Journal of Machine Learning Research 27 (2026) 1-37 Submitted 4/25; Revised 8/26; Published 8/26

Residual connections provably mitigate oversmoothing in graph neural networks

Ziang Chen ZIANG@MIT.EDU
Department of Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Zhengjiang Lin LINZJ@MIT.EDU
Department of Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Shi Chen SCHEN636@MIT.EDU
Department of Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Yury Polyanskiy YP@MIT.EDU
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Philippe Rigollet RIGOLLET@MATH.MIT.EDU
Department of Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Editor: Lorenzo Rosasco

Abstract

Graph neural networks (GNNs) have achieved remarkable empirical success in processing and representing graph-structured data across various domains. However, a significant challenge known as "oversmoothing" persists, where vertex features become nearly indistinguishable in deep GNNs, severely restricting their expressive power and practical utility. In this work, we analyze the asymptotic oversmoothing behavior of deep GNNs with and without residual connections by deriving explicit convergence rates for a normalized vertex similarity measure. Our analytical framework is grounded in the multiplicative ergodic theorem. Furthermore, we demonstrate that adding residual connections effectively mitigates or prevents oversmoothing across several broad families of parameter distributions. The theoretical findings are strongly supported by numerical experiments.

Keywords: Deep graph neural networks, oversmoothing, residual connection, multiplicative ergodic theorem

1. Introduction

Graph neural networks (GNNs) (Scarselli et al., 2008; Wu et al., 2020; Zhou et al., 2020; Kipf and Welling, 2017; Veličković et al., 2018) have achieved significant empirical success in

©2026 Ziang Chen, Zhengjiang Lin, Shi Chen, Yury Polyanskiy, and Philippe Rigollet.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/25-0890.html.

learning and representing graph data, with broad applications across fields such as physics (Shlomi et al., 2020), bioinformatics (Zhang et al., 2021), finance (Wang et al., 2022), electronic engineering (Liao et al., 2021; He et al., 2021; Lee et al., 2022), and operations research (Gasse et al., 2019).

Among many types of GNN architectures, the most foundational one is the message-passing graph neural network (MP-GNN), which employs a message-passing mechanism (Gilmer et al., 2017) to update each vertex's feature by aggregating information from its neighbors. Specifically, given an undirected and unweighted graph $G = (V, E)$, where $V = \{v_1, \dots, v_n\}$ is the vertex set and $E \subseteq V \times V$ represents the edges, the vertex features $\{x_i(t) : i = 1, \dots, n\}$ at the t-th layer are computed as:

$$x_i(t+1) = f^{(t)}\left(x_i(t), \text{AGGREGATE}\left(\left\{\!\left\{g^{(t)}(x_j(t)) : j \in \mathcal{N}_i\right\}\!\right\}\right)\right), \tag{1}$$

where $f^{(t)}$ and $g^{(t)}$ are trainable functions of local updates, AGGREGATE is an aggregation function (e.g., sum, average, max, attention), $\mathcal{N}_i = \{j \in \{1, \dots, n\} : (v_i, v_j) \in E\}$ is the index set of neighbors of v_i, and $\{\!\{\dots\}\!\}$ denotes a multiset allowing repeated elements.

The widespread adoption and practical success of MP-GNNs can be attributed to two key features. The first is their permutation-invariant/equivariant property, which ensures that when vertices are relabeled, the vertex features undergo a corresponding relabeling, preserving the MP-GNN's outputs on isomorphic graphs up to a permutation. The second is their efficient and scalable implementation, where vertex features are updated according to (1) using only local functions and neighborhood information from a small subgraph. Together, these properties make MP-GNNs particularly well-suited for graph learning tasks, allowing them to handle graphs of varying sizes and scale effectively across large datasets.

Despite their strengths, MP-GNNs face certain limitations that constrain their applications. A primary challenge is the *oversmoothing* phenomenon in deep GNNs, which is well documented empirically (Rusch et al., 2023; Zhang et al., 2023; Chen et al., 2020; Li et al., 2018; Oono and Suzuki, 2020; Wu et al., 2023b). More specifically, vertex features tend to become increasingly similar as the network depth grows. Although researchers have proposed various mitigation strategies including residual connections, attention mechanisms, and normalization layers (Wu et al., 2023a; Chen et al., 2020; Rusch et al., 2023; Scholkemper et al., 2025; Dezoort and Hanin, 2025), oversmoothing remains a persistent challenge. This phenomenon indicates that deep MP-GNNs struggle to maintain vertex distinctiveness even when the underlying properties differ significantly. The resulting homogeneity in vertex features creates numerical difficulties during training, often forcing the use of shallow architectures and thereby limiting MP-GNNs' expressiveness and, in turn, hinders their performance in complex, large-scale applications.

This work focuses on understanding and mitigating oversmoothing in deep GNNs. Specifically, we conduct a theoretical investigation into the oversmoothing phenomenon in deep MP-GNNs, offering practical insights and guidance.

We analyze the asymptotic behavior of oversmoothing in two widely used families of MP-GNN architectures. The first architecture is given by

$$x_i(t+1) = \sum_{j=1}^{n} P_{ij}^{(t)} W^{(t)} \sigma\left(x_j(t)\right), \tag{nrs}$$

where $P_{ij}^{(t)} \geq 0$ and $P_{ij}^{(t)} > 0$ if and only if $j \in \mathcal{N}_i$. This architecture encompasses graph convolutional networks (GCNs) (Kipf and Welling, 2017), where the aggregation coefficient matrix $P^{(t)} = P$ is time-independent and is given by $P = D^{-1}A$ or $P = D^{-1/2}AD^{-1/2}$, with $A \in \{0,1\}^{n\times n}$ as the symmetric adjacency matrix and $D \in \mathbb{R}^{n\times n}$ as the degree matrix, a diagonal matrix whose diagonal entries represent the vertex degrees, and graph attention networks (GATs) (Veličković et al., 2018), where $P^{(t)}$ is time-varying and is determined by the attention mechanism on the graph.

The second architecture employs a residual connection, and is defined as

$$x_i(t+1) = x_i(t) + \alpha \sum_{j=1}^{n} P_{ij}^{(t)} W^{(t)} \sigma\left(x_j(t)\right). \tag{rs}$$

In (nrs) and (rs), $W^{(t)} \in \mathbb{R}^{d\times d}$ is a matrix of learnable parameters with d being the dimension of vertex feature, $P_{ij}^{(t)}$ represents the coefficients in information aggregation, $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function, and $\alpha > 0$ is a constant. By examining these architectures, we aim to provide a deeper understanding of oversmoothing and strategies to alleviate its impact. We note that our analytical framework would be focusing on constant aggregation case $P_{ij}^{(t)} = P$ while we still use the notation $P_{ij}^{(t)}$ in (nrs) and (rs) for discussion and comparison.

Related work. Several works in the literature have analyzed the oversmoothing phenomenon in deep GNNs from a theoretical perspective.

For GNNs without residual connections, Li et al. (2018); Keriven (2022) prove that for GCNs with time-independent $P^{(t)} = P$, the vertex features converge to a subspace of identical or highly correlated features. This analysis is further refined in Oono and Suzuki (2020); Cai and Wang (2020), showing that the convergence occurs at an exponential rate, determined by the eigengap of the normalized graph Laplacian. Additionally, non-asymptotic analysis for GCNs is provided in Wu et al. (2023b). For GATs without residual connections and with time-varying $P^{(t)}$, Wu et al. (2023a) demonstrate that they also suffer from oversmoothing at an exponential rate. Moreover, the upper bound of the convergence rate established in Wu et al. (2023a) is shown to be greater than or equal to the rate derived for GCNs in Oono and Suzuki (2020); Cai and Wang (2020). While this observation pertains only to upper bounds that could be loose, it points to the potential mitigation of oversmoothing in GATs even in the absence of a residual connection.

Previous works have also analyzed GNNs with residual connections. Scholkemper et al. (2025) demonstrate that oversmoothing can be prevented by introducing a residual connection between each hidden layer and the *initial* layer which amounts to adding a constant drift $x_i(0)$ to the right-hand side of (nrs). The nature of this remedy is in sharp contrast to (rs), where the residual connection links consecutive layers as in classical ResNets (He et al., 2016).

Closer to our setting in (rs), Dezoort and Hanin (2025) investigate the oversmoothing phenomenon in residual GNNs when σ is a ReLU activation and the entries of $W^{(t)}$ are i.i.d. Gaussian. They prove that for a deep residual GNN with T message-passing layers, the oversmoothing issue can be prevented if the stepsize α is chosen as small, say $\alpha = \Theta(1/T)$.

Our contributions. We define a normalized vertex similarity measure $\mu(x) \in [0,1]$ for $x = (x_1, \ldots, x_n) \in \mathbb{R}^{d\times n}$, where $\mu(x) = 0$ indicates that all vertex features are identical, and analyze the asymptotic behavior of $\mu(x_{\mathsf{nrs}}(t))$ and $\mu(x_{\mathsf{rs}}(t))$, where $\{x_{\mathsf{nrs}}(t)\}_{t\in\mathbb{N}}$ and $\{x_{\mathsf{rs}}(t)\}_{t\in\mathbb{N}}$ are generated by (nrs) and (rs), respectively.

We establish a new and unified analytical framework that rigorously demonstrates that residual connections mitigate or prevent oversmoothing in deep GNNs. In particular, our main contributions are summarized informally as follows:

- *Non-residual dynamics:* we prove that $\mu(x_{\mathsf{nrs}}(t))^{1/2t} \to \max_{\lambda\in\mathrm{spec}(P)\setminus\{1\}} |\lambda| < 1$ as $t \to +\infty$ almost surely, and the limit is *independent* of the distribution of $W^{(t)}$.
- *Residual dynamics:* We show that almost surely, $\lim_{t\to\infty} \mu(x_{\mathsf{rs}}(t))^{1/2t}$ is lower bounded by a constant depending on the spectrum of P and the distribution of $W^{(t)}$. The lower bound is achieved if P is diagonalizable in $\mathbb{C}$. Furthermore, this bound can be explicitly computed or estimated for several commonly used distributions of $W^{(t)}$.
- *Comparison of non-residual and residual dynamics:* For several distributions of $W^{(t)}$, we rigorously demonstrate that $\lim_{t\to\infty} \mu(x_{\mathsf{rs}}(t))^{1/2t} > \lim_{t\to\infty} \mu(x_{\mathsf{nrs}}(t))^{1/2t}$, and in some cases, $\lim_{t\to\infty} \mu(x_{\mathsf{rs}}(t))^{1/2t} = 1$, which indicates that $\mu(x_{\mathsf{rs}}(t)$ can only converge to zero at a subexponential rate. This shows that residual connections effectively mitigate the oversmoothing issue.

Our analysis relies on three main assumptions: (i) $\sigma = \mathrm{Id}$, (ii) $P^{(t)} \equiv P \in \mathbb{R}^{n\times n}$ is a nonnegative matrix for which the Perron-Frobenius theorem applies with leading eigenvector $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^n$ associated to the leading eigenvalue assumed equal to 1, and (iii) $W^{(t)},\ t \in \mathbb{N}$ are sampled i.i.d. from some probability distribution on $\mathbb{R}^{d\times d}$. Note that (ii) holds when P is the probability transition matrix of a simple random walk on a graph that is irreducible and aperiodic, as in Norris (1998). A significant strength of our analysis is that it extends readily to asymmetric matrices P with complex spectrum.

Our proof techniques draw inspiration from the multiplicative ergodic theorem (Oseledets, 1968; Arnold, 1998), which precisely characterizes the asymptotic behavior of linear random dynamical systems or products of random matrices.

Organization. The rest of this paper will be organized as follows. We state our main theory on the asymptotic oversmoothing rate and their applications for several commonly used distributions in Section 2 . All proofs are presented in Section 3 and numerical experiments are conducted in Section 4. Section 5 concludes the paper.

2. Main Results

This section presents our main results. In Section 2.1, we establish the asymptotic oversmoothing rates for (nrs) and (rs). In Section 2.2, we compare the oversmoothing rates of (nrs) and (rs) for several specific distributions.

2.1 Asymptotic oversmoothing rate

We will use the (normalized) vertex similarity measure introduced below to quantify the degree of oversmoothing.

Definition 1 (Vertex similarity measure) *For $x = (x_1, \ldots, x_n) \in \mathbb{R}^{d\times n}$, we define (normalized) vertex similarity measure as*

$$\mu(x) = \frac{\sum_{i=1}^n \|x_i - \bar{x}\|_2^2}{\sum_{i=1}^n \|x_i\|_2^2}, \tag{2}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$.

It is straightforward to check that $\mu(x) \in [0,1]$ for any $x \in \mathbb{R}^{d\times n}\backslash\{0\}$ and that $\mu(x) = 0$ if and only if $x_1 = \cdots = x_n$. Next, we outline the assumptions used in our analysis. The first assumption is that $\sigma = \mathrm{Id}$, which simplifies (nrs) and (rs) to linear dynamics.

Assumption 1 (Activation function) *We assume that the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is the identity map, i.e., the dynamics in* (nrs) *and* (rs) *are linear.*

The second assumption concerns the aggregation coefficients $P^{(t)}$.

Assumption 2 (Aggregation coefficients) *We assume that the aggregation coefficients $P^{(t)} \in \mathbb{R}^{n\times n}$ satisfy the followings:*

(i) $P^{(t)} = P$ *is a constant matrix in* $t \in \mathbb{N}$.

(ii) P *is a primitive matrix:* $\exists\ k \geq 1$ *such that for any* i, j, $P_{ij} \geq 0$ *and* $P_{ij}^k > 0$.

(iii) $P_{ij} > 0$ *if and only if* $j \in \mathcal{N}_i$, *and for any* $i \in \{1, \ldots, n\}$, *it holds that*

$$\sum_{j=1}^n P_{ij} = \sum_{j\in\mathcal{N}_i} P_{ij} = 1. \tag{3}$$

Assumption 2 states that $P^{(t)} = P \in \mathbb{R}^{n\times n}$ is a probability transition matrix of an irreducible aperiodic Markov chain on the underlying graph G. Moreover, for a nonnegative matrix $P \geq 0$ with $P_{ij} > 0$ if and only if $j \in \mathcal{N}_i$, the primitivity of P, or equivalently the irreducibility and aperiodicity of P, can be ensured by specific properties of the graph G: namely, G is connected and contains at least one odd-length cycle.

We also note that though the primitivity of P implies the connectivity of the graph G, our theory can extend directly to disconnected graphs, since P is block-diagonal across connected components and the dynamics decouple. The analysis therefore applies independently on each component whose block in P satisfies Assumption 2.

Remark 2 *Suppose that Assumption 2 holds. By the Perron-Frobenius theorem, the spectral radius of the matrix P is 1, and $\lambda_1 = 1$ is an eigenvalue with multiplicity being one. One can see from* (3) *that $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^n$ is an eigenvector corresponding to $\lambda_1 = 1$. Moreover, the magnitudes of all other eigenvalues of P are strictly smaller than 1, i.e.,*

$$|\lambda| < 1, \quad \forall\ \lambda \in \mathrm{spec}(P)\backslash\{1\},$$

where $\mathrm{spec}(P) \subseteq \mathbb{C}$ is the spectrum of P. Note that we do not assume that P is symmetric and our analysis accounts for the presence of complex eigenvalues in P. In comparison, prior works (Oono and Suzuki, 2020; Cai and Wang, 2020; Dezoort and Hanin, 2025) assume symmetry or a real spectrum for P.

We also assume that the weight matrices $W_1, W_2, \ldots$ are i.i.d. from some ensemble.

Assumption 3 (Weight matrices) *We assume that $W^{(t)}$, $t \in \mathbb{N}$ are i.i.d. drawn from a Borel probability measure $\mathbb{P}_W$ over $\mathbb{R}^{d\times d}$ that satisfies that*

$$\mathbb{E}\left[\max\left\{\log\|W^{(t)}\|_2, 0\right\}\right] < +\infty. \tag{4}$$

Moreover, we require that the probability distribution $\mathbb{P}_W$ satisfies some non-degenerate conditions. To state them rigorously, we present the next proposition that will be proved in Section 3.1.

Proposition 3 *Suppose that Assumption 3 holds. The followings are true.*

(i) There exists a constant $R(\mathbb{P}_W) \geq 0$ depending on the probability distribution $\mathbb{P}_W$, such that

$$\lim_{t\to+\infty}\left\|W^{(t-1)}\cdots W^{(1)}W^{(0)}\right\|_2^{1/t} = R(\mathbb{P}_W), \quad \text{a.s.} \tag{5}$$

(ii) For any $\beta \in \mathbb{C}$, there exists a constant $R(\beta, \mathbb{P}_W) \geq 0$ depending on β and the probability distribution $\mathbb{P}_W$, such that

$$\lim_{t\to+\infty}\left\|\left(\mathrm{Id}+\beta W^{(t-1)}\right)\cdots\left(\mathrm{Id}+\beta W^{(1)}\right)\left(\mathrm{Id}+\beta W^{(0)}\right)\right\|_2^{1/t} = R(\beta, \mathbb{P}_W), \quad \text{a.s.} \tag{6}$$

Our last set of assumptions is on the constants in Proposition 3.

Assumption 4 *The constant $R(\mathbb{P}_W)$ defined in Proposition 3 (i) satisfies that $R(\mathbb{P}_W) > 0$.*

Assumption 5 *The spectrum of aggregation coefficient matrix P and the constants in Proposition 3 (ii) satisfy that $\max_{\lambda\in\mathrm{spec}(P)} R(\alpha\lambda, \mathbb{P}_W) > 0$, where $\alpha > 0$ is defined in* (rs).

With the assumptions outlined above, we are now prepared to present our main results. The first result characterizes the asymptotic oversmoothing rate of deep GNNs described by (nrs) without residual connections.

Theorem 4 (Asymptotic oversmoothing rate of deep non-residual GNNs) *Consider the GNN dynamics $\{x_{\mathrm{nrs}}(t)\}_{t\in\mathbb{N}}$ generated by* (nrs)*, and suppose that Assumptions 1, 2, 3, and 4 hold. With probability one, we have for almost every $x_{\mathrm{nrs}}(0) \in \mathbb{R}^{d\times n}$ that*

$$\lim_{t\to+\infty}\mu(x_{\mathrm{nrs}}(t))^{1/2t} = \max_{\lambda\in\mathrm{spec}(P)\setminus\{1\}}|\lambda| < 1. \tag{7}$$

Theorem 4 establishes that $\mu(x_{\mathrm{nrs}}(t))$ converges to 0 at an exponential rate determined solely by the spectrum of P, independent of the distribution of $W^{(t)}$. Specifically, the convergence rate is governed by the eigenvalue λ with the second largest magnitude. Notably, Theorem 4 remains valid even when P has complex eigenvalues. The proof of Theorem 4 is provided in Section 3.3, following an introduction to linear random dynamical systems and to the multiplicative ergodic theorem in Section 3.1.

In the special case of GCNs where $P = D^{-1}A$, the gap $1 - \max_{\lambda \in \mathrm{spec}(P)\setminus\{1\}} |\lambda|$ corresponds to the second smallest eigenvalue of the normalized graph Laplacian $\Delta = I - D^{-1/2}AD^{-1/2}$. Moreover, recall that for graphs satisfying Assumption 2, all eigenvalues of Δ lie in $[0, 2)$, and there is exactly one eigenvalue equal to 0. Hence, Theorem 4 confirms the findings of Oono and Suzuki (2020); Cai and Wang (2020), with the difference that we offer greater precision by establishing an exact convergence rate in the case of an identity activation function, whereas Oono and Suzuki (2020); Cai and Wang (2020) present upper bounds for nonlinear activations (ReLU and Leaky-ReLU). Our sharp characterization is necessary to demonstrate that residual dynamics improve on this limit.

Our second main theorem characterizes the asymptotic oversmoothing rate of deep residual GNNs as described by (rs).

Theorem 5 (Asymptotic oversmoothing rate of deep residual GNNs) *Consider the GNN dynamics $\{x_{\mathrm{rs}}(t)\}_{t\in\mathbb{N}}$ generated by (rs), and suppose that Assumptions 1, 2, 3, and 5 hold. With probability one, we have for almost every $x_{\mathrm{rs}}(0) \in \mathbb{R}^{d\times n}$ that*

$$\lim_{t\to+\infty} \mu(x_{\mathrm{rs}}(t))^{1/2t} \geq \frac{\max\limits_{\lambda\in\mathrm{spec}(P)\setminus\{1\}} R(\alpha\lambda, \mathbb{P}_W)}{\max\limits_{\lambda\in\mathrm{spec}(P)} R(\alpha\lambda, \mathbb{P}_W)}. \tag{8}$$

Moreover, if P is diagonalizable in $\mathbb{C}$, then the inequality above becomes an equality.

Theorem 5 establishes a lower bound for $\lim_{t\to+\infty} \mu(x_{\mathrm{rs}}(t))^{1/2t}$, which depends on the distribution $\mathbb{P}_W$, the spectrum of P, and the step size α in the residual connection. This lower bound is achieved when P is diagonalizable in $\mathbb{C}$. The proof of Theorem 5 is presented in Section 3.4 and is also based on linear random dynamical systems and the multiplicative ergodic theorem.

From Theorem 4 and Theorem 5, one can conclude that the oversmoothing issue is mitigated if the lower bound in (8) is strictly greater than $\max_{\lambda\in\mathrm{spec}(P)\setminus\{1\}} |\lambda|$. In Section 2.2, we discuss several commonly used probability distributions $\mathbb{P}_W$ for which the lower bound in (8) can be explicitly computed or estimated. In these cases, it is either strictly greater than $\max_{\lambda\in\mathrm{spec}(P)\setminus\{1\}} |\lambda|$ or equal to 1.

Dezoort and Hanin (2025) also consider residual aggregation operators $P_s = (1-s)I + sP$ for $s \in (0, 1]$, which interpolate between the identity and the base aggregator P. Whenever P satisfies Assumption 2, so does P_s (after adjoining self-loops to the neighborhoods $\mathcal{N}_i$), and

$$\mathrm{spec}(P_s) = \big\{(1 - s) + s\lambda : \lambda \in \mathrm{spec}(P)\big\}.$$

Theorems 4 and 5 therefore apply verbatim with P replaced by P_s. Thus s is an additional architectural handle on the asymptotic rate: in the non-residual dynamics it acts only through the second spectral radius of P_s (still independently of $\mathbb{P}_W$), whereas in the residual dynamics it reshapes the spectrum fed into the map $R(\cdot, \mathbb{P}_W)$ and complements the step size α. Residual aggregation (identity mixed into the graph operator) and residual connections (identity mixed into the feature update) are therefore distinct, complementary mechanisms; the case $s = 1$ recovers Theorem 4 and Theorem 5 as stated.

To conclude this section, we note that if we do not assume the row sums of P are equal to 1, as in (3), our theory still holds for a modified vertex similarity measure, given by

$$\mu(x) = \frac{\|x - x\pi_1\pi_1^\top\|_F^2}{\|x\|_F^2}, \tag{9}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and $\pi_1 \in \mathbb{R}^n$ is the eigenvector corresponding to the leading eigenvalue of P, with $\|\pi_1\|_2 = 1$ and strictly positive entries, as guaranteed by the Perron-Frobenius theorem. The proofs almost follow the same lines. If (3) is assumed, then $\pi_1 = \frac{1}{\sqrt{n}}(1, \dots, 1)^\top$, and (9) coincides with (2). If the entries of π_1 are not identical, then $\mu(x) = 0$ does not imply that $x_1, \dots, x_n$ are equal but rather that they all lie on a one-dimensional subspace of $\mathbb{R}^d$.

2.2 Weight matrices

Comparing Theorems 4 and 5 indicates that the distribution $\mathbb{P}_W$ of the weight matrix controls the oversmoothing of residual GNNs whereas it has no impact on their nonresidual counterparts. In this section, we examine the effect of this distribution in more details in the context of specific choices for $\mathbb{P}_W$. All the cases considered here indicate that residual connections effectively mitigate oversmoothing over a wide class of distributions, and all the proofs are deferred to Section 3.5. In comparison to Dezoort and Hanin (2025), our analysis encompasses a broader class of parameter distributions, not limited to i.i.d. Gaussian.

2.2.1 Deterministic

We begin by considering the case where $W^{(t)} = W$ is deterministic, i.e., $\mathbb{P}_W$ is a point mass at $W \in \mathbb{R}^{d\times d}$. In this case, equality holds in (8) even if $\mathbb{P}$ is not diagonalizable in $\mathbb{C}$.

Theorem 6 *Suppose that Assumptions 1, 2, 3, and 5 hold with $W^{(t)} = W$ almost surely, where W is a matrix with deterministic entries. Let $\{x_{\mathrm{rs}}(t)\}_{t\in\mathbb{N}}$ be generated by* (rs)*. Then we have for almost every $x_{\mathrm{rs}}(0) \in \mathbb{R}^{d\times n}$ that*

$$\lim_{t\to+\infty} \mu(x_{\mathrm{rs}}(t))^{1/2t} = \frac{\max\limits_{\lambda\in\mathrm{spec}(P)\setminus\{1\}} \rho(\mathrm{Id} + \alpha\lambda W)}{\max\limits_{\lambda\in\mathrm{spec}(P)} \rho(\mathrm{Id} + \alpha\lambda W)},$$

where $\rho(A) = \max\limits_{\lambda\in\mathrm{spec}(A)} |\lambda|$ is the spectral radius of a matrix A.

The following theorem characterizes several cases with deterministic $W^{(t)}$ in which the oversmoothing problem is provably mitigated or prevented.

Theorem 7 *Suppose that Assumptions 1, 2, 3, 4, and 5 hold with $W^{(t)} = W$ almost surely, and in addition that P has a real spectrum. Let $\{x_{\mathrm{nrs}}(t)\}_{t\in\mathbb{N}}$ and $\{x_{\mathrm{rs}}(t)\}_{t\in\mathbb{N}}$ be generated by* (nrs) *and* (rs)*, respectively. We also assume that $\rho(I + \alpha W)$ is achieved by $|1 + \alpha\mu_0|$ for $\mu_0 \in \mathrm{spec}(W)$.*

(i) If $\max_{\lambda\in\mathrm{spec}(P)\backslash\{1\}}|\lambda|$ *is achieved by* $\lambda_0\in\mathrm{spec}(P)\backslash\{1\}$ *and* $1+\lambda_0(2\alpha\mathrm{Re}(\mu_0)+1)>0$*, then for almost every* $x_{\mathrm{nrs}}(0)\in\mathbb{R}^{d\times n}$ *and* $x_{\mathrm{rs}}(0)\in\mathbb{R}^{d\times n}$*,*

$$\lim_{t\to+\infty}\mu(x_{\mathrm{rs}}(t))^{1/2t}>\max_{\lambda\in\mathrm{spec}(P)\backslash\{1\}}|\lambda|=\lim_{t\to+\infty}\mu(x_{\mathrm{nrs}}(t))^{1/2t}.$$

(ii) If $2\mathrm{Re}(\mu_0)+\alpha(\lambda+1)|\mu_0|^2\le 0$ *for some* $\lambda\in\mathrm{spec}(P)\backslash\{1\}$*, then for almost every* $x_{\mathrm{rs}}(0)\in\mathbb{R}^{d\times n}$*,*

$$\lim_{t\to+\infty}\mu(x_{\mathrm{rs}}(t))^{1/2t}=1.$$

We remark that the condition that P has a real spectrum holds for GCNs with $P=D^{-1}A$ since $\mathrm{spec}(D^{-1}A)=\mathrm{spec}(D^{-1/2}AD^{-1/2})\subset(-1,1]$.

2.2.2 Ginibre ensemble

Next, consider the case where the entries of $W^{(t)}$ are i.i.d. Gaussian with a mean of 0 and variance τ^2 (Ginibre ensemble) as in Dezoort and Hanin (2025). The following theorem provides a characterization of the lower bound in (8) in terms of one-dimensional normal distribution and chi-squared distribution.

Theorem 8 *Suppose that Assumptions 1, 2, 3, and 5 hold with entries of* $W^{(t)}$ *i.i.d. drawn from* $\mathcal{N}(0,\tau^2)$*, and in addition that* P *has a real spectrum. Let* $\{x_{\mathrm{rs}}(t)\}_{t\in\mathbb{N}}$ *be generated by* (rs)*. Then with probability one, for almost every* $x_{\mathrm{rs}}(0)\in\mathbb{R}^{d\times n}$*,*

$$\lim_{t\to+\infty}\mu(x_{\mathrm{rs}}(t))^{1/2t}\ge\frac{\max_{\lambda\in\mathrm{spec}(P)\backslash\{1\}}\exp\left(\frac{1}{2}\mathbb{E}\log\left((1+\alpha\lambda\tau\xi)^2+\alpha^2\lambda^2\tau^2\chi^2_{d-1}\right)\right)}{\max_{\lambda\in\mathrm{spec}(P)}\exp\left(\frac{1}{2}\mathbb{E}\log\left((1+\alpha\lambda\tau\xi)^2+\alpha^2\lambda^2\tau^2\chi^2_{d-1}\right)\right)},\tag{10}$$

where $\xi\sim\mathcal{N}(0,1)$ *is a one-dimensional standard real Gaussian distribution, and* χ^2_{d-1} *is a chi-squared random variable with* $d-1$ *degrees of freedom that is independent of* ξ*. Additionally, if* P *is diagonalizable in* $\mathbb{C}$*, then the equality in* (10) *holds.*

Based on the lower bound (10), it can be proved that $\lim_{t\to+\infty}\mu(x_{\mathrm{rs}}(t))^{1/2t}$ is strictly larger than $\lim_{t\to+\infty}\mu(x_{\mathrm{nrs}}(t))^{1/2t}$, for any $\alpha>0$ and $\tau>0$.

Theorem 9 *Suppose that Assumptions 1, 2, 3, 4, and 5 hold with entries of* $W^{(t)}$ *i.i.d. drawn from* $\mathcal{N}(0,\tau^2)$*, and in addition that* P *has a real spectrum. Let* $\{x_{\mathrm{nrs}}(t)\}_{t\in\mathbb{N}}$ *and* $\{x_{\mathrm{rs}}(t)\}_{t\in\mathbb{N}}$ *be generated by* (nrs) *and* (rs)*, respectively. It holds for all* $\alpha>0$ *and* $\tau>0$ *that, with probability one, for almost every* $x_{\mathrm{nrs}}(0)\in\mathbb{R}^{d\times n}$ *and* $x_{\mathrm{rs}}(0)\in\mathbb{R}^{d\times n}$*,*

$$\lim_{t\to+\infty}\mu(x_{\mathrm{rs}}(t))^{1/2t}>\max_{\lambda\in\mathrm{spec}(P)\backslash\{1\}}|\lambda|=\lim_{t\to+\infty}\mu(x_{\mathrm{nrs}}(t))^{1/2t}.$$

Our analysis for Ginibre ensembles differs from Dezoort and Hanin (2025) in several senses. Firstly, we impose no requirement for α to depend on the number of layers, whereas Dezoort and Hanin (2025) carefully select α based on the predetermined depth of GNNs. Secondly, Dezoort and Hanin (2025) consider the ReLU activation, while our analysis is for linear dynamics.

2.2.3 Bounded norm

Useful quantitative lower bounds can be extracted under minimal assumptions on the distribution of weight matrices. In this part, we examine compactly supported distributions for which $\lim_{t\to+\infty}\mu(x_{\mathrm{rs}}(t))^{1/2t}$ can be proved as strictly greater than $\lim_{t\to+\infty}\mu(x_{\mathrm{nrs}}(t))^{1/2t}$, provided that α is sufficiently small.

Theorem 10 *Suppose that Assumptions 1, 2, 3, 4, and 5 hold with $\|W^{(t)}\| \le r_W$ almost surely. Let $\{x_{\mathrm{nrs}}(t)\}_{t\in\mathbb{N}}$ and $\{x_{\mathrm{rs}}(t)\}_{t\in\mathbb{N}}$ be generated by (nrs) and (rs), respectively. Then it holds for any $\alpha \le 1/r_W$ that, with probability one, for almost every $x_{\mathrm{rs}}(0)\in\mathbb{R}^{d\times n}$,*

$$\lim_{t\to+\infty}\mu(x_{\mathrm{rs}}(t))^{1/2t} \ge \frac{1-\alpha r_W \min_{\lambda\in\mathrm{spec}(P)\setminus\{1\}}|\lambda|}{1+\alpha r_W}.$$

As a corollary, if

$$\frac{1-\max\limits_{\lambda\in\mathrm{spec}(P)\setminus\{1\}}|\lambda|}{1+\min\limits_{\lambda\in\mathrm{spec}(P)\setminus\{1\}}|\lambda|} > \frac{\alpha r_W}{1+\alpha r_W},$$

then with probability one, we have for almost every $x_{\mathrm{nrs}}(0)\in\mathbb{R}^{d\times n}$ and $x_{\mathrm{rs}}(0)\in\mathbb{R}^{d\times n}$ that

$$\lim_{t\to+\infty}\mu(x_{\mathrm{rs}}(t))^{1/2t} > \lim_{t\to+\infty}\mu(x_{\mathrm{nrs}}(t))^{1/2t}.$$

Remark 11 *The lower bound in Theorem 10 should be interpreted as a conservative, distribution-free guarantee rather than as a sharp tuning rule for the residual stepsize α. The proof uses only the almost-sure operator-norm bound $\|W^{(t)}\|_2 \le r_W$, and therefore ignores finer information about the distribution of the weights. As a result, the bound may be far from optimal for common initialization schemes. For such initialization schemes, distribution-specific results can provide sharper guidance. As an example, in the Ginibre/Gaussian case, Theorem 9 shows that for every $\alpha > 0$ and $\tau > 0$, the asymptotic vertex-similarity rate of the residual GCN is strictly larger than that of the corresponding non-residual GCN. Similar comparisons can be obtained for other common i.i.d. initialization distributions, such as uniform initializations, but require separate arguments. Thus, Theorem 10 should be viewed as a robust sufficient condition under a minimal bounded-norm assumption, not as a prescription for optimizing α or r_W.*

2.2.4 Simultaneously diagonalizable

Finally, we consider the scenario where all matrices in the support of $\mathbb{P}_W$ are simultaneously diagonalizable.

Assumption 6 *We assume that $W^{(t)} \sim \mathbb{P}_W$ satisfies*

$$W^{(t)} = Q^{-1}\,diag\left(w_1^{(t)},\dots,w_d^{(t)}\right)Q,$$

where $Q\in\mathbb{R}^{d\times d}$ is a fixed invertible matrix and $w_i^{(t)}$, $t\in\mathbb{N}$ are i.i.d. drawn from some Borel probability distribution $\mathbb{P}_i$ over $\mathbb{R}$ satisfying

$$\log\left|1+\alpha\lambda w_i^{(t)}\right| \in L^1(\mathbb{P}_i), \quad \forall\ \lambda\in\mathrm{spec}(P).$$

for $i=1,\dots,d$.

Under Assumption 6, the following theorem relates the constant $R(\alpha\lambda, \mathbb{P}_W)$ to the marginal distributions of $w_i^{(t)}$.

Theorem 12 *Suppose that Assumptions 1, 2, 3, 5, and 6 hold. Let $\{x_{\mathrm{rs}}(t)\}_{t\in\mathbb{N}}$ be generated by* (rs)*. Then with probability one, we have for almost every $x_{\mathrm{rs}}(0) \in \mathbb{R}^{d\times n}$ that*

$$\lim_{t\to+\infty} \mu(x_{\mathrm{rs}}(t))^{1/2t} \geq \frac{\max\limits_{\lambda\in\mathrm{spec}(P)\setminus\{1\}} \max\limits_{1\leq i\leq d} \exp\left(\mathbb{E}\left[\log\left|1+\alpha\lambda w_i^{(t)}\right|\right]\right)}{\max\limits_{\lambda\in\mathrm{spec}(P)} \max\limits_{1\leq i\leq d} \exp\left(\mathbb{E}\left[\log\left|1+\alpha\lambda w_i^{(t)}\right|\right]\right)}. \tag{11}$$

Additionally, if P is diagonalizable in $\mathbb{C}$, then the equality in (11) *holds.*

Furthermore, when P has a real spectrum and the distribution $\mathbb{P}_1, \ldots, \mathbb{P}_d$ are all symmetric, then $\mu(x_{\mathrm{rs}}(t))^{1/2t}$ converges to 1 almost surely as stated in the next theorem.

Theorem 13 *Suppose that assumptions made in Theorem 12 are satisfied. Suppose in addition that P has a real spectrum and that $\mathbb{P}_i$ is symmetric, i.e., $\mathbb{P}_i(B) = \mathbb{P}_i(-B)$ for any Borel subset $B \subseteq \mathbb{R}$, for $i = 1, 2, \ldots, d$. Let $\{x_{\mathrm{rs}}(t)\}_{t\in\mathbb{N}}$ be generated by* (rs)*. If $|w_i^{(t)}| \leq r_i$ almost surely, $i = 1, \ldots, d$, and $\alpha \max_{1\leq i\leq d} r_i < 1$, it holds with probability one that, for almost every $x_{\mathrm{rs}}(0) \in \mathbb{R}^{d\times n}$,*

$$\lim_{t\to+\infty} \mu(x_{\mathrm{rs}}(t))^{1/2t} = 1.$$

3. Proofs

We collect all proofs in this section.

3.1 Preliminary on Linear Random Dynamical System

This section reviews some definitions and fundamental results for linear random dynamical systems; see Arnold (1998) and references therein for more details.

Definition 14 (Metric dynamical system) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A metric dynamical system on $(\Omega, \mathcal{F}, \mathbb{P})$ is a family of maps $\{\theta(t) : \Omega \to \Omega\}_{t\in\mathbb{N}}$ satisfying that*

(i) The mapping $\mathbb{N} \times \Omega \to \Omega,\ (t, \omega) \mapsto \theta(t)\omega$ is measurable.

(ii) It holds that

$$\theta(0) = \mathrm{Id}_\Omega \quad \textit{and} \quad \theta(t+s) = \theta(t) \circ \theta(s), \quad \forall\ s, t \in \mathbb{N}. \tag{12}$$

(iii) $\theta(t)$ is $\mathbb{P}$-preserving for any $t \in \mathbb{N}$, i.e.,

$$\mathbb{P}[\theta(t)^{-1}B] = \mathbb{P}[B], \quad \forall\ B \in \mathcal{F},\ t \in \mathbb{N}.$$

In the above definition, the index $t \in \mathbb{N}$ serves as the notion of one-sided discrete time. Though this paper focuses only on time represented by non-negative integers, we remark that in general the metric dynamical system is defined with a time set $\mathbb{T}$ that can be a

semigroup with $0 \in \mathbb{T}$, including two-sided discrete time $\mathbb{T} = \mathbb{Z}$, one-sided continuous time $\mathbb{T} = \mathbb{R}_{\geq 0}$, and two-sided continuous time $\mathbb{T} = \mathbb{R}$.

In the case of $\mathbb{T} = \mathbb{N}$, one can set $\theta = \theta(1)$ and (12) implies that $\theta(t)$ is the t-fold composition of θ, namely, $\theta(t) = \theta^t = \theta \circ \cdots \circ \theta$.

Definition 15 (Linear random dynamical system) *Let $\{\theta^t : \Omega \to \Omega\}_{t\in\mathbb{N}}$ be a metric dynamical system on $(\Omega, \mathcal{F}, \mathbb{P})$ and let $A : \Omega \to \mathbb{R}^{m\times m}$ be measurable. A linear random dynamical system on $\mathbb{R}^m$ over $\{\theta^t\}_{t\in\mathbb{N}}$, induced by A, is the measurable map*

$$\begin{aligned} \Phi : \mathbb{N} \times \Omega \times \mathbb{R}^m \to &\quad \mathbb{R}^m, \\ (t, \omega, u) &\quad \mapsto \Phi(t, \omega)u, \end{aligned}$$

where

$$\Phi(t, \omega) = A(\theta^{t-1}\omega) \cdots A(\theta\omega)A(\omega). \tag{13}$$

We have slightly abused the notation by writing $\Phi(t, \omega, u) = \Phi(t, \omega)u$. The asymptotic behavior of a linear random dynamical system can be characterized precisely by the celebrated multiplicative ergodic theorem in the following.

Theorem 16 (Multiplicative ergodic theorem (Arnold, 1998, Theorem 3.4.1)) *Let $\{\theta^t : \Omega \to \Omega\}_{t\in\mathbb{N}}$ be a metric dynamical system on $(\Omega, \mathcal{F}, \mathbb{P})$ and let $A : \Omega \to \mathbb{R}^{m\times m}$ be measurable. Consider a linear random dynamical system $\Phi(t, w)$ as in* (13)*. Suppose that*

$$\max\left\{\log \|A(\cdot)\|_2, 0\right\} \in L^1(\Omega, \mathcal{F}, \mathbb{P}).$$

There exists $\widetilde{\Omega} \in \mathcal{F}$ with $\theta\widetilde{\Omega} \subseteq \widetilde{\Omega}$ and $\mathbb{P}(\widetilde{\Omega}) = 1$, such that the followings hold for any $\omega \in \widetilde{\Omega}$:

(i) It holds that the limit

$$\Lambda(\omega) = \lim_{t\to\infty} \left(\Phi(t, \omega)^\top \Phi(t, \omega)\right)^{1/2t},$$

exists and is a positive semidefinite matrix.

(ii) Suppose $\Lambda(\omega)$ has $q(\omega)$ distinct eigenvalues, which are ordered as $e^{\nu_1(\omega)} > \cdots > e^{\nu_{q(\omega)}(\omega)} \geq 0$. Denote $E_i(\omega)$ as the corresponding eigenspace, with dimension $d_i(\omega) > 0$, for $i = 1, \ldots, q(\omega)$. Then the functions $q(\cdot)$, $\nu_i(\cdot)$, and $d_i(\cdot)$, $i = 1, \ldots, p(\cdot)$, are all measurable and θ-invariant on $\widetilde{\Omega}$:

$$q(\theta\omega) = q(\omega),\ \nu_i(\theta\omega) = \nu_i(\omega),\ \text{and } d_i(\theta\omega) = d_i(\omega).$$

(iii) Set $U_i(\omega) = \bigoplus_{j\geq i} E_j(\omega)$, $i = 1, \ldots, q(\omega)$, and $U_{q(\omega)+1}(\omega) = \{0\}$. Then it holds that

$$\lim_{t\to\infty} \frac{1}{t} \log \|\Phi(t, \omega)u\|_2 = \nu_i(\omega), \quad \forall\ u \in U_i(\omega)\backslash U_{i+1}(\omega), \tag{14}$$

for $i = 1, \ldots, q(\omega)$. The maps $E_i(\cdot)$ and $U_i(\cdot)$ from $\widetilde{\Omega}$ to the Grassmannian manifold are measurable.

(iv) It holds that

$$A(\omega)U_i(\omega) \subseteq U_i(\theta\omega), \quad i = 1, \dots, q(\omega).$$

(v) When $(\Omega, \mathcal{F}, \mathbb{P}, \theta)$ is ergodic, i.e., every $B \in \mathcal{F}$ with $\theta^{-1}B = B$ satisfies $\mathbb{P}(B) = 0$ or $\mathbb{P}(B) = 1$, the functions $q(\cdot)$, $\nu_i(\cdot)$, and $d_i(\cdot)$, $i = 1, \dots, p(\cdot)$, are constants on $\widetilde{\Omega}$.

In Theorem 16, $\nu_1(\omega) > \cdots > \nu_{q(\omega)}(\omega)$ are known as Lyapunov exponents and it is possible that $\lambda_{q(\omega)}(\omega) = -\infty$. Moreover, the filtration

$$\{0\} = U_{q(\omega)+1}(\omega) \subseteq U_{q(\omega)}(\omega) \subseteq \cdots \subseteq U_1(\omega) = \mathbb{R}^m$$

is known as the Oseledets filtration and Theorem 16 is also named as Oseledets theorem. In particular, V.I. Oseledets proves the first multiplicative ergodic theorem in Oseledets (1968) and since then, there has been a rich literature in this field, see e.g., Raghunathan (1979); Ruelle (1979); Walters (1993).

Remark 17 *Although Theorem 16 is stated for only systems on $\mathbb{R}^m$, we remark that the same result also holds for complex space $\mathbb{C}^m$ as mentioned in Arnold (1998, Remark 3.4.10).*

We prove the following general proposition, which will be used frequently in the proof of results in Section 2.

Proposition 18 *Adopting the notation in Theorem 16 in the case of $\mathbb{C}^m$. Assume that $\mathbb{C}^m = Q_1 \oplus \cdots \oplus Q_k$, where each Q_i is almost surely a $\Phi(t,\omega)$-invariant $\mathbb{C}$-linear subspace. That is, with probability one, for any $t \in \mathbb{N}$ and any $u_i \in Q_i$, we have that $\Phi(t,\omega)u_i \in Q_i$. We define $\nu_1(Q_i, \omega)$ as the largest Lyapunov exponent of the restriction of $\Phi(t,\omega)$ on Q_i. Then, for any fixed $\omega \in \widetilde{\Omega}$, there exists a proper $\mathbb{C}$-linear subspace of $\mathbb{C}^m$, $U(\omega)$, such that*

$$\lim_{t\to+\infty} \frac{1}{t} \log \|\Phi(t,\omega)u\|_2 = \max_{1\le i\le k} \nu_1(Q_i, \omega), \quad \forall\ u \in \mathbb{C}^m \backslash U(\omega).$$

Proof *Without loss of generality, we assume that $\nu_1(Q_1, \omega) = \max_{1\le i\le k} \nu_1(Q_i, \omega)$. We then let $U(\omega) \subsetneqq Q_1$ be the corresponding $U_2(\omega)$ in (14), part (iii) of Theorem 16, when we consider the restriction of $\Phi(t,\omega)$ on Q_1. $U(\omega)$ is also a $\mathbb{C}$-linear subspace of $\mathbb{C}^m$.*

For any $u \in \mathbb{C}^m$, we can write $u = \sum_{i=1}^k u_i$ with $u_i \in Q_i$ for each i. We are going to prove that if $u_1 \notin U(\omega)$, then

$$\lim_{t\to+\infty} \frac{1}{t} \log \left(\sum_{i=1}^k \|\Phi(t,\omega)u_i\|_2 \right) = \nu_1(Q_1, \omega).$$

First, by the definition of $U(\omega)$ and (14), we see that

$$\lim_{t\to+\infty} \frac{1}{t} \log \left(\|\Phi(t,\omega)u_1\|_2 \right) = \nu_1(Q_1, \omega). \tag{15}$$

On the other hand, by our assumption, we see that for any $i \in \{1, \dots, k\}$,

$$\lim_{t\to+\infty} \frac{1}{t} \log \left(\|\Phi(t,\omega)u_i\|_2 \right) \le \nu_1(Q_1, \omega).$$

Hence, fix any $\epsilon > 0$, there is a $T(\epsilon)$, such that when $t > T(\epsilon)$, we have that

$$\frac{1}{t}\log\|\Phi(t,\omega)u_i\|_2 \le \nu_1(Q_1,\omega) + \epsilon,$$

for any $i \in \{1,\dots,k\}$. So, when $t > T(\epsilon)$,

$$\sum_{i=1}^{k}\|\Phi(t,\omega)u_i\|_2 \le k \cdot exp\left(t(\nu_1(Q_1,\omega)+\epsilon)\right),$$

and then

$$\limsup_{t\to+\infty}\frac{1}{t}\log\left(\sum_{i=1}^{k}\|\Phi(t,\omega)u_i\|_2\right) \le \nu_1(Q_1,\omega)+\epsilon. \tag{16}$$

Because $\epsilon > 0$ is arbitrary, together with (15), we can conclude that

$$\lim_{t\to+\infty}\frac{1}{t}\log\left(\sum_{i=1}^{k}\|\Phi(t,\omega)u_i\|_2\right) = \nu_1(Q_1,\omega).$$

Finally, we see that there is a constant $C > 0$ only depending on the decomposition $\mathbb{C}^m = Q_1 \oplus \cdots \oplus Q_k$, such that for any $v \in \mathbb{C}^m$ uniquely written as $v = \sum_{i=1}^{k} v_i$ with $v_i \in Q_i$ for each i, we have that

$$C^{-1}\|v\|_2 \le \sum_{i=1}^{k}\|v_i\|_2 \le C\|v\|_2.$$

Then, if $u_1 \notin U(\omega)$, we conclude that

$$\begin{aligned}\lim_{t\to+\infty}\frac{\log\left(\sum_{i=1}^{k}\|\Phi(t,\omega)u_i\|_2\right)-\log C}{t} &\le \lim_{t\to+\infty}\frac{\log\|\Phi(t,\omega)u\|_2}{t}\\ &\le \lim_{t\to+\infty}\frac{\log\left(\sum_{i=1}^{k}\|\Phi(t,\omega)u_i\|_2\right)+\log C}{t},\end{aligned}$$

and in both sides, the limits are $\nu_1(Q_1,\omega) = \max_{1\le i\le k}\nu_1(Q_i,\omega)$. ∎

Next, because of Assumption 3 and the assumptions in Theorem 16, we can work with $(\Omega,\mathcal{F},\mathbb{P})$ that is the product probability space of $(\mathbb{R}^{d\times d},\mathcal{B}(\mathbb{R}^{d\times d}),\mathbb{P}_W)_{l\in\mathbb{N}}$, where $\mathcal{B}(\mathbb{R}^{d\times d})$ is the Borel σ-algebra of $\mathbb{R}^{d\times d}$. Specifically, $\Omega = (\mathbb{R}^{d\times d})^{\mathbb{N}}$ and a sample $\omega \in \Omega$ is a sequence of matrices in $\mathbb{R}^{d\times d}$, namely,

$$\omega = (W^{(0)}, W^{(1)}, \dots) \in \Omega. \tag{17}$$

We consider the metric dynamical system $(\Omega,\mathcal{F},\mathbb{P})$ induced by the shift operator $\theta : \Omega \to \Omega$ that maps a sample ω as in (17) to

$$\theta\omega = (W^{(1)}, W^{(2)}, \dots),$$

which is clearly $\mathbb{P}$-preserving. Additionally, $(\Omega, \mathcal{F}, \mathbb{P}, \theta)$ is ergodic by the Kolmogorov's zero-one law.

With the preparations above, Proposition 3 can be proved straightforwardly.

Proof [Proof of Proposition 3] (i) (5) is a direct corollary of Theorem 16 and the ergodicity of $(\Omega, \mathcal{F}, \mathbb{P}, \theta)$. In particular, $\log R(\mathbb{P}_W)$, where $R(\mathbb{P}_W)$ is the limit in (5), is the leading Lyapunov exponent of the linear random dynamical system induced by $A(\omega) = W$ with W being the first matrix in the sequence ω.
(ii) To prove (6), one needs to verify the condition (4) for $\mathrm{Id} + \beta W^{(t)}$ provided Assumption 3. We use $\mathbb{1}(\cdot)$ as the indicator function. Then, we see that

$$\begin{aligned}
\mathbb{E}\left[\max\left\{\log\|\mathrm{Id} + \beta W^{(t)}\|_2, 0\right\}\right] &\leq \mathbb{E}\left[\log\left(1 + |\beta|\|W^{(t)}\|_2\right)\right] \\
&\leq \mathbb{E}\left[\log(1+|\beta|)\mathbb{1}\left(\|W^{(t)}\|_2 \leq 1\right)\right] + \mathbb{E}\left[\log\left((1+|\beta|)\|W^{(t)}\|_2\right)\mathbb{1}\left(\|W^{(t)}\|_2 \geq 1\right)\right] \\
&\leq 2\log(1+|\beta|) + \mathbb{E}\left[\log\|W^{(t)}\|_2 \cdot \mathbb{1}\left(\|W^{(t)}\|_2 \geq 1\right)\right] \\
&= 2\log(1+|\beta|) + \mathbb{E}\left[\max\left\{\log\|W^{(t)}\|_2, 0\right\}\right] < \infty.
\end{aligned}$$

Therefore, one can apply Theorem 16 to the linear random dynamical system induced by $A(\omega) = I + \beta W$, with W being the first matrix in the sequence ω, and (6) follows directly from Theorem 16 and the ergodicity of $(\Omega, \mathcal{F}, \mathbb{P}, \theta)$. In particular, the largest Lyapunov exponent of this linear random dynamical system is $\log R(\beta, \mathbb{P}_W)$, where $R(\beta, \mathbb{P}_W)$ is the limit in (6). ■

Ergodic theory has appeared in analyses of oversmoothing, but for a different mathematical object than the one used in our work. Wu et al. (2023a) analyze attention-based GNNs using product of aggregation matrices, where a sequence of matrices $\{M^{(k)}\}$ is ergodic if $\lim_{t\to\infty} B\prod_{k=0}^{t} M^{(k)} = 0$, with B being the orthogonal projection onto the complement of $\mathrm{span}\{\mathbf{1}\}$. This is used to prove exponential oversmoothing upper bounds with joint spectral radius. By contrast, we rely on the multiplicative ergodic theorem (Theorem 16) for products of i.i.d. random matrices. This yields almost-sure asymptotic oversmoothing rates which is sharp in many cases. To the best of our knowledge, such Oseledets-type analysis has not appeared in prior GNN oversmoothing theory, which has used spectral properties of fixed graph operators (Oono and Suzuki, 2020; Cai and Wang, 2020) or ergodicity / joint spectral radius of time-varying aggregations (Wu et al., 2023a) rather than Lyapunov exponents of random matrix product.

3.2 Preliminary on Tensor Product Space

This section sets up more preparations for presenting the proofs of Theorem 4 and Theorem 5.

For a $y \in \mathbb{C}^d$ and a $\varphi \in \mathbb{C}^n$, we use the notation

$$y \otimes \varphi \coloneqq y\varphi^\top$$

to denote the $d \times n$ dimensional matrix $y\varphi^\top$, which can also be identified with $\mathbb{C}^{d\times n}$. On the other hand, any element in $\mathbb{C}^{d\times n}$ can be written as a finite linear combination of some

$y_i \otimes \varphi_i$'s for $y_i \in \mathbb{C}^d$ and $\varphi_i \in \mathbb{C}^n$. That is, for any $x \in \mathbb{C}^{d\times n}$, one can assume that x has a form of

$$x = \sum_i y_i \otimes \varphi_i.$$

We remark that, in general, the above linear combinations of $y_i \otimes \varphi_i$ may not be unique. To get a unique decomposition, we need to fix a basis of either $\mathbb{C}^d$ or $\mathbb{C}^n$.

Suppose that P has p Jordan blocks with eigenvlues $\lambda_1, \ldots, \lambda_p$ and sizes $n_1, \ldots, n_p$. By Remark 2, without loss of generalization we can assume that

$$1 = \lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_p|, \quad \text{and} \quad n_1 = 1.$$

Denote $V_1, \ldots, V_p \subseteq \mathbb{C}^n$ as the corresponding generalized left eigenspaces for λ_i of P. For each V_i, $i \in \{1, \ldots, p\}$, there exists a basis of V_i, $\varphi_{i,1}, \ldots, \varphi_{i,n_i}$, satisfying that

$$P\varphi_{i,j} = \lambda_i \varphi_{i,j} + \varphi_{i,j+1},\ j = 1, \ldots, n_i - 1, \quad \text{and} \quad P\varphi_{i,n_i} = \lambda_i \varphi_{i,n_i}. \tag{18}$$

In particular, according to Remark 2, we have that $V_1 = \mathrm{span}_{\mathbb{C}}\{\varphi_{1,1}\}$, with $\varphi_{1,1} = \mathbf{1}_n \coloneqq (1, \ldots, 1)^\top \in \mathbb{C}^n$. After fixing the basis $\{\varphi_{i,j}\}$ of $\mathbb{C}^n$, we have the decomposition

$$\mathbb{C}^{d\times n} = \mathbb{C}^d \otimes \mathbb{C}^n = (\mathbb{C}^d \otimes V_1) \oplus \cdots \oplus (\mathbb{C}^d \otimes V_p). \tag{19}$$

That is, for any $x \in \mathbb{C}^{d\times n}$, there is a unique series of $\{y_{i,j}\}$, such that

$$x = \sum_{i=1}^{p} \sum_{j=1}^{n_i} y_{i,j} \otimes \varphi_{i,j}. \tag{20}$$

For any $x \in \mathbb{C}^{d\times n}$ with x uniquely written in the form of (20), we can define a norm $\|x\|_D$ by

$$\|x\|_D \coloneqq \sum_{i=1}^{p} \sum_{j=1}^{n_i} \|y_{i,j}\|_2, \tag{21}$$

where the norm $\|y_{i,j}\|_2$ is the usual L^2-norm for a vector in $\mathbb{C}^d$. Recall that the Frobenius norm of x is defined by, if we write $x = (x_1, \ldots, x_n)$ for x_i's in $\mathbb{C}^d$, then

$$\|x\|_F \coloneqq \left(\sum_{i=1}^{n} \|x_i\|_2^2 \right)^{\frac{1}{2}},$$

where the norm $\|x_i\|_2$ is also the usual L^2-norm for a vector in $\mathbb{C}^d$. Because the correspondence between x and the series $\{y_{i,j}\}$ forms a linear isomorphism on $\mathbb{C}^{d\times n}$, we know that the norm $\|x\|_D$ is equivalent to the norm $\|x\|_F$. More precisely, we have the following proposition.

Proposition 19 *There is a constant $C_P > 0$ depending on the matrix P, such that for any $x \in \mathbb{C}^{d\times n}$, we have that*

$$\frac{1}{C_P}\|x\|_F \leq \|x\|_D \leq C_P \|x\|_F.$$

We omit the proof of this proposition. But for our purpose to estimate the vertex similarity measure as defined in Definition 1, we need the following definition and proposition.

Definition 20 *For any $x \in \mathbb{C}^{d\times n}$ uniquely written in the form of (20), we define*

$$\widetilde{x} \coloneqq x - y_{1,1} \otimes \varphi_{1,1} = \sum_{i=2}^{p} \sum_{j=1}^{n_i} y_{i,j} \otimes \varphi_{i,j}.$$

In Definition 1, we defined the quantity $\sum_{i=1}^{n} \|x_i - \bar{x}\|_2^2$ with $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n} x\varphi_{1,1} \in \mathbb{C}^d$. This quantity equals to $\|x - \bar{x} \otimes \varphi_{1,1}\|_F^2$. We remark that, in general, $y_{1,1}$ for x may not necessarily equal to $\bar{x}$, and hence $\|x - \bar{x} \otimes \varphi_{1,1}\|_F$ may not be the same as $\|\widetilde{x}\|_F$. On the other hand, these two quantities are equivalent in the following sense:

Proposition 21 *There is a constant $C_P > 0$ depending on the matrix P, such that for any $x \in \mathbb{C}^{d\times n}$,*

$$\frac{1}{C_P}\|\widetilde{x}\|_F \leq \|x - \bar{x} \otimes \varphi_{1,1}\|_F \leq C_P \|\widetilde{x}\|_F. \tag{22}$$

Proof We define a linear map T on $\mathbb{C}^{d\times n}$ by $T(x) \coloneqq x - \bar{x} \otimes \varphi_{1,1}$. Then we see that the kernel of T is $\mathrm{Ker}(T) = \mathbb{C}^d \otimes V_1$. Hence, T induces a linear isomorphism from $\mathbb{C}^{d\times n}/\mathrm{Ker}(T)$ to $\mathrm{Im}(T)$. Because we have the unique decomposition (19) and (20), the norm $\|\widetilde{x}\|_F$ can induce a norm on the quotient space $\mathbb{C}^{d\times n}/\mathrm{Ker}(T)$. Hence, $\|\widetilde{x}\|_F$ is equivalent to $\|T(x)\|_F$ up to a constant depending on T. This finishes the proof of (22). ■

3.3 Proof of Theorem 4

This section proves Theorem 4.

3.3.1 Random GNN dynamical system on tensor product space

We consider a linear random dynamical system $\Phi(t,\omega)$ on $\mathbb{C}^{d\times n}$ over $(\Omega, \mathcal{F}, \mathbb{P}, \theta)$, induced by

$$\begin{aligned} A(\omega): \qquad \mathbb{C}^{d\times n} = \mathbb{C}^d \otimes \mathbb{C}^n \qquad &\to \qquad \mathbb{C}^{d\times n} = \mathbb{C}^d \otimes \mathbb{C}^n, \\ x = \sum_{i=1}^{p}\sum_{j=1}^{n_i} y_{i,j} \otimes \varphi_{i,j} &\mapsto \sum_{i=1}^{p}\sum_{j=1}^{n_i} (W y_{i,j}) \otimes (P\varphi_{i,j}), \end{aligned} \tag{23}$$

where W is the first matrix in the sequence ω. This linear map is well-defined, as if we regard x as a $d \times n$ dimensional matrix, then $A(\omega)x = WxP^\top$. Then, we see that the dynamics (nrs) with σ being the identity map is precisely driven by $\Phi(t,w)$. More precisely, if we write that, by abusing our notations,

$$x_{\mathrm{nrs}}(0) = \sum_{i=1}^{p}\sum_{j=1}^{n_i} y_{i,j}(0) \otimes \varphi_{i,j}, \tag{24}$$

then for $t \in \mathbb{N}$,

$$\begin{aligned} x_{\mathrm{nrs}}(t) &= \Phi(t,\omega)x(0) = A(\theta^{t-1}\omega)\cdots A(\theta\omega)A(\omega)x(0) \\ &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(W^{(t-1)}\cdots W^{(1)}W^{(0)}y_{i,j}(0)\right)\otimes P^t\varphi_{i,j}. \end{aligned} \tag{25}$$

For notation purpose, we let $W^{(-1)}$ be the identity matrix.

By the definition of the action of $A(\omega)$ in (23) and (18), one can easily see that $\mathbb{C}^d \otimes V_i$ is an invariant subspace of $A(\omega)$ and $\Phi(t,\omega)$ for any $i \in \{1,\ldots,p\}$ and any $t \in \mathbb{N}, \omega \in \Omega$. Furthermore, by induction, one can show that for each $i \in \{1,\ldots,p\}, j \in \{1,\ldots,n_i\}$,

$$P^t\varphi_{i,j} = \sum_{s=0}^{n_i-j}\lambda_i^{t-s}\binom{t}{s}\varphi_{i,j+s}. \tag{26}$$

Hence, we see that (25) becomes

$$x_{\mathrm{nrs}}(t) = \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(\sum_{s=1}^{j}\lambda_i^{t-(j-s)}\binom{t}{j-s}W^{(t-1)}\cdots W^{(1)}W^{(0)}y_{i,s}(0)\right)\otimes\varphi_{i,j}. \tag{27}$$

To simplify our notations, we define

$$y_{i,j}(t) \coloneqq \sum_{s=1}^{j}\lambda_i^{t-(j-s)}\binom{t}{j-s}W^{(t-1)}\cdots W^{(1)}W^{(0)}y_{i,s}(0). \tag{28}$$

For notation purpose, we let $\binom{0}{i} = 0$ if $i > 0$ and let $\binom{j}{0} = 1$ if $j \geq 0$.

The forms of (27) and (28) suggests us to consider the following linear random dynamical system on $\mathbb{C}^d$: for any $y \in \mathbb{C}^d$, we define $\Phi_W(t,\omega)y$ by

$$\Phi_W(t,\omega)y \coloneqq W^{(t-1)}\cdots W^{(1)}W^{(0)}y.$$

Theorem 16 applies for $\Phi_W(t,\omega)$ since (4) is assumed. Denote $\nu_1^W > \cdots > \nu_{q^W}^W \geq -\infty$ as the ordered Lyapunov exponents of $\Phi_W(t,\omega)$ on $\mathbb{C}^d$ as in Theorem 16, which are almost surely constants on Ω according to the part (v) of Theorem 16 since $(\Omega,\mathcal{F},\mathbb{P},\theta)$ is ergodic. We then have that for any $\omega \in \widetilde{\Omega}$,

$$\lim_{t\to+\infty}\frac{1}{t}\log\|\Phi_W(t,\omega)y\|_2 = \nu_i^W, \quad \forall\, y \in U_i^W(\omega)\backslash U_{i+1}^W(\omega), \quad i \in \{1,\ldots,q^W\}. \tag{29}$$

In particular, we see that $U_1^W = \mathbb{C}^d$.

Next, similar to Definition 20, we can define

$$\widetilde{x}_{\mathrm{nrs}}(t) \coloneqq \widetilde{x_{\mathrm{nrs}}(t)} = \sum_{i=2}^{p}\sum_{j=1}^{n_i}y_{i,j}(t)\otimes\varphi_{i,j}. \tag{30}$$

We remark that under this definition, we have that $\widetilde{x}_{\mathrm{nrs}}(t) = \Phi(t,\omega)(\widetilde{x_{\mathrm{nrs}}(0)})$. We have the following lemma for $\|\widetilde{x}_{\mathrm{nrs}}(t)\|_F$:

Lemma 22 *Fix an $\omega \in \widetilde{\Omega}$. For any $x(0)$ in the form of (24), if $y_{2,1}(0) \in U_1^W(\omega) \backslash U_2^W(\omega) = \mathbb{C}^d \backslash U_2^W(\omega)$, then for the corresponding $x(t)$ in (27), we have that*

$$\lim_{t\to+\infty} \frac{1}{t} \log \|\widetilde{x}_{\mathrm{nrs}}(t)\|_F = \log|\lambda_2| + \nu_1^W. \tag{31}$$

Proof We are going to prove that $\lim_{t\to+\infty} \frac{1}{t} \log \|\widetilde{x}_{\mathrm{nrs}}(t)\|_D = \log|\lambda_2| + \nu_1^W$ first. According to (21), (28), and (30), we see that

$$\|\widetilde{x}_{\mathrm{nrs}}(t)\|_D = \sum_{i=2}^{p} \sum_{j=1}^{n_i} \|y_{i,j}(t)\|_2 = |\lambda_2|^t \|\Phi_W(t,\omega) y_{2,1}(0)\|_2 + \sum_{j=2}^{n_2} \|y_{2,j}(t)\|_2 + \sum_{i=3}^{p} \sum_{j=1}^{n_i} \|y_{i,j}(t)\|_2.$$

Because we assumed that $y_{2,1}(0) \in \mathbb{C}^d \backslash U_2^W(\omega)$ and by (29), we first see that

$$\lim_{t\to+\infty} \frac{1}{t} \log\left(|\lambda_2|^t \|\Phi_W(t,\omega) y_{2,1}(0)\|_2\right) = \log|\lambda_2| + \nu_1^W. \tag{32}$$

On the other hand, (29) also implies that $\lim_{t\to+\infty} \frac{1}{t} \log \|\Phi_W(t,\omega) y\|_2 \leq \nu_1^W$ for any $y \in \mathbb{C}^d$. Fix any $\epsilon > 0$, there is a $T(\epsilon)$, such that when $t > T(\epsilon)$, we have that

$$\frac{1}{t} \log \|\Phi_W(t,\omega) y_{i,j}(0)\|_2 \leq \nu_1^W + \epsilon,$$

for any $i \in \{2, \ldots, p\}$ and $j \in \{1, \ldots, n_i\}$. By (28), we see that for any $i \in \{2, \ldots, p\}$ and $j \in \{1, \ldots, n_i\}$,

$$\|y_{i,j}(t)\|_2 \leq \sum_{s=1}^{j} |\lambda_i|^{t-(j-s)} \binom{t}{j-s} \|\Phi_W(t,\omega) y_{i,s}(0)\|_2 \leq n|\lambda_2|^t t^n \exp\left(t(\nu_1^W + \epsilon)\right).$$

So, when $t > T(\epsilon)$,

$$\|\widetilde{x}_{\mathrm{nrs}}(t)\|_D \leq n^2 |\lambda_2|^t t^n \exp\left(t(\nu_1^W + \epsilon)\right),$$

and then

$$\limsup_{t\to+\infty} \frac{1}{t} \log \|\widetilde{x}_{\mathrm{nrs}}(t)\|_D \leq \log|\lambda_2| + \nu_1^W + \epsilon. \tag{33}$$

Because $\epsilon > 0$ is arbitrary, together with (32), we can conclude that

$$\lim_{t\to+\infty} \frac{1}{t} \log \|\widetilde{x}_{\mathrm{nrs}}(t)\|_D = \log|\lambda_2| + \nu_1^W.$$

Then, by Proposition 19, we see that

$$\lim_{t\to+\infty} \frac{\log \|\widetilde{x}_{\mathrm{nrs}}(t)\|_D - \log C_P}{t} \leq \lim_{t\to+\infty} \frac{1}{t} \log \|\widetilde{x}_{\mathrm{nrs}}(t)\|_F \leq \lim_{t\to+\infty} \frac{\log \|\widetilde{x}_{\mathrm{nrs}}(t)\|_D + \log C_P}{t},$$

and in both sides, the limits are $\log|\lambda_2| + \nu_1^W$. ■

Similar to the proof for Lemma 22, one can directly get the following lemma for $\|x_{\mathrm{nrs}}(t)\|_F$ and $\mu(x_{\mathrm{nrs}}(t))$.

Lemma 23 *Fix an $\omega \in \widetilde{\Omega}$. For any $x_{\mathrm{nrs}}(0)$ in the form of (24), if $y_{1,1}(0) \in U_1^W(\omega) \backslash U_2^W(\omega) = \mathbb{C}^d \backslash U_2^W(\omega)$, then for the corresponding $x_{\mathrm{nrs}}(t)$ in (27), we have that*

$$\lim_{t\to+\infty} \frac{1}{t} \log \|x_{\mathrm{nrs}}(t)\|_F = \log|\lambda_1| + \nu_1^W = \nu_1^W. \tag{34}$$

If we further assume that $y_{2,1}(0) \in \mathbb{C}^d \backslash U_2^W(\omega)$ and $\nu_1^W > -\infty$, then

$$\lim_{t\to+\infty} \frac{1}{2t} \log \mu(x_{\mathrm{nrs}}(t)) = \log|\lambda_2|. \tag{35}$$

Proof The proof for (34) is the same as the proof for Lemma 22. For (35), we notice that

$$\mu(x_{\mathrm{nrs}}(t)) = \frac{\|x_{\mathrm{nrs}}(t) - \overline{x_{\mathrm{nrs}}(t)} \otimes \varphi_{1,1}\|_F^2}{\|x_{\mathrm{nrs}}(t)\|_F^2},$$

as discussed after Definition 20. By Proposition 21, we see that

$$\begin{aligned} \lim_{t\to+\infty} \frac{\log \|\widetilde{x}_{\mathrm{nrs}}(t)\|_F - \log C_P}{t} &\le \lim_{t\to+\infty} \frac{1}{t} \log \|x_{\mathrm{nrs}}(t) - \overline{x_{\mathrm{nrs}}(t)} \otimes \varphi_{1,1}\|_F \\ &\le \lim_{t\to+\infty} \frac{\log \|\widetilde{x}_{\mathrm{nrs}}(t)\|_F + \log C_P}{t}, \end{aligned}$$

and in both sides, the limits are $\log|\lambda_2| + \nu_1^W$ by Lemma 22. Hence, we can conclude (35). ■

3.3.2 Proof of Theorem 4

We notice that Lemma 23 with $\nu_1^W = \log R(\mathbb{P}_W)$ already implies Theorem 4 for almost every $x_{\mathrm{nrs}}(0) \in \mathbb{C}^{d\times n}$, because if we use the form (20), for any fixed $\omega \in \widetilde{\Omega}$, the set

$$\left\{ x \in \mathbb{C}^{d\times n} \mid y_{1,1} \in U_2^W(\omega) \text{ or } y_{2,1} \in U_2^W(\omega) \right\} \tag{36}$$

is the union of two proper $\mathbb{C}$-linear subspaces of $\mathbb{C}^{d\times n}$, and hence has zero Lebesgue measure in $\mathbb{C}^{d\times n}$. We can also get the conclusion for almost every $x(0) \in \mathbb{R}^{d\times n}$ once we have the following fact:

Lemma 24 *Assume that $U \subsetneq \mathbb{C}^{d\times n}$ is a proper $\mathbb{C}$-linear subspace of $\mathbb{C}^{d\times n}$, then $U \cap \mathbb{R}^{d\times n}$ is a proper $\mathbb{R}$-linear subspace of $\mathbb{R}^{d\times n}$.*

Proof Let $\{e_i\}_{i=1}^{d\times n}$ be the standard coordinates vectors of $\mathbb{R}^{d\times n}$, that is, 1 on the i-th entry and 0 on the other entries. If $\mathbb{R}^{d\times n} = U \cap \mathbb{R}^{d\times n}$, then we have that $\{e_i\}_{i=1}^{d\times n} \subseteq U$. As a consequence,

$$\mathbb{C}^{d\times n} = \mathrm{span}_{\mathbb{C}}\{e_i\}_{i=1}^{d\times n} \subseteq U, \tag{37}$$

which is a contradiction. ■

3.4 Proof of Theorem 5

This section proves Theorem 5.

3.4.1 Random residual GNN dynamical system on tensor product space

The linear random dynamical system $\Phi(t,\omega)$ associated to (rs) is defined on $\mathbb{C}^{d\times n}$ over $(\Omega,\mathcal{F},\mathbb{P},\theta)$, induced by

$$\begin{aligned} A(\omega): \qquad \mathbb{C}^{d\times n}=\mathbb{C}^d\otimes\mathbb{C}^n \qquad &\to \qquad \mathbb{C}^{d\times n}=\mathbb{C}^d\otimes\mathbb{C}^n,\\ x=\sum_{i=1}^{p}\sum_{j=1}^{n_i} y_{i,j}\otimes\varphi_{i,j} &\mapsto x+\alpha\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Wy_{i,j})\otimes(P\varphi_{i,j}), \end{aligned} \tag{38}$$

where we still used the unique decomposition (20) for any $x\in\mathbb{C}^{d\times n}$, and W is the first matrix in the sequence ω. This linear map is well-defined, as if we regard x as a $d\times n$ dimensional matrix, then $A(\omega)x=x+\alpha WxP^\top$. Then, we see that the dynamics (rs) with σ being the identity map is precisely driven by $\Phi(t,w)$. For any $x_{\rm rs}(0)\in\mathbb{C}^{d\times n}$, $x_{\rm rs}(t)=\Phi(t,\omega)x_{\rm rs}(0)=A(\theta^{t-1}\omega)\cdots A(\theta\omega)A(\omega)x_{\rm rs}(0)$ in the dynamics (rs). For notation purpose, we write

$$x_{\rm rs}(t)=\sum_{i=1}^{p}\sum_{j=1}^{n_i} y_{i,j}(t)\otimes\varphi_{i,j},\quad x_{\rm rs}(0)=\sum_{i=1}^{p}\sum_{j=1}^{n_i} y_{i,j}(0)\otimes\varphi_{i,j}. \tag{39}$$

By the definition of the action of $A(\omega)$ in (38) and (18), one can easily see that $\mathbb{C}^d\otimes V_i$ is an invariant subspace of $A(\omega)$ and $\Phi(t,\omega)$ for any $i\in\{1,\dots,p\}$ and any $t\in\mathbb{N},\omega\in\Omega$. Hence, for each $i\in\{1,\dots,p\}$, let $\nu_1(\mathbb{C}^d\otimes V_i)$ be the largest Lyapunov exponent of $\Phi(t,\omega)$ restricted on $\mathbb{C}^d\otimes V_i$, which is a constant for $\omega\in\widetilde{\Omega}$ by ergodicity and part (v) of Theorem 16.

Next, similar to (30), we still define

$$\widetilde{x}_{\rm rs}(t)\coloneqq\widetilde{x_{\rm rs}(t)}=\sum_{i=2}^{p}\sum_{j=1}^{n_i} y_{i,j}(t)\otimes\varphi_{i,j}. \tag{40}$$

We remark that under this definition, we have that $\widetilde{x}_{\rm rs}(t)=\Phi(t,\omega)\widetilde{(x_{\rm rs}(0))}$.

Lemma 25 *Fix an $\omega\in\widetilde{\Omega}$. There is a set $U(\omega)$ consisting of the union of two proper $\mathbb{C}$-linear subspace of $\mathbb{C}^{d\times n}$, such that for any $x_{\rm rs}(0)\in\mathbb{C}^{d\times n}\backslash U(\omega)$, we have that*

$$\lim_{t\to+\infty}\frac{1}{t}\log\|x_{\rm rs}(t)\|_F=\max_{1\le i\le p}\nu_1(\mathbb{C}^d\otimes V_i),\quad \lim_{t\to+\infty}\frac{1}{t}\log\|\widetilde{x}_{\rm rs}(t)\|_F=\max_{2\le i\le p}\nu_1(\mathbb{C}^d\otimes V_i). \tag{41}$$

As a corollary, if $\max_{1\le i\le p}\nu_1(\mathbb{C}^d\otimes V_i)>-\infty$,

$$\lim_{t\to+\infty}\frac{1}{2t}\log\mu(x_{\rm rs}(t))=\min\left\{\max_{2\le i\le p}\nu_1(\mathbb{C}^d\otimes V_i)-\nu_1(\mathbb{C}^d\otimes V_1),0\right\}. \tag{42}$$

Proof (41) is a direct application of Proposition 18, and the proof of (42) is the same as the proof of (35) in Lemma 23. ■

3.4.2 Proof of Theorem 5

Now, we are going to use Lemma 25 and (38) to estimate $\nu_1(\mathbb{C}^d \otimes V_i)$'s for each $i \in \{1, \ldots, p\}$. Again, assume that $x(t)$ has the form (39).

First, (18) gives that $P\varphi_{i,n_i} = \lambda_i \varphi_{i,n_i}$. Hence, $A(\omega)(y_{i,n_i} \otimes \varphi_{i,n_i}) = ((\mathrm{Id} + \alpha\lambda_i W) y_{i,n_i}) \otimes \varphi_{i,n_i}$. Hence, apply Theorem 16 and Proposition 3 (ii) to the linear random dynamic system $A_i(\omega)y = (\mathrm{Id} + \alpha\lambda_i W)y$, we see that there is a proper $\mathbb{C}$-linear subspace $U_i(\omega)$ of $\mathbb{C}^d$, such that if $y_{i,n_i}(0) \in \mathbb{C}^d \backslash U_i(\omega)$, then

$$\begin{aligned} &\lim_{t\to+\infty} \frac{1}{t} \log \|y_{i,n_i}(t) \otimes \varphi_{i,n_i}\|_F = \lim_{t\to+\infty} \frac{1}{t} \log \|y_{i,n_i}(t) \otimes \varphi_{i,n_i}\|_D \\ &= \lim_{t\to+\infty} \frac{1}{t} \log \left\| \left(\mathrm{Id} + \alpha\lambda_i W^{(t-1)}\right) \cdots \left(\mathrm{Id} + \alpha\lambda_i W^{(1)}\right) \left(\mathrm{Id} + \alpha\lambda_i W^{(0)}\right) y_{i,n_i}(0) \right\|_2 \\ &= \log R(\alpha\lambda_i, \mathbb{P}_W), \end{aligned} \tag{43}$$

where the first equality is by Proposition 19 and the last equality is by Proposition 3 (ii). Because for any $i \in \{1, \ldots, p\}$ and any $t \in \mathbb{N}, \omega \in \Omega$, $\mathbb{C}^d \otimes V_i$ is an invariant subspace of $A(\omega)$ and $\Phi(t, \omega)$, we see that $\nu_1(\mathbb{C}^d \otimes V_i) \geq \log R(\alpha\lambda_i, \mathbb{P}_W)$, and if $n_i = 1$, we have that $\nu_1(\mathbb{C}^d \otimes V_i) = \log R(\alpha\lambda_i, \mathbb{P}_W)$.

Hence, by (42), because $\lambda_1 = 1$, we see that

$$\begin{aligned} \lim_{t\to+\infty} \frac{1}{2t} \log \mu(x_{\mathrm{rs}}(t)) &= \min \left\{ \max_{2\leq i\leq p} \nu_1(\mathbb{C}^d \otimes V_i) - \nu_1(\mathbb{C}^d \otimes V_1), 0 \right\} \\ &= \min \left\{ \max_{2\leq i\leq p} \nu_1(\mathbb{C}^d \otimes V_i) - \log R(\alpha, \mathbb{P}_W), 0 \right\} \\ &\geq \min \left\{ \max_{2\leq i\leq p} \log R(\alpha\lambda_i, \mathbb{P}_W) - \log R(\alpha, \mathbb{P}_W), 0 \right\}, \end{aligned}$$

where the last inequality is an equality if $n_i = 1$ for all $i \in \{1, \ldots, p\}$, that is, if P is diagonalizable in $\mathbb{C}$. This finishes the proof of Theorem 5.

3.5 Proofs for Section 2.2

This section collects proofs for results in Section 2.2.

3.5.1 Proofs of Theorem 6 and Theorem 7

Proof [Proof of Theorem 6] Again, because of (38), we see that $\mathbb{C}^d \otimes V_i$'s are invariant subspaces of the action $A : x \mapsto x + \alpha W x P^\top$ for any $i \in \{1, \ldots, p\}$. According to Lemma 25, we only need to estimate $\nu_1(\mathbb{C}^d \otimes V_i)$ for each $i \in \{1, \ldots, p\}$. We fix an $i \in \{1, \ldots, p\}$.

Hence, as in (39), we assume that

$$x_{\mathrm{rs}}(t) = \sum_{j=1}^{n_i} y_{i,j}(t) \otimes \varphi_{i,j}, \quad x_{\mathrm{rs}}(0) = \sum_{j=1}^{n_i} y_{i,j}(0) \otimes \varphi_{i,j}.$$

We can explicitly compute $y_{i,j}(t)$'s in the following: by induction, one can first show that

$$x_{\mathrm{rs}}(t) = \sum_{k=0}^{t} \binom{t}{k} \alpha^k W^k x_{\mathrm{rs}}(0) (P^k)^\top = \sum_{k=0}^{t} \binom{t}{k} \alpha^k \sum_{j=1}^{n_i} (W^k y_{i,j}(0)) \otimes (P^k \varphi_{i,j}).$$

By (26), we see that the right hand side of the above equation equals to

$$\begin{aligned}
&\sum_{k=0}^{t}\binom{t}{k}\alpha^k\sum_{j=1}^{n_i}\left[(W^k y_{i,j}(0))\otimes\sum_{s=j}^{n_i}\lambda_i^{k-(s-j)}\binom{k}{s-j}\varphi_{i,s}\right]\\
&\quad=\sum_{k=0}^{t}\binom{t}{k}\alpha^k\sum_{s=1}^{n_i}\left[\sum_{j=1}^{s}\lambda_i^{k-(s-j)}\binom{k}{s-j}W^k y_{i,j}(0)\right]\otimes\varphi_{i,s}\\
&\quad=\sum_{s=1}^{n_i}\left[\sum_{j=1}^{s}\sum_{k=0}^{t}\binom{t}{k}\alpha^k\lambda_i^{k-(s-j)}\binom{k}{s-j}W^k y_{i,j}(0)\right]\otimes\varphi_{i,s}\\
&\quad=\sum_{s=1}^{n_i}\left[\sum_{j=1}^{s}\sum_{k=s-j}^{t}\binom{t}{k}\alpha^k\lambda_i^{k-(s-j)}\binom{k}{s-j}W^k y_{i,j}(0)\right]\otimes\varphi_{i,s},
\end{aligned}$$

where we used the notation that $\binom{a}{b}=0$ if $b<a$. We notice that

$$\binom{t}{k}\binom{k}{s-j}=\frac{t!}{(t-k)!k!}\frac{k!}{(k-(s-j))!(s-j)!}=\binom{t}{s-j}\binom{t-(s-j)}{k-(s-j)}.$$

So,

$$\begin{aligned}
y_{i,s}(t)&=\sum_{j=1}^{s}\sum_{k=s-j}^{t}\binom{t}{k}\alpha^k\lambda_i^{k-(s-j)}\binom{k}{s-j}W^k y_{i,j}(0)\\
&=\sum_{j=1}^{s}(\alpha W)^{s-j}\binom{t}{s-j}\sum_{k=s-j}^{t}\binom{t-(s-j)}{k-(s-j)}(\alpha\lambda_i W)^{k-(s-j)}y_{i,j}(0)\\
&=\sum_{j=1}^{s}(\alpha W)^{s-j}\binom{t}{s-j}(\mathrm{Id}+\alpha\lambda_i W)^{t-(s-j)}y_{i,j}(0).
\end{aligned}$$

Hence, by (21), we see that

$$\begin{aligned}
\|x_{\mathrm{rs}}(t)\|_D&\le\sum_{s=1}^{n_i}\|y_{i,s}(t)\|_2\le n_i t^{n_i}\max_{1\le s\le n_i}\sum_{j=1}^{s}\|\alpha W\|_2^{s-j}\|(\mathrm{Id}+\alpha\lambda_i W)^{t-(s-j)}\|_2\|y_{i,j}(0)\|_2\\
&\le C(n_i,\alpha,\lambda,W,\{\|y_{i,j}(0)\|_2\})\cdot t^{n_i}\|(\mathrm{Id}+\alpha\lambda_i W)^{t-n_i}\|_2,
\end{aligned}$$

where we used $C(n_i,\alpha,\lambda,W,\{\|y_{i,j}(0)\|_2\})$ to denote a constant depending on $n_i,\alpha,\lambda,\|W\|_2$, and $\{\|y_{i,j}(0)\|_2\}$'s, which is independent of t. Hence,

$$\begin{aligned}
\nu_1(\mathbb{C}^d\otimes V_i)&=\limsup_{t\to+\infty}\frac{1}{t}\log\|x(t)\|_D\le\limsup_{t\to+\infty}\frac{1}{t}\log\|(\mathrm{Id}+\alpha\lambda_i W)^{t-n_i}\|_2\\
&=\log\rho(\mathrm{Id}+\alpha\lambda_i W),
\end{aligned}$$

where the last equality is because of Gelfand's formula. Together with the proof for Theorem 5, we see that for each $i\in\{1,\dots,p\}$, $\nu_1(\mathbb{C}^d\otimes V_i)=\rho(\mathrm{Id}+\alpha\lambda_i W)$. This fact together

with Lemma 25 finishes the proof. ■

Proof [Proof of Theorem 7] (i) Take $\lambda_0 \in \mathrm{spec}(P)\backslash\{1\} \subseteq (-1,1)$ achieving $\max_{\lambda\in\mathrm{spec}(P)\backslash\{1\}} |\lambda|$. It follows from $1+\lambda_0(2\alpha\mathrm{Re}\mu_0+1)>0$ that

$$(1+\alpha\lambda_0\mathrm{Re}(\mu_0))^2-(\lambda_0+\alpha\lambda_0\mathrm{Re}(\mu_0))^2=(1-\lambda_0)(1+\lambda_0(2\alpha\mathrm{Re}\mu_0+1))>0,$$

which implies that

$$\begin{aligned}\rho(I+\alpha\lambda_0 W) &\geq |1+\alpha_0\lambda_0\mu_0| = \left((1+\alpha\lambda_0\mathrm{Re}(\mu_0))^2+(\alpha\lambda_0\mathrm{Im}(\mu_0))^2\right)^{1/2}\\ &> \left((\lambda_0+\alpha\lambda_0\mathrm{Re}(\mu_0))^2+(\alpha\lambda_0\mathrm{Im}(\mu_0))^2\right)^{1/2} = \lambda_0\cdot|1+\alpha\mu_0| = \lambda_0\cdot\rho(I+\alpha W),\end{aligned}$$

where the inequality is guaranteed by $\alpha\mathrm{Re}(\mu_0)+1\geq 0$. Therefore, if

$$\max_{\lambda\in\mathrm{spec}(P)\backslash\{1\}}\rho(I+\alpha\lambda W)\geq\rho(I+\alpha W),$$

then,

$$\begin{aligned}\lim_{t\to+\infty}\mu(x_{\mathrm{rs}}(t))^{1/2t} &\geq \frac{\max_{\lambda\in\mathrm{spec}(P)\backslash\{1\}}\rho(I+\alpha\lambda W)}{\max_{\lambda\in\mathrm{spec}(P)}\rho(I+\alpha\lambda W)}=1\\ &> \max_{\lambda\in\mathrm{spec}(P)\backslash\{1\}}|\lambda| = \lim_{t\to+\infty}\mu(x_{\mathrm{nrs}}(t))^{1/2t}.\end{aligned}$$

Otherwise,

$$\begin{aligned}\lim_{t\to+\infty}\mu(x_{\mathrm{rs}}(t))^{1/2t} &\geq \frac{\max_{\lambda\in\mathrm{spec}(P)\backslash\{1\}}\rho(I+\alpha\lambda W)}{\max_{\lambda\in\mathrm{spec}(P)}\rho(I+\alpha\lambda W)}\geq\frac{\rho(I+\alpha\lambda_0 W)}{\rho(I+\alpha W)}\\ &> \lambda_0=\max_{\lambda\in\mathrm{spec}(P)\backslash\{1\}}|\lambda| = \lim_{t\to+\infty}\mu(x_{\mathrm{nrs}}(t))^{1/2t}.\end{aligned}$$

(ii) Take a $\lambda\in\mathrm{spec}(P)\backslash\{1\}\subseteq(-1,1)$ with $2\mathrm{Re}\mu_0+\alpha(\lambda+1)|\mu_0|^2\leq 0$, we see that

$$\begin{aligned}&\rho(I+\alpha\lambda W)^2-\rho(I+\alpha W)^2\\ &\quad\geq\left((1+\alpha\lambda\mathrm{Re}(\mu_0))^2+(\alpha\lambda\mathrm{Im}(\mu_0))^2\right)-\left((1+\alpha\mathrm{Re}(\mu_0))^2+(\alpha\mathrm{Im}(\mu_0))^2\right)\\ &\quad=2\alpha(\lambda-1)\mathrm{Re}\mu_0+\alpha^2(\lambda^2-1)|\mu_0|^2=\alpha(\lambda-1)\left(2\mathrm{Re}\mu_0+\alpha(\lambda+1)|\mu_0|^2\right)\geq 0,\end{aligned}$$

which implies that

$$\rho(I+\alpha\lambda W)\geq\rho(I+\alpha W).$$

We can thus conclude that

$$\lim_{t\to+\infty}\mu(x_{\mathrm{rs}}(t))^{1/2t}\geq\frac{\max_{\lambda\in\mathrm{spec}(P)\backslash\{1\}}\rho(I+\alpha\lambda W)}{\max_{\lambda\in\mathrm{spec}(P)}\rho(I+\alpha\lambda W)}=1>\max_{\lambda\in\mathrm{spec}(P)\backslash\{1\}}|\lambda|,$$

which completes the proof together with Theorem 4. ■

3.5.2 Proofs of Theorem 8 and Theorem 9

We first present a proposition that implies Theorem 8 immediately.

Proposition 26 *Assume that Assumption 3 holds and in addition that $W^{(t)}$'s entries are drawn i.i.d. from $\mathcal{N}(0,\tau^2)$. Then, for $\beta \in \mathbb{R}$, the $R(\beta, \mathbb{P}_W)$ defined in Proposition 3 (ii) satisfies that*

$$\log R(\beta, \mathbb{P}_W) = \frac{1}{2}\mathbb{E}\log\left((1+\beta\tau\xi)^2 + \beta^2\tau^2\chi^2_{d-1}\right),$$

where $\xi \sim \mathcal{N}(0,1)$ is a 1-dimensional standard real Gaussian distribution, and χ^2_{d-1} is a chi-squared distribution with $d-1$ degrees of freedom that is independent of ξ.

Proof [Proof of Theorem 8] Theorem 8 is a direct corollary of Theorem 5 and Proposition 26. ■

We then present the proof of Proposition 26, with the main tool being Proposition 2.1 in Cohen and Newman (1984). For the convenience of the readers, we adopt the statement of this proposition in the settings of our Proposition 3 (ii).

Theorem 27 (Cohen and Newman (1984, Proposition 2.1)) *Suppose Assumption 3 holds. If $\beta \in \mathbb{R}$ and $\|(\mathrm{Id} + \beta W^{(t)})x\|_2$ has a distribution which does not depend on $x \in \mathbb{S}^{d-1} \subseteq \mathbb{R}^d$, where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$, then*

$$\log R(\beta, \mathbb{P}_W) = \mathbb{E}\log\|(\mathrm{Id} + \beta W^{(t)})e_1\|_2,$$

where $e_1 = (1, 0, \ldots, 0)^\top \in \mathbb{R}^d$.

Proof [Proof of Proposition 26] By Theorem 27, we need to verify that $\|\mathrm{Id} + \beta W^{(t)}x\|_2$ has a distribution which does not depend on $x \in \mathbb{S}^{d-1}$. We notice that for any $x \in \mathbb{S}^{d-1}$, there is an orthonormal matrix $Q \in \mathbb{R}^{d\times d}$ such that $x = Qe_1$, where $e_1 = (1, 0, \ldots, 0)^\top \in \mathbb{S}^{d-1}$. We are going to show that the entries of $W^{(t)}Q$, which we denote as $\{Z^{(t)}_{ij}\}$'s, are also i.i.d. Gaussian distributions $\mathcal{N}(0,\tau^2)$. By definition, we see that $Z^{(t)}_{ij} = \sum_{k=1}^d W^{(t)}_{ik}Q_{kj}$, which is a linear combination of Gaussian distributions, and hence $Z^{(t)}_{ij}$ is also a Gaussian distribution. Also, for $(i_1, j_1), (i_2, j_2) \in \{1, \ldots, d\} \times \{1, \ldots, d\}$, we have that

$$\mathbb{E}\left[Z^{(t)}_{i_1j_1}Z^{(t)}_{i_2j_2}\Big)\right] = \mathbb{E}\left[\sum_{k,l=1}^d W^{(t)}_{i_1k}Q_{kj_1}W^{(t)}_{i_2l}Q_{lj_2}\right] = \sum_{k,l=1}^d Q_{kj_1}Q_{lj_2}\mathbb{E}\left[W^{(t)}_{i_1k}W^{(t)}_{i_2l}\right].$$

If $i_1 \neq i_2$, then $\mathbb{E}\left[W^{(t)}_{i_1k}W^{(t)}_{i_2l}\right] = 0$ for any $k, l \in \{1, \ldots, d\}$ because $W^{(t)}_{i_1k}$ is independent of $W^{(t)}_{i_2l}$. If $i_1 = i_2$, then $\mathbb{E}\left[W^{(t)}_{i_1k}W^{(t)}_{i_2l}\right] = \tau^2\delta_{kl}$, and then $\mathbb{E}\left[Z^{(t)}_{i_1j_1}Z^{(t)}_{i_1j_2}\right] = \tau^2\sum_{k,l}Q_{kj_1}Q_{lj_2}\delta_{kl} = \tau^2\sum_k Q_{kj_1}Q_{kj_2}$. Because Q is an orthonormal matrix, its columns consist of an orthonormal basis of $\mathbb{R}^d$. Hence, $\mathbb{E}\left[Z^{(t)}_{i_1j_1}Z^{(t)}_{i_1j_2}\right] = \tau^2\sum_k Q_{kj_1}Q_{kj_2} = \tau^2\delta_{j_1j_2}$. We can then conclude

that the entries of $W^{(t)}Q$ are also i.i.d. standard Gaussian distributions $\mathcal{N}(0,\tau^2)$. Repeating the same arguments yields that the entries of $Q^\top W^{(t)}Q$ are still i.i.d. standard Gaussian distributions $\mathcal{N}(0,\tau^2)$, which implies that the distribution of $\|(\mathrm{Id}+\beta W^{(t)})x\|_2 = \|(\mathrm{Id}+\beta Q^\top W^{(t)}Q)e_1\|_2$ is the same as the distribution of $\|(\mathrm{Id}+\beta W^{(t)})e_1\|_2$ and is hence independent of $x \in \mathbb{S}^{d-1}$.

Therefore, we can apply Theorem 27 and conclude that

$$\log R(\beta, \mathbb{P}_W) = \mathbb{E}\log\|(\mathrm{Id}+\beta W^{(t)})e_1\|_2 = \frac{1}{2}\mathbb{E}\log\left((1+\beta\tau\xi)^2+\beta^2\tau^2\chi^2_{d-1}\right),$$

where ξ is a 1-dimensional standard Gaussian distribution $\mathcal{N}(0,1)$, and χ^2_{d-1} is a chi-squared distribution with $d-1$ degrees of freedom that is independent of ξ. ■

Next, we present the proof of Theorem 9, for which we need the following lemma.

Lemma 28 *Let $\xi \sim \mathcal{N}(0,1)$ and fix $b \geq 0$. The function*

$$f(a) := \mathbb{E}\log\left((a+\xi)^2+b\right)$$

is strictly increasing on $(0,+\infty)$.

Proof It can be computed that

$$f(a) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \log\left((a+\eta)^2+b\right)e^{-\frac{\eta^2}{2}}\mathrm{d}\eta,$$

and that

$$\begin{aligned}
f'(a) &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\frac{2(a+\eta)}{(a+\eta)^2+b}e^{-\frac{\eta^2}{2}}\mathrm{d}\eta = \frac{\sqrt{2}}{\sqrt{\pi}}\int_{\mathbb{R}}\frac{\eta}{\eta^2+b}e^{-\frac{(\eta-a)^2}{2}}\mathrm{d}\eta \\
&= \frac{\sqrt{2}}{\sqrt{\pi}}\left(\int_0^{+\infty}\frac{\eta}{\eta^2+b}e^{-\frac{(\eta-a)^2}{2}}\mathrm{d}\eta + \int_0^{+\infty}\frac{-\eta}{\eta^2+b}e^{-\frac{(-\eta-a)^2}{2}}\mathrm{d}\eta\right) \\
&= \frac{\sqrt{2}}{\sqrt{\pi}}\int_0^{+\infty}\frac{\eta}{\eta^2+b}\left(e^{-\frac{(\eta-a)^2}{2}} - e^{-\frac{(\eta+a)^2}{2}}\right)\mathrm{d}\eta > 0,
\end{aligned}$$

which completes the proof. ■

Proof [Proof of Theorem 9] By using Lemma 28, one can compute for any $\lambda \in \mathrm{spec}(P) \subseteq (-1,1)$ that

$$\begin{aligned}
&\mathbb{E}\log\left((1+\alpha\lambda\tau\xi)^2+\alpha^2\lambda^2\tau^2\xi^2_{d-1}\right) = \mathbb{E}\log\left((1+\alpha|\lambda|\tau\xi)^2+\alpha^2\lambda^2\tau^2\chi^2_{d-1}\right) \\
&\quad= \mathbb{E}_{\chi_{d-1}}\mathbb{E}_\xi\log\left((1+\alpha|\lambda|\tau\xi)^2+\alpha^2\lambda^2\tau^2\chi^2_{d-1}\right) \\
&\quad= \log(\alpha^2\lambda^2\tau^2) + \mathbb{E}_{\chi_{d-1}}\mathbb{E}_\xi\log\left(\left(\frac{1}{\alpha|\lambda|\tau}+\xi\right)^2+\chi^2_{d-1}\right) \\
&\quad> \log(\alpha^2\lambda^2\tau^2) + \mathbb{E}_{\chi_{d-1}}\mathbb{E}_\xi\log\left(\left(\frac{1}{\alpha\tau}+\xi\right)^2+\chi^2_{d-1}\right)
\end{aligned}$$

$$= 2\log|\lambda| + \mathbb{E}_{\chi_{d-1}}\mathbb{E}_{\xi}\log\left((1+\alpha\tau\xi)^2 + \alpha^2\tau^2\chi_{d-1}^2\right).$$

Therefore, we have

$$\exp\left(\frac{1}{2}\mathbb{E}\log\left((1+\alpha\lambda\tau\xi)^2 + \alpha^2\lambda^2\tau^2\xi_{d-1}^2\right)\right) > |\lambda|\exp\left(\frac{1}{2}\mathbb{E}\log\left((1+\alpha\tau\xi)^2 + \alpha^2\tau^2\xi_{d-1}^2\right)\right),$$

which combined with (10) leads to that

$$\begin{aligned}\lim_{t\to+\infty}\mu(x_{\mathrm{rs}}(t))^{1/2t} &\geq \frac{\max_{\lambda\in\mathrm{spec}(P)\setminus\{1\}}\exp\left(\frac{1}{2}\mathbb{E}\log\left((1+\alpha\lambda\tau\xi)^2 + \alpha^2\lambda^2\tau^2\chi_{d-1}^2\right)\right)}{\max_{\lambda\in\mathrm{spec}(P)}\exp\left(\frac{1}{2}\mathbb{E}\log\left((1+\alpha\lambda\tau\xi)^2 + \alpha^2\lambda^2\tau^2\chi_{d-1}^2\right)\right)} \\ &> \max_{\lambda\mathrm{spec}(P)\setminus\{1\}}|\lambda| = \lim_{t\to+\infty}\mu(x_{\mathrm{nrs}}(t))^{1/2t}.\end{aligned}$$

■

3.5.3 Proof of Theorem 10

For any $\lambda \in \mathrm{spec}(P)$, one can estimate that

$$\begin{aligned}R(\alpha\lambda, \mathbb{P}_W) &= \lim_{t\to+\infty}\left\|\left(I+\alpha\lambda W^{(t-1)}\right)\cdots\left(I+\alpha\lambda W^{(1)}\right)\left(I+\alpha\lambda W^{(0)}\right)\right\|^{1/t} \\ &\leq 1+\alpha|\lambda|r_W \leq 1+\alpha r_W.\end{aligned}$$

Moreover, for $\lambda \in \mathrm{spec}(P)\setminus\{1\}$ and $\alpha \leq 1/r_W$, it follows from

$$\left\|(I+\alpha\lambda W^{(t)})y\right\| \geq \|y\| - \alpha|\lambda|\left\|W^{(t)}y\right\| \geq (1-\alpha|\lambda|r_W)\|y\|,$$

that

$$\left\|\left(I+\alpha\lambda W^{(t-1)}\right)\cdots\left(I+\alpha\lambda W^{(1)}\right)\left(I+\alpha\lambda W^{(0)}\right)y\right\| \geq (1-\alpha|\lambda|r_W)^t\|y\|.$$

One can thus conclude that

$$R(\alpha\lambda, \mathbb{P}_W) = \lim_{t\to+\infty}\left\|\left(I+\alpha\lambda W^{(t-1)}\right)\cdots\left(I+\alpha\lambda W^{(1)}\right)\left(I+\alpha\lambda W^{(0)}\right)\right\|^{1/t} \geq 1-\alpha|\lambda|r_W,$$

and hence that

$$\lim_{t\to+\infty}\mu(x_{\mathrm{rs}}(t))^{1/2t} \geq \frac{\max_{\lambda\in\mathrm{spec}(P)\setminus\{1\}}R(\alpha\lambda,\mathbb{P}_W)}{\max_{\lambda\in\mathrm{spec}(P)}R(\alpha\lambda,\mathbb{P}_W)} \geq \frac{1-\alpha r_W\min_{\lambda\in\mathrm{spec}(P)\setminus\{1\}}|\lambda|}{1+\alpha r_W}.$$

If we further assume that

$$\alpha r_W\left(\min_{\lambda\in\mathrm{spec}(P)\setminus\{1\}}|\lambda| + \max_{\lambda\in\mathrm{spec}(P)\setminus\{1\}}|\lambda|\right) < 1 - \max_{\lambda\in\mathrm{spec}(P)\setminus\{1\}}|\lambda|,$$

then we notice that

$$\frac{1-\alpha r_W\min_{\lambda\in\mathrm{spec}(P)\setminus\{1\}}|\lambda|}{1+\alpha r_W} > \max_{\lambda\in\mathrm{spec}(P)\setminus\{1\}}|\lambda|.$$

This together with Theorem 4 finishes the proof.

3.5.4 Proofs of Theorem 12 and Theorem 13

Proof [Proof of Theorem 12] It can be computed directly that

$$\begin{aligned}\log R(\alpha\lambda, \mathbb{P}_W) &= \lim_{t\to+\infty} \frac{1}{t} \log \left\| \left(I + \alpha\lambda W^{(t-1)}\right) \cdots \left(I + \alpha\lambda W^{(1)}\right) \left(I + \alpha\lambda W^{(0)}\right) \right\| \\ &= \max_{1\le i\le d} \lim_{t\to+\infty} \frac{1}{t} \sum_{j=0}^{t-1} \log \left|1 + \alpha\lambda w_i^{(j)}\right| = \max_{1\le i\le d} \mathbb{E}\left[\log \left|1 + \alpha\lambda w_i^{(t)}\right|\right],\end{aligned}$$

where we used Proposition 18 and the last step is from the strong law of large numbers. The Theorem 12 follows directly from Theorem 5. ■

Proof [Proof of Theorem 13] For any $\lambda \in \mathrm{spec}(P)\backslash\{1\} \subseteq (-1, 1)$ and any $i \in \{1, \ldots, d\}$, because $\alpha \max_{1\le i\le d} r_i < 1$ and $\mathbb{P}_i$ is symmetric, we have that

$$\begin{aligned}&\mathbb{E}\left[\log \left|1 + \alpha\lambda w_i^{(t)}\right|\right] = \mathbb{E}\left[\log\left(1 + \alpha\lambda w_i^{(t)}\right)\right] \\ &\quad= \frac{1}{2}\mathbb{E}\left[\log\left(1 + \alpha\lambda w_i^{(t)}\right)\right] + \frac{1}{2}\mathbb{E}\left[\log\left(1 - \alpha\lambda w_i^{(t)}\right)\right] = \frac{1}{2}\mathbb{E}\left[\log\left(1 - \alpha^2\lambda^2 (w_i^{(t)})^2\right)\right] \\ &\quad\ge \frac{1}{2}\mathbb{E}\left[\log\left(1 - \alpha^2 (w_i^{(t)})^2\right)\right] = \frac{1}{2}\mathbb{E}\left[\log\left(1 + \alpha w_i^{(t)}\right)\right] + \frac{1}{2}\mathbb{E}\left[\log\left(1 - \alpha w_i^{(t)}\right)\right] \\ &\quad= \mathbb{E}\left[\log\left(1 + \alpha w_i^{(t)}\right)\right] = \mathbb{E}\left[\log\left|1 + \alpha w_i^{(t)}\right|\right].\end{aligned}$$

This implies that

$$\lim_{t\to+\infty} \mu(x_{\mathrm{rs}}(t))^{1/2t} \ge \frac{\max\limits_{\lambda\in\mathrm{spec}(P)\backslash\{1\}} \max\limits_{1\le i\le d} \exp\left(\mathbb{E}\left[\log\left|1 + \alpha\lambda w_i^{(t)}\right|\right]\right)}{\max\limits_{\lambda\in\mathrm{spec}(P)} \max\limits_{1\le i\le d} \exp\left(\mathbb{E}\left[\log\left|1 + \alpha\lambda w_i^{(t)}\right|\right]\right)} = 1.$$

The proof is thus completed. ■

4. Numerical Experiments

In this section, we validate our theoretical results and discuss some practical insights via numerical experiments.

4.1 Experimental setup

Our experiments utilize three citation network datasets: Cora, CiteSeer, and PubMed (Yang et al., 2016). In these datasets, vertices represent publications, and edges denote citation relationships. Features are encoded as binary bag-of-words representations based on the corpus vocabulary, where a feature value of 1 signifies the presence of a word in the paper and 0 indicates its absence. Each vertex is labeled with a class corresponding to the category of the publication. For all three datasets, the graphs are treated as undirected, with symmetric adjacency matrices A. We use the dataset versions provided in the PyTorch Geometric (PyG) library (Fey and Lenssen, 2019).

For each dataset, we implement GCNs and residual GCNs with $P = D^{-1}A$, considering the linear activation $\sigma = \mathrm{Id}$ matching Assumption 1 and two nonlinear activation functions: ReLU and LeakyRuLU, with the negative slope in LeakyRuLU being 0.8. ReLU and LeakyReLU have the positive homogeneity property, matching in some sense the scale-equivariant structure of the normalized node similarity measure. Other nonlinear activations such as tanh can change normalized dynamics and are not covered by our theory; we leave a systematic study of such nonlinearities to future work. In all models, each message-passing layer, i.e., (nrs) or (rs), is configured with 32 hidden dimensions, preceded by an embedding MLP layer for vertex feature encoding and followed by an output MLP layer for classification. We set $\alpha = 0.1$ in all residual connections.

The properties of the datasets are summarized in Table 1.

Dataset	#Vertices	#Edges	#Features	#Classes	#CC	$\lambda_{2,\mathrm{LCC}}(P)$
Cora	2708(2485)	5278(5069)	1433	7	78	0.99638
Citeseer	3327(2120)	4552(3679)	3703	6	438	0.99874
PubMed	19717	44324	500	3	1	0.99052

Table 1: Properties of citation network datasets: Cora, CiteSeer, and PubMed. We use the versions available in PyG. Abbreviations: CC = connected components, LCC = largest connected components. The numbers in parentheses indicate the numbers of vertices and edges in the LCC of each graph. $\lambda_{2,\mathrm{LCC}}(P) = \max_{\lambda \in \mathrm{spec}(P|_{\mathrm{LCC}}) \setminus \{1\}} |\lambda|$ is the second-largest magnitude of eigenvalues of P on LCC.

4.2 Vertex similarity of initialized GCNs and residual GCNs

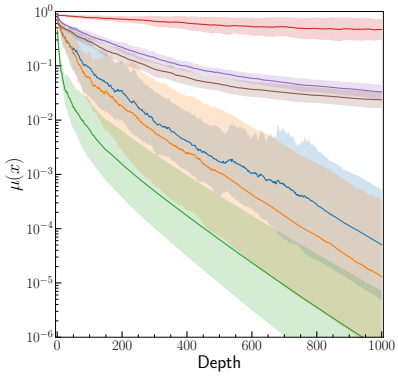
We begin by comparing the vertex similarity at initialization with the standard method in the PyG library. Specifically, the weights are initialized using the Xavier uniform distribution (Glorot and Bengio, 2010).

In Figure 1, we illustrate the evolution of both the normalized vertex similarity measure, $\mu(x(t))$, as defined in Theorem 1, and the unnormalized vertex similarity measure, $\tilde{\mu}(x(t))$, defined as

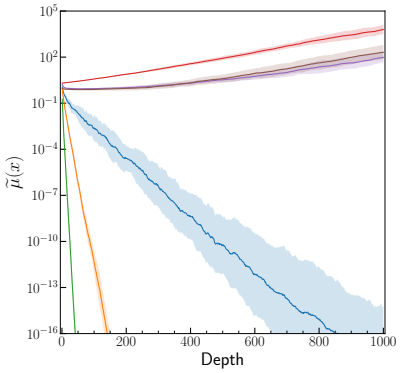
$$\tilde{\mu}(x) = \sum_{i=1}^{n} \|x_i - \bar{x}\|_2^2, \tag{44}$$

on the largest connected component of each graph as the input $x(t)$ propagates through the model with 1000 message-passing layers. Each experiment is repeated 15 times. The solid lines in Figure 1 represent $\exp(\mathbb{E}[\log \mu(x(t))])$ and $\exp(\mathbb{E}[\log \tilde{\mu}(x(t))])$, where the expectation is taken over the 15 independent trials. The shaded regions indicate the standard deviation across these trials.

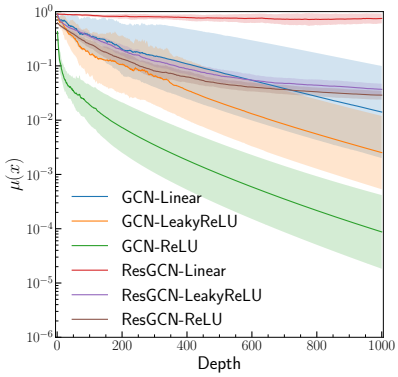
As shown in Figure 1, the normalized vertex similarity measure, $\mu(x(t))$, for GCNs decays at a faster rate compared to residual GCNs across all settings, corroborating our theoretical findings in Section 2. Furthermore, the benefit of residual connections becomes more pronounced when $\lambda_{2,\mathrm{LCC}}(P)$ is smaller. This observation aligns with our theory, as Theorem 4 establishes that $\mu(x(t))$ of deep GCNs diminishes more rapidly when $\lambda_{2,\mathrm{LCC}}(P)$ is smaller. We also plot the unnormalized vertex similarity measure, $\tilde{\mu}(x(t))$, as defined in



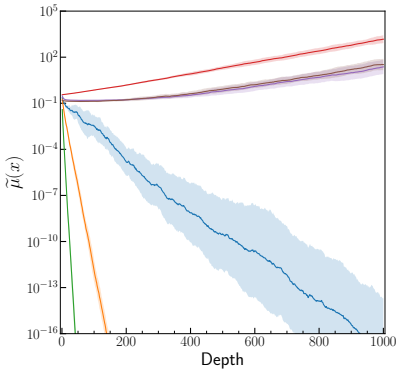
(a) normalized vertex similarity on Cora



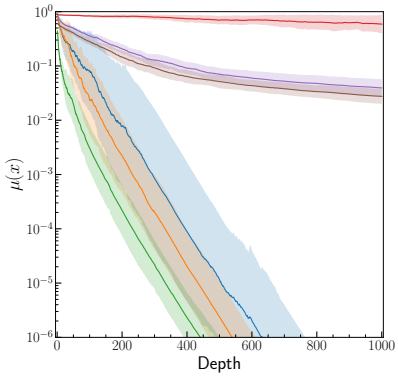
(b) unnormalized vertex similarity on Cora



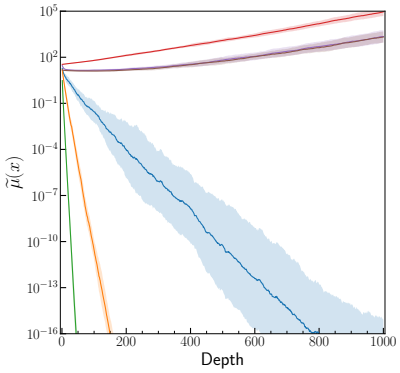
(c) normalized vertex similarity on CiteSeer



(d) unnormalized vertex similarity on CiteSeer



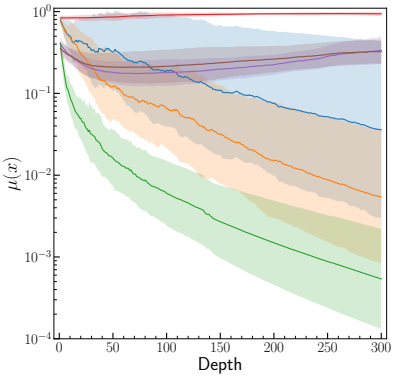
(e) normalized vertex similarity on PubMed



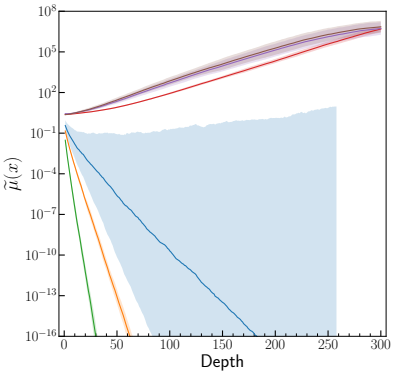
(f) unnormalized vertex similarity on PubMed

Figure 1: Vertex similarity measure of initialized GCNs and residual GCNs on the largest connected component

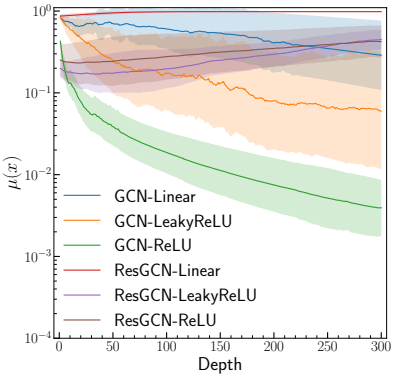
(44). As seen in Figure 1, $\tilde{\mu}(x(t))$ converges to zero exponentially fast for GCNs, while for residual GCNs, it remains away from zero.



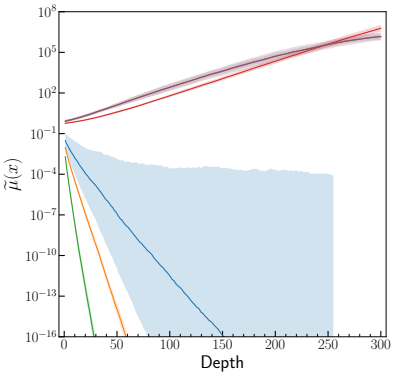
(a) normalized vertex similarity on Cora



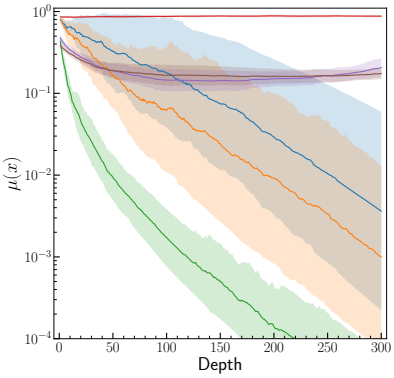
(b) unnormalized vertex similarity on Cora



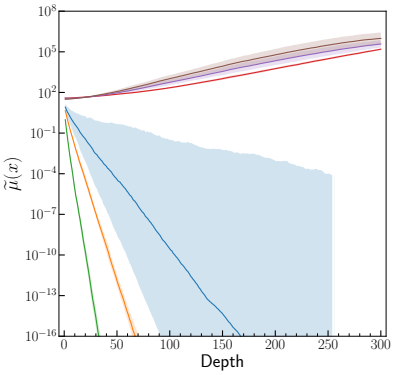
(c) normalized vertex similarity on CiteSeer



(d) unnormalized vertex similarity on CiteSeer



(e) normalized vertex similarity on PubMed



(f) unnormalized vertex similarity on PubMed

Figure 2: Vertex similarity measure of trained GCNs and residual GCNs on the largest connected component

4.3 Vertex similarity and performance of trained GCNs and residual GCNs

In this subsection, we present the vertex similarity and performance of GCNs and residual GCNs after training. We utilize the "full split" mode in the PyG library to generate train, validation, and test masks for each dataset. Specifically, 500 vertices are used for validation, 1200 vertices for test, following the setup in Chen et al. (2018), and all remaining vertices

are used for training. The cross-entropy loss is employed, and the Adam optimizer (Kingma, 2014) is used to minimize the loss, with a fixed learning rate of 1×10^{-5} and weight decay of 5×10^{-4}. Each model is trained for 3000 epochs, and all experiments are performed with 15 independent trials. The models are trained on an Nvidia RTX 4090 GPU.

Figure 2 shows the vertex similarity measure for trained GCNs and residual GCNs of 300 message-passing layers, with all other settings consistent with Figure 1. The observations are similar to those in Figure 1. In particular, the normalized vertex similarity measure $\mu(x(t))$ for GCNs decays more quickly than for residual GCNs. Additionally, the unnormalized vertex similarity measure $\tilde{\mu}(x(t))$ for GCNs converges exponentially to 0, while it remains away from 0 for residual GCNs.

Next, we train GCNs and residual GCNs with $0, 20, 40, \ldots, 300$ message-passing layers, and compare their performance on the training, validation, and test sets. The training loss and training classification accuracy are shown in Figure 3, while the classification accuracies on the validation and test sets are displayed in Figure 4. In both figures, solid lines represent the average values, and shaded regions indicate the standard deviation.

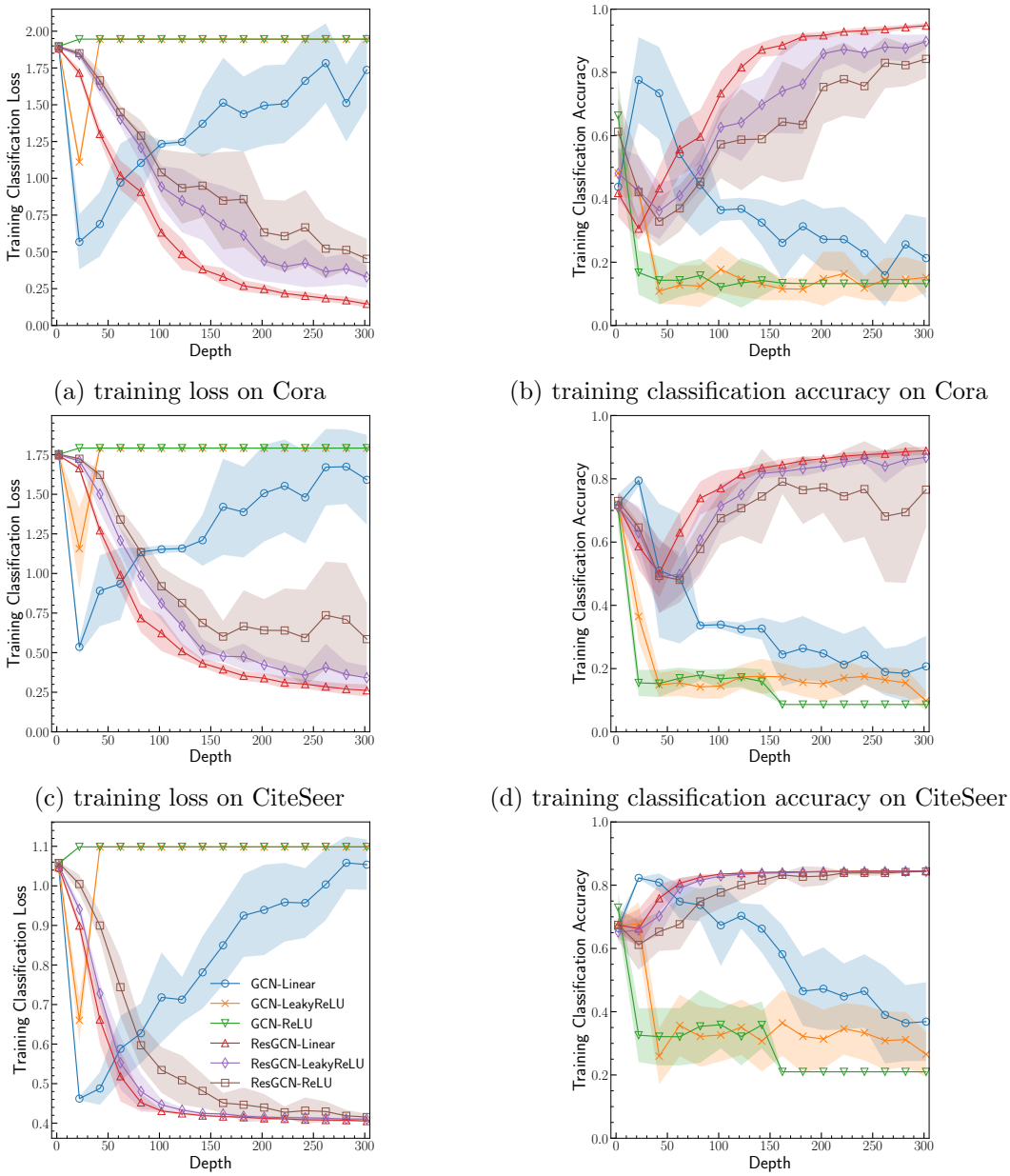
As shown in Figure 3, residual GCNs generally achieve lower training loss and higher training accuracy as the model depth increases. In contrast, deeper GCNs experience larger training loss and lower training accuracy. Similar trends are observed in the validation and test accuracies, as displayed in Figure 4. Furthermore, deep residual GCNs consistently outperform their non-residual counterparts, highlighting the effectiveness of residual connections in building and training deep graph machine learning models.

We also briefly comment on the choice of activation. In both residual and non-residual models, ReLU and LeakyReLU in general lead to faster decay of μ and slightly worse classification accuracy than the identity activation. This empirical gap is not covered by our analytical framework and might be due to the extra coordinatewise contraction from ReLU/LeakyReLU.

5. Conclusion

This work establishes the asymptotic oversmoothing rates for deep GNNs with and without residual connections using the multiplicative ergodic theorem. Under suitable assumptions, we show that the normalized vertex similarity of deep non-residual GNNs converges to zero at an exponential rate determined by the second-largest eigenvalue magnitude of the aggregation coefficient matrix. Furthermore, we precisely characterize the asymptotic behavior of the normalized vertex similarity of deep residual GNNs, proving that it either avoids exponential oversmoothing or exhibits a significantly slower rate compared to deep non-residual GNNs across several broad classes of parameter distributions. These findings highlight that incorporating residual connections effectively mitigates or prevents the oversmoothing issue. Our theoretical results are strongly corroborated by numerical experiments.

Let us also comment on the limitations and future works. First, the current analysis relies on the linearity of the dynamics, which does not account for nonlinear activations. In particular, one could try analytical tools from nonlinear random dynamical system and there exists some extension of the multiplicative ergodic theorem in the nonlinear setting, say the stable/unstable manifold theorem. However, such extension holds in a random local neighborhood of the equilibrium, making it hard to obtain a global oversmoothing



(a) training loss on Cora

(b) training classification accuracy on Cora

(c) training loss on CiteSeer

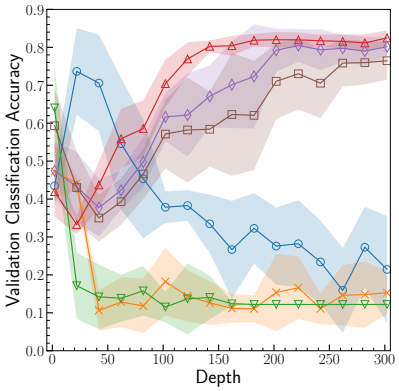
(d) training classification accuracy on CiteSeer

(e) training loss on PubMed

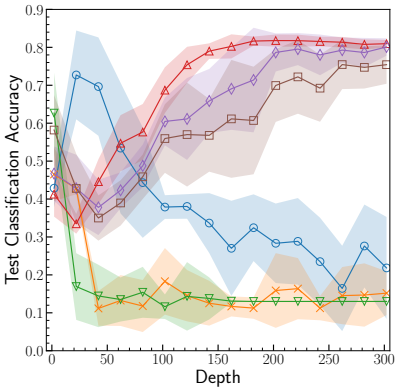
(f) training classification accuracy on PubMed

Figure 3: Training loss and training classification accuracy of GCNs and residual GCNs

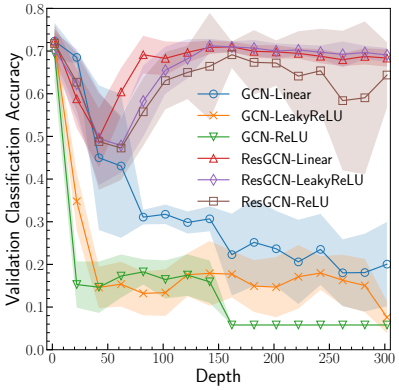
rate as in the present work. Second, the assumption that the parameters of each layer are independently drawn from the same distribution may not hold for trained neural networks. A potential idea is to start from the setting that trained weights are within a certain distance from initialization, which is still limited though. Finally, our analytical framework assumes constant aggregation coefficients across layers, whereas, in practice, different aggregations could be used in different layers. One technical difficulty is also that it breaks the linear



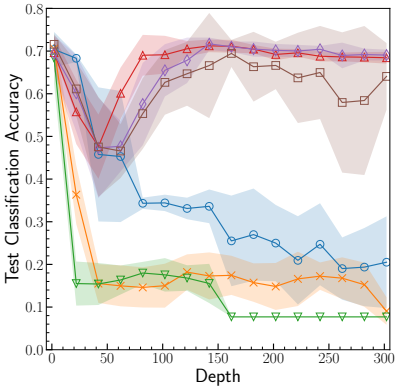
(a) validation classification accuracy on Cora



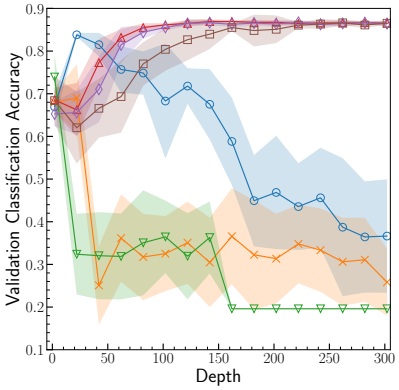
(b) test classification accuracy on Cora



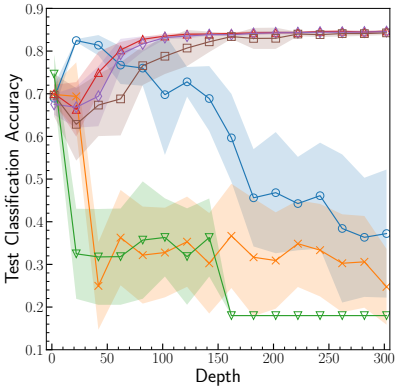
(c) validation classification accuracy on CiteSeer



(d) test classification accuracy on CiteSeer



(e) validation classification accuracy on PubMed



(f) test classification accuracy on PubMed

Figure 4: Validation and test classification accuracy of GCNs and residual GCNs

random dynamical system structure; for instance, attention-based aggregations make each layer nonlinear in the parameters. These limitations and challenges present important directions for future research.

Acknowledgments

Z.C. thanks Runzhong Wang for the helpful discussions on oversmoothing and was supported in part by NSF under DMS-2509011. Y.P. is supported in part by NSF under CCF-2131115. P.R. is supported by NSF grant DMS-2509011.

References

Ludwig Arnold. *Random dynamical systems*. Springer monographs in mathematics. Springer, New York, 1998. ISBN 3540637583.

Chen Cai and Yusu Wang. A note on over-smoothing for graph neural networks. *arXiv preprint arXiv:2006.13318*, 2020.

Deli Chen, Yankai Lin, Wei Li, Peng Li, Jie Zhou, and Xu Sun. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3438–3445, 2020.

Jie Chen, Tengfei Ma, and Cao Xiao. Fastgcn: fast learning with graph convolutional networks via importance sampling. *arXiv preprint arXiv:1801.10247*, 2018.

Joel E. Cohen and Charles M. Newman. The stability of large random matrices and their products. *Ann. Probab.*, 12:283–310, 1984. ISSN 0091-1798. doi: 10.1214/aop/1176993291.

Gage Dezoort and Boris Hanin. Principles for initialization and architecture selection in graph neural networks with relu activations. *SIAM Journal on Mathematics of Data Science*, 7(1):1–27, 2025.

Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Maxime Gasse, Didier Chételat, Nicola Ferroni, Laurent Charlin, and Andrea Lodi. Exact combinatorial optimization with graph convolutional neural networks. *Advances in neural information processing systems*, 32, 2019.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Shiwen He, Shaowen Xiong, Yeyu Ou, Jian Zhang, Jiaheng Wang, Yongming Huang, and Yaoxue Zhang. An overview on the application of graph neural networks in wireless networks. *IEEE Open Journal of the Communications Society*, 2:2547–2565, 2021.

Nicolas Keriven. Not too little, not too much: a theoretical analysis of graph (over) smoothing. *Advances in Neural Information Processing Systems*, 35:2268–2281, 2022.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Mengyuan Lee, Guanding Yu, Huaiyu Dai, and Geoffrey Ye Li. Graph neural networks meet wireless communications: Motivation, applications, and future directions. *IEEE Wireless Communications*, 29(5):12–19, 2022.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Wenlong Liao, Birgitte Bak-Jensen, Jayakrishnan Radhakrishna Pillai, Yuelong Wang, and Yusen Wang. A review of graph neural networks and their applications in power systems. *Journal of Modern Power Systems and Clean Energy*, 10(2):345–360, 2021.

James R Norris. *Markov chains*. Number 2. Cambridge university press, 1998.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020.

Valery Iustinovich Oseledets. A multiplicative ergodic theorem. liapunov characteristic numbers for dynamical systems. *Trans. Moscow Math. Soc.*, 19:197–231, 1968.

Madabusi S Raghunathan. A proof of Oseledec's multiplicative ergodic theorem. *Israel Journal of Mathematics*, 32(4):356–362, 1979.

David Ruelle. Ergodic theory of differentiable dynamical systems. *Publications mathématiques. Institut des hautes études scientifiques*, 50(1):27–58, 1979. ISSN 1618-1913.

T Konstantin Rusch, Michael M Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks. *SAM Research Report*, 2023, 2023.

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

Michael Scholkemper, Xinyi Wu, Ali Jadbabaie, and Michael Schaub. Residual connections and normalization can provably prevent oversmoothing in gnns. In *International Conference on Learning Representations*, volume 2025, pages 43581–43610, 2025.

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

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

Peter Walters. A dynamical proof of the multiplicative ergodic theorem. *Transactions of the American Mathematical Society*, 335(1):245–257, 1993.

Jianian Wang, Sheng Zhang, Yanghua Xiao, and Rui Song. A review on graph neural network methods in financial applications. *Journal of Data Science*, 20(2), 2022.

Xinyi Wu, Amir Ajorlou, Zihui Wu, and Ali Jadbabaie. Demystifying oversmoothing in attention-based graph neural networks. *Advances in Neural Information Processing Systems*, 37, 2023a.

Xinyi Wu, Zhengdao Chen, William Wei Wang, and Ali Jadbabaie. A non-asymptotic analysis of oversmoothing in graph neural networks. In *The Eleventh International Conference on Learning Representations*, 2023b.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, pages 40–48. PMLR, 2016.

Xiao-Meng Zhang, Li Liang, Lin Liu, and Ming-Jing Tang. Graph neural networks and their current applications in bioinformatics. *Frontiers in genetics*, 12:690049, 2021.

Xu Zhang, Yonghui Xu, Wei He, Wei Guo, and Lizhen Cui. A comprehensive review of the oversmoothing in graph neural networks. In *CCF Conference on Computer Supported Cooperative Work and Social Computing*, pages 451–465. Springer, 2023.

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