{N}_\delta(\Omega)$ be the number of δ-mesh cubes that intersect Ω. The upper box-counting dimensions of Ω is defined as

$$\overline{\dim}_{\mathrm{B}}\Omega := \overline{\lim_{\delta\to 0}} \frac{\log \mathcal{N}_\delta(\Omega)}{-\log\delta}. \tag{10}$$

It is clear that $\overline{\dim}_{\mathrm{B}}(\Omega) \in [0,2]$ for any non-empty bounded subset Ω of $\mathbb{R}^2$. As a simple example, the straight line $\{(x,0) : x \in [0,1]\}$ has an upper box-counting dimension of 1.

Theorem 3.4 (Rates for unweighted graphs, proof in Appendix A.6)**.** *Suppose that AS0, AS1 and AS3 hold. Let $\mathbf{W} : I^2 \to \{0,1\}$ be a graphon for unweighted graphs with $b := \overline{\dim}_{\mathrm{B}}(\partial\mathbf{W}^+) \in [1,2)$. Let the adjacency matrices and node feature matrices of graphs $\{\mathcal{G}_n\}$ be generated according to* (8) *and* (9)*, respectively. Let $T \in \mathbb{R}^+$. Let $\mathbf{X}$ be the solution of Graphon-NDE* (2) *and $\mathbf{X}_{\mathcal{G}_n}$ be the induced graphon function of the solution $\boldsymbol{X}_{\mathcal{G}_n}$ of GNDE* (3)*. Then for any $\epsilon \in (0, 2-b)$, there exists a positive integer $N_{\epsilon,\mathbf{W}}$ (depending on ϵ and $\mathbf{W}$) such that when $n > N_{\epsilon,\mathbf{W}}$, it holds that*

$$\|\mathbf{X} - \mathbf{X}_{\mathcal{G}_n}\|_{C([0,T];L^2(I;\mathbb{R}^{1\times F}))} \leq \frac{\widetilde{C}}{n^{1-\frac{b+\epsilon}{2}}}, \tag{11}$$

where $\widetilde{C}$ is a constant independent of n with explicit formula provided in equation (36). *As a result, for any $n_1, n_2 > N_{\epsilon,\mathbf{W}}$, it holds that*

$$\|\mathbf{X}_{\mathcal{G}_{n_1}} - \mathbf{X}_{\mathcal{G}_{n_2}}\|_{C([0,T];L^2(I;\mathbb{R}^{1\times F}))} \leq \widetilde{C}\left(\frac{1}{n_1^{1-\frac{b+\epsilon}{2}}} + \frac{1}{n_2^{1-\frac{b+\epsilon}{2}}}\right). \tag{12}$$

Discussion. The $\epsilon > 0$ in Theorem 3.4 is a pre-specified parameter that can be chosen arbitrarily small, making the convergence rate in Theorem 3.4 *almost* $\mathcal{O}\left(1/n^{1-b/2}\right)$. In contrast to the rate for weighted graphs established in Theorem 3.3, the rate for unweighted graphs relies on the complexity of the boundary $\partial\mathbf{W}^+$, measured by its upper box-counting dimension b. The more intricate $\partial\mathbf{W}^+$ is, leading to the larger value of b, the poorer the convergence rate becomes. For boundaries with box-counting dimension $b = 1$ (e.g., smooth curves or piecewise linear segments), convergence is relatively fast at rate $\mathcal{O}(1/n^{0.5})$. For boundaries with greater fractal complexity, where $b \in (1, 2)$ (e.g., moderately irregular or self-similar structures such as the hexaflake), convergence slows to $\mathcal{O}(1/n^c)$ for some $c \in (0, 0.5)$. We note that numerical experiments (see HSBM (hierarchical stochastic block model) and hexaflake graphons in Figure 1) suggest that our theoretical rate for unweighted graphs may be *pessimistic*, reflecting a worst-case scenario. Empirically, faster convergence rates are observed. In addition, we find that the HSBM graphon appears to yield faster convergence than the hexaflake, likely due to its smaller box-counting dimension. This observation is consistent with the trend indicated in Theorem 3.4, where a larger box-counting dimension corresponds to a slower convergence rate.

We remark that the graphons for unweighted graphs are discontinuous and prior studies on GNNs [Ruiz et al., 2021a,b, Morency and Leus, 2021, Maskey et al., 2023] lack convergence rates for this case. In contrast, our result goes beyond GNNs and establishes trajectory-wise rates for GNDEs over unweighted graphs, using a novel analysis based on the box-counting dimension.

3.4 Implications

Size transferability bounds. Estimates (7) and (12) provide quantitative bounds on how GNDE solutions differ when defined over structurally similar graphs of different sizes (n_1 and n_2), assuming shared convolutional filters. These bounds offer theoretical insight into the size transferability of GNDEs, quantifying how solution trajectories remain consistent as the graph scales. In particular, they highlight the role of graph structure (e.g., graphon property) and model smoothness (e.g., convolutional filters and activation functions) in ensuring reliable transferability across graph sizes. Our analysis implies that size transferability becomes more challenging for irregular graphs.

Two-scale convergence of discretized GNDEs. Discretized GNDEs can be obtained by applying numerical solvers to GNDEs, resulting in novel constructions of GNNs with residual connections. Despite their practical importance, no convergence analysis for these discretized GNDEs exists in the current literatures. Our convergence results show that GNDE solutions over size-n graphs converge uniformly in time to a Graphon-NDE solution with rate $\mathcal{O}(n^{-\alpha})$, with α dependent on regularity of graphons. To ensure that such convergence behavior carries over to discretized GNDEs used in practice, we also need to control the numerical solver error. Specifically, if a solver with global error $\mathcal{O}(h^p)$ is used, then to preserve the overall convergence to the graphon limit, we need to require $h^p \ll n^{-\alpha}$. This setup reflects a *two-scale convergence*: as both the graph size increases and the time step decreases, the discretized numerical solutions of GNDEs will converge to the Graphon-NDE solution. In practice, this informs the choice of solver: for smooth GNDEs, high-order explicit methods (e.g., RK4) suffice, while stiff dynamics may call for implicit solvers to control long-term error growth. This principle ensures that the discretized model remains consistent across graph sizes and time resolutions.

References

Charu C. Aggarwal. *Recommender systems*, volume 1. Springer, 2016.

Raghu Arghal, Eric Lei, and Shirin Saeedi Bidokhti. Robust graph neural networks via probabilistic lipschitz constraints. In *Learning for Dynamics and Control Conference*, pages 1073–1085. PMLR, 2022.

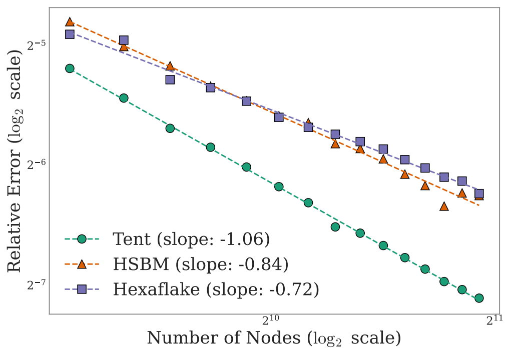

Figure 1: **Convergence rates of GNDE solutions.** Relative errors between GNDE and Graphon-NDE solutions on graphs sampled from three graphons: (1) *Tent graphon* (Lipschitz), matching $\mathcal{O}(1/n)$ rate in Theorem 3.3, (2) *HSBM graphon* (box counting dimension 1) and (3) *Hexaflake graphon* (fractal boundary with box counting dimension 1.77). The HSBM graphon yields faster convergence than the hexaflake, consistent with the trend indicated in Theorem 3.4. More details are in Appendix B.

Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *Proceedings of the national academy of sciences*, 101 (11):3747–3752, 2004.

Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. *Advances in neural information processing systems*, 29, 2016.

Chen Cai and Yusu Wang. Convergence of invariant graph networks. In *International Conference on Machine Learning*, pages 2457–2484. PMLR, 2022.

Ricky T. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Sajid M. Choudhury and M. A. Matin. Effect of fss ground plane on second iteration of hexaflake fractal patch antenna. In *2012 7th International Conference on Electrical and Computer Engineering*, pages 694–697, 2012. doi: 10.1109/ICECE.2012.6471645.

Matthieu Cordonnier, Nicolas Keriven, Nicolas Tremblay, and Samuel Vaiter. Convergence of message passing graph neural networks with generic aggregation on large random graphs. *arXiv preprint arXiv:2304.11140*, 2023.

Harry Crane and Walter Dempsey. A framework for statistical network modeling. *arXiv preprint arXiv:1509.08185*, 2015.

George Dasoulas, Kevin Scaman, and Aladin Virmaux. Lipschitz normalization for self-attention layers with application to graph neural networks. In *International Conference on Machine Learning*, pages 2456–2466. PMLR, 2021.

Austin Derrow-Pinion, Jennifer She, David Wong, Oliver Lange, Todd Hester, Luis Perez, Marc Nunkesser, Seongjae Lee, Xueying Guo, Brett Wiltshire, et al. Eta prediction with graph neural networks in google maps. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 3767–3776, 2021.

John R. Dormand and Peter J. Prince. A family of embedded Runge-Kutta formulae. *Journal of Computational and Applied Mathematics*, 6(1):19–26, 1980.

Sever S. Dragomir. Some Gronwall type inequalities and applications. *Science Direct Working Paper*, 0(S1574-0358):04, 2003.

David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28, 2015.

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

Kenneth Falconer. *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons, 2014.

Michelle Girvan and Mark E. Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

Daniel Herbst and Stefanie Jegelka. Higher-order graphon neural networks: Approximation and cut distance. *arXiv preprint arXiv:2503.14338*, 2025.

Paul W. Holland, Kathryn B. Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Svante Janson. Graphons, cut norm and distance, couplings and rearrangements. *arXiv preprint arXiv:1009.2376*, 2010.

Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N. Oltvai, and A-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

Henry Kenlay, Dorina Thano, and Xiaowen Dong. On the stability of graph convolutional neural networks under edge rewiring. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8513–8517. IEEE, 2021a.

Henry Kenlay, Dorina Thanou, and Xiaowen Dong. Interpretable stability bounds for spectral graph filters. In *International conference on machine learning*, pages 5388–5397. PMLR, 2021b.

Nicolas Keriven and Samuel Vaiter. What functions can graph neural networks compute on random graphs? the role of positional encoding. *Advances in Neural Information Processing Systems*, 36: 11823–11849, 2023.

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

Nicolas Keriven, Alberto Bietti, and Samuel Vaiter. On the universality of graph neural networks on large random graphs. *Advances in Neural Information Processing Systems*, 34:6960–6971, 2021.

Sanjukta Krishnagopal and Luana Ruiz. Graph neural tangent kernel: Convergence on large graphs. In *International Conference on Machine Learning*, pages 17827–17841. PMLR, 2023.

Thien Le and Stefanie Jegelka. Limits, approximation and size transferability for gnns on sparse graphs via graphops. *Advances in Neural Information Processing Systems*, 36, 2024.

Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

Ron Levie, Wei Huang, Lorenzo Bucci, Michael Bronstein, and Gitta Kutyniok. Transferability of spectral graph convolutional neural networks. *Journal of Machine Learning Research*, 22(272): 1–59, 2021.

Zewen Liu, Xiaoda Wang, Bohan Wang, Zijie Huang, Carl Yang, and Wei Jin. Graph ODEs and beyond: A comprehensive survey on integrating differential equations with graph neural networks. *arXiv preprint arXiv:2503.23167*, 2025.

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

Sohir Maskey, Ron Levie, and Gitta Kutyniok. Transferability of graph neural networks: an extended graphon approach. *Applied and Computational Harmonic Analysis*, 63:48–83, 2023.

Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

Matthew W. Morency and Geert Leus. Graphon filters: Graph signal processing in the limit. *IEEE Transactions on Signal Processing*, 69:1740–1754, 2021.

Mark E. Newman. Analysis of weighted networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 70(5):056131, 2004.

AI Perov. K voprosu o strukture integral'noı voronki. *Nauc. Dokl. Vysšeıi Školy. Ser FMN*, 2, 1959.

Michael Poli, Stefano Massaroli, Junyoung Park, Atsushi Yamashita, Hajime Asama, and Jinkyoo Park. Graph neural ordinary differential equations. *arXiv preprint arXiv:1911.07532*, 2019.

Luana Ruiz, Luiz Chamon, and Alejandro Ribeiro. Graphon neural networks and the transferability of graph neural networks. *Advances in Neural Information Processing Systems*, 33:1702–1712, 2020.

Luana Ruiz, Luiz F. Chamon, and Alejandro Ribeiro. Graphon signal processing. *IEEE Transactions on Signal Processing*, 69:4961–4976, 2021a.

Luana Ruiz, Luiz F. Chamon, and Alejandro Ribeiro. Graphon filters: Signal processing in very large graphs. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 1050–1054. IEEE, 2021b.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International conference on machine learning*, pages 8459–8468. PMLR, 2020.

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

Larry Schumaker. *Spline functions: basic theory*. Cambridge university press, 2007.

A Supplemental Materials: Theory

A.1 Graph Limits

In this section, we provide more details of graphons as graph limits and present the formal definition for the convergence of a sequence of graph-feature pairs to a graphon-feature pair.

We begin with the concept of a sequence of graphs converging to a graphon in the sense of *homomorphism density* [Lovász, 2012]. A *motif* $\mathcal{F}$ is an arbitrary simple graph. A *homomorphism* from a motif $\mathcal{F}$ to a simple unweighted graph $\mathcal{G}$ is an adjacency-preserving mapping $\phi : V(\mathcal{F}) \to V(\mathcal{G})$, meaning $(i, j) \in E(\mathcal{F})$ implies $(\phi(i), \phi(j)) \in E(\mathcal{G})$, and the *homomorphism number* $\hom(\mathcal{F}, \mathcal{G})$ refers to the total number of homomorphisms from $\mathcal{F}$ to $\mathcal{G}$. The *homomorphism density* $t(\mathcal{F}, \mathcal{G})$ is defined as the ratio of $\hom(\mathcal{F}, \mathcal{G})$ and $|V(\mathcal{G})|^{|V(\mathcal{F})|}$, which represents the probability of a random mapping $\phi : V(\mathcal{F}) \to V(\mathcal{G})$ being a homomorphism. The notion of homomorphism density can be similarly extended to the case of $\mathcal{G}$ being weighted graphs [Lovász, 2012]

$$t(\mathcal{F}, \mathcal{G}) = \frac{\hom(\mathcal{F}, \mathcal{G})}{|V(\mathcal{G})|^{|V(\mathcal{F})|}} = \frac{\sum_{\phi} \prod_{(i,j)\in E(\mathcal{F})} [\boldsymbol{W}_{\mathcal{G}}]_{\phi(i)\phi(j)}}{|V(\mathcal{G})|^{|V(\mathcal{F})|}}. \tag{13}$$

The homomorphism density from a motif to a graphon is generalized via integrals. We define the homomorphism density from a motif $\mathcal{F}$ to a graphon $\mathbf{W}$, denoted by $t(\mathcal{F}, \mathbf{W})$, as

$$t(\mathcal{F}, \mathbf{W}) := \int_{I^{|V(\mathcal{F})|}} \prod_{(i,j)\in E(\mathcal{F})} \mathbf{W}(u_i, u_j) \prod_{i\in V(\mathcal{F})} du_i. \tag{14}$$

We say that a sequence of graphs $\{\mathcal{G}_n\}$ converges to the graphon $\mathbf{W}$ in the sense of homomorphism density if, for any motif $\mathcal{F}$, it holds that

$$\lim_{n\to\infty} t(\mathcal{F}, \mathcal{G}_n) = t(\mathcal{F}, \mathbf{W}).$$

In the sense of homomorphism density, every graphon is a limit object of some convergent graph sequence, and conversely, every convergent graph sequence converges to a unique graphon [Lovász, 2012]. Thus, a graphon represents a family of graphs that approximate a same underlying structure, even if their sizes differ. By categorizing graphs into such "graphon families", graphons allow for easier and more structured analysis of graph sequences, providing a robust framework for studying large-scale networks.

In the following, we review the relation of convergence in homomorphism density and cut norm. The *cut norm* of a graphon $\mathbf{W}$ is defined by

$$\|\mathbf{W}\|_{\square} := \sup_{S,S' \subseteq I} \left| \int_{S \times S'} \mathbf{W}(x,y)\, dx\, dy \right|, \tag{15}$$

where the supremum is taken over all subsets S and S' of I. The cut norm measures the maximum discrepancy in the graphon over any pair of subsets. Let $\mathcal{G}$ be a graph with adjacency matrix $\boldsymbol{W}_{\mathcal{G}} \in \mathbb{R}^{|V(\mathcal{G})| \times |V(\mathcal{G})|}$. We recall that the induced graphon $\mathbf{W}_{\mathcal{G}}$ is defined by

$$\mathbf{W}_{\mathcal{G}}(u,v) := \sum_{i,j \in [|V(\mathcal{G})|]} [\boldsymbol{W}_{\mathcal{G}}]_{ij} \chi_{I_i}(u) \chi_{I_j}(v), \quad u,v \in I. \tag{16}$$

The following result from Lovász [2012] states that the convergence of graphs in terms of homomorphism density implies convergence in the cut norm of induced graphons, up to some permutations.

Lemma A.1. *Let $\{\mathcal{G}_n\}$ be a sequence of graphs with adjacency matrices $\{\boldsymbol{W}_{\mathcal{G}_n}\}$. Suppose that $\{\mathcal{G}_n\}$ converges to a graphon $\mathbf{W}$ in the sense of homomorphism density. Then, there exists a sequence $\{\pi_n\}$ of permutations such that* $\lim_{n\to\infty} \|\mathbf{W}_{\pi_n(\mathcal{G}_n)} - \mathbf{W}\|_{\square} = 0$.

Given a sequence $\{\mathcal{G}_n\}$ of graphs converging to a graphon $\mathbf{W}$ in the sense of homomorphism density, we introduce a set of the permutation sequences $\{\pi_n\}$ such that the permuted induced graphons $\mathbf{W}_{\pi_n(\mathcal{G}_n)}$ converge under the cut norm to the graphon $\mathbf{W}$, that is,

$$\mathfrak{P} := \left\{ \{\pi_n\} : \lim_{n\to\infty} \|\mathbf{W}_{\pi_n(\mathcal{G}_n)} - \mathbf{W}\|_{\square} = 0 \right\}. \tag{17}$$

It is clear that the set $\mathfrak{P}$ is not empty due to Lemma A.1.

To formulate the definition of graph-feature pairs converging to graphon-feature pair, we need the convergence of induced graphon feature functions. For a graph $\mathcal{G}$ with node feature matrix $\boldsymbol{Z}_{\mathcal{G}} \in \mathbb{R}^{|V(\mathcal{G})| \times F}$, we recall that the induced graphon feature function $\mathbf{Z}_{\mathcal{G}} : I \to \mathbb{R}^{1\times F}$ is defined by

$$\mathbf{Z}_{\mathcal{G}}(u) := \sum_{i \in [|V(\mathcal{G})|]} [\boldsymbol{Z}_{\mathcal{G}}]_{i,:} \chi_{I_i}(u),\ u \in I. \tag{18}$$

We adopt the following definition of graph-feature pairs converging to a graphon-feature pair, introduced in Ruiz et al. [2021a].

Definition 1. *Let $\{\mathcal{G}_n\}$ be a sequence of graphs with adjacency matrices $\{\boldsymbol{W}_{\mathcal{G}_n}\}$ and graph node feature matrices $\{\boldsymbol{Z}_{\mathcal{G}_n}\}$. Suppose that $\{\mathcal{G}_n\}$ converges to a graphon $\mathbf{W}$ in the sense of homomorphism density. Let $\mathbf{Z} \in L^2(I;\mathbb{R}^{1\times F})$ be a graphon feature function. We say that $\{(\mathcal{G}_n, \boldsymbol{Z}_{\mathcal{G}_n})\}$ converges to $(\mathbf{W},\mathbf{Z})$ if there exists a sequence of permutations $\{\pi_n\} \in \mathfrak{P}$ such that $\lim_{n\to\infty} \|\mathbf{Z}_{\pi_n(\mathcal{G}_n)} - \mathbf{Z}\|_{L^2(I;\mathbb{R}^{1\times F})} = 0$, where the set $\mathfrak{P}$ is defined by* (17).

A.2 Function Spaces

The function space $L^p(I;\mathbb{R}^{1\times F})$ consists of all L^p-integrable vector valued functions mapping I to $\mathbb{R}^{1\times F}$, where $p \in [1,\infty]$ and F denotes the number of features. The norm in $L^p(I;\mathbb{R}^{1\times F})$ is defined by $\|\mathbf{Z}\|_{L^p(I;\mathbb{R}^{1\times F})} := (\sum_{f\in[F]} \|Z_f\|^2_{L^p(I)})^{1/2}$ for $\mathbf{Z} = [Z_f : f \in [F]]$. By $\int_I \mathbf{Z}(u)du$ we denote the entry-wise integral $\left[\int_I Z_f(u)du : f \in [F]\right]$. Let Ω be a subset of $\mathbb{R}^+$ and $p \in [1,\infty]$, the Banach space $C(\Omega; L^p(I;\mathbb{R}^{1\times F}))$ is composed of vector-valued functions $\mathbf{X} = [X_f : f \in [F]] : I \times \Omega \to \mathbb{R}^{1\times F}$ satisfying that for each $t \in \Omega$, $\mathbf{X}(\cdot,t) \in L^p(I;\mathbb{R}^{1\times F})$; for each $u \in I$ and $f \in [F]$, $X_f(u,\cdot)$ is continuous on Ω; and with finite norm $\|\mathbf{X}\|_{C(\Omega;L^p(I;\mathbb{R}^{1\times F}))} := \sup_{t\in\Omega} \|\mathbf{X}(\cdot,t)\|_{L^p(I;\mathbb{R}^{1\times F})}$. By $C^1(\Omega; L^p(I;\mathbb{R}^{1\times F}))$ we denote a subspace of $C(\Omega; L^p(I;\mathbb{R}^{1\times F}))$, in which the vector-valued function $\mathbf{X} = [X_f : f \in [F]]$ satisfies that for each $f \in [F]$ and $u \in I$, $X_f(u,\cdot)$ is continuously differentiable.

A.3 Proof of Theorem 3.1

Prior to the detailed proof of Theorem 3.1, we present several useful observations. Under the assumption AS0, for $T > 0$, we define a constant

$$h_T := \sup_{t\in[0,T]} \max_{f,g\in[F],\ell\in[L],k\in\mathbb{Z}_K} \left| \boldsymbol{h}_{fgk}^{(\ell,t)} \right|. \tag{19}$$

Lemma A.2. *Let $T > 0$ and $\mathbf{X} \in C([0,T]; L^\infty(I;\mathbb{R}^{1\times F}))$. Suppose that AS0 and AS1 hold. Then, for $p \in [1,\infty]$, $\ell \in [L]$ and $t \in [0,T]$, it holds that*

$$\left\| \mathbf{X}^{(\ell,t)} \right\|_{L^p(I;\mathbb{R}^{1\times F})} \le FKh_T \left\| \mathbf{X}^{(\ell-1,t)} \right\|_{L^p(I;\mathbb{R}^{1\times F})},$$

where h_T is defined in (19)*.*

Proof. Note that the updating rule of Graphon-NN gives

$$\mathbf{X}_f^{(\ell,t)} = \sigma\left(\sum_{g=1}^{F} \sum_{k=0}^{K-1} \boldsymbol{h}_{fgk}^{(\ell,t)} T_{\mathbf{W}}^k \mathbf{X}_g^{(\ell-1,t)} \right), \quad f \in [F], \ell \in [L], t \in [0,T].$$

It follows that

$$\left\| \mathbf{X}_f^{(\ell,t)} \right\|_{L^p(I)} \le h_T \left(\sum_{k=0}^{K-1} \|T_{\mathbf{W}}\|_{L^p(I)\to L^p(I)}^k \right) \left\| \sum_{g=1}^{F} \mathbf{X}_g^{(\ell-1,t)} \right\|_{L^p(I)} \le h_T K\sqrt{F} \left\| \mathbf{X}^{(\ell-1,t)} \right\|_{L^p(I;\mathbb{R}^{1\times F})},$$

in which the first inequality is due to AS0, AS1 and triangle inequality; the second is according to the fact of $\|T_{\mathbf{W}}\|_{L^p(I)\to L^p(I)} \le \|\mathbf{W}\|_{L^\infty(I^2)} \le 1$ and the norm defined in $L^p(I;\mathbb{R}^{1\times F})$. The desired result immediately follows by rewriting the norm of $\mathbf{X}^{(\ell,t)}$. ☐

Proposition A.3. *Suppose that AS0 and AS1 hold. Let $T > 0$ and $\mathbf{X}, \widetilde{\mathbf{X}} \in C([0,T]; L^\infty(I;\mathbb{R}^{1\times F}))$. Then for all $t \in [0,T]$, it holds that*

$$\left\| \Phi(\mathbf{W}; \mathbf{X}(\cdot,t); \mathbf{H}(t)) - \Phi(\mathbf{W}; \widetilde{\mathbf{X}}(\cdot,t); \mathbf{H}(t)) \right\|_{L^\infty(I;\mathbb{R}^{1\times F})} \le (FKh_T)^L \left\| \mathbf{X}(\cdot,t) - \widetilde{\mathbf{X}}(\cdot,t) \right\|_{L^\infty(I;\mathbb{R}^{1\times F})}.$$

Proof. According to the normalized Lipschitz continuity of activation function σ, similarly to the proof of Lemma A.2 with $p = \infty$, we have

$$\left\| \mathbf{X}^{(\ell,t)} - \widetilde{\mathbf{X}}^{(\ell,t)} \right\|_{L^\infty(I;\mathbb{R}^{1\times F})} \le FKh_T \left\| \mathbf{X}^{(\ell-1,t)} - \widetilde{\mathbf{X}}^{(\ell-1,t)} \right\|_{L^\infty(I;\mathbb{R}^{1\times F})}. \tag{20}$$

Recall the notations $\mathbf{X}(\cdot,t) = \mathbf{X}^{(0,t)}$, $\Phi(\mathbf{W}; \mathbf{X}(\cdot,t); \mathbf{H}(t)) = \mathbf{X}^{(L,t)}$ (similar for $\widetilde{\mathbf{X}}$). The desired result follows from recursively applying (20). ☐

Proof of Theorem 3.1. The proof is based on the Banach contraction mapping principle. Let $T > 0$ be arbitrary but fixed, and $0 < \tau < \frac{1}{2(FKh_T)^L}$. We define a subspace $\mathcal{S}_{\mathbf{Z}}$ of $C([0,\tau]; L^\infty(I;\mathbb{R}^{1\times F}))$, associated with τ, by

$$\mathcal{S}_{\mathbf{Z}} := \left\{ \mathbf{X} : \mathbf{X} \in C([0,\tau]; L^\infty(I;\mathbb{R}^{1\times F})), \mathbf{X}(\cdot,0) = \mathbf{Z} \right\}.$$

Moreover, we define an integral operator $\mathcal{K} : \mathcal{S}_{\mathbf{Z}} \to \mathcal{S}_{\mathbf{Z}}$ by

$$[\mathcal{K}\mathbf{X}](u,t) := \mathbf{Z}(u) + \int_0^t \Phi(\mathbf{W}; \mathbf{X}(u,s); \mathbf{H}(s)) ds. \tag{21}$$

It follows that we can rewrite the initial value problem (2) as the fixed point equation $\mathbf{X} = \mathcal{K}\mathbf{X}$. We show below that the operator $\mathcal{K}$ is a contraction. For any $\mathbf{X}, \widetilde{\mathbf{X}} \in \mathcal{S}_{\mathbf{Z}}$, according to the definition of norm in $C([0,\tau]; L^\infty(I;\mathbb{R}^{1\times F}))$, we have

$$\begin{aligned}
\|\mathcal{K}\mathbf{X} - \mathcal{K}\widetilde{\mathbf{X}}\|_{\mathcal{S}_{\mathbf{Z}}} &= \sup_{t\in[0,\tau]} \|\mathcal{K}\mathbf{X}(\cdot,t) - \mathcal{K}\widetilde{\mathbf{X}}(\cdot,t)\|_{L^\infty(I;\mathbb{R}^{1\times F})} \\
&= \sup_{t\in[0,\tau]} \left\| \int_0^t \Phi(\mathbf{W}; \mathbf{X}(\cdot,s); \mathbf{H}(s)) - \Phi(\mathbf{W}; \widetilde{\mathbf{X}}(\cdot,s); \mathbf{H}(s)) ds \right\|_{L^\infty(I;\mathbb{R}^{1\times F})} \\
&\le \tau \sup_{t\in[0,\tau]} \left\| \Phi(\mathbf{W}; \mathbf{X}(\cdot,t); \mathbf{H}(t)) - \Phi(\mathbf{W}; \widetilde{\mathbf{X}}(\cdot,t); \mathbf{H}(t)) \right\|_{L^\infty(I;\mathbb{R}^{1\times F})}.
\end{aligned} \tag{22}$$

It follows from Lemma A.3 that

$$\left\|\Phi(\mathbf{W};\mathbf{X}(\cdot,t);\mathbf{H}(t))-\Phi(\mathbf{W};\widetilde{\mathbf{X}}(\cdot,t);\mathbf{H}(t))\right\|_{L^\infty(I;\mathbb{R}^{1\times F})}\le(FKh_T)^L\|\mathbf{X}(\cdot,t)-\widetilde{\mathbf{X}}(\cdot,t)\|_{L^\infty(I;\mathbb{R}^{1\times F})}.$$

By substituting the above estimate into (22), we obtain that

$$\begin{aligned}\|\mathcal{K}\mathbf{X}-\mathcal{K}\widetilde{\mathbf{X}}\|_{\mathcal{S}_\mathbf{Z}}&\le\tau(FKh_T)^L\sup_{t\in[0,\tau]}\|\mathbf{X}(\cdot,t)-\widetilde{\mathbf{X}}(\cdot,t)\|_{L^\infty(I;\mathbb{R}^{1\times F})}\\&=\tau(FKh_T)^L\|\mathbf{X}-\widetilde{\mathbf{X}}\|_{\mathcal{S}_\mathbf{Z}}\le\frac{1}{2}\|\mathbf{X}-\widetilde{\mathbf{X}}\|_{\mathcal{S}_\mathbf{Z}}\end{aligned}$$

where the last inequality follows from the definition of τ. Therefore, the operator $\mathcal{K}$ is a contraction. By the Banach contraction mapping principle, there exists a unique solution $\widehat{\mathbf{X}}\in\mathcal{S}_\mathbf{Z}$ of the initial value problem (2). Taking $\widehat{\mathbf{X}}(\tau)$ as the initial condition, we repeat the argument to extend the solution to $[0,2\tau]$. In such a way, we can keep doing until the solution extends to $[0,T]$, and get a unique solution $\mathbf{X}\in C([0,T];L^\infty(I;\mathbb{R}^{1\times F}))$. According to AS0 and AS1, it follows that $\Phi(\mathbf{W};\mathbf{X}(u,\cdot);\mathbf{H}(\cdot))$ is continuous, that is, the integrand in (21) is continuous. Therefore, by fundamental theorem of calculus, we see that $\mathcal{K}\mathbf{X}$ is continuously differentiable about the second variable t. As $\mathcal{K}\mathbf{X}=\mathbf{X}$, we conclude that $\mathbf{X}\in C^1([0,T];L^\infty(I;\mathbb{R}^{1\times F}))$. This completes the proof. □

A.4 Stability Analysis of Graphon-NDEs

To lay a foundation for the subsequent proofs of the convergence result (Theorem 3.2) and also the convergence rate results (Theorems 3.3 and 3.4), this section focuses on the stability analysis of Graphon-NDEs. We proceed with several technical lemmas.

Lemma A.4. *Let T_1 and T_2 be two bounded linear operators on $L^2(I)$. Let k be a given positive integer. If $\|T_1\|_{L^2(I)\to L^2(I)}\le 1$ and $\|T_2\|_{L^2(I)\to L^2(I)}\le 1$, then $\left\|T_1^k-T_2^k\right\|_{L^2(I)\to L^2(I)}\le k\|T_1-T_2\|_{L^2(I)\to L^2(I)}$.*

Lemma A.5 (Stability of Graphon-NNs)**.** *Let $T>0$, $\mathbf{X},\widetilde{\mathbf{X}}\in C([0,T];L^\infty(I;\mathbb{R}^{1\times F}))$, and graphons $\mathbf{W},\widetilde{\mathbf{W}}$. If AS0 and AS1 hold, then for any $t\in[0,T]$, it holds that*

$$\begin{aligned}&\left\|\Phi\left(\widetilde{\mathbf{W}};\widetilde{\mathbf{X}}(\cdot,t);\mathbf{H}(t)\right)-\Phi\left(\mathbf{W};\mathbf{X}(\cdot,t);\mathbf{H}(t)\right)\right\|_{L^2(I;\mathbb{R}^{1\times F})}\\&\le(FKh_T)^L\left(\left\|\widetilde{\mathbf{X}}(\cdot,t)-\mathbf{X}(\cdot,t)\right\|_{L^2(I;\mathbb{R}^{1\times F})}+LK\left\|T_{\widetilde{\mathbf{W}}}-T_\mathbf{W}\right\|_{L^2(I)\to L^2(I)}\|\mathbf{X}\|_{C([0,T];L^2(I;\mathbb{R}^{1\times F}))}\right).\end{aligned}$$

Proof. Recall that for $f\in[F],\ell\in[L],t\in[0,T]$, the updating rule of Graphon-NN gives

$$\mathbf{X}_f^{(\ell,t)}=\sigma\left(\sum_{g=1}^F\sum_{k=0}^{K-1}\boldsymbol{h}_{fgk}^{(\ell,t)}T_\mathbf{W}^k\mathbf{X}_g^{(\ell-1,t)}\right),\quad\widetilde{\mathbf{X}}_f^{(\ell,t)}=\sigma\left(\sum_{g=1}^F\sum_{k=0}^{K-1}\boldsymbol{h}_{fgk}^{(\ell,t)}T_{\widetilde{\mathbf{W}}}^k\widetilde{\mathbf{X}}_g^{(\ell-1,t)}\right).$$

Then by the triangle inequality and similar argument as in the proof of Lemma A.2, we obtain

$$\begin{aligned}\left\|\widetilde{\mathbf{X}}_f^{(\ell,t)}-\mathbf{X}_f^{(\ell,t)}\right\|_{L^2(I)}\le&\sqrt{F}Kh_T\left\|\widetilde{\mathbf{X}}^{(\ell-1,t)}-\mathbf{X}^{(\ell-1,t)}\right\|_{L^2(I;\mathbb{R}^{1\times F})}\\&+\sqrt{F}h_T\left(\sum_{k=0}^{K-1}\left\|T_{\widetilde{\mathbf{W}}}^k-T_\mathbf{W}^k\right\|_{L^2(I)\to L^2(I)}\right)\left\|\mathbf{X}^{(\ell-1,t)}\right\|_{L^2(I;\mathbb{R}^{1\times F})}.\end{aligned}$$

It follows from Lemma A.4 that

$$\sum_{k=0}^{K-1}\left\|T_{\widetilde{\mathbf{W}}}^k-T_\mathbf{W}^k\right\|_{L^2(I)\to L^2(I)}\le K^2\left\|T_{\widetilde{\mathbf{W}}}-T_\mathbf{W}\right\|_{L^2(I)\to L^2(I)}.$$

Therefore,

$$\begin{aligned}\left\|\widetilde{\mathbf{X}}^{(\ell,t)}-\mathbf{X}^{(\ell,t)}\right\|_{L^2(I;\mathbb{R}^{1\times F})}\le&FKh_T\left\|\widetilde{\mathbf{X}}^{(\ell-1,t)}-\mathbf{X}^{(\ell-1,t)}\right\|_{L^2(I;\mathbb{R}^{1\times F})}\\&+FK^2h_T\left\|T_{\widetilde{\mathbf{W}}}-T_\mathbf{W}\right\|_{L^2(I)\to L^2(I)}\left\|\mathbf{X}^{(\ell-1,t)}\right\|_{L^2(I;\mathbb{R}^{1\times F})}.\end{aligned}$$

Then a recursion argument gives

$$\left\|\widetilde{\mathbf{X}}^{(L,t)} - \mathbf{X}^{(L,t)}\right\|_{L^2(I;\mathbb{R}^{1\times F})} \leq (FKh_T)^L \left\|\widetilde{\mathbf{X}}^{(0,t)} - \mathbf{X}^{(0,t)}\right\|_{L^2(I;\mathbb{R}^{1\times F})} + FK^2 h_T \left\|T_{\widetilde{\mathbf{W}}} - T_{\mathbf{W}}\right\|_{L^2(I)\to L^2(I)} \sum_{\ell=0}^{L-1} (FKh_T)^{L-1-\ell} \left\|\mathbf{X}^{(\ell,t)}\right\|_{L^2(I;\mathbb{R}^{1\times F})}.$$

Note that by Lemma A.2, we have $\left\|\mathbf{X}^{(\ell,t)}\right\|_{L^2(I;\mathbb{R}^{1\times F})} \leq (FKh_T)^\ell \left\|\mathbf{X}^{(0,t)}\right\|_{L^2(I;\mathbb{R}^{1\times F})}$. Hence,

$$\left\|\widetilde{\mathbf{X}}^{(L,t)} - \mathbf{X}^{(L,t)}\right\|_{L^2(I;\mathbb{R}^{1\times F})} \leq (FKh_T)^L \left\|\widetilde{\mathbf{X}}^{(0,t)} - \mathbf{X}^{(0,t)}\right\|_{L^2(I;\mathbb{R}^{1\times F})} + LK (FKh_T)^L \left\|T_{\widetilde{\mathbf{W}}} - T_{\mathbf{W}}\right\|_{L^2(I)\to L^2(I)} \left\|\mathbf{X}^{(0,t)}\right\|_{L^2(I;\mathbb{R}^{1\times F})}.$$

Note that $\mathbf{X}^{(0,t)} = \mathbf{X}(\cdot,t)$, $\mathbf{X}^{(L,t)} = \Phi(\mathbf{W};\mathbf{X}(\cdot,t);\mathbf{H}(t))$ (similar for $\widetilde{\mathbf{X}}$) and norm $\|\mathbf{X}\|_{C([0,T];L^2(\mathbb{R}^{1\times F}))}$ is defined as the supremum of $\|\mathbf{X}(\cdot,t)\|_{L^2(I;\mathbb{R}^{1\times F})}$ about $t\in[0,T]$. Therefore, the above inequality implies the desired result. $\square$

The following result is a special case of Perov [1959] (also see Theorem 21 in Dragomir [2003]).

Lemma A.6 (Generalized Grönwall's inequality). *Let a, b and c be non-negative constants. Let $u(t)$ be a non-negative function that satisfies the integral inequality $u(t) \leq c + \int_0^t \left(au(s) + bu^{\frac{1}{2}}(s)\right) ds$, then we have $u(t) \leq \left(c^{\frac{1}{2}}\exp(at/2) + \frac{\exp(at/2)-1}{a}b\right)^2$.*

Now given a sequence of graphons $\{\mathbf{W}_n\}$ and (bounded) input feature functions $\{\mathbf{Z}_n\}$, we consider the following Graphon-NDEs

$$\begin{aligned}\frac{\partial}{\partial t}\mathbf{X}_n(u,t) &= \Phi(\mathbf{W}_n;\mathbf{X}_n(u,t);\mathbf{H}(t)),\\ \mathbf{X}_n(u,0) &= \mathbf{Z}_n(u).\end{aligned} \tag{23}$$

We note that Theorem 3.1 guarantees the existence and uniqueness of the solution $\mathbf{X}_n$ of (23). We establish in the following that the error between solutions of (2) and (23) is bounded above by a linear combination of the initial feature error and graphon error.

Theorem A.7 (Stability of Graphon-NDEs). *Suppose that AS0 and AS1 hold. Let $\mathbf{X}$ and $\mathbf{X}_n$ denote the solutions of* (2) *and* (23)*, respectively. Then it holds that*

$$\|\mathbf{X}_n - \mathbf{X}\|_{C([0,T];L^2(I;\mathbb{R}^{1\times F}))} \leq P\|\mathbf{Z}_n - \mathbf{Z}\|_{L^2(I;\mathbb{R}^{1\times F})} + Q\|T_{\mathbf{W}_n} - T_{\mathbf{W}}\|_{L^2(I)\to L^2(I)}, \tag{24}$$

where

$$P := \exp\left(T\,(FKh_T)^L\right), \quad Q := (P-1)LK\,\|\mathbf{X}\|_{C([0,T];L^2(I;\mathbb{R}^{1\times F}))}. \tag{25}$$

Proof. Denote $\Delta = \mathbf{X}_n - \mathbf{X}$. Taking the difference between (23) and (2), we have

$$\begin{aligned}\frac{\partial}{\partial t}\Delta(u,t) &= \Phi(\mathbf{W}_n;\mathbf{X}_n(u,t);\mathbf{H}(t)) - \Phi(\mathbf{W};\mathbf{X}(u,t);\mathbf{H}(t)),\\ \Delta(u,0) &= \mathbf{Z}_n(u) - \mathbf{Z}(u).\end{aligned}$$

It follows that

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|\Delta(\cdot,t)\|^2_{L^2(I;\mathbb{R}^{1\times F})} &= \left|\int_I \frac{\partial\Delta(u,t)}{\partial t}\,(\Delta(u,t))^\top du\right|\\ &= \left|\int_I \left(\Phi\left(\mathbf{W}_n;\mathbf{X}_n(u,t);\mathbf{H}(t)\right) - \Phi\left(\mathbf{W};\mathbf{X}(u,t);\mathbf{H}(t)\right)\right)(\Delta(u,t))^\top du\right|\\ &\leq \|\Phi(\mathbf{W}_n;\mathbf{X}_n(\cdot,t);\mathbf{H}(t)) - \Phi(\mathbf{W};\mathbf{X}(\cdot,t);\mathbf{H}(t))\|_{L^2(I;\mathbb{R}^{1\times F})}\,\|\Delta(\cdot,t)\|_{L^2(I;\mathbb{R}^{1\times F})}.\end{aligned}$$

According to Lemma A.5, we have

$$\begin{aligned}&\|\Phi(\mathbf{W}_n;\mathbf{X}_n(\cdot,t);\mathbf{H}(t))-\Phi(\mathbf{W};\mathbf{X}(\cdot,t);\mathbf{H}(t))\|_{L^2(I;\mathbb{R}^{1\times F})}\\&\leq \underbrace{(FKh_T)^L}_{\text{denoted by } a/2}\|\Delta(\cdot,t)\|_{L^2(I;\mathbb{R}^{1\times F})}+\underbrace{LK\,(FKh_T)^L\,\|T_{\mathbf{W}_n}-T_{\mathbf{W}}\|_{L^2(I)\to L^2(I)}\,\|\mathbf{X}\|_{C([0,T];L^2(I;\mathbb{R}^{1\times F}))}}_{\text{denoted by } b/2}.\end{aligned}$$

Let $\delta(t):=\|\Delta(\cdot,t)\|^2_{L^2(I;\mathbb{R}^{1\times F})}$. Then the above estimates lead to

$$\begin{aligned}\frac{d}{dt}\delta(t)&\leq a\delta(t)+b\sqrt{\delta(t)},\\ \delta(0)&=\|\mathbf{Z}_n-\mathbf{Z}\|^2_{L^2(I;\mathbb{R}^{1\times F})}.\end{aligned}$$

Let $s\in[0,T]$ be arbitrary but fixed. We integrate above $[0,s]$ about the variable t, and get

$$\delta(s)\leq\delta(0)+\int_0^s\left(a\delta(t)+b\sqrt{\delta(t)}\right)dt.$$

We then apply the generalized Grönwall's inequality (Lemma A.6), and get

$$\delta(s)\leq\left(\sqrt{\delta(0)}\exp(as/2)+\frac{\exp(as/2)-1}{a}b\right)^2.$$

By noting $s\leq T$ and plugging definitions of a, b and δ into the above inequality, we obtain

$$\|\Delta(\cdot,s)\|_{L^2(I;\mathbb{R}^{1\times F})}\leq P\,\|\mathbf{Z}_n-\mathbf{Z}\|_{L^2(I;\mathbb{R}^{1\times F})}+Q\,\|T_{\mathbf{W}_n}-T_{\mathbf{W}}\|_{L^2(I)\to L^2(I)},$$

with P and Q defined in (25). Since s is arbitrary in $[0,T]$, we take the supremum about s over $[0,T]$ for the above inequality, and immediately get (24) by recalling the norm defined in $C([0,T];L^2(I;\mathbb{R}^{1\times F}))$. □

A.5 Proof of Theorem 3.2

Proof of Theorem 3.2. By the assumption of $\{(\mathcal{G}_n,\boldsymbol{Z}_{\mathcal{G}_n})\}$ converging to $(\mathbf{W},\mathbf{Z})$ in the sense of Definition 1, there exists a sequence $\{\pi_n\}$ of permutations such that

$$\lim_{n\to\infty}\|\mathbf{W}_{\pi_n(\mathcal{G}_n)}-\mathbf{W}\|_{\square}=0,\quad \lim_{n\to\infty}\|\mathbf{Z}_{\pi_n(\mathcal{G}_n)}-\mathbf{Z}\|_{L^2(I;\mathbb{R}^{1\times F})}=0. \tag{26}$$

We denote $\mathbf{W}_n:=\mathbf{W}_{\pi_n(\mathcal{G}_n)}$ and $\mathbf{Z}_n:=\mathbf{Z}_{\pi_n(\mathcal{G}_n)}$. It is known (Lemma E.6. in Janson [2010]) that $\lim_{n\to\infty}\|\mathbf{W}_n-\mathbf{W}\|_{\square}=0$ if and only if $\lim_{n\to\infty}\|T_{\mathbf{W}_n}-T_{\mathbf{W}}\|_{L^2(I)\to L^2(I)}=0$. Therefore, (26) implies

$$\lim_{n\to\infty}\|T_{\mathbf{W}_n}-T_{\mathbf{W}}\|_{L^2(I)\to L^2(I)}=0,\quad \lim_{n\to\infty}\|\mathbf{Z}_n-\mathbf{Z}\|_{L^2(I;\mathbb{R}^{1\times F})}=0. \tag{27}$$

Then the desired result immediately follows from Theorem A.7. □

A.6 Proof of Theorems 3.3 and 3.4

Proof of Theorem 3.3. Recall that $u_i:=(i-1)/n$, $I_i:=[u_i,u_{i+1})$, for each $i\in[n]$. According to definition $\mathbf{W}_n$ of (16) with (4), we have

$$\|\mathbf{W}-\mathbf{W}_n\|^2_{L^2(I^2)}=\sum_{i,j\in[n]}\int_{I_i\times I_j}|\mathbf{W}(u,v)-\mathbf{W}(u_i,u_j)|^2\,dudv.$$

According to AS2, we obtain that

$$\|\mathbf{W}-\mathbf{W}_n\|^2_{L^2(I^2)}\leq A_1^2\sum_{i,j\in[n]}\int_{I_i\times I_j}(|u-u_i|+|v-u_j|)^2\,dudv. \tag{28}$$

For each $i,j\in[n]$, direct computation gives $\int_{I_i\times I_j}(|u-u_i|+|v-u_j|)^2\,dudv=\frac{7}{6n^4}$, which combining with (28) gives

$$\|\mathbf{W}-\mathbf{W}_n\|^2_{L^2(I^2)}\leq A_1^2\frac{7}{6n^2}. \tag{29}$$

Denote $\mathbf{Z} = [Z_f : f \in [F]]$ and $\mathbf{Z}_n = [(Z_n)_f : f \in [F]]$. According to definition $\mathbf{Z}_n$ of (18) with (5), we have

$$\|\mathbf{Z} - \mathbf{Z}_n\|^2_{L^2(I;\mathbb{R}^{1\times F})} = \sum_{f\in[F]} \|Z_f - (Z_n)_f\|^2_{L^2(I)} = \sum_{f\in[F]}\sum_{j\in[n]} \int_{I_j} |Z_f(u) - Z_f(u_j)|^2 du. \quad (30)$$

It follows from AS3 that for each $f \in [F]$ and $j \in [n]$,

$$\int_{I_j} |Z_f(u) - Z_f(u_j)|^2 du \leq A_2^2 \int_{I_j} (u - u_j)^2 du = \frac{A_2^2}{3}\frac{1}{n^3}.$$

Therefore, from (30), we get

$$\|\mathbf{Z} - \mathbf{Z}_n\|^2_{L^2(I;\mathbb{R}^{1\times F})} \leq \frac{A_2^2 F}{3}\frac{1}{n^2}. \quad (31)$$

Recall we have established in Theorem A.7 that

$$\|\mathbf{X}_n - \mathbf{X}\|_{C([0,T];L^2(I;\mathbb{R}^{1\times F}))} \leq P\|\mathbf{Z}_n - \mathbf{Z}\|_{L^2(I;\mathbb{R}^{1\times F})} + Q\|T_{\mathbf{W}_n} - T_{\mathbf{W}}\|_{L^2(I)\to L^2(I)},$$

which combining with estimates (29) and (31) and the fact of

$$\|T_{\mathbf{W}_n} - T_{\mathbf{W}}\|_{L^2(I)\to L^2(I)} \leq \|\mathbf{W}_n - \mathbf{W}\|_{L^2(I^2)},$$

further implies

$$\|\mathbf{X}_n - \mathbf{X}\|_{C([0,T];L^2(I;\mathbb{R}^{1\times F}))} \leq PA_2\sqrt{\frac{F}{3}}\frac{1}{n} + QA_1\sqrt{\frac{7}{6}}\frac{1}{n} \leq \frac{C}{n},$$

where C is defined by

$$C := \exp\left(T\left(FKh_T\right)^L\right)\left(A_2\sqrt{\frac{F}{3}} + LK\,\|\mathbf{X}\|_{C([0,T];L^2(I;\mathbb{R}^{1\times F}))}\,A_1\sqrt{\frac{7}{6}}\right). \quad (32)$$

This completes the proof of (6). The estimate (7) can be immediately obtained from (6) and the triangle inequality. $\square$

Lemma A.8. *Suppose that $\Omega \subset \mathbb{R}^d$ and $f \in L^2(\Omega)$. Let $|\Omega|$ be the volume of Ω. Then the constant function $h(u) := \frac{1}{|\Omega|}\int_\Omega f(u)du$, $u \in \Omega$, is the best constant approximation of f, i.e.,* $\inf\{\|f - c\|_{L^2(\Omega)} : c \in \mathbb{R}\} = \|f - h\|_{L^2(\Omega)}$.

Proof of Theorem 3.4. We begin with estimating $\|\mathbf{W} - \mathbf{W}_n\|_{L^2(I^2)}$. Recall that $\mathcal{N}_\delta(\partial\mathbf{W}^+)$ denotes the number of δ-mesh cubes that intersect $\partial\mathbf{W}^+$. We set $\delta = 1/n$. Recall that $\mathbf{W}_n$ is defined by (16) with adjacency matrix generated by (8). It follows that

$$\|\mathbf{W} - \mathbf{W}_n\|^2_{L^2(I^2)} = \int_I |\mathbf{W}(u,v) - \mathbf{W}_n(u,v)|^2 dudv \leq \mathcal{N}_{1/n}(\partial\mathbf{W}^+)\frac{1}{n^2}. \quad (33)$$

According to definition (10) of upper box-counting dimension, for any $\epsilon \in (0, 2-b)$, there exists $N_{\epsilon,\mathbf{W}} \in \mathbb{N}$ such that when $n > N_{\epsilon,\mathbf{W}}$, $\frac{\log \mathcal{N}_{1/n}(\partial\mathbf{W}^+)}{-\log(1/n)} < b + \epsilon$. Therefore, $\mathcal{N}_{1/n}(\partial\mathbf{W}^+) \leq n^{b+\epsilon}$, which combining with (33) yields

$$\|\mathbf{W} - \mathbf{W}_n\|_{L^2(I^2)} \leq n^{-(1-\frac{b+\epsilon}{2})}. \quad (34)$$

We next estimate $\|\mathbf{Z} - \mathbf{Z}_n\|_{L^2(I;\mathbb{R}^{1\times F})}$. Recall that $\mathbf{Z}_n$ is the induced graphon feature function associated with the graph feature matrix generated in the way of (9). Let $\mathbf{Z}'_n$ be the induced graphon feature function associated with the graph feature matrix generated in the way of (5). It has been shown in the proof of Theorem 3.3 that, with assumption AS3, $\|\mathbf{Z} - \mathbf{Z}'_n\|_{L^2(I;\mathbb{R}^{1\times F})} \leq A_2\sqrt{\frac{F}{3}}\frac{1}{n}$. According to Lemma A.8, we know that $\|\mathbf{Z} - \mathbf{Z}_n\|_{L^2(I;\mathbb{R}^{1\times F})} \leq \|\mathbf{Z} - \mathbf{Z}'_n\|_{L^2(I;\mathbb{R}^{1\times F})}$. Therefore,

$$\|\mathbf{Z} - \mathbf{Z}_n\|_{L^2(I;\mathbb{R}^{1\times F})} \leq A_2\sqrt{\frac{F}{3}}\frac{1}{n}. \quad (35)$$

With a similar argument in the proof of Theorem 3.3, by Theorem A.7 and estimates (34) and (35), we have

$$\|\mathbf{X}_n - \mathbf{X}\|_{C([0,T];L^2(I;\mathbb{R}^{1\times F}))} \leq PA_2\sqrt{\frac{F}{3}}\frac{1}{n} + Qn^{-(1-\frac{b+\epsilon}{2})} \leq \frac{\widetilde{C}}{n^{1-\frac{b+\epsilon}{2}}},$$

where $\widetilde{C}$ is defined by

$$\widetilde{C} := \exp\left(T\left(FKh_T\right)^L\right)\left(A_2\sqrt{\frac{F}{3}} + LK\|\mathbf{X}\|_{C([0,T];L^2(I;\mathbb{R}^{1\times F}))}\right). \tag{36}$$

This proves (11). The estimate (12) can be obtained from (11) and the triangle inequality.

□

B Supplemental Materials: Synthetic Numerical Experiments

Graphons. We utilize three distinct graphons for experimental verification. To explore the weighted graph convergence behavior detailed in Theorem 3.3, we use the tent graphon

$$\mathbf{W}(u,v) = 1 - |u - v|, \tag{37}$$

which is symmetric, continuous, and Lipschitz on the unit square and thus fulfills our desired conditions. To explore the unweighted graph convergence behavior of Theorem 3.4, we use two $\{0,1\}$-valued graphons. First, to showcase multiscale community structure common in scientific applications, we consider the hierarchical stochastic block model (HSBM) graphon [Holland et al., 1983, Crane and Dempsey, 2015], where the box-counting dimension of the support is 1 with a controllable grid size parameter to increase the overall recursive complexity. Second, we consider the hexaflake fractal, a Sierpiński n-gon–based construction that has been used in certain practical design applications [Choudhury and Matin, 2012], as a graphon with box-counting dimension of $\frac{\log(7)}{\log(3)}$ or about 1.77. We illustrate these graphons in Figure 2.

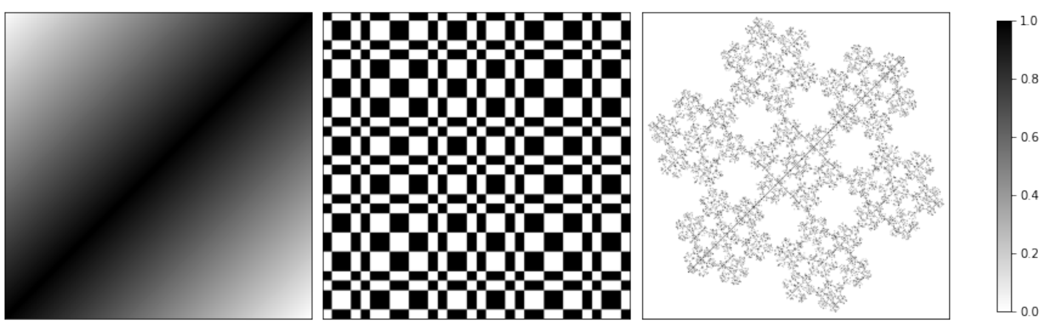

Figure 2: Tent (left), HSBM (center), and Hexaflake (right) graphon visualizations.

Experiment setup. We use GNDEs parameterized with a two-layer GNN, based on the models of [Poli et al., 2019], where both the feature and hidden dimensions are 1, sharing the same *constant filters* with entries bounded in $[-1,1]$. The initial conditions are random Fourier polynomials of degree $D = 10$, defined by $\mathbf{Z}(u) := \sum_{k=1}^{D} a_k \cos(2\pi k u) + b_k \sin(2\pi k u)$, where a_k and b_k are independently sampled from a uniform distribution, creating diverse and smooth signals over graph nodes. The subgraphs and their associated input features are obtained as in Section 3.3.1 for the tent graphon and Section 3.3.2 for the HSBM and hexaflake graphons. We conduct 100 randomly initialized trials, each with random weight initialization for the associated model and random features. We plot mean and standard deviation for the experiment results in Figure 3. All experiments were performed locally on a single Nvidia A4000 GPU. Evaluation is relatively fast, with all experiments completed over the course of a few hours.

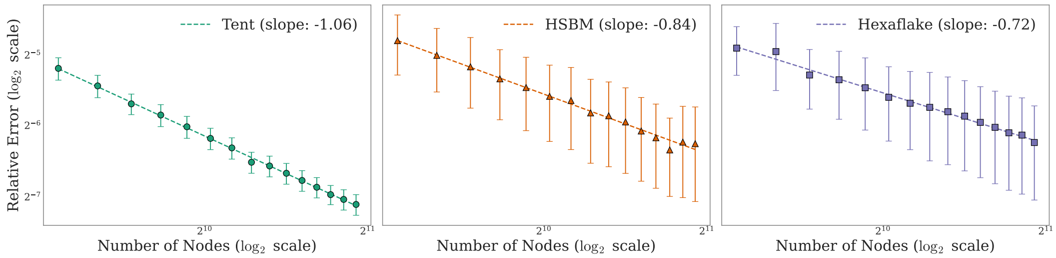

Figure 3: Tent (left), HSBM (center), and Hexaflake (right) graphon convergence with error bars displayed.

Evaluation. To approximate the graphon solution $\mathbf{X}$, we use a reference graph with $N_{\text{largest}} = 5000$ nodes. We present the log-log convergence plot of $\max_t \frac{\|\boldsymbol{X}_n(t) - \boldsymbol{X}_{5000}(t)\|_2}{\|\boldsymbol{X}_{5000}(t)\|_2}$ for number of nodes n ranging from 550 to 1950 with a step size of 100. This approximates the maximal relative error over all $t \in [0, 1]$ of GNDE evolution, though $t = 1$ is the timepoint with maximal error in most runs. We evolve GNDEs through the Dormand-Prince method of order 5 [Dormand and Prince, 1980]. We plot the resulting curves in Figure 1.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper acknowledges its limitations and outlines directions for future work in the section of Limitations and Future Directions.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: The paper states all theoretical assumptions, labeled AS0–AS3. Each main theoretical result is accompanied by a formal and complete proof provided in Appendix A.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: All relevant details for reproducing the exact results are discussed in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation,

it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: Reference the Supplemental Material. Code for reproducing experimental results is provided.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

Justification: While our experiment does not consider training and optimization, relevant parameter choices are presented in Appendix B.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: Additional plots with error bars included are shown in Figure 3.

Guidelines:

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

8. **Experiments compute resources**

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

Answer: [Yes]

Justification: We have included the details in Appendix B.

Guidelines:

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

9. **Code of ethics**

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

Answer: [Yes]

Justification: This research conforms with the NeurIPS Code of Ethics.

Guidelines:

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

10. **Broader impacts**

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

    Answer: [NA]

    Justification: The paper presents foundational theoretical results without direct application or deployment, and thus does not raise any foreseeable societal impact.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [NA]

    Justification: The work does not release any models or datasets with potential for misuse and poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: GNDE model code is adapted from the torchgde library under the MIT license, and the associated paper is cited in the Numerical Experiments section of our Supplemental Materials.

    Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: Easy instructions for running the code in the Supplemental Materials are included in the files.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [NA]

    Justification: The paper does not involve human subjects or crowdsourced data and therefore does not require IRB approval.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [NA]

    Justification: The core method in this work does not involve LLMs as components.

    Guidelines:

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