Rank Collapse Causes Over-Smoothing and Over-Correlation in Graph Neural Networks

Andreas Roth[1][0000−0002−0515−7635]($\boxtimes$) and Thomas Liebig[1,2][0000−0002−9841−1101]

[1] TU Dortmund University, Germany
[2] Lamarr Institute for Machine Learning and Artificial Intelligence
{andreas.roth,thomas.liebig}@tu-dortmund.de

Abstract. Our study reveals new theoretical insights into over-smoothing and feature over-correlation in deep graph neural networks. We show the prevalence of invariant subspaces, demonstrating a fixed relative behavior that is unaffected by feature transformations. Our work clarifies recent observations related to convergence to a constant state and a potential over-separation of node states, as the amplification of subspaces only depends on the spectrum of the aggregation function. In linear scenarios, this leads to node representations being dominated by a low-dimensional subspace with an asymptotic convergence rate independent of the feature transformations. This causes a rank collapse of the node representations, resulting in over-smoothing when smooth vectors span this subspace, and over-correlation even when over-smoothing is avoided. Guided by our theory, we propose a sum of Kronecker products as a beneficial property that can provably prevent over-smoothing, over-correlation, and rank collapse. We empirically extend our insights to the non-linear case, demonstrating the inability of existing models to capture linearly independent features.

Keywords: Graph Neural Networks · Rank Collapse · Expressivity

1 Introduction

Despite the great success of Graph Neural Networks (GNNs) for homophilic tasks [1–4], their performance for more complex and heterophilic tasks is largely unsatisfying [5]. Two leading causes are over-smoothing [6, 7] and feature over-correlation [8], resulting in primarily shallow models with small receptive fields. The conditions under which over-smoothing and over-correlation occur and in which cases they can be prevented are well-understood in some instances but not in the general case. Some works argue that all node representations converge to a constant state and have discarded the interaction with feature transformations in their analysis [9–11]. Other works have proven that over-smoothing only occurs when the feature transformations satisfy some constraints [6, 12, 13], and choosing suitable parameters can mitigate over-smoothing and even cause an over-separation of node representations [6, 12, 13]. Over-correlation refers to

all feature columns becoming strongly correlated, which was only recently empirically observed [8]. While Jin et al. [8] showed that this occurs even when over-smoothing is prevented, its theoretical understanding is limited.

This work clears up different views on over-smoothing and provides a theoretical investigation of the underlying reason behind both over-smoothing and over-correlation. We show that graph convolutions of a common type induce invariant subspaces, each demonstrating distinct predefined behaviors. Critically, this behavior only depends on the spectrum of the aggregation function and not on the learnable feature transformations or the initial node features. When considering the limit in the linear case, a low-dimensional subspace dominates the results, leading to over-smoothing when this subspace aligns with smooth signals. More importantly, node representations suffer from rank collapse for all aggregation functions, explaining the effect of over-correlation. This severely limits the expressivity of deep GNNs, as we cannot optimize these models for tasks that require multiple independent features, which we empirically extend to the non-linear case. We propose a simple but provably more expressive family of models based on the sum of Kronecker products (SKP) that can maintain the rank of its features.

We summarize our key contributions as follows:

- We establish the presence of invariant subspaces, characterized by fixed relative behaviors that remain unaffected by feature transformations, across all graph convolutions of a common type. This discovery allows us to lift constraints on existing proofs related to over-smoothing, providing enhanced insights into this phenomenon.
- Our work bridges the gap between rank collapse, over-smoothing, and over-correlation, as these are caused by a low-dimensional subspace dominating the node representations.
- Theoretically and empirically, we demonstrate the effectiveness of utilizing a sum of Kronecker products (SKP) to counteract this limited expressivity of deep GNNs.

2   Preliminaries

Notation We consider a graph $G = (\mathcal{V}, \mathcal{E})$ consisting of a set of n nodes $\mathcal{V} = \{v_1, \ldots, v_n\}$ and a set of edges $\mathcal{E}$. The adjacency matrix $\mathbf{A} \in \{0, 1\}^{n \times n}$ has binary entries indicating whether an edge between two nodes exists or not. We assume graphs to be irreducible and aperiodic. We denote the set of nodes neighboring node v_i as $\mathcal{N}_i = \{v_j | a_{ij} = 1\}$. The degree matrix $\mathbf{D} \in \mathbb{N}^{n \times n}$ is a diagonal matrix with each entry $d_{ii} = |\mathcal{N}_i|$ representing the number of neighboring nodes. For a given matrix $\mathbf{M}$, we denote its eigenvalues with $\lambda_1^{\mathbf{M}}, \ldots, \lambda_n^{\mathbf{M}}$ that are sorted with decreasing absolute value $|\lambda_1^{\mathbf{M}}| \geq \cdots \geq |\lambda_n^{\mathbf{M}}|$. A matrix is vectorized $vec(\mathbf{M})$ by stacking its columns into a single vector. The identity with d dimensions is denoted by $\mathbf{I}_d \in \mathbb{R}^{d \times d}$.

Graph neural networks Given a graph G and a graph signal $\mathbf{X} \in \mathbb{R}^{n \times d}$ representing d features at each node, the goal of a graph neural network (GNN) is to find node representations that can be used effectively within node classification, edge prediction or various other challenges. Within these, graph convolutional operations repeatedly update the state of each node by combining the state of all nodes with the states from their neighbors [14].

We consider graph convolutions that iteratively transform the previous node states of the form

$$\mathbf{X}^{(k+1)} = \tilde{\mathbf{A}} \mathbf{X}^{(k)} \mathbf{W}^{(k)}. \tag{1}$$

At each iteration k, we consider distinct feature transformations $\mathbf{W}^{(k)} \in \mathbb{R}^{d \times d}$ of the node representations and a homogeneous neighbor aggregation represented by $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times n}$. Popular instantiations covered by this notation include the graph convolutional network (GCN) [15], the graph isomorphism network (GIN) [16], and the graph attention network (GAT) [17] with a single head.

3   Related Work

Over-smoothing in GNNs Over-smoothing arises when node representations $\mathbf{X}^{(k)}$ exhibit excessive similarity to one another as the number of layers k increases. As our study extends prior analyses, we first familiarize readers with the presently available theoretical insights concerning over-smoothing when considering the linear case. Li et al. [9] connected over-smoothing in GCNs to a special form of Laplacian smoothing when ignoring the feature transformation. While some methods similarly ignore the feature transformation [10, 11], its role is more complicated. The pioneering work of Oono and Suzuki [6] showed that the feature transformation must not be ignored by bounding the distance

$$d_{\mathcal{M}}(\mathbf{AXW}) \leq \lambda_2^{\mathbf{A}} \sigma_1^{\mathbf{W}} d_{\mathcal{M}}(\mathbf{X}) \tag{2}$$

of the representations to a smooth subspace $\mathcal{M}$ that is induced by the dominant eigenvector $\mathbf{v}_1$. The bound uses the second largest eigenvalue $\lambda_2^{\mathbf{A}}$ and the largest singular value $\sigma_1^{\mathbf{W}}$ of $\mathbf{W}$. Intuitively, each aggregation step $\mathbf{A}$ reduces this distance, while $\mathbf{W}$ can increase the distance arbitrarily. Thus, they consequently claim potential over-separation when $\sigma_1^{\mathbf{W}} > \frac{1}{\lambda_2^{\mathbf{A}}}$, which refers to node representations differing strongly. As an interpretable metric to determine the smoothness of a graph signal, Cai and Wang [12] introduced the Dirichlet energy

$$E(\mathbf{X}) = tr(\mathbf{X}^T \boldsymbol{\Delta} \mathbf{X}) = \frac{1}{2} \sum_{(i,j) \in \mathcal{E}} \left\| \frac{\mathbf{x}_i}{\sqrt{d_i}} - \frac{\mathbf{x}_j}{\sqrt{d_j}} \right\|_2^2 \tag{3}$$

using the symmetrically normalized graph Laplacian $\boldsymbol{\Delta} = \mathbf{I}_n - \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}$. A low energy value corresponds to similar neighboring node states. Similarly to Oono and Suzuki [6], they provided the bound

$$E(\mathbf{AXW}) \leq \left(\lambda_2^{\mathbf{A}} \right)^2 \left(\sigma_1^{\mathbf{W}} \right)^2 E(\mathbf{X}) \tag{4}$$

(a) Dirichlet energy converges to zero.   (b) Dirichlet energy does not converge.
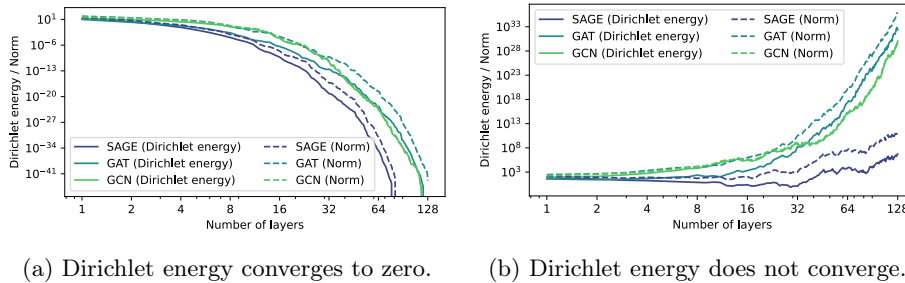
Fig. 1: Comparison of the Dirichlet energy $E_{rw}(\mathbf{X}^{(l)})$ and the norm $||\mathbf{X}^{(l)}||_F^2$ of the representations after l layers for three commonly used models (Graph Convolutional Network (GCN) [15], Graph Attention Network (GAT) [17], Graph-SAGE [28]) on the Cora dataset [29]. For 1a, we used random weights, for 1b, we scaled these weights by a factor of 2.

for each convolution and prove an exponential convergence in the limit. As their proof again only holds in case $\sigma_1^{\mathbf{W}} \leq \frac{1}{\lambda_2^{\mathbf{A}}}$, they similarly claim potential over-separation. Zhou et al. [13] provide a lower bound on the energy to show that the Dirichlet energy can go to infinite. All of these works conclude that over-smoothing only occurs with high probability and propose to find suitable feature transformations $\mathbf{W}$ to trade-off between over-smoothing and over-separation [6, 12, 13, 18]. Di Giovanni et al. [19] and Maskey et al. [20] studied the case when all feature transformations are restricted to be the same symmetric matrix. It remains unclear which feature transformations reduce or increase the Dirichlet energy in the unconstrained case. Another line of work found node representations to converge to a constant state [11, 21–23], though it remains unclear for which aggregation functions and feature transformations this holds. While various methods to mitigate over-smoothing have been proposed, the underlying issue persists [24–27].

Over-correlation in GNNs Another notable issue, termed feature over-correlation, has been empirically observed in a recent analysis by Jin et al. [8]. As the model depth increases, the node features become excessively correlated, thereby limiting the performance of deep GNNs. The study highlights that while over-smoothing leads to over-correlation, over-correlation also arises independently. However, the theory behind over-correlation remains unexplored, which could guide the design of new methods to prevent this phenomenon.

4   Over-Smoothing and Convergence to a Constant State

We begin by clarifying distinctions in the current comprehension of over-smoothing. Recent research [11, 21–23] defines over-smoothing as the exponential convergence to a constant state using the Dirichlet energy $E_{rw}(\mathbf{X}) = tr(\mathbf{X}^T \mathbf{\Delta}_{rw} \mathbf{X})$

based on the random walk Laplacian $\mathbf{\Delta}_{rw} = \mathbf{I}_n - \mathbf{D}^{-1}\mathbf{A}$ as the constant state corresponds to its nullspace. We validate this empirically for various GNNs in Figure 1a. In contrast, other studies provided theoretical insights into convergence toward the dominant eigenvector of the aggregation function [6, 12, 13], which is not constant for the GCN. As two linearly independent vectors only intersect at the origin, we compute the distance to the origin $||\mathbf{X}||_F^2$, illustrated in Figure 1a. Detailed information is provided in Appendix B, and the supplementary material contains a reproducible implementation. We observe a close alignment between both metrics, indicating that the norm overshadows potential insights from the Dirichlet energy. Analogous to the Dirichlet energy (see Eq. 4), the norm is likewise bounded by the largest singular value of the feature transformation:

Proposition 1. *(Node representations vanish.) Let $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times n}$ be symmetric with maximum absolute eigenvalue $|\lambda_1^{\tilde{\mathbf{A}}}| = 1$, $\mathbf{W} \in \mathbb{R}^{d \times d}$ be any matrix with maximum singular value $\sigma_1^{\mathbf{W}}$, and ϕ a component-wise non-expansive mapping satisfying $\phi(\mathbf{0}) = \mathbf{0}$. Then,*

$$||\phi(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W})||_F \leq \sigma_1^{\mathbf{W}} \cdot ||\mathbf{X}||_F . \tag{5}$$

We present all proofs in Appendix A. When $\sigma_1^{\mathbf{W}} < 1$ for all $\mathbf{W}$ this implies convergence to the zero matrix: $\lim_{l \to \infty} \mathbf{X}^{(l)} = \mathbf{0}$. With all node states close to zero, the Dirichlet energy becomes also zero, irrespective of the chosen graph Laplacian. This elucidates deviations between directions claiming either convergence of the GCN to constant columns [11, 21, 22] or values proportional to each node's degree [6, 12, 13]. Repeating this experiment for scaled feature transformations to prevent vanishing norms, we observe the Dirichlet energy also avoids convergence to zero, as shown in Figure 1b.

Since the norm of the node representations heavily influences the Dirichlet energy, we question whether assessing the energy of an unnormalized state provides sufficient insights into over-smoothing. Other metrics, such as MAD [18] and SMV [30], already incorporate feature normalization to quantify over-smoothing. Several theoretical studies base their findings about over-smoothing on assumptions on the singular values of the feature transformations [6, 12, 13, 23], which might oversimplify the cases. Our insights prompt a deeper exploration of over-smoothing and its interaction with feature transformations, aiming to better comprehend its causes, mitigate its impacts, and advance the development of more powerful GNNs.

5 Understanding Graph Convolutions using Invariant Subspaces

We delve into the underlying cause for over-smoothing and over-correlation while analyzing the role of feature transformations in these phenomena. We consider the linearized update function from Eq. 1 employing any symmetric aggregation

function $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times n}$ and time-inhomogeneous feature transformations $\mathbf{W}^{(k)} \in \mathbb{R}^{d \times d}$ for each layer k. Leveraging the vectorized form

$$\mathrm{vec}(\tilde{\mathbf{A}}\mathbf{X}^{(k)}\mathbf{W}^{(k)}) = (\mathbf{W}^{(k)^T} \otimes \tilde{\mathbf{A}})\,\mathrm{vec}(\mathbf{X}^{(k)}) = \mathbf{T}^{(k)}\,\mathrm{vec}(\mathbf{X}^{(k)}) \qquad (6)$$

allows us to combine the aggregation and transformation steps through the Kronecker product $\mathbf{T}^{(k)} = (\mathbf{W}^{(k)^T} \otimes \tilde{\mathbf{A}}) \in \mathbb{R}^{nd \times nd}$. Further information about the Kronecker product can be found in Appendix A. For clarity, we will omit the transpose.

Initially, we demonstrate that for a fixed aggregation $\tilde{\mathbf{A}}$, all transformations $\mathbf{T}^{(k)}$ induce the same invariant subspaces that only depend on the eigenvectors of $\tilde{\mathbf{A}}$. Formally, the vectorization operation enables the decomposition of vectorized node representations

$$\mathbf{T}^{(k)}\,\mathrm{vec}(\mathbf{X}^{(k)}) = \mathbf{T}^{(k)}\mathbf{S}\mathbf{c} = \sum_{i=1}^{m} \mathbf{T}^{(k)}\mathbf{S}_{(i)}\mathbf{c}_{(i)} \qquad (7)$$

into a linear combination $\mathbf{c} \in \mathbb{R}^{nd}$ of basis vectors $\mathbf{S} \in \mathbb{R}^{nd \times nd}$. We further split these into a sum of n invariant components across disjoint subspaces $\mathcal{Q}_i = \mathrm{span}(\mathbf{S}_{(i)}) \subset \mathbb{R}^{nd}$ with their direct sum $\bigoplus_{i=1}^{n} \mathcal{Q}_i = \mathbb{R}^{nd}$ covering the entire space. The linearity of $\mathbf{T}^{(k)}$ allows us to apply the transformation on each subspace separately. We construct our bases as $\mathbf{S}_{(i)} = \mathbf{I}_d \otimes \mathbf{v}_i \in \mathbb{R}^{n \times d}$, utilizing the eigenvectors $\mathbf{v}_i$ of $\tilde{\mathbf{A}}$, as these are invariant to any $\mathbf{T}^{(k)}$:

Lemma 1. *(The subspaces are invariant to any $\mathbf{T}^{(k)}$.) Let $\mathbf{T} = \mathbf{W} \otimes \tilde{\mathbf{A}}$ with $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times n}$ symmetric with eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and $\mathbf{W} \in \mathbb{R}^{d \times d}$ any matrix. Consider the subspaces $\mathcal{Q}_i = \mathrm{span}(\mathbf{I}_d \otimes \mathbf{v}_i)$ for $i \in [n]$. Then,*

$$\forall\, i \in [n]\colon \mathbf{z} \in \mathcal{Q}_i \Rightarrow \mathbf{T}\mathbf{z} \in \mathcal{Q}_i\,.$$

This discovery is pivotal in our investigation, enabling us to dissect each subspace individually and relate their differences.

5.1   Relating Dynamics in Subspaces

Our investigation now delves into the effect $\mathbf{T}^{(k)}$ has on each these subspaces. Previous work considered the impact of graph convolutions on coefficients $\mathbf{c}$, deriving coarse bounds based on the singular values of $\mathbf{W}$ were found [6, 12, 13] (see Section 3). We instead analyze how $\mathbf{T}^{(k)}$ alters the basis vectors $\mathbf{S}_{(i)}$ of each subspace while maintaining the coefficients $\mathbf{c}$ constant, leading to a more streamlined analysis of the underlying process. When applying only the aggregation function $\tilde{\mathbf{A}}$, each basis $\mathbf{S}_{(i)}$ gets scaled by the corresponding eigenvalue $\lambda_i^{\tilde{\mathbf{A}}}$. By construction, all transformations $\mathbf{W}^{(k)}$ act the same on each subspace, nullifying when assessing the relative norm change amongst pairs of subspaces:

Theorem 1. *(The relative behavior is fixed.) Let $\mathbf{T} = \mathbf{W} \otimes \tilde{\mathbf{A}}$ with $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times n}$ symmetric with eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and eigenvalues $\lambda_1^{\tilde{\mathbf{A}}}, \ldots, \lambda_n^{\tilde{\mathbf{A}}}$ and $\mathbf{W} \in$*

$\mathbb{R}^{d \times d}$ *any matrix. Consider the bases* $\mathbf{S}_{(i)} = \mathbf{I}_d \otimes \mathbf{v}_i$ *and* $\mathbf{S}_{(j)} = \mathbf{I}_d \otimes \mathbf{v}_j$ *for* $i, j \in [n]$. *Then,*

$$\frac{\|\mathbf{T}\mathbf{S}_{(i)}\|_F}{\|\mathbf{T}\mathbf{S}_{(j)}\|_F} = \frac{|\lambda_i^{\tilde{\mathbf{A}}}|}{|\lambda_j^{\tilde{\mathbf{A}}}|} . \tag{8}$$

We underline the significance of this key property. As the right-hand side is independent of $\mathbf{W}$ with equality, the amplification or reduction of individual subspaces cannot be affected by any feature transformation or the graph signal. The outcome is predefined exclusively by the eigenvalues of the aggregation function. The functions learnable by each operation are inherently restricted.

We spotlight the distinctions between our research on over-smoothing and previous studies. Prior work bounded the effect of each operation using the maximal singular value $\sigma_1^{\mathbf{W}}$, while we provide a statement that holds exactly and does not rely on $\mathbf{W}$. Our statement offers deeper insights into the behavior of GNNs rather than only presenting a single value for the energy.

5.2   Implications in the Limit

Considering the case where graph convolutions are repeatedly applied, we extend our results directly to the iterated case and different transformations $\mathbf{T}^{(k)}$ at each layer k, yielding the dominance of fixed subspaces:

Proposition 2. *(Fixed subspaces dominate.) Let* $\mathbf{T}^{(k)} = \mathbf{W}^{(k)} \otimes \tilde{\mathbf{A}}$ *with* $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times n}$ *symmetric with eigenvectors* $\mathbf{v}_1, \ldots, \mathbf{v}_n$ *and eigenvalues* $\lambda_1^{\tilde{\mathbf{A}}}, \ldots, \lambda_n^{\tilde{\mathbf{A}}}$ *and* $\mathbf{W}^{(k)} \in \mathbb{R}^{d \times d}$ *any matrix. Consider the bases* $\mathbf{S}_{(i)} = \mathbf{I}_d \otimes \mathbf{v}_i$ *for* $i \in [n]$. *Then,*

$$\lim_{l \to \infty} \frac{\|\mathbf{T}^{(l)} \ldots \mathbf{T}^{(1)} \mathbf{S}_{(i)}\|_F}{\max_p \|\mathbf{T}^{(l)} \ldots \mathbf{T}^{(1)} \mathbf{S}_{(p)}\|_F} = \lim_{l \to \infty} \frac{|\lambda_i^{\tilde{\mathbf{A}}}|^l \cdot \|\mathbf{S}_{(i)}\|_F}{\max_p |\lambda_p^{\tilde{\mathbf{A}}}|^l \cdot \|\mathbf{S}_{(p)}\|_F} = \begin{cases} 1, & \text{if} |\lambda_i^{\tilde{\mathbf{A}}}| = |\lambda_1^{\tilde{\mathbf{A}}}| \\ 0, & \text{otherwise} \end{cases}$$

with convergence rate $\frac{|\lambda_i^{\tilde{\mathbf{A}}}|}{|\lambda_1^{\tilde{\mathbf{A}}}|}$.

As depth increases, only signals corresponding to the largest eigenvalue of the aggregation function significantly influence representations, while those in other subspaces become negligible. The feature transformations only affect the scale of all subspaces by a shared scalar. As previous studies did not have insights into this dominance of a subspace, they claimed potential over-separation of node representations when this scalar diverges to infinity [6, 12, 13]. Contrarily, considering the normalized representations $\frac{\mathbf{X}^{(l)}}{\|\mathbf{X}^{(l)}\|_F}$ seems to be necessary given our insights.

We now assume that $|\lambda_2^{\tilde{\mathbf{A}}}| < |\lambda_1^{\tilde{\mathbf{A}}}|$ is strict, so that a single subspace dominates. While representations may not converge as each $\mathbf{W}^{(l)}$ can introduce changes, they converge to the set $\mathcal{Q}_1$. This type of convergence and its properties were introduced by [31].

Theorem 2. *(Representations converge to a fixed set.) Let $\mathbf{X}^{(k+1)} = \tilde{\mathbf{A}}\mathbf{X}^{(k)}\mathbf{W}^{(k)}$ with $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times n}$ symmetric with eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and $\mathbf{W}^{(k)} \in \mathbb{R}^{d \times d}$ any matrix. Consider the subspaces $\mathcal{Q}_i = \mathrm{span}(\mathbf{I}_d \otimes \mathbf{v}_i)$ for $i \in [n]$. Then,*

$$\forall\, \epsilon > 0.\, \exists\, N \in \mathbb{N}.\, \forall\, l \in \mathbb{N}\ with\ l > N.\, \exists\, \mathbf{m}^{(l)} \in \mathcal{Q}_1 : \left\| \frac{\mathbf{X}^{(l)}}{\|\mathbf{X}^{(l)}\|_F} - \mathbf{m}^{(l)} \right\|_2 < \epsilon.$$

We emphasize that smoothing has yet to be part of our analysis. Our analysis and proofs are more general, relying solely on the symmetry of the aggregation function $\tilde{\mathbf{A}}$. Smoothing occurs only for specific aggregations for which $\mathcal{Q}_1$ consists of smooth signals. The symmetrically normalized aggregation function $\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ has this property, as its dominant eigenvector is $\mathbf{v}_1 = \mathbf{D}^{\frac{1}{2}}\mathbf{1}$ and $\lambda_1^{\tilde{\mathbf{A}}}$ is unique [32]. These novel insights enable us to frame our findings using the well-established Dirichlet energy of the normalized node representations, which corresponds to the Rayleigh quotient $E(\frac{\mathbf{X}}{\|\mathbf{X}\|_F}) = \frac{tr(\mathbf{X}^T \boldsymbol{\Delta} \mathbf{X})}{\|\mathbf{X}\|_F^2}$ [12]. The key property needed is the equivalence of the dominating subspace $\mathcal{Q}_1$ and the nullspace of $\mathbf{I}_d \otimes \boldsymbol{\Delta}$, allowing us to provide the novel proof for unconstrained feature transformations:

Proposition 3. *(Over-smoothing happens for all $\mathbf{W}^{(k)}$.) Let $\mathbf{X}^{(k+1)} = \tilde{\mathbf{A}}\mathbf{X}^{(k)}\mathbf{W}^{(k)}$ with $\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ and $\mathbf{W}^{(k)} \in \mathbb{R}^{d \times d}$ any matrix. We consider $E(\mathbf{H}) = tr(\mathbf{H}^T \boldsymbol{\Delta} \mathbf{H})$ for $\boldsymbol{\Delta} = \mathbf{I} - \tilde{\mathbf{A}}$. Then,*

$$\lim_{l \to \infty} E\left(\frac{\mathbf{X}^{(l)}}{\|\mathbf{X}^{(l)}\|_F} \right) = 0 \tag{9}$$

with convergence rate $\left(\frac{\lambda_2^{\tilde{\mathbf{A}}}}{\lambda_1^{\tilde{\mathbf{A}}}} \right)^2$ independently of all $\mathbf{W}^{(k)}$.

We note that Proposition 3 can be similarly applied to the sum aggregation $\mathbf{A}$. This emphasizes the importance of considering the normalized signal to determine over-smoothing.

5.3   Layer-wise Bounds are Insufficient

It remains open whether existing layer-wise bounds on the Dirichlet energy [12, 13] would similarly benefit from normalized representations as their assumption for over-smoothing, i.e., $\sigma_1^{\mathbf{W}} < \frac{1}{\lambda_2^{\tilde{\mathbf{A}}}}$, would be satisfied when applying a normalized feature transformation $\frac{\mathbf{W}}{\|\mathbf{W}\|_F}$. However, these statements do not benefit from normalization, and the Dirichlet energy of the normalized representations can increase after applying a graph convolution:

Proposition 4. *(Layer-wise bounds are insufficient.) There exist $\mathbf{X} \in \mathbb{R}^{n \times d}, \mathbf{W} \in \mathbb{R}^{d \times d}$, for which*

$$E\left(\frac{\mathbf{X}\mathbf{W}}{\|\mathbf{X}\mathbf{W}\|_F} \right) > E\left(\frac{\mathbf{X}}{\|\mathbf{X}\|_F} \right).$$

This underscores the need for a novel approach to analyze over-smoothing.

5.4   Extending the Analysis to Arbitrary Aggregations

We now expand our analysis to encompass all possible aggregation matrices and the graph structures they induce. While prior investigations (e.g., [6, 12, 13]) focused on specific types of aggregations that were not easily generalizable, our approach provides a more versatile framework. Here, we drop the assumption of a symmetric graph and allow for arbitrary edge weights between any pair of nodes, a concept recently proposed to account for attraction or repulsion between nodes [5, 33–35]. We find that Theorem 1 can be extended to arbitrary aggregation matrices $\tilde{\mathbf{A}}$, revealing that the dominating signal for these graph convolutions depends solely on the dominating signal of $\tilde{\mathbf{A}}$:

Proposition 5. *(Signal amplification only depends on $\tilde{\mathbf{A}}$.) Let two bases be* $\mathbf{S}_{(i)} = (\mathbf{I}_d \otimes \mathbf{P}_{(i)}) \in \mathbb{R}^{nd \times qd}$ *and* $\mathbf{S}_{(j)} = (\mathbf{I}_d \otimes \mathbf{P}_{(j)}) \in \mathbb{R}^{nd \times qd}$ *for any* $\mathbf{P}_{(i)} \in \mathbb{R}^{n \times q}, \mathbf{P}_{(j)} \in \mathbb{R}^{n \times r}$ *with* $q, r \leq n$. *Further let* $\mathbf{T} = \mathbf{W} \otimes \tilde{\mathbf{A}} \in \mathbb{R}^{nd \times nd}$ *consisting of any* $\mathbf{W} \in \mathbb{R}^{d \times d}$ *and any* $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times n}$. *Then,*

$$\frac{||\mathbf{T}\mathbf{S}_{(i)}||_F}{||\mathbf{T}\mathbf{S}_{(j)}||_F} = \frac{||\tilde{\mathbf{A}}\mathbf{P}_{(i)}||_F}{||\tilde{\mathbf{A}}\mathbf{P}_{(j)}||_F} \,.$$

This result illustrates that the dominating signal remains unaffected by feature transformations and relies solely on the dominant eigenvalues of $\tilde{\mathbf{A}}$. Importantly, invariant bases always exist, as given by the Jordan normal form [36]. Preventing over-smoothing can therefore be achieved by selecting or learning an aggregation matrix $\tilde{\mathbf{A}}$ with a well-suited spectrum.

5.5   Stochastic Aggregation Functions

Our findings can be readily extended to cases where we have knowledge about the dominance of a particular signal for the aggregation function itself. This applies to all row-stochastic aggregations, also referred to as the weighted mean, where the values in each row of $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times n}$ are non-negative and sum up to one. This analysis covers models such as the Graph Attention Network (GAT) [17, 37] with a single attention head. Edge weights may be dynamically computed based on their adjacent nodes, so we consider the case of time-inhomgeneous aggregation matrices $\tilde{\mathbf{A}}(0) \neq \tilde{\mathbf{A}}(1) \neq \ldots \neq \tilde{\mathbf{A}}(l)$. Assuming the underlying graph is ergodic, each $\tilde{\mathbf{A}}^{(k)}$ possesses the same right-eigenvector $\mathbf{p}_1 = \mathbf{1}$ with corresponding eigenvalue $\lambda_1^{\tilde{\mathbf{A}}^{(k)}} = 1$. All other eigenvalues are strictly less in absolute value as given by the Perron-Frobenius theorem [38] and stochastic processes [39]. Given a minimum edge weight bound $\epsilon > 0$, we build on the insight from Wu et al. [23] that any product $\Pi_{k=0}^{\infty} \tilde{\mathbf{A}}^{(k)}$ also converges to a matrix with constant rows. This subspace dominates the representations as given by Proposition 5, and we show over-smoothing using the Dirichlet energy:

Proposition 6. *(GAT and Graph Transformer over-smooth.) Let* $\mathbf{X}^{(k+1)} = \tilde{\mathbf{A}}^{(k)}\mathbf{X}^{(k)}\mathbf{W}^{(k)}$ *where all* $\tilde{\mathbf{A}}^{(k)} \in \mathbb{R}^{n \times n}$ *row-stochastic, representing the same ergodic graph, and each non-zero entry* $(\tilde{\mathbf{A}}^{(k)})_{pq} \geq \epsilon$ *for some* $\epsilon > 0$ *and all* $i \in [h]$,

$k \in \mathbb{N}$, $p, q \in [n]$. We further allow all $\mathbf{W}_i^{(k)} \in \mathbb{R}^{d \times d}$ to be any matrices. We consider $E(\mathbf{X}) = tr(\mathbf{X}^T \boldsymbol{\Delta}_{rw} \mathbf{X})$ using the random walk Laplacian $\boldsymbol{\Delta}_{rw} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}$. Then,

$$\lim_{l \to \infty} E\left(\frac{\mathbf{X}^{(l)}}{\|\mathbf{X}^{(l)}\|_F}\right) = 0\,.$$

6   Rank Collapse and Over-Correlation

We have seen that the choice of the aggregation function leads to over-smoothing, which can therefore be prevented by selecting an aggregation with a different spectrum. However, the representations are always dominated by some low-dimensional subspace $\mathcal{Q}_1 = \text{span}(\mathbf{I}_d \otimes \mathbf{V}_1)$ based on some matrix $\mathbf{V}_1 \in \mathbb{R}^{n \times j}$ and the algebraic multiplicity j of all eigenvalues with the maximal absolute value $|\lambda_1^{\tilde{\mathbf{A}}}|$. However, node representations in $\mathcal{Q}_1$ cannot exceed the rank of j:

Theorem 3. *(The rank collapses.) Let $\mathcal{Q}_i = \text{span}(\mathbf{I}_d \otimes \mathbf{V}_i)$ for any matrix $\mathbf{V}_i \in \mathbb{R}^{n \times j}$ with $j \leq n$. Then,*

$$\forall \ \text{vec}(\mathbf{X}) \in \mathcal{Q}_i \colon \text{rank}(\mathbf{X}) \leq j\,. \tag{10}$$

This implies that, regardless of the aggregation function chosen, the representations collapse to a low-rank state bounded by the algebraic multiplicity j of the dominant eigenvalue. This poses a significant challenge for the expressiveness of deep graph neural networks, as node representations can only contain, at most, j relevant features. It also leads to a perfect correlation among features if $\text{rank}(\mathbf{V}_i) = 1$, e.g., when the dominant eigenvalue is unique. [8] empirically observed this phenomenon and termed it over-correlation. Due to the exponential convergence of the normalized state towards $\mathcal{Q}_1$, over-correlation of all features occurs at an exponential rate. This study reveals that the fundamental issue underlying over-smoothing, over-correlation, and rank collapse is the same. These findings also establish a connection between these phenomena and recent theoretical insights on deep Transformer models, where rank collapse has also been independently observed [40, 41].

7   Preventing Rank Collapse with a Sum of Kronecker Products

The core issue leading to rank collapse is the fact that the transformation matrix $\mathbf{T} = \mathbf{W} \otimes \mathbf{A}$ can be factored into a single Kronecker product. To counteract rank collapse and consequently over-smoothing and over-correlation, we need to construct a $\mathbf{T}$ that cannot be decomposed into a single Kronecker product. We highlight that any matrix

$$\mathbf{T} = (\mathbf{W}_1 \otimes \mathbf{A}_1) + (\mathbf{W}_2 \otimes \mathbf{A}_2) + \cdots + (\mathbf{W}_p \otimes \mathbf{A}_p) \tag{11}$$

can be decomposed into a finite sum of Kronecker products (SKP), as concatenation is a special case [42, 43]. While the minimum number of terms p required is not universally determined [42], we show that at most d terms are sufficient to amplify arbitrary signals across d columns:

Theorem 4. *(Any subspace can get amplified.) Let $\mathbf{e}_i \in \mathbb{R}^d$ be the canonical basis with a single 1 at position i. For any columns $\mathbf{s}_1, \ldots, \mathbf{s}_d \in \mathbb{R}^n$ and the induced subspace $\mathbf{S} = \begin{bmatrix} \mathbf{e}_1 \otimes \mathbf{s}_1 & \ldots & \mathbf{e}_d \otimes \mathbf{s}_d \end{bmatrix} \in \mathbb{R}^{nd \times d}$, there exists an $\mathbf{T} = \sum_1^d (\mathbf{W}_i \otimes \tilde{\mathbf{A}}_i)$ such that for all orthogonal bases $\mathbf{S}' = \begin{bmatrix} \mathbf{e}_1 \otimes \mathbf{s}'_1 & \ldots & \mathbf{e}_d \otimes \mathbf{s}'_d \end{bmatrix} \in \mathbb{R}^{nd \times d}$*

$$\frac{\|\mathbf{TS}\|_F}{\|\mathbf{TS}'\|_F} > \frac{\|\mathbf{S}\|_F}{\|\mathbf{S}'\|_F}. \tag{12}$$

This is most directly achieved when the employed transformations filter the columns used by each aggregation. Even using a sum of just two terms for $\mathbf{T}$ improves the expressiveness of deep GNNs. The SKP does not add computational complexity, as we would not compute Kronecker products, i.e., $(\mathbf{W}_1 \otimes \tilde{\mathbf{A}}_1) + \cdots + (\mathbf{W}_p \otimes \tilde{\mathbf{A}}_p)) \operatorname{vec}(\mathbf{X}) = \operatorname{vec}(\tilde{\mathbf{A}}_1 \mathbf{X} \mathbf{W}_1 + \cdots + \tilde{\mathbf{A}}_p \mathbf{X} \mathbf{W}_p)$. It is important to note that the concept of a SKP is not tied to a specific model, but rather a guiding principle that helps design methods with favorable properties. This understanding also provides theoretical insights into the success of many existing methods that can be understood as SKPs, including residual connections [44–46], mixing aggregation functions [47] or transforming signals from incoming and outgoing edges differently [48]. In addition, we briefly discuss some important methods:

Multi-Head Graph Attention Networks Graph attention networks (GATs) [17] introduced the concept of using multiple attention heads to filter relevant neighboring information. We denote pre-activated attention coefficients for p heads as $\tilde{\mathbf{A}}_1, \ldots, \tilde{\mathbf{A}}_p \in \mathbb{R}^{n \times n}$ and the row-wise softmax operation by $\tau(\cdot)$. Then, GAT can be expressed as the SKP

$$\mathbf{T} = \frac{1}{p}(\mathbf{W}_1 \otimes \tau(\tilde{\mathbf{A}}_1)) + \cdots + \frac{1}{p}(\mathbf{W}_p \otimes \tau(\tilde{\mathbf{A}}_p)), \tag{13}$$

where separate feature transformations $\mathbf{W}_1, \ldots, \mathbf{W}_p$ for each attention head. However, rank collapse may persist, since all products of aggregation matrices share the same dominating subspace $\mathcal{Q}_1 = (\mathbf{I}_d \otimes \mathbf{v}_1)$ (see Proposition 3).

Discrete Convolutions The discrete convolution, as commonly employed in convolutional neural networks(CNN) [49], is an SKP that can avoid rank collapse. For instance, consider a 3×3 kernel, and translate this operation into a graph operation where each pixel corresponds to a node. Edges are formed between adjacent pixels, and each neighboring state is transformed differently based on their relative direction. We express this as an SKP

$$\mathbf{T} = \frac{1}{9}(\mathbf{W}_1 \otimes \tilde{\mathbf{A}}_1) + \frac{1}{9}(\mathbf{W}_2 \otimes \tilde{\mathbf{A}}_2) + \ldots \frac{1}{9}(\mathbf{W}_9 \otimes \tilde{\mathbf{A}}_9) \tag{14}$$

with each aggregation $\mathbf{A}_1, \ldots, \mathbf{A}_9$ containing edges for a single relative direction.
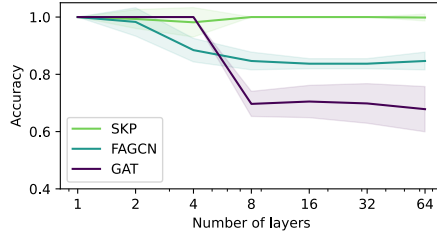
Fig. 2: Mean accuracies (bold) and standard deviations (shaded) across 50 random ER graphs for different numbers of layers.
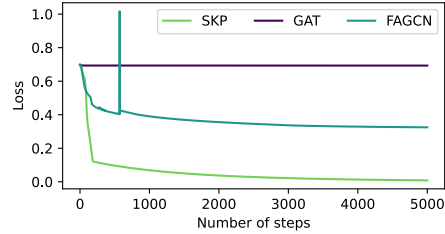
Fig. 3: Loss progression during parameter optimization for one graph with each model having eight convolutional layers.

8   Empirical Validation

Our conducted empirical validation aims to confirm whether our theoretical insights hold true in the non-linear case of finite depth. The theory states that models utilizing a single Kronecker product suffer from rank collapse, limiting their learnability of tasks that require multiple independent features per node. Given any graph, if the feature space collapses to two or fewer dimensions, the representations of any four nodes always form a quadrilateral in that plane. Linear decision boundaries cannot classify points on opposite sides together. The features extracted by SKP can be linearly independent (see Theorem 6.1), so tasks of this form are trivial when accessing another dimension. Our experimental setting follows this idea: Given a random Erdős–Rényi (ER) graph [50] with 20 nodes, we consider a 3-class classification task, where we evaluate the performance for four random nodes, for which we assign each pair to the same class for one of the tasks. We consider the following three model families:

$$\mathbf{T}_{\text{GAT}}^{(k)} = \frac{1}{2}(\mathbf{W}_1^{(k)} \otimes \tau(\tilde{\mathbf{A}}_1^{(k)})) + \frac{1}{2}(\mathbf{W}_2^{(k)} \otimes \tau(\tilde{\mathbf{A}}_2^{(k)})), \qquad (15)$$

$$\mathbf{T}_{\text{FAGCN}}^{(k)} = \mathbf{W}_1^{(k)} \otimes \tilde{\mathbf{A}}_1^{(k)}, \qquad (16)$$

$$\mathbf{T}_{\text{SKP}}^{(k)} = \frac{1}{2}(\mathbf{W}_1^{(k)} \otimes \tilde{\mathbf{A}}_1^{(k)}) + \frac{1}{2}(\mathbf{W}_2^{(k)} \otimes \tilde{\mathbf{A}}_2^{(k)}), \qquad (17)$$

where the edge weights $\tilde{\mathbf{A}}_1^{(k)}, \tilde{\mathbf{A}}_2^{(k)} \in \mathbb{R}^{20 \times 20}$, and the feature transformations $\mathbf{W}_1^{(k)}, \mathbf{W}_2^{(k)} \in \mathbb{R}^{6 \times 6}$ are randomly initialized and directly optimized, so that they can represent any models of these forms. To validate that non-linear models suffer from rank collapse, ReLU activations are applied after each $\mathbf{T}^{(k)}$. $\mathbf{T}_{\text{GAT}}^{(k)}$ extends our theory for row-stochastic models activated by a row-wise softmax τ to the multi-head case and $\mathbf{T}_{\text{FAGCN}}^{(k)}$ allows arbitrary aggregations including negative edge weights, as in FAGCN [34] and several other methods [5, 33, 35]. $\mathbf{T}_{\text{SKP}}^{(k)}$ is a mix of both directions, as the edge weights are not normalized to avoid matching spectrums. The update function is $\text{vec}(\mathbf{X}^{(k+1)}) = \phi(\mathbf{T}^{(k)} \text{vec}(\mathbf{X}^{(k)}))$ with ϕ being

the ReLU activation function. After k such layers, the node representations are mapped to class probabilities by separate affine transformations and a sigmoid activation for each task. All model parameters are optimized to minimize the binary cross-entropy of the four nodes of interest using Adam [51] until convergence, i.e., the loss is not decreased for 500 steps. Node states are initialized to be normally distributed, and we consider the best fit of three random initializations for each graph. Our reproducible implementation is based on PyTorch [52], and further details are provided in Appendix B.

We report the mean accuracies and standard deviations over 50 random graphs with an edge probability of 0.2 for $k \in [1, 2, 4, 8, 16, 32, 64]$ layers in Figure 2. GAT cannot fit the data for a single graph to 100% accuracy for 8 or more layers, while FAGCN does not reach perfect accuracy for four or more layers. With increased depth, GAT converges to a lower accuracy than FAGCN, as it is dominated by smooth signals, while the dominating eigenspace of FAGCN can be arbitrary. Contrarily, SKP fits the data to perfect accuracy for 341 out of 350 graphs, though it is occasionally stuck in local minima. We provide results for the time-homogeneous and linear case in Appendix B, but find the results to be similar. We additionally visualize the optimization process for the considered models with eight convolutional layers in Figure 3. While SKP converges to a loss close to zero, due to rank collapse, FAGCN converges to a loss of 0.32, and GAT converges to a loss of 0.69. This experiment serves as proof of concept that even non-linear GNNs suffer from rank collapse.

9   Conclusion

Our work shows that rank collapses is the underlying cause of over-smoothing and over-correlation in GNNs. We provided novel proofs that the dominating subspace of a graph convolution only depends on the dominating eigenspace of the aggregation function and is independent of feature transformations. Our work also emphasizes the importance of feature normalization when determining and quantifying over-smoothing and over-correlation, as we show that over-separation is merely an artifact of the Dirichlet energy on unnormalized node representations. To mitigate the limited expressivity due to this rank collapse of node representations, we propose the sum of Kronecker products (SKP) as a general property models should exhibit that can provably prevent rank collapse. We empirically confirm this behavior for non-linear GNNs in a multi-class classification task, which typical existing methods cannot fit, but an SKP can.

The main limitation of our work is that the limit case only applies to linearized GNNs, and our empirical validation extends and confirms our theory for the non-linear case only for synthetic tasks. Our insights show that future methods should aim to avoid rank collapse instead of dealing with over-smoothing or over-correlation. Novel metrics to quantify the degree of rank collapse in graph neural networks need to be designed, and these need to consider the normalized feature representation, as the feature magnitude may otherwise overshadow the insights.

Bibliography

[1] Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. In *International Conference on Learning Representations*, 2018.

[2] Hannes Stärk, Octavian Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. Equibind: Geometric deep learning for drug binding structure prediction. In *International Conference on Machine Learning*, pages 20503–20521. PMLR, 2022.

[3] Ahmed El-Kishky, Michael Bronstein, Ying Xiao, and Aria Haghighi. Graph-based representation learning for web-scale recommender systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 4784–4785, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. https://doi.org/10.1145/3534678.3542598.

[4] Andreas Roth and Thomas Liebig. Forecasting unobserved node states with spatio-temporal graph neural networks. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 740–747. IEEE, 2022.

[5] Yujun Yan, Milad Hashemi, Kevin Swersky, Yaoqing Yang, and Danai Koutra. Two sides of the same coin: Heterophily and oversmoothing in graph convolutional neural networks. *arXiv preprint arXiv:2102.06462*, 2021.

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

[7] Hoang Nt and Takanori Maehara. Revisiting graph neural networks: All we have is low-pass filters. *arXiv preprint arXiv:1905.09550*, 2019.

[8] Wei Jin, Xiaorui Liu, Yao Ma, Charu Aggarwal, and Jiliang Tang. Feature overcorrelation in deep graph neural networks: A new perspective. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 709–719, 2022.

[9] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI conference on artificial intelligence*, 2018.

[10] Nicolas Keriven. Not too little, not too much: a theoretical analysis of graph (over) smoothing. *arXiv preprint arXiv:2205.12156*, 2022.

[11] T Konstantin Rusch, Ben Chamberlain, James Rowbottom, Siddhartha Mishra, and Michael Bronstein. Graph-coupled oscillator networks. In *International Conference on Machine Learning*, pages 18888–18909. PMLR, 2022.

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

[13] Kaixiong Zhou, Xiao Huang, Daochen Zha, Rui Chen, Li Li, Soo-Hyun Choi, and Xia Hu. Dirichlet energy constrained learning for deep graph

neural networks. *Advances in Neural Information Processing Systems*, 34: 21834–21846, 2021.

[14] 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.

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

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

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

[18] 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.

[19] Francesco Di Giovanni, James Rowbottom, Benjamin P Chamberlain, Thomas Markovich, and Michael M Bronstein. Graph neural networks as gradient flows. *arXiv preprint arXiv:2206.10991*, 2022.

[20] Sohir Maskey, Raffaele Paolino, Aras Bacho, and Gitta Kutyniok. A fractional graph laplacian approach to oversmoothing. *arXiv preprint arXiv:2305.13084*, 2023.

[21] T Konstantin Rusch, Benjamin P Chamberlain, Michael W Mahoney, Michael M Bronstein, and Siddhartha Mishra. Gradient gating for deep multi-rate learning on graphs. *arXiv preprint arXiv:2210.00513*, 2022.

[22] T Konstantin Rusch, Michael M Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks. *arXiv preprint arXiv:2303.10993*, 2023.

[23] Xinyi Wu, Amir Ajorlou, Zihui Wu, and Ali Jadbabaie. Demystifying oversmoothing in attention-based graph neural networks. *arXiv preprint arXiv:2305.16102*, 2023.

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

[25] Lingxiao Zhao and Leman Akoglu. Pairnorm: Tackling oversmoothing in gnns. In *International Conference on Learning Representations*, 2020.

[26] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *International Conference on Machine Learning*, pages 1725–1735. PMLR, 2020.

[27] Andreas Roth and Thomas Liebig. Transforming pagerank into an infinite-depth graph neural network. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part II*, pages 469–484. Springer, 2023.

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

[29] Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3:127–163, 2000.

[30] Meng Liu, Hongyang Gao, and Shuiwang Ji. Towards deeper graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 338–348, 2020.

[31] Akira Iwasa, Masaru Kada, and Shizuo Kamo. Preservation of a convergence of a sequence to a set. In *Topology Proceedings*, volume 44, page 97, 2014.

[32] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[33] Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized pagerank graph neural network. In *International Conference on Learning Representations*, 2020.

[34] Deyu Bo, Xiao Wang, Chuan Shi, and Huawei Shen. Beyond low-frequency information in graph convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3950–3957, 2021.

[35] Cristian Bodnar, Francesco Di Giovanni, Benjamin Paul Chamberlain, Pietro Lio, and Michael M Bronstein. Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in gnns. In *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*, 2022.

[36] Gilbert Strang. *Linear algebra and its applications*. Belmont, CA: Thomson, Brooks/Cole, 2006.

[37] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.

[38] Oskar Perron. Zur theorie der matrices. *Mathematische Annalen*, 64(2): 248–263, 1907.

[39] Robert G Gallager. Discrete stochastic processes. *Journal of the Operational Research Society*, 48(1):103–103, 1997.

[40] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pages 2793–2803. PMLR, 2021.

[41] Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *Advances in Neural Information Processing Systems*, 35:27198–27211, 2022.

[42] Charles F Van Loan and Nikos Pitsianis. *Approximation with Kronecker products*. Springer, 1993.

[43] Jian Cao, Marc G Genton, David E Keyes, and George M Turkiyyah. Sum of kronecker products representation and its cholesky factorization for spatial covariance matrices from large grids. *Computational Statistics & Data Analysis*, 157:107165, 2021.

[44] 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.

[45] Xavier Bresson and Thomas Laurent. Residual gated graph convnets. *arXiv preprint arXiv:1711.07553*, 2017.

[46] Guohao Li, Matthias Müller, Bernard Ghanem, and Vladlen Koltun. Training graph neural networks with 1000 layers. In *International conference on machine learning*, pages 6437–6449. PMLR, 2021.

[47] Eran Rosenbluth, Jan Toenshoff, and Martin Grohe. Some might say all you need is sum. *arXiv preprint arXiv:2302.11603*, 2023.

[48] Emanuele Rossi, Bertrand Charpentier, Francesco Di Giovanni, Fabrizio Frasca, Stephan Günnemann, and Michael Bronstein. Edge directionality improves learning on heterophilic graphs. *arXiv preprint arXiv:2305.10498*, 2023.

[49] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. https://doi.org/10.1162/neco.1989.1.4.541.

[50] P. Erdös and A. Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290, 1959.

[51] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[52] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

A   Mathematical details

In this section, we provide the details of our approach and all of our statements.

A.1   Basic Operations

We start by listing the most important properties used throughout our formal proofs.

Kronecker Product. The Kronecker product for any two matrices $\mathbf{A} \in \mathbb{R}^{p \times q}$, $\mathbf{B} \in \mathbb{R}^{r \times s}$ is denoted as

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \dots & a_{1q}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{p1}\mathbf{B} & \dots & a_{pq}\mathbf{B} \end{bmatrix}. \tag{18}$$

The importance of the Kronecker product for our work stems from its powerful properties. We briefly present the most relevant here. First, a vectorized matrix product

$$\text{vec}(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W}) = (\mathbf{W}^T \otimes \tilde{\mathbf{A}}) \, \text{vec}(\mathbf{X}) \tag{19}$$

of any matrices $\tilde{\mathbf{A}}, \mathbf{X}, \mathbf{W}$ with matching shapes can be written using the Kronecker product. The Kronecker product of two orthogonal matrices results in an orthogonal matrix, allowing us to rewrite any vector as a linear combination

$$\text{vec}(\mathbf{X}) = (\mathbf{U} \otimes \mathbf{V})\mathbf{c} \tag{20}$$

using the singular value decomposition $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{N}^T$ and eigendecomposition $\tilde{\mathbf{A}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$. The Kronecker product also satisfies the mixed-product property

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{A}\mathbf{C}) \otimes (\mathbf{B}\mathbf{D}). \tag{21}$$

Dirichlet Energy The standard interpretation for the Dirichlet energy is the sum of differences in representations for adjacent nodes. Another interpretation we mainly use is based on the decomposition of the signal into eigenvectors of the graph Laplacian $\mathbf{\Delta} = \mathbf{I} - \tilde{\mathbf{A}}$ using $\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$. Utilizing the Kronecker product, only the signals not belonging to eigenvalue $\lambda_1^{\tilde{\mathbf{A}}} = 1$ are summed up and weighted by the corresponding eigenvalue of $\mathbf{\Delta}$:

$$E(\mathbf{X}) = tr(\mathbf{X}^T \mathbf{\Delta}\mathbf{X}) \tag{22}$$

$$= \text{vec}(\mathbf{X})^T \text{vec}(\mathbf{\Delta}\mathbf{X}). \tag{23}$$

We then use the decomposition $vec(\mathbf{X}) = (\mathbf{I}_d \otimes \mathbf{V})\mathbf{c}$ as described in Eq. **??**, based on the eigenvectors $\mathbf{V}$ of $\tilde{\mathbf{A}}$ and the identity matrix $\mathbf{I}_d$, leading to

$$E(\mathbf{X}) = \mathbf{c}^T(\mathbf{I}_d \otimes \mathbf{V})^T(\mathbf{I}_d \otimes \mathbf{\Delta})(\mathbf{I}_d \otimes \mathbf{V})\mathbf{c} \tag{24}$$

$$= \mathbf{c}^T(\mathbf{I}_d \otimes \mathbf{V}^T\mathbf{\Delta}\mathbf{V})\mathbf{c}. \tag{25}$$

As $\boldsymbol{\Delta} = \mathbf{V}(\mathbf{I}_n - \boldsymbol{\Lambda})\mathbf{V}$ has the same eigenvectors and shifted eigenvalues as $\tilde{\mathbf{A}}$, we then write the statement as a sum of coefficients that are weighted by their corresponding eigenvalue of the graph Laplacian:

$$E(\mathbf{X}) = \mathbf{c}^T(\mathbf{I}_d \otimes (\mathbf{I}_n - \boldsymbol{\Lambda}^{\boldsymbol{\Delta}})\mathbf{c} \tag{26}$$

$$= \sum_{l,r=1}^{n,d} (1 - \lambda_r^{\tilde{\mathbf{A}}})c_{l,r}^2 . \tag{27}$$

Frobenius Norm The squared Frobenius norm has a similar interpretation, the coefficients are just not weighted by eigenvalues:

$$\begin{aligned} ||\mathbf{X}||_F^2 &= tr(\mathbf{X}^T\mathbf{X}) \\ &= \mathbf{c}^T(\mathbf{U} \otimes \mathbf{V})^T(\mathbf{U} \otimes \mathbf{V})\mathbf{c} \\ &= \sum_{l,r=1}^{n,d} c_{l,r}^2 \end{aligned} \tag{28}$$

An important property of the Frobenius norm in conjunction with the Kronecker product is the following:

$$\|\mathbf{A} \otimes \mathbf{B}\|_F = \|\mathbf{A}\|_F \cdot \|\mathbf{B}\|_F \tag{29}$$

A.2 Proof of Proposition 4.1

Proposition. *(Node representations vanish.) Let $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times n}$ be symmetric with maximum absolute eigenvalue $|\lambda_1^{\tilde{\mathbf{A}}}| = 1$, $\mathbf{W} \in \mathbb{R}^{d \times d}$ be any matrix with maximum singular value $\sigma_1^{\mathbf{W}}$, and ϕ a component-wise non-expansive mapping satisfying $\phi(\mathbf{0}) = \mathbf{0}$. Then,*

$$||\phi(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W})||_F \le \sigma_1^{\mathbf{W}} \cdot ||\mathbf{X}||_F . \tag{30}$$

Proof. The key property we use is that the non-expansive property of $\phi(\cdot)$ implies the Lipschitz continuity $||\phi(\mathbf{X}) - \phi(\mathbf{Y})|| \le ||\mathbf{X} - \mathbf{Y}||$:

$$||\phi(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W})||_F = ||\phi(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W}) - \phi(\mathbf{0})||_F \tag{31}$$

$$\le ||\tilde{\mathbf{A}}\mathbf{X}\mathbf{W} - \mathbf{0}||_F \tag{32}$$

$$= ||\tilde{\mathbf{A}}\mathbf{X}\mathbf{W}||_F . \tag{33}$$

We then use common bounds on the norm of the matrix product for symmetric matrices using the maximum eigenvalue $\lambda_1^{\tilde{\mathbf{A}}}$ and for an arbitrary matrix based on the maximum singular value $\sigma_1^{\mathbf{W}}$, resulting in

$$||\tilde{\mathbf{A}}\mathbf{X}\mathbf{W}||_F \le |\lambda_1^{\tilde{\mathbf{A}}}|\sigma_1^{\mathbf{W}}||\mathbf{X}||_F \tag{34}$$

$$= \sigma_1^{\mathbf{W}}||\mathbf{X}||_F , \tag{35}$$

using the assumption $|\lambda_1^{\tilde{\mathbf{A}}}| = 1$.

Proof of Lemma 5.1

Lemma. *(The subspaces are invariant to any $\mathbf{T}^{(k)}$.) Let $\mathbf{T} = \mathbf{W} \otimes \tilde{\mathbf{A}}$ with $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times n}$ symmetric with eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and $\mathbf{W} \in \mathbb{R}^{d \times d}$ any matrix. Consider the subspaces $\mathcal{Q}_i = \text{span}(\mathbf{I}_d \otimes \mathbf{v}_i)$ for $i \in [n]$. Then,*

$$\forall\, i \in [n] : \mathbf{z} \in \mathcal{Q}_i \implies \mathbf{T}\mathbf{z} \in \mathcal{Q}_i\,.$$

Proof. We express $\mathbf{z}$ as a linear combination $(\mathbf{I}_d \otimes \mathbf{v}_i)\mathbf{c} = \mathbf{z} \in \mathcal{Q}_i$ of the given basis vectors. Then,

$$
\begin{aligned}
(\mathbf{W} \otimes \tilde{\mathbf{A}})(\mathbf{I}_d \otimes \mathbf{v}_i)\mathbf{c} &= (\mathbf{W}\mathbf{I}_d \otimes \tilde{\mathbf{A}}\mathbf{v}_i)\mathbf{c} \\
&= (\mathbf{I}_d \otimes \mathbf{v}_i)(\mathbf{W} \otimes \lambda_i^{\tilde{\mathbf{A}}}\mathbf{I}_n)\mathbf{c} \\
&= (\mathbf{I}_d \otimes \mathbf{v}_i)\mathbf{c}' \in \mathcal{Q}_i
\end{aligned}
\tag{36}
$$

using some new coefficients $\mathbf{c}' = (\mathbf{W} \otimes \lambda_i^{\tilde{\mathbf{A}}}\mathbf{I}_n)\mathbf{c}$. $\qquad\square$

Proof of Theorem 5.2

Theorem. *(The relative behavior is fixed.) Let $\mathbf{T} = \mathbf{W} \otimes \tilde{\mathbf{A}}$ with $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times n}$ symmetric with eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and eigenvalues $\lambda_1^{\tilde{\mathbf{A}}}, \ldots, \lambda_n^{\tilde{\mathbf{A}}}$ and $\mathbf{W} \in \mathbb{R}^{d \times d}$ any matrix. Consider the bases $\mathbf{S}_{(i)} = \mathbf{I}_d \otimes \mathbf{v}_i$ and $\mathbf{S}_{(j)} = \mathbf{I}_d \otimes \mathbf{v}_j$ for $i, j \in [n]$. Then,*

$$\frac{\|\mathbf{T}\mathbf{S}_{(i)}\|_F}{\|\mathbf{T}\mathbf{S}_{(j)}\|_F} = \frac{|\lambda_i^{\tilde{\mathbf{A}}}|}{|\lambda_j^{\tilde{\mathbf{A}}}|}\,. \tag{37}$$

Proof.

$$
\begin{aligned}
\frac{\|\mathbf{T}\mathbf{S}_{(i)}\|}{\|\mathbf{T}\mathbf{S}_{(j)}\|} &= \frac{\left\|\left(\mathbf{W} \otimes \tilde{\mathbf{A}}\right)(\mathbf{I}_d \otimes \mathbf{v}_i)\right\|}{\left\|\left(\mathbf{W} \otimes \tilde{\mathbf{A}}\right)(\mathbf{I}_d \otimes \mathbf{v}_j)\right\|} \\[2mm]
&= \frac{\left\|\left(\mathbf{W} \otimes \lambda_i^{\tilde{\mathbf{A}}}\mathbf{v}_i\right)\right\|}{\left\|\left(\mathbf{W} \otimes \lambda_j^{\tilde{\mathbf{A}}}\mathbf{v}_j\right)\right\|} \\[2mm]
&= \frac{|\lambda_i^{\tilde{\mathbf{A}}}| \cdot \|\mathbf{W}\| \cdot \|\mathbf{v}_i\|}{|\lambda_j^{\tilde{\mathbf{A}}}| \cdot \|\mathbf{W}\| \cdot \|\mathbf{v}_j\|} \\[2mm]
&= \frac{|\lambda_i^{\tilde{\mathbf{A}}}|}{|\lambda_j^{\tilde{\mathbf{A}}}|}
\end{aligned}
\tag{38}
$$

$\square$

Proof of Proposition 5.3

Proposition. *(Fixed subspaces dominate.) Let $\mathbf{T}^{(k)} = \mathbf{W}^{(k)} \otimes \tilde{\mathbf{A}}$ with $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times n}$ symmetric with eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and eigenvalues $\lambda_1^{\tilde{\mathbf{A}}}, \ldots, \lambda_n^{\tilde{\mathbf{A}}}$ and $\mathbf{W}^{(k)} \in \mathbb{R}^{d \times d}$ any matrix. Consider the bases $\mathbf{S}_{(i)} = \mathbf{I}_d \otimes \mathbf{v}_i$ for $i \in [n]$. Then,*

$$\lim_{l \to \infty} \frac{\|\mathbf{T}^{(l)} \ldots \mathbf{T}^{(1)} \mathbf{S}_{(i)}\|_F}{\max_p \|\mathbf{T}^{(l)} \ldots \mathbf{T}^{(1)} \mathbf{S}_{(p)}\|_F} = \lim_{l \to \infty} \frac{|\lambda_i^{\tilde{\mathbf{A}}}|^l \cdot \|\mathbf{S}_{(i)}\|_F}{\max_p |\lambda_p^{\tilde{\mathbf{A}}}|^l \cdot \|\mathbf{S}_{(p)}\|_F} = \begin{cases} 1, & \text{if} |\lambda_i^{\tilde{\mathbf{A}}}| = |\lambda_1^{\tilde{\mathbf{A}}}| \\ 0, & \text{otherwise} \end{cases}$$

with convergence rate $\frac{|\lambda_i^{\tilde{\mathbf{A}}}|}{|\lambda_1^{\tilde{\mathbf{A}}}|}$.

Proof.

$$\lim_{l \to \infty} \frac{\|\mathbf{T}^{(l)} \ldots \mathbf{T}^{(1)} \mathbf{S}_{(i)}\|}{\max_p \|\mathbf{T}^{(l)} \ldots \mathbf{T}^{(1)} \mathbf{S}_{(p)}\|} = \lim_{l \to \infty} \frac{||(\mathbf{W}^{(l)} \otimes \tilde{\mathbf{A}}) \ldots (\mathbf{W}^{(1)} \otimes \tilde{\mathbf{A}})(\mathbf{I}_d \otimes \mathbf{v}_i)||}{\max_p ||(\mathbf{W}^{(l)} \otimes \tilde{\mathbf{A}}) \ldots (\mathbf{W}^{(1)} \otimes \tilde{\mathbf{A}})(\mathbf{I}_d \otimes \mathbf{v}_p)||}$$

$$= \lim_{l \to \infty} \frac{||(\mathbf{W}^{(l)} \ldots \mathbf{W}^{(1)} \otimes \left(\lambda_i^{\tilde{\mathbf{A}}}\right)^l \mathbf{v}_i)||}{\max_p ||(\mathbf{W}^{(l)} \ldots \mathbf{W}^{(1)} \otimes \left(\lambda_p^{\tilde{\mathbf{A}}}\right)^l \mathbf{v}_p)||}$$

$$= \lim_{l \to \infty} \frac{|\lambda_i^{\tilde{\mathbf{A}}}|^l \cdot ||\mathbf{v}_i||}{\max_p ||\lambda_p^{\tilde{\mathbf{A}}}|^l| \cdot ||\mathbf{v}_p||}$$

$$= \begin{cases} 1, & \text{if} |\lambda_i^{\tilde{\mathbf{A}}}| = |\lambda_1^{\tilde{\mathbf{A}}}| \\ 0, & \text{otherwise} \end{cases}$$

$$\tag{39}$$

$\square$

Proof of Theorem 5.4

Theorem. *(Representations converge to a fixed set.) Let $\mathbf{X}^{(k+1)} = \tilde{\mathbf{A}} \mathbf{X}^{(k)} \mathbf{W}^{(k)}$ with $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times n}$ symmetric with eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and $\mathbf{W}^{(k)} \in \mathbb{R}^{d \times d}$ any matrix. Consider the subspaces $\mathcal{Q}_i = \mathrm{span}(\mathbf{I}_d \otimes \mathbf{v}_i)$ for $i \in [n]$. Then,*

$$\forall \epsilon > 0, \exists N \in \mathbb{N}, \forall l \in \mathbb{N} \, with \, l > N, \exists \mathbf{m}^{(l)} \in \mathcal{Q}_1 : \left\| \frac{\mathbf{X}^{(l)}}{\|\mathbf{X}^{(l)}\|_F} - \mathbf{m}^{(l)} \right\|_2 < \epsilon.$$

Proof. Let $\mathbf{\Lambda}$ be the matrix of eigenvalues of $\tilde{\mathbf{A}}$ and the singular value composition of $\mathbf{W}^{(l)} \ldots \mathbf{W}^{(1)} = \mathbf{U} \mathbf{\Sigma} \mathbf{N}^T$. Then,

$$\left\| \frac{\mathbf{X}^{(l)}}{\|\mathbf{X}^{(l)}\|} - \mathbf{m}^{(l)} \right\| = \left\| \frac{(\mathbf{W}^{(l)} \otimes \tilde{\mathbf{A}}) \ldots (\mathbf{W}^{(1)} \otimes \tilde{\mathbf{A}})(\mathbf{N}_d \otimes \mathbf{V})\mathbf{c}}{||(\mathbf{W}^{(l)} \otimes \tilde{\mathbf{A}}) \ldots (\mathbf{W}^{(1)} \otimes \tilde{\mathbf{A}})(\mathbf{N}_d \otimes \mathbf{V})\mathbf{c}||} - \mathbf{m}^{(l)} \right\|$$

$$= \left\| \frac{(\mathbf{W}^{(l)} \ldots \mathbf{W}^{(1)} \otimes \mathbf{V})(\mathbf{N}_d \otimes \mathbf{\Lambda}^l)\mathbf{c}}{||(\mathbf{W}^{(l)} \ldots \mathbf{W}^{(1)} \otimes \mathbf{V})(\mathbf{N}_d \otimes \mathbf{\Lambda}^l)\mathbf{c}||} - \mathbf{m}^{(l)} \right\| \tag{40}$$

$$= \left\| \frac{(\mathbf{W}^{(l)} \ldots \mathbf{W}^{(1)} \otimes \mathbf{V}) \left(\mathbf{N}_d \otimes \frac{\mathbf{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}}}\right) \mathbf{c}}{|| \left(\mathbf{W}^{(l)} \ldots \mathbf{W}^{(1)} \otimes \mathbf{V}\right) \left(\mathbf{N}_d \otimes \frac{\mathbf{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}^l}}\right) \mathbf{c}||} - \mathbf{m}^{(l)} \right\|$$

We simplify the notation and set $\mathbf{P}^{(l)} = \mathbf{W}^{(l)} \ldots \mathbf{W}^{(1)}$. We now choose $\mathbf{m}^{(l)} = \dfrac{\left(\mathbf{P}^{(l)} \otimes \mathbf{V}\right)\left(\mathbf{N}_d \otimes \frac{\mathbf{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}^l}}\right)\mathbf{c}'}{||\left(\mathbf{P}^{(l)} \otimes \mathbf{V}\right)\left(\mathbf{N}_d \otimes \frac{\mathbf{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}^l}}\right)\mathbf{c}||}$ by only replacing the $\mathbf{c}$ in the numerator by $\mathbf{c}'$ containing the same values for coefficients corresponding to $\mathcal{Q}_1$ and zeros for other subspaces, i.e., $c'_{1,r} = c_{1,r}\, \forall r$ and $c'_{i,r} = 0\, \forall i > 1, \forall r$. Continuing,

$$
\begin{aligned}
&= \left\| \frac{\left(\frac{\mathbf{P}^{(l)}}{\sigma_1^{\mathbf{P}^{(l)}}} \otimes \mathbf{V}\right)}{\left\| \left(\frac{\mathbf{P}^{(l)}}{\sigma_1^{\mathbf{P}^{(l)}}} \otimes \mathbf{V}\right)\left(\mathbf{N}_d \otimes \frac{\mathbf{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}^l}}\right)\mathbf{c}\right\|} \left(\left(\mathbf{N}_d \otimes \frac{\mathbf{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}^l}}\right)\mathbf{c} - \mathbf{c}'\right) \right\| \\
&\leq \frac{\left\| \frac{\mathbf{P}^{(l)}}{\sigma_1^{\mathbf{P}^{(l)}}} \otimes \mathbf{V}\right\|}{\left\| \left(\frac{\mathbf{P}^{(l)}}{\sigma_1^{\mathbf{P}^{(l)}}} \otimes \mathbf{V}\right)\left(\mathbf{N}_d \otimes \frac{\mathbf{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}^l}}\right)\mathbf{c}\right\|} \left\|\left(\mathbf{N}_d \otimes \frac{\mathbf{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}^l}}\right)\mathbf{c} - \mathbf{c}'\right\| \\
&\leq \frac{\sqrt{nd}}{c_{1,1}} \left(\frac{\lambda_2}{\lambda_1}\right)^l \max_{r,i} c_{r,i} \sqrt{(n-1)\cdot d}
\end{aligned}
\tag{41}
$$

Which is smaller than any ϵ for l large enough and $\cos\theta > 0$. For the last step, we provide the details for the upper bounds separately for each term:

$$
\begin{aligned}
\left\| \frac{\mathbf{P}^{(l)}}{\sigma_1^{\mathbf{P}^{(l)}}} \otimes \mathbf{V}\right\| &= \sqrt{tr\left(\left(\frac{\mathbf{P}^{(l)}}{\sigma_1^{\mathbf{P}^{(l)}}} \otimes \mathbf{V}\right)^T \left(\frac{\mathbf{W}^{(l)} \ldots \mathbf{W}^{(1)}}{\sigma_1^{\mathbf{P}^{(l)}}} \otimes \mathbf{V}\right)\right)} \\
&= \sqrt{tr\left(\frac{\left(\mathbf{P}^{(l)}\right)^T \mathbf{P}^{(l)}}{\left(\sigma_1^{\mathbf{P}^{(l)}}\right)^2} \otimes \mathbf{I}_n\right)} \\
&= \sqrt{\sum_{l,r} \frac{\sigma_l^2}{\sigma_1^2}} \leq \sqrt{nd}
\end{aligned}
\tag{42}
$$

$$\frac{1}{\left\|\left(\frac{\mathbf{P}^{(l)}}{\sigma_1^{\mathbf{P}^{(l)}}} \otimes \mathbf{V}\right)\left(\mathbf{N}_d \otimes \left(\frac{\mathbf{\Lambda}}{\lambda_1^{\tilde{\mathbf{A}}}}\right)^l\right)\mathbf{c}\right\|} = \frac{1}{\sqrt{\left(\left(\frac{\mathbf{P}^{(l)}}{\sigma_1^{\mathbf{P}^{(l)}}} \otimes \mathbf{V}\right)\left(\mathbf{N}_d \otimes \frac{\mathbf{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}l}}\right)\mathbf{c}\right)^T \left(\frac{\mathbf{P}^{(l)}}{\sigma_1^{\mathbf{P}^{(l)}}} \otimes \mathbf{V}\right)\left(\mathbf{N}_d \otimes \left(\frac{\mathbf{\Lambda}}{\lambda_1^{\tilde{\mathbf{A}}}}\right)^l\right)\mathbf{c}}}$$

$$= \frac{1}{\sqrt{\mathbf{c}^T \left(\mathbf{N}\frac{\left(\mathbf{P}^{(l)}\right)^T \mathbf{P}^{(l)}}{\left(\sigma_1^{\mathbf{P}^{(l)}}\right)^2}\mathbf{N} \otimes \left(\frac{\mathbf{\Lambda}}{\lambda_1^{\tilde{\mathbf{A}}}}\right)^{2l}\right)\mathbf{c}}}$$

$$= \frac{1}{\sqrt{\sum_{l,r}\left(\frac{\sigma_l^{\mathbf{P}^{(l)}}}{\sigma_1^{\mathbf{P}^{(l)}}}\right)^2 \mathbf{c}_{l,r}^2 \left(\frac{\lambda_r}{\lambda_1}\right)^{2l}}}$$

$$\leq \frac{1}{c_{1,1}} \tag{43}$$

$$\left\|\left(\mathbf{N}_d \otimes \frac{\mathbf{\Lambda}^{\mathbf{A}^l}}{\lambda_1^{\mathbf{A}}}\right)\mathbf{c} - \mathbf{c}'\right\| = \sqrt{\sum_{r,i}\left(\left(\frac{\lambda_i}{\lambda_1}\right)^l c_{r,i} - \left(\frac{\lambda_i}{\lambda_1}\right)^l c'_{r,i}\right)^2}$$

$$\leq \sqrt{\sum_{r=1,i=2}^{d,n}\left(\left(\frac{\lambda_i}{\lambda_1}\right)^l c_{r,i}\right)^2}$$

$$\leq \left(\frac{\lambda_2}{\lambda_1}\right)^l \sqrt{\sum_{r=1,i=2}^{n,d} c_{r,i}^2}$$

$$\leq \left(\frac{\lambda_2}{\lambda_1}\right)^l \sqrt{(n-1)\cdot d \cdot \max_{r,i} c_{r,i}^2} = \left(\frac{\lambda_2}{\lambda_1}\right)^l \max_{r,i} c_{r,i}\sqrt{(n-1)\cdot d} \tag{44}$$

$$\square$$

Proof of Proposition 5.5

Proposition. *(Over-smoothing happens for all $\mathbf{W}^{(k)}$.) Let $\mathbf{X}^{(k+1)} = \tilde{\mathbf{A}}\mathbf{X}^{(k)}\mathbf{W}^{(k)}$ with $\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ and $\mathbf{W}^{(k)} \in \mathbb{R}^{d\times d}$ any matrix. We consider $E(\mathbf{H}) = tr(\mathbf{H}^T\mathbf{\Delta}\mathbf{H})$ for $\mathbf{\Delta} = \mathbf{I} - \tilde{\mathbf{A}}$. Then,*

$$\lim_{l\to\infty} E\left(\frac{\mathbf{X}^{(l)}}{\|\mathbf{X}^{(l)}\|_F}\right) = 0 \tag{45}$$

with convergence rate $\left(\frac{\lambda_2^{\tilde{\mathbf{A}}}}{\lambda_1^{\tilde{\mathbf{A}}}}\right)^2$ independently of all $\mathbf{W}^{(k)}$.

Proof. We again use the singular value decomposition $\mathbf{W}^{(1)}\ldots\mathbf{W}^{(l)} = \mathbf{P} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{N}^T$

$$
\begin{aligned}
\lim_{l\to\infty} E\left(\frac{\mathbf{X}^{(l)}}{||\mathbf{X}^{(l)}||}\right) &= \lim_{l\to\infty} \frac{\text{tr}(\mathbf{X}^{(l)}\boldsymbol{\Delta}\mathbf{X}^{(l)})}{||\mathbf{X}^{(l)}||^2} \\
&= \lim_{l\to\infty} \frac{\mathbf{c}(\mathbf{U}^T\otimes\mathbf{V}^T)(\mathbf{W}^{(1)}\otimes\tilde{\mathbf{A}}^T)\ldots(\mathbf{W}^{(l)}\otimes\tilde{\mathbf{A}})(\mathbf{I}_d\otimes\tilde{\boldsymbol{\Delta}})(\mathbf{W}^{(l)^T}\otimes\tilde{\mathbf{A}})(\mathbf{W}^{(1)^T}\otimes\tilde{\mathbf{A}})(\mathbf{U}\otimes\mathbf{V})}{||\mathbf{X}^{(l)}||^2} \\
&= \lim_{l\to\infty} \frac{\mathbf{c}(\mathbf{U}^T\mathbf{W}^{(1)}\ldots\mathbf{W}^{(l)}\mathbf{W}^{(l)^T}\ldots\mathbf{W}^{(1)^T}\mathbf{U}\otimes\mathbf{V}^T\tilde{\mathbf{A}}^l\tilde{\boldsymbol{\Delta}}\tilde{\mathbf{A}}^l\mathbf{V})\mathbf{c}}{||\mathbf{X}^{(l)}||^2} \\
&= \lim_{l\to\infty} \frac{\mathbf{c}(\boldsymbol{\Sigma}^2\otimes\boldsymbol{\Lambda}^l\mathbf{V}^T\tilde{\boldsymbol{\Delta}}\mathbf{V}\boldsymbol{\Lambda}^l)\mathbf{c}}{\mathbf{c}(\boldsymbol{\Sigma}^2\otimes\boldsymbol{\Lambda}^l\mathbf{V}^T\mathbf{V}\boldsymbol{\Lambda}^l)\mathbf{c}} \\
&= \lim_{l\to\infty} \frac{\mathbf{c}(\frac{\boldsymbol{\Sigma}^2}{\sigma_1^2}\otimes\frac{\boldsymbol{\Lambda}^l(\mathbf{I}_d-\boldsymbol{\Lambda})\boldsymbol{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}}\lambda_1^{\tilde{\mathbf{A}}}})\mathbf{c}}{\mathbf{c}(\frac{\boldsymbol{\Sigma}^2}{\sigma_1^2}\otimes\frac{\boldsymbol{\Lambda}^l\boldsymbol{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}}\lambda_1^{\tilde{\mathbf{A}}}})\mathbf{c}} \\
&= 0
\end{aligned}
\tag{46}
$$

The last equality is true because $(1-\lambda_1^{\mathbf{A}})=0$ $\square$

A.3   Proof of Proposition 5.6

Proposition. *There exist $\mathbf{X}\in\mathbb{R}^{n\times d}, \mathbf{W}\in\mathbb{R}^{d\times d}$, for which*

$$
E\left(\frac{\mathbf{X}\mathbf{W}}{||\mathbf{X}\mathbf{W}||_F}\right) > E\left(\frac{\mathbf{X}}{||\mathbf{X}||_F}\right)
$$

Proof. We choose the case when $\sigma_d = \epsilon$ is a small number. When the signal $\mathbf{X}$ has a large value $c_{1,d} = k > 1$, the denominator gets reduced by more than the numerator, as the corresponding eigenvalue is $1-\lambda_1^{\mathbf{A}} = 0$. More formally, we choose $\mathbf{W} = \mathbf{U}\boldsymbol{\Sigma}^{\mathbf{W}}\mathbf{V}^T$ to have singular values $\sigma_1 = \ldots\sigma_{d-1} = 1, \sigma_d = \epsilon$ for some ϵ to be determined. We choose $vec(\mathbf{X}) = (\mathbf{U}\otimes\mathbf{V})\mathbf{c}$ with $c_{1,d} = k$ and all other $c_{l,r} = 1$. In this case, the above inequality reduces to

$$
\begin{aligned}
E\left(\frac{\mathbf{X}}{||\mathbf{X}||}\right) &= \frac{\sum_{l,r=1}^{n,d-1}(1-\lambda_l^{\mathbf{A}}) + \sum_{l=2}^{n}(1-\lambda_l^{\mathbf{A}})}{\sum_{l,r=1}^{n,d-1}1 + k + \sum_{l=2}^{n}1} \\
&< \frac{\sum_{l,r=1}^{n,d-1}(1-\lambda_l^{\mathbf{A}}) + \sum_{l=2}^{n}(1-\lambda_l^{\mathbf{A}})\epsilon^2}{\sum_{l,r=1}^{n,d-1}1 + k\epsilon^2 + \sum_{l=2}^{n}\epsilon^2} = E\left(\frac{\mathbf{X}\mathbf{W}}{||\mathbf{X}\mathbf{W}||}\right)
\end{aligned}
$$

Solving for ϵ determines the values for which the inequality is fulfilled:

$$\frac{\sum_{l,r=1}^{n,d-1}(1-\lambda_l^{\mathbf{A}}) + \sum_{l=2}^{n}(1-\lambda_l^{\mathbf{A}})}{\sum_{l,r=1}^{n,d-1}1 + k + \sum_{l=2}^{n}1} < \frac{\sum_{l,r=1}^{n,d-1}(1-\lambda_l^{\mathbf{A}}) + \sum_{l=2}^{n}(1-\lambda_l^{\mathbf{A}})\epsilon^2}{\sum_{l,r=1}^{n,d-1}1 + k\epsilon^2 + \sum_{l=2}^{n}\epsilon^2}$$

$$\Leftrightarrow (\sum_{l,r=1}^{n,d-1}(1-\lambda_l^{\mathbf{A}}) + \sum_{l=2}^{n}(1-\lambda_l^{\mathbf{A}}))(\sum_{l,r=1}^{n,d-1}1 + k\epsilon^2 + \sum_{l=2}^{n}\epsilon^2) < (\sum_{l,r=1}^{n,d-1}(1-\lambda_l^{\mathbf{A}}) + \sum_{l=2}^{n}(1-\lambda_l^{\mathbf{A}})\epsilon^2)(\sum_{l,r=1}^{n,d-1}1 + k + \sum_{l=2}^{n})$$

$$\Leftrightarrow \sum_{l,r=1}^{n,d-1}(1-\lambda_l^{\mathbf{A}})k\epsilon^2 + \sum_{l,r=1}^{n,d-1}(1-\lambda_l^{\mathbf{A}})\sum_{l=2}^{n}\epsilon^2) + \sum_{l=2}^{n}(1-\lambda_l^{\mathbf{A}})\sum_{l,r=1}^{n,d-1}1$$

$$< \sum_{l,r=1}^{n,d-1}(1-\lambda_l^{\mathbf{A}})k + \sum_{l,r=1}^{n,d-1}(1-\lambda_l^{\mathbf{A}})\sum_{l=2}^{n}1 + \sum_{l=2}^{n}(1-\lambda_l^{\mathbf{A}})\epsilon^2\sum_{l,r=1}^{n,d-1}1$$

$$\Leftrightarrow \epsilon^2(\sum_{l,r=1}^{n,d-1}(1-\lambda_l^{\mathbf{A}})k + \sum_{l,r=1}^{n,d-1}(1-\lambda_l^{\mathbf{A}})\sum_{l=2}^{n}) - \sum_{l=2}^{n}(1-\lambda_l^{\mathbf{A}})\sum_{l,r=1}^{n,d-1}1)$$

$$< \sum_{l,r=1}^{n,d-1}(1-\lambda_l^{\mathbf{A}})k + \sum_{l,r=1}^{n,d-1}(1-\lambda_l^{\mathbf{A}})\sum_{l=2}^{n}1 - \sum_{l=2}^{n}(1-\lambda_l^{\mathbf{A}})\sum_{l,r=1}^{n,d-1}1$$

$$\Leftrightarrow \epsilon^2 < 1$$

In the last step, we divide by $(\sum_{l,r=1}^{n,d-1}(1-\lambda_l^{\mathbf{A}})k + \sum_{l,r=1}^{n,d-1}(1-\lambda_l^{\mathbf{A}})\sum_{l=2}^{n}) - \sum_{l=2}^{n}(1-\lambda_l^{\mathbf{A}})\sum_{l,r=1}^{n,d-1}1)$, which is positive for $k > 1$. $\qquad\square$

A.4   Proof of Proposition 5.7

Proposition. *(Signal amplification only depends on $\tilde{\mathbf{A}}$.) Let two bases be $\mathbf{S}_{(i)} = (\mathbf{I}_d \otimes \mathbf{P}_{(i)}) \in \mathbb{R}^{nd \times qd}$ and $\mathbf{S}_{(j)} = (\mathbf{I}_d \otimes \mathbf{P}_{(j)}) \in \mathbb{R}^{nd \times qd}$ for any $\mathbf{P}_{(i)} \in \mathbb{R}^{n \times q}, \mathbf{P}_{(j)} \in \mathbb{R}^{n \times r}$ with $q, r \leq n$. Further let $\mathbf{T} = \mathbf{W} \otimes \tilde{\mathbf{A}} \in \mathbb{R}^{nd \times nd}$ consisting of any $\mathbf{W} \in \mathbb{R}^{d \times d}$ and any $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times n}$. Then,*

$$\frac{||\mathbf{T}\mathbf{S}_{(i)}||_F}{||\mathbf{T}\mathbf{S}_{(j)}||_F} = \frac{||\tilde{\mathbf{A}}\mathbf{P}_{(i)}||_F}{||\tilde{\mathbf{A}}\mathbf{P}_{(j)}||_F} .$$

Proof. We again use the property $||\mathbf{A} \otimes \mathbf{B}||_F = ||\mathbf{A}||_F \cdot ||\mathbf{B}||_F$ and basic properties of the Kronecker product:

$$\frac{||\mathbf{T}\mathbf{S}_{(i)}||_F}{||\mathbf{T}\mathbf{S}_{(j)}||_F} = \frac{||(\mathbf{W} \otimes \tilde{\mathbf{A}})(\mathbf{I}_d \otimes \mathbf{P}_{(i)})||_F}{||(\mathbf{W} \otimes \tilde{\mathbf{A}})(\mathbf{I}_d \otimes \mathbf{P}_{(j)})||_F} \tag{47}$$

$$= \frac{||\mathbf{W}||_F \cdot ||\tilde{\mathbf{A}}\mathbf{P}_{(i)}||_F}{||\mathbf{W}|| \cdot ||\tilde{\mathbf{A}}\mathbf{P}_{(j)}||_F} \tag{48}$$

$$= \frac{||\tilde{\mathbf{A}}\mathbf{P}_{(i)}||_F}{||\tilde{\mathbf{A}}\mathbf{P}_{(j)}||_F} \tag{49}$$

$$\square$$

Proof of Proposition 5.8

Proposition 7. *(GAT and Graph Transformer over-smooth.) Let* $\mathbf{X}^{(k+1)} = \tilde{\mathbf{A}}^{(k)}\mathbf{X}^{(k)}\mathbf{W}^{(k)}$ *where all* $\tilde{\mathbf{A}}^{(k)} \in \mathbb{R}^{n \times n}$ *row-stochastic, representing the same ergodic graph, and each non-zero entry* $(\tilde{\mathbf{A}}^{(k)})_{pq} \geq \epsilon$ *for some* $\epsilon > 0$ *and all* $i \in [h]$, $k \in \mathbb{N}$, $p, q \in [n]$. *We further allow all* $\mathbf{W}_i^{(k)} \in \mathbb{R}^{d \times d}$ *to be any matrices. We consider* $E(\mathbf{X}) = tr(\mathbf{X}^T \mathbf{\Delta}_{rw} \mathbf{X})$ *using the random walk Laplacian* $\mathbf{\Delta}_{rw} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}$. *Then,*

$$\lim_{l \to \infty} E\left(\frac{\mathbf{X}^{(l)}}{\|\mathbf{X}^{(l)}\|_F}\right) = 0.$$

Proof. The proof is mostly analogous to our proof of Theorem 3 but relies on the fact that $\Pi_{l=0}^{\infty}\tilde{\mathbf{A}}^{(l)} = \mathbf{1}(\mathbf{y}^{(l)})^T$ for all $i_l \in [h]$ and some $\mathbf{y}^{(l)} \in \mathbb{R}^n$ as established by Wu et al. [23]. The intuition behind the statement builds on the ergodicity and minimum edge weight ϵ, which combined result in each pairwise edge strength being larger than ϵ^s after a finite number of steps s. The product of two such matrices reduces the maximum and increases the minimum, which results in constant states in the limit.

Let $\mathbf{\Delta}_{rw} = \mathbf{V}(\mathbf{I}_n - \Lambda^{\tilde{\mathbf{A}}})\mathbf{V}^T$. As the first eigenvector $\mathbf{v}_1$ is constant, we can write the limit state $\Pi_{l=0}^{\infty}\tilde{\mathbf{A}}^{(l)} = \mathbf{V}\begin{bmatrix} \mathbf{y}_1^{(l)} & \cdots & \mathbf{y}_n^{(l)} \\ q_{21}^{(l)} & \cdots & q_{2n}^{(l)} \\ \vdots & \ddots & \vdots \\ q_{n2}^{(l)} & \cdots & q_{nn}^{(l)} \end{bmatrix}\mathbf{V}^T = \mathbf{V}\mathbf{Q}^{(l)}\mathbf{V}^T$ using the same eigenbasis and each q_{kl} converging to zero. For notational simplicity we define $\mathbf{R} = \Pi_{l=0}^{\infty}\mathbf{W}^{(l)}$. Decomposing the state $vec(\mathbf{X}) = (\mathbf{I}_d \otimes \mathbf{V})\mathbf{c}$ using the eigenbasis of $\mathbf{\Delta}_{rw}$, the Dirichlet energy then simplifies to

$$\lim_{l \to \infty} E\left(\frac{\mathbf{V}\mathbf{Q}^{(l)}\mathbf{V}^T\mathbf{X}^{(0)}\mathbf{R}}{\|\mathbf{V}\mathbf{Q}^{(l)}\mathbf{V}^T\mathbf{X}^{(0)}\mathbf{R}\|_F}\right) = \lim_{l \to \infty} \frac{\mathbf{c}(\mathbf{R}^T\mathbf{R} \otimes \mathbf{V}^T\mathbf{V}\mathbf{Q}^{(l)}\mathbf{V}^T\mathbf{V}(\mathbf{I}_n - \Lambda^{\tilde{\mathbf{A}}})\mathbf{V}^T\mathbf{V}\mathbf{Q}^{(l)}\mathbf{V}^T\mathbf{V})\mathbf{c}}{\mathbf{c}(\mathbf{R}^T\mathbf{R} \otimes \mathbf{V}^T\mathbf{V}\mathbf{Q}^{(l)}\mathbf{V}^T\mathbf{V}\mathbf{Q}^{(l)}\mathbf{V}^T\mathbf{V})\mathbf{c}}$$

$$= \lim_{l \to \infty} \frac{\mathbf{c}(\mathbf{R}^T\mathbf{R} \otimes \mathbf{Q}^{(l)^T}(\mathbf{I}_n - \Lambda^{\tilde{\mathbf{A}}})\mathbf{Q}^{(l)})\mathbf{c}}{\mathbf{c}(\mathbf{R}^T\mathbf{R} \otimes \mathbf{Q}^{(l)^T}\mathbf{Q}^{(l)})\mathbf{c}}$$

We then use the fact that $\lim_{l \to \infty} \mathbf{Q}^{(l)^T}(\mathbf{I}_n - \Lambda^{\tilde{\mathbf{A}}})\mathbf{Q}^{(l)} = \mathbf{0}$ as the columns of $\mathbf{Q}^{(l)}$ and the rows of $\mathbf{Q}^{(l)^T}$ converge to zero and the corresponding eigenvalue of $(\mathbf{I}_n - \Lambda^{\tilde{\mathbf{A}}})$ being also zero, leaving all entries to converge to zero. We then simply need to factor out the growth of R, so that this convergence holds:

$$\lim_{l \to \infty} \frac{\mathbf{c}(\mathbf{R}^T\mathbf{R} \otimes \mathbf{Q}^{(l)^T}(\mathbf{I}_n - \Lambda^{\tilde{\mathbf{A}}})\mathbf{Q}^{(l)})\mathbf{c}}{\mathbf{c}(\mathbf{R}^T\mathbf{R} \otimes \mathbf{Q}^{(l)^T}\mathbf{Q}^{(l)})\mathbf{c}} = \pm \lim_{l \to \infty} \frac{\mathbf{c}(\frac{\mathbf{R}^T\mathbf{R}}{|\lambda_1^{\mathbf{R}^T\mathbf{R}}|} \otimes \mathbf{Q}^{(l)^T}(\mathbf{I}_n - \Lambda^{\tilde{\mathbf{A}}})\mathbf{Q}^{(l)})\mathbf{c}}{\mathbf{c}(\frac{\mathbf{R}^T\mathbf{R}}{|\lambda_1^{\mathbf{R}^T\mathbf{R}}|} \otimes \mathbf{Q}^{(l)^T}\mathbf{Q}^{(l)})\mathbf{c}} = 0$$

(50)

$\square$

A.5   Over-correlation

Proof of Theorem 6.1

Theorem. *(The rank collapses.) Let $\mathcal{Q}_i = \mathrm{span}(\mathbf{I}_d \otimes \mathbf{V}_i)$ for any matrix $\mathbf{V}_i \in \mathbb{R}^{n \times j}$ with $j \leq n$. Then,*

$$\forall \, \mathrm{vec}(\mathbf{X}) \in \mathcal{Q}_i : \mathrm{rank}(\mathbf{X}) \leq j \tag{51}$$

Proof. We rewrite $vec(\mathbf{X}) = (\mathbf{I}_d \otimes \mathbf{v}_i)\mathbf{c}$ as a linear combination $\mathbf{c}$ of the basis vectors:

$$\mathbf{X} = \mathrm{vec}^{-1}((\mathbf{I}_d \otimes \mathbf{V}_i)\mathbf{c}) = \mathbf{V}_i \, \mathrm{vec}^{-1}(\mathbf{c}^T)\mathbf{I}_d = \mathbf{V}_i \, \mathrm{vec}^{-1}(\mathbf{c}^T) \tag{52}$$

using the inverse vectorize operation vec^{-1}. The statement holds given $\mathrm{rank}(\mathrm{vec}^{-1}(\mathbf{c}^T)) \leq \min(j, d) \leq j$, $\mathrm{rank}(\mathbf{V}_i) \leq j$ and Sylvester's rank inequality. $\qquad\square$

A.6   Sum of Kronecker products

A.7   Proof of Theorem 7.1

Theorem. *(Any subspace can get amplified.) Let $\mathbf{e}_i \in \mathbb{R}^d$ be the canonical basis with a single 1 at position i. For any columns $\mathbf{s}_1, \ldots, \mathbf{s}_d \in \mathbb{R}^n$ and the induced subspace $\mathbf{S} = \begin{bmatrix} \mathbf{e}_1 \otimes \mathbf{s}_1 \; \ldots \; \mathbf{e}_d \otimes \mathbf{s}_d \end{bmatrix} \in \mathbb{R}^{nd \times d}$, there exists an $\mathbf{T} = \sum_1^d (\mathbf{W}_i \otimes \tilde{\mathbf{A}}_i)$ such that for all orthogonal bases $\mathbf{S}' = \begin{bmatrix} \mathbf{e}_1 \otimes \mathbf{s}'_1 \; \ldots \; \mathbf{e}_d \otimes \mathbf{s}'_d \end{bmatrix} \in \mathbb{R}^{nd \times d}$*

$$\frac{\|\mathbf{TS}\|_F}{\|\mathbf{TS}'\|_F} > \frac{\|\mathbf{S}\|_F}{\|\mathbf{S}'\|_F} \, . \tag{53}$$

Proof. We choose each $\tilde{\mathbf{A}}_i = \mathbf{V}^{(i)} \mathbf{\Lambda} \mathbf{V}^{(i)^T}$ to be symmetric with dominant eigenvectors $\mathbf{v}_1^{(i)} = \mathbf{s}_i$ and shared eigenvalues $|\lambda_1| > |\lambda_2^{\tilde{\mathbf{A}}}| > |\lambda_d^{\tilde{\mathbf{A}}}| > 0$. Further, $\mathbf{W}_i = \mathrm{diag}(\mathbf{e}_i)$ where the diag operation creates a matrix with the entries of its arguments along the diagonal. Thus, all columns are independent and $\mathbf{T}$ is a block-diagonal matrix with eigenvectors being $\mathbf{e}_i \otimes \mathbf{v}^{(j)}$ with corresponding eigenvalue $\lambda_j^{\tilde{\mathbf{A}}}$. The eigenspace corresponding to $\lambda_1^{\tilde{\mathbf{A}}}$ are all $\mathbf{e}_i \otimes \mathbf{s}_i$ for all i, i.e., $\mathrm{span}(\mathbf{S})$. Any orthogonal column $\mathbf{s}'_i = \sum_{k=2}^d \mathbf{v}_2^{(k)} c_2^{(k)}$ can be written as a linear combination of the other eigenvectors. Thus,

$$\frac{\|\mathbf{TS}\|_F}{\|\mathbf{TS}'\|_F} \geq \frac{|\lambda_1|}{|\lambda_2|} \frac{\|\mathbf{S}\|_F}{\|\mathbf{S}'\|_F} > \frac{\|\mathbf{S}\|_F}{\|\mathbf{S}'\|_F} \tag{54}$$

$$\square$$

B   Experimental details

B.1   Convergence to a Constant State

For this experiment, we consider the Cora dataset provided by Pytorch-Geometric [52], consisting of 2708 nodes and 5429. For initialize 128 layers of GAT, GCN, and GraphSAGE using their default initialization and no bias. After each layer, we use a ReLU activation. After each ReLU activation, we track both the squared norm $||\mathbf{X}||_F^2$ and the Dirichlet energy $E(\mathbf{X}) = tr(\mathbf{X}^T \mathbf{\Delta}_{rw} \mathbf{X})$ using the random walk Laplacian $\mathbf{\Delta}_{rw} = \mathbf{I}_n - \mathbf{D}^{-1}\mathbf{A}$, as the constant vector is in its nullspace, as suggested by Rusch et al. [11].

B.2   Empirical Validation

Here, we describe the details of our considered task for our empirical validation in Section 8. We randomly generate an Erdos–Rény (ER) graph, that consists of 20 nodes and an edge probability of 0.2. We then select four its nodes at random, denoted by $V = (v_1, v_2, v_3, v_4)$. The 3-class classification task consists of three tasks for each node, where all pairs of nodes belong to the same class exactly once. Precisely, the target for all nodes is $\mathbf{Y} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 & 0 \end{bmatrix}$.

For loss, optimization, and accuracy calculation, we only consider the labels of the four nodes and ignore all remaining outputs.

If each feature is constant across nodes, all nodes are classified the same, resulting in a 50% accuracy. If the representations have rank one and all feature vectors are on a line, representations need to be adjacent in order to be classified correctly, which cannot be fulfilled for all pairs simultaneously. Since our GNNs are finite-depth, representations are not normalized, and the maximum eigenvalue may have geometric multiplicity larger than one, the models can achieve higher accuracy.

Training details Node representations $\mathbf{X}^{(0)} \in \mathbb{R}^{4 \times 6}$ are randomly initialized from a normal distribution $x_{ij} \sim N(0, 1)$ with 0 mean and standard deviation of 1. Aggregations and feature transformation are randomly initialized such that the norm of resulting node representations is typically not close to zero or to infinity so that the gradients do not vanish or explode. Precisely, for $\tilde{\mathbf{A}}_1^{(k)} \in \mathbb{R}^{20 \times 20}$, $\tilde{\mathbf{A}}_2^{(k)} \in \mathbb{R}^{20 \times 20}$ the weight of each generated edge (i, j) is sampled from a normal distribution $\tilde{a}_{ij} \sim \mathcal{N}(\frac{1}{|N_j|}, 0.05)$ with mean as one over the number of incoming nodes and standard deviation 0.05. Similarly, the feature

transformations $\mathbf{W}_1^{(k)} \in \mathbb{R}^{6\times6}$, $\mathbf{W}_2^{(k)} \in \mathbb{R}^{6\times6}$ are sampled from a normal distribution $\tilde{w}_{ij} \sim \mathcal{N}(\frac{1}{3}, 0.05)$ with mean as one over the number of features and standard deviation 0.05. Using $\mathbf{T}^{(k)}$ as one of the considered graph convolutions, the update function is $\text{vec}(\mathbf{X}^{(k+1)}) = \phi(\mathbf{T}^{(k)}\text{vec}(\mathbf{X}^{(k)}))$ with ϕ as the ReLU activation function. After l iterations, we use these node representations for our three classification tasks using the affine transformation

$$\hat{\mathbf{Y}} = \sigma(\mathbf{X}^{(l)}\mathbf{W}_c + \mathbf{b}) \tag{55}$$

with σ as sigmoid activation, a feature transformation $\mathbf{W}_c \in \mathbb{R}^{6\times3}$ and a feature-wise bias term $\mathbf{b} \in \mathbb{R}^3$. We evaluated all three described update functions for $l \in [1, 2, 4, 8, 16, 32, 64, 128]$ layers. For each method and each number of layers, variables are randomly initialized and optimized with binary cross-entropy using the Adam optimizer until the loss does not decrease for 500 steps. Each experiment is executed three times for each graph, of which the best achieved accuracy is considered. We then repeat this process for 50 random graphs and report the mean accuracy and its standard deviation. The reproducible experiments are added as supplementary material.

We additionally run the same experimental setting for a linearized version (ϕ as identity function) and a version that also reuses the same aggregation across all layers $\tilde{\mathbf{A}}_1^{(1)} = \cdots = \tilde{\mathbf{A}}_1^{(k)}$, $\tilde{\mathbf{A}}_2^{(1)} = \cdots = \tilde{\mathbf{A}}_2^{(k)}$. All configurations were run sequentially on a single CPU. The entire runtime was around 30 hours.

# of layers	1	2	4	8	16	32	64	128
GAT_{HL}	100 ± 0	100 ± 0	82 ± 19	71 ± 6	72 ± 6	70 ± 6	67 ± 7	63 ± 8
GAT_L	100 ± 0	100 ± 0	100 ± 0	70 ± 4	71 ± 6	70 ± 7	68 ± 8	65 ± 8
GAT	100 ± 0	100 ± 0	100 ± 0	70 ± 4	71 ± 6	70 ± 7	68 ± 8	65 ± 8
FAGCN_{HL}	100 ± 0	100 ± 0	83 ± 0	83 ± 0	83 ± 0	83 ± 0	83 ± 0	82 ± 4
FAGCN_L	100 ± 0	100 ± 0	83 ± 0	83 ± 0	83 ± 0	83 ± 0	83 ± 0	82 ± 4
FAGCN	100 ± 0	98 ± 5	88 ± 4	85 ± 3	84 ± 2	84 ± 2	85 ± 3	82 ± 4
SKP_{HL}	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	94 ± 12
SKP_L	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	96 ± 10
SKP	100 ± 0	99 ± 3	98 ± 5	100 ± 0	100 ± 0	100 ± 0	100 ± 1	95 ± 10

Table 1: Maximum and mean±standard deviations accuracies for all considered scenarios. Subscript L denotes the linearized version, subscript HL denotes the version that additionally uses homogeneous aggregation matrices.

Detailed results We provide additional numerical results in Table 1, including the standard deviation of all settings. If the representations were constant, only a 50% accuracy could be achieved, so there is slightly more relevant information in all features, even for the GAT-like model. However, the representations are quickly converging to a low-rank state, so both GAT and FAGCN do not solve

this task a single time for eight or more layers. All models are slightly more unstable with increased depth, which is mainly a vanishing or exploding gradient issue. As this was historically an issue in other domains, similar methods can be used to solve this, e.g., normalization layers.