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

Andreas Roth (✉)
TU Dortmund University, Germany
andreas.roth@tu-dortmund.de

Thomas Liebig
TU Dortmund University, Germany,
Lamarr Institute for Machine Learning
and Artificial Intelligence
thomas.liebig@tu-dortmund.de

Abstract

Our study reveals new theoretical insights into over-smoothing and feature over-correlation in graph neural networks. Specifically, we demonstrate that with increased depth, node representations become dominated by a low-dimensional subspace that depends on the aggregation function but not on the feature transformations. For all aggregation functions, the rank of the node representations collapses, resulting in over-smoothing for particular aggregation functions. Our study emphasizes the importance for future research to focus on rank collapse rather than over-smoothing. Guided by our theory, we propose a sum of Kronecker products as a beneficial property that provably prevents over-smoothing, over-correlation, and rank collapse. We empirically demonstrate the shortcomings of existing models in fitting target functions of node classification tasks.

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].The conditions under which over-smoothing and over-correlation occur and in which cases they can be prevented are not well-understood 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 even cause an over-separation of node representations [6, 12, 13]. Over-correlation refers to all feature columns becoming overly correlated, which was only recently empirically observed [8]. While Jin et al. [8] highlighted the occurrence 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 common graph convolutions 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 these models cannot fit target functions that require multiple signals to be amplified, which we empirically confirm also for 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 provide a shared theoretical background of over-smoothing and over-correlation by showing the dominance of a fixed subspace induced by the aggregation function, allowing us to provide general proofs for both phenomena.

Roth et al., Rank Collapse Causes Over-Smoothing and Over-Correlation in Graph Neural Networks. *Proceedings of the Second Learning on Graphs Conference (LoG 2023)*, PMLR 231, Virtual Event, November 27–30, 2023.

- While our insights confirm that over-smoothing can be solved by choosing an appropriate aggregation function, we identify the phenomenon of a rank collapse of node representations as the crucial challenge for GNNs going forward.
- To counteract this limited expressivity, we propose the utilization of a sum of Kronecker products (SKP) as a general property that provably solves rank collapse, which we empirically confirm.

2 Preliminaries

Notation. We consider a graph $G = (\mathcal{V}, \mathcal{E})$ consisting of a set of n nodes $\mathcal{V} = \{v_1, \dots, v_n\}$ 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 $\lambda_1^{\mathbf{M}}, \dots, \lambda_n^{\mathbf{M}}$ that are sorted with decreasing absolute value $|\lambda_1^{\mathbf{M}}| \geq \dots \geq |\lambda_n^{\mathbf{M}}|$. 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 outline 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

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

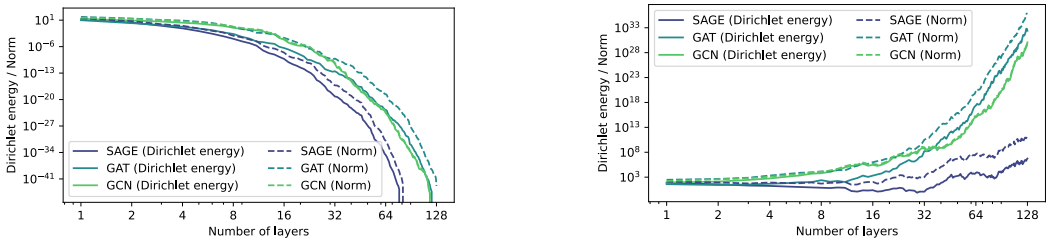
of the representations to a smooth subspace $\mathcal{M}$ that is induced by the dominant eigenvector $\mathbf{v}_1$. The bound uses the second largest eigenvalue $\lambda_2^{\mathbf{A}}$ and the largest singular value $\sigma_1^{\mathbf{W}}$ of $\mathbf{W}$. Intuitively, each aggregation step $\mathbf{A}$ 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

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

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

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

for each convolution and prove an exponential convergence in the limit. As their proof again only holds in case $\sigma_1^{\mathbf{W}} \leq 1/\lambda_2^{\mathbf{A}}$, they similarly claim potential over-separation. Zhou et al. [13] provide



(a) Dirichlet energy converges to zero. **(b)** Dirichlet energy does not converge.

Figure 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], GraphSAGE [33]) on the Cora dataset [34]. For (a), we used random weights, for (b), we scaled these weights by a factor of 2.

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. Jin et al. [8] empirically revealed an excessive correlation among node features with increased model depth, as measured by the Pearson correlation coefficient. This over-correlation results in redundant information and limits the performance of deep GNNs. The study emphasizes that over-smoothing always leads to over-correlation, but that over-correlation also arises independently. Similarly, Guo et al. [28] empirically observed that feature columns in GNNs and Transformers have the tendency to converge to non-smooth states characterized by low effective rank, where most singular values are close to zero. This phenomenon, which they refer to as *dimensional collapse*, has also been observed in the context of contrastive learning [29, 30]. To address over-correlation, some works suggest incorporating it into the loss function [8, 31], while others propose combining the results of message-passing with decorrelated signals from the previous state [28, 32]. However, the underlying theory behind over-correlation remains largely unexplored, especially concerning feature transformations.

4 Over-Smoothing and Convergence to a Constant State

We start 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. In contrast, other studies provided theoretical insights into convergence toward the dominant eigenvector of the aggregation function [6, 12, 13], which is not always constant, e.g., for the GCN [15]. We attribute these differences to the norm of $\mathbf{X}^{(k)}$ overshadowing 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 4.1. *(Node representations vanish.) Let $\tilde{\mathbf{A}} \in \mathbb{R}^{n\times n} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$, $\mathbf{W} \in \mathbb{R}^{d\times d}$ 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 $\sigma_1^{\mathbf{W}^{(k)}} < 1$ for all $\mathbf{W}^{(k)}$ this implies convergence to the zero matrix.With all node states close to zero, the Dirichlet energy becomes also zero.

We compare the Dirichlet energy and the norm in Figure 1, and observe a close alignment between both metrics. While we confirm that the Dirichlet energy converges to zero, our findings show that this solely results from the norm converging to zero. 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 1. 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].

Since the norm of the node representations overshadows the Dirichlet energy, it seems insufficient to assess the energy of the unnormalized state. Other metrics, such as MAD [18] and SMV [35], already incorporate feature normalization to quantify over-smoothing. Furthermore, recent studies already consider the Dirichlet energy of the normalized state for studying over-smoothing [19, 20]. Our insights prompt a deeper exploration of over-smoothing and its interaction with feature transformations, aiming to comprehend its causes better, 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

$$\operatorname{vec}(\tilde{\mathbf{A}}\mathbf{X}^{(k)}\mathbf{W}^{(k)}) = (\mathbf{W}^{(k)^T} \otimes \tilde{\mathbf{A}}) \operatorname{vec}(\mathbf{X}^{(k)}) = \mathbf{T}^{(k)} \operatorname{vec}(\mathbf{X}^{(k)}) \tag{6}$$

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

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

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 $\mathcal{Q}_i = \operatorname{span}(\mathbf{S}_{(i)}) \subset \mathbb{R}^{nd}$ 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 5.1. *(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 $\mathbf{v}_1, \dots, \mathbf{v}_n$ and $\mathbf{W} \in \mathbb{R}^{d\times d}$ any matrix. Consider the subspaces $\mathcal{Q}_i = \operatorname{span}(\mathbf{I}_d \otimes \mathbf{v}_i)$ 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 5.2. *(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.

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 5.3. *(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 the bases in all subspaces are the same, the distribution of the signals might be heavily skewed. For example, the signal in $\mathcal{Q}_1$ might align with the nullspace of $\mathbf{W}^{(l)} \ldots \mathbf{W}^{(1)}$, while the signal in other subspaces might not. However, the probability of $\mathbf{W}^{(l)} \ldots \mathbf{W}^{(1)}$ satisfying this property converges to zero at an exponential rate with respect to the Lebesgue measure. This allows us to formally show the convergence of the node representations to $\mathcal{Q}_1$ in probability:

Theorem 5.4. *(Representations converge in probability 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* $\mathcal{Q}_i = \operatorname{span}(\mathbf{I}_d \otimes \mathbf{v}_i)$ *for* $i \in [n]$*. Then,*

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

Thus, the probability of avoiding convergence to $\mathcal{Q}_1$ exponentially converges to zero with increased model depth. While the chances always remain non-zero, finding these solutions is particularly challenging for an optimization method based on gradient descent. The heavily constrained solution space directly limits the model's expressivity. In addition, any slight change in the input can cause the signal to no longer align with the null space of the feature transformations, making a method with this property meaningless for practical applications.

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 [36]. 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 $E(\mathbf{X}/||\mathbf{X}||_F) = tr(\mathbf{X}^T\mathbf{\Delta}\mathbf{X})/||\mathbf{X}||_F^2$ [12]. The Dirichlet energy of the normalized state was also proposed by Di Giovanni et al. [19]. The key property needed is the equivalence of the dominating subspace $\mathcal{Q}_1$ and the null space of $\mathbf{I}_d \otimes \mathbf{\Delta}$, allowing us to provide the novel proof for feature transformations in probability:

Proposition 5.5. *(Over-smoothing happens for all $\mathbf{W}^{(k)}$ in probability.) Let $\mathbf{X}^{(k+1)} = \tilde{\mathbf{A}}\mathbf{X}^{(k)}\mathbf{W}^{(k)}$ with $\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$, $\mathbf{W}^{(k)} \in \mathbb{R}^{d\times d}$, and $E(\mathbf{X}) = tr(\mathbf{X}^T\mathbf{\Delta}\mathbf{X})$ for $\mathbf{\Delta} = \mathbf{I} - \tilde{\mathbf{A}}$. Then,*

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

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

5.3 Extending the Analysis to Arbitrary Aggregations

We now 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, 37–39]. 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. We find that Theorem 5.2 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 5.6. *(informal) For any matrices $\tilde{\mathbf{A}} \in \mathbb{R}^{n\times n}$ and $\mathbf{W} \in \mathbb{R}^{d\times d}$, the relative amplification of subspaces only depends on $\tilde{\mathbf{A}}$ and not on $\mathbf{W}$.*

Importantly, invariant bases always exist, as given by the Jordan normal form [40]. Over-smoothing is prevented by selecting or learning an aggregation matrix $\tilde{\mathbf{A}}$ with a well-suited spectrum.

5.4 Stochastic Aggregation Functions

Our findings can be readily extended to cases where we know about the dominance of a particular signal for the aggregation function. 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, 41] with a single attention head. Edge weights may be dynamically computed based on their adjacent nodes, so we consider the case of time-inhomogeneous 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 [42] and stochastic processes [43]. 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 5.6, and we show over-smoothing using the Dirichlet energy in probability to exclude the cases where the nullspace of $\mathbf{W}^{(1)}\ldots\mathbf{W}^{(l)}$ aligns with the signal in the dominating subspace:

Proposition 5.7. *(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 with minimum non-zero value $> \epsilon$ for some $\epsilon > 0$ and $\mathbf{W}_i^{(k)} \in \mathbb{R}^{d\times d}$. Then,*

$$\operatorname*{plim}_{l\to\infty} E_{\text{rw}}\left(\frac{\mathbf{X}^{(l)}}{\|\mathbf{X}^{(l)}\|_F}\right) = 0\,.$$

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 $|\lambda_1^{\tilde{\mathbf{A}}}|$. Node representations in $\mathcal{Q}_1$ cannot exceed the rank of j:

Theorem 6.1. *(informal) For all $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times n}$ and for $\mathbf{W}^{(l)} \dots \mathbf{W}^{(1)}$ in probability, in the limit $l \to \infty$, the rank of $\mathbf{X}^{(l)}$ is bounded the joint algebraic multiplicity of all eigenvalues of $\tilde{\mathbf{A}}$ with magnitude $|\lambda_1|$.*

This implies that, regardless of the aggregation function chosen, the representations collapse in probability 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 $\mathrm{rank}(\mathbf{V}_i) = 1$, e.g., when the dominant eigenvalue is unique. Jin et al. [8] empirically observed this phenomenon and termed it over-correlation. This study reveals that rank collapse is the fundamental issue underlying over-smoothing and over-correlation. These findings also establish a connection between these phenomena and recent theoretical insights on deep Transformer models, where rank collapse was also independently observed [44, 45].

7 Preventing Rank Collapse with a Sum of Kronecker Products

The core issue leading to rank collapse is the fact that the transformation matrix $\mathbf{T} = \mathbf{W} \otimes \mathbf{A}$ is composed of 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

$$\mathbf{T} = (\mathbf{W}_1 \otimes \mathbf{A}_1) + (\mathbf{W}_2 \otimes \mathbf{A}_2) + \cdots + (\mathbf{W}_p \otimes \mathbf{A}_p) \tag{10}$$

can be decomposed into a finite sum of Kronecker products (SKP), as concatenation is a special case [46, 47]. As each term can amplify a different signal per feature column, we show that d terms are sufficient to amplify arbitrary signals across d columns:

Theorem 7.1. *(informal) A sum of d Kronecker products $\mathbf{T} = \sum_1^d (\mathbf{W}_i \otimes \tilde{\mathbf{A}}_i)$ can amplify independent signals for at least d different columns.*

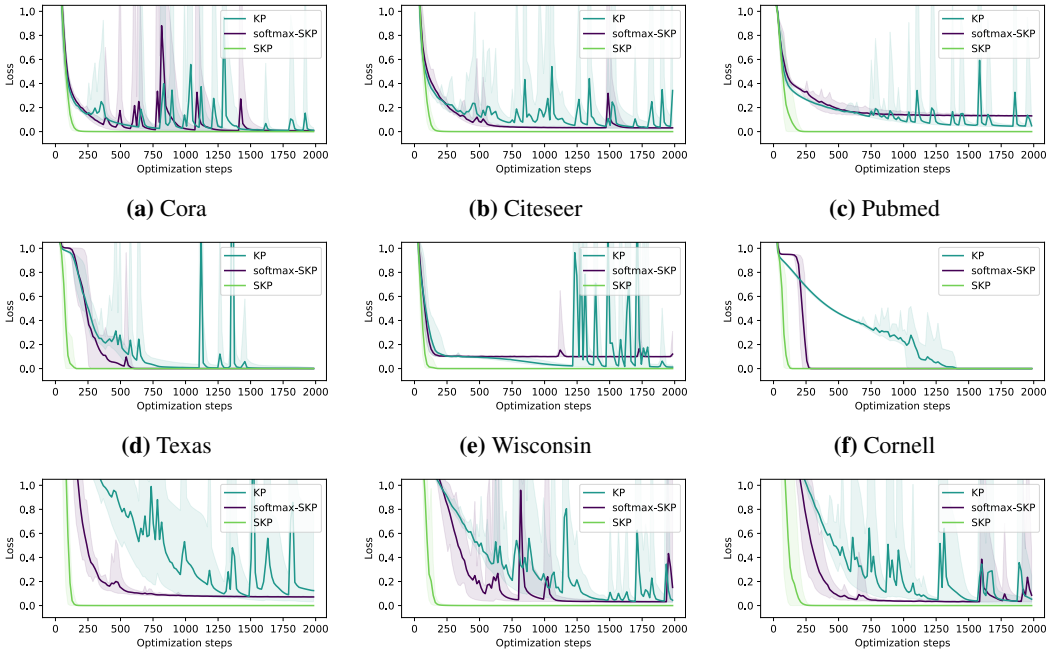
The crucial advantage of an SKP is that each $\mathbf{W}_i$ can control the degree to which that term contributes to each feature separately for each column. The SKP does not add computational complexity, as we would not compute Kronecker products, i.e., $(\mathbf{W}_1 \otimes \tilde{\mathbf{A}}_1) + \cdots + (\mathbf{W}_p \otimes \tilde{\mathbf{A}}_p))\,\mathrm{vec}(\mathbf{X}) = \mathrm{vec}(\tilde{\mathbf{A}}_1 \mathbf{X} \mathbf{W}_1 + \cdots + \tilde{\mathbf{A}}_p \mathbf{X} \mathbf{W}_p)$. It is important to note that the concept of an 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 discrete convolutions [48], residual connections [49–51], multi-head graph attention networks [17], mixing aggregation functions [52] or transforming signals from incoming and outgoing edges differently [53]. However, Theorem 7.1 relies on each aggregation matrix having a different dominating eigenvector to ensure the amplification of different signals across feature columns.

8 Empirical Validation

Our theoretical statements show that the current message-passing frameworks suffer from rank collapse, which makes learning complex functions in deep graph neural networks challenging. To empirically confirm this phenomenon also for the non-linear case of finite depth, we now evaluate the ability of message-passing frameworks to fit the target function of nine common benchmark datasets for node classification.

Datasets. Three homophilic citation networks that generally benefit from smoothing are considered, namely the citation datasets Cora [54], Citeseer [54], and Pubmed [54]. The other six datasets are heterophilic, of which three are webpage datasets, namely Texas [55], Wisconsin [55], Cornell [55], two are Wikipedia networks, namely Chameleon [56], and Squirrel [56], and one is the actor co-occurrence network Film [57]. For each dataset, we select the largest strongly connected component and optimize the cross-entropy loss for all nodes in order to compare each model's ability to fit the desired function. We do not utilize self-loops.

Considered Methods. Our theory motivates us to evaluate three different message-passing frameworks: A single Kronecker product, a softmax-activated sum of Kronecker products, and any sum of



(a) Cora **(b)** Citeseer **(c)** Pubmed

(d) Texas **(e)** Wisconsin **(f)** Cornell

(g) Film **(h)** Chameleon **(i)** Squirrel

Figure 2: Mean optimization loss across 10 runs for various node classification tasks with 8 layers of message-passing. The shaded region indicates minimum and maximum loss values.

Kronecker products. A single Kronecker product

$$\mathbf{T}_{\mathrm{KP}}^{(k)} = \mathbf{W}_1^{(k)} \otimes \tilde{\mathbf{A}}_1^{(k)} \tag{11}$$

includes most current message-passing frameworks, such as the GCN [15], mean aggregation as in GraphSAGE [33], sum aggregation as in GIN [16], negative edge weights as in FAGCN [38] and many more [5, 37, 39]. To cover all of these methods in our experiments, the weights of the available edges in $\tilde{\mathbf{A}}_1$ are directly learned from the data. The second framework

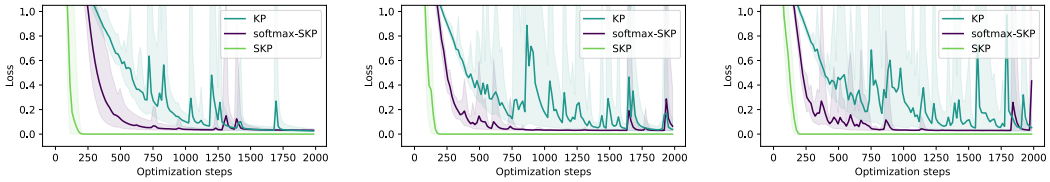
$$\mathbf{T}_{\mathrm{softmax-SKP}}^{(k)} = \frac{1}{2}(\mathbf{W}_1^{(k)} \otimes \tau(\tilde{\mathbf{A}}_1^{(k)})) + \frac{1}{2}(\mathbf{W}_2^{(k)} \otimes \tau(\tilde{\mathbf{A}}_2^{(k)})) \tag{12}$$

extends our proof on time-inhomogeneous softmax-activated aggregation function (Prop. 5.7) to the multi-head case, i.e., a sum of Kronecker products for which each aggregation function is activated by a softmax function τ over each node's neighbors. The weights of all available edges in $\tilde{\mathbf{A}}_1$ and $\tilde{\mathbf{A}}_2$ are learned directly before applying τ, which causes the spectrum of both matrices to align and Theorem 7.1 on improved signal amplification to not hold for this model. This represents the multi-head versions of GAT [17], GATv2 [41], TransformerConv [58] and other potential attention mechanisms. Following our theoretical insights on achieving a more powerful message-passing framework, our third approach is based on a sum of Kronecker products

$$\mathbf{T}_{\mathrm{SKP}}^{(k)} = \mathbf{W}_1^{(k)} \otimes \tilde{\mathbf{A}}_1^{(k)} + \mathbf{W}_2^{(k)} \otimes \tilde{\mathbf{A}}_2^{(k)}\,, \tag{13}$$

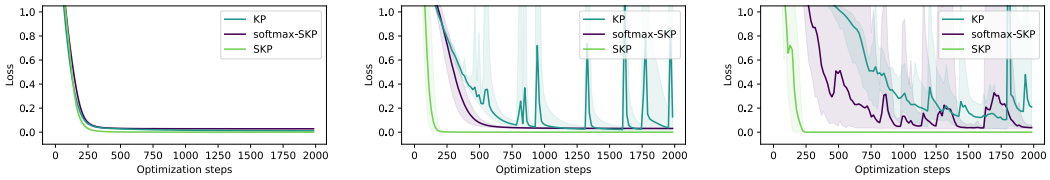
where the edge weights are not activated to avoid matching spectra between terms. Based on Theorem 7.1, this amplifies separate signals per channel and node.

Experimental Setup. Our full model starts by encoding the given node features into a d-dimensional vector using a linear layer and a ReLU activation. We then apply the described message-passing schemes $\mathrm{vec}(\mathbf{X}^{(k+1)}) = \phi(\mathbf{T}^{(k)}\,\mathrm{vec}(\mathbf{X}^{(k)})$ for l layers, and apply the ReLU activation ϕ after each layer to validate that non-linear models also suffer from rank collapse. After l such layers, the node representations are mapped to class probabilities by an affine transformation and a softmax activation. All model parameters are optimized to minimize the cross-entropy using Adam [59] for 2000 steps.



(a) $d = 16$ **(b)** $d = 64$ **(c)** $d = 128$

Figure 3: Optimization loss for the Squirrel dataset when varying the feature dimension d.



(a) $l = 1$ **(b)** $l = 4$ **(c)** $l = 32$

Figure 4: Optimization loss for the Squirrel dataset when varying the number of layers l.

Each experiment is repeated for ten random initializations. Our reproducible implementation is based on PyTorch [60] and PyTorch Geometric [61][1]. We refer to Appendix B for all details.

Results. We report the mean, minimal, and maximal achieved loss values over ten runs for $l = 8$ layers, a feature dimension of $d = 32$, and all considered datasets in Figure 2. The SKP fits all datasets almost perfectly within less than 250 steps and decreases almost monotonically. Contrarily, KP and softmax-SKP require significantly more steps to converge and remain visibly unstable for most datasets. We ascribe this to both models' inability to amplify different signals per node and channel. In most cases, softmax-SKP also converges to a higher loss than the SKP. To better understand the reasons behind these shortcomings, we provide additional experiments. In Figure 3, we evaluate the benefit of additional parameters by setting the feature dimension to $d \in \{16, 64, 128\}$. However, the model size does not help to fit the data better, as the only difference we observe is an increased instability during optimization for larger models. This strengthens our theory of a more fundamental challenge, which the SKP is able to solve. Differences in optimization for varying model depth of $l \in \{1, 4, 32\}$ are displayed in Figure 4. Matching our theory, fitting the data becomes more challenging with increased depth for both the SKP and FAGCN, while the SKP is almost unaffected by the model depth. We provide an additional experiment in Appendix B.2.1 that confirms further shortcomings for multi-class classification settings using a synthetic dataset. These results confirm the severity of rank collapse for common tasks while also extending our theoretical insights to the non-linear case with multiple heads.

9 Conclusion

Our work shows that rank collapse of node representations is the underlying cause of over-smoothing and over-correlation in GNNs. We provided the theoretical foundation that a rank collapse of node representations occurs independently of the chosen aggregation function and the learned feature transformations. To mitigate these fundamental shortcomings of current message-passing frameworks, we propose the sum of Kronecker products (SKP) as a general property that models should exhibit to provably prevent rank collapse. We empirically confirm this behavior for non-linear GNNs on nine node classification tasks, which currently employed message-passing frameworks struggle to fit. The SKP easily fits the data even for 32 layers. 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 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.

[1] Our implementation is available at `https://github.com/roth-andreas/rank_collapse`

Acknowledgements

This research has been funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence and by the Federal Ministry of Education and Research of Germany under grant no. 01IS22094E WEST-AI.

References

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

[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. doi: 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. 1

[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. 1, 6, 8

[6] Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020. 1, 2, 3, 4, 5, 6

[7] Hoang Nt and Takanori Maehara. Revisiting graph neural networks: All we have is low-pass filters. *arXiv preprint arXiv:1905.09550*, 2019. 1

[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. 1, 3, 7

[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. 1, 2

[10] Nicolas Keriven. Not too little, not too much: a theoretical analysis of graph (over) smoothing. *arXiv preprint arXiv:2205.12156*, 2022. 2

[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. 1, 2, 3, 4, 21

[12] Chen Cai and Yusu Wang. A note on over-smoothing for graph neural networks. *arXiv preprint arXiv:2006.13318*, 2020. 1, 2, 3, 4, 5, 6

[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. 1, 2, 3, 4, 5, 6

[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. 2

[15] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017. 2, 3, 8

[16] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018. 2, 8

[17] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. 2, 3, 6, 7, 8

[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. 3, 4

[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. 3, 4, 6

[20] Sohir Maskey, Raffaele Paolino, Aras Bacho, and Gitta Kutyniok. A fractional graph laplacian approach to oversmoothing. *arXiv preprint arXiv:2305.13084*, 2023. 3, 4

[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. 3, 4

[22] T Konstantin Rusch, Michael M Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks. *arXiv preprint arXiv:2303.10993*, 2023. 4

[23] Xinyi Wu, Amir Ajorlou, Zihui Wu, and Ali Jadbabaie. Demystifying oversmoothing in attention-based graph neural networks. *arXiv preprint arXiv:2305.16102*, 2023. 3, 6, 20

[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. 3

[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. 3

[28] Xiaojun Guo, Yifei Wang, Tianqi Du, and Yisen Wang. Contranorm: A contrastive learning perspective on oversmoothing and beyond. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[29] Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. 3

[30] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *10th International Conference on Learning Representations, ICLR 2022*, 2022. 3

[31] Kai Guo, Kaixiong Zhou, Xia Hu, Yu Li, Yi Chang, and Xin Wang. Orthogonal graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3996–4004, 2022. 3

[32] Hua Liu, Haoyu Han, Wei Jin, Xiaorui Liu, and Hui Liu. Enhancing graph representations learning with decorrelated propagation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1466–1476, 2023. 3

[33] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017. 3, 8

[34] 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. 3

[35] 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. 4

[36] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007. 6

[37] Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized pagerank graph neural network. In *International Conference on Learning Representations*, 2020. 6, 8

[38] 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. 8

[39] Cristian Bodnar, Francesco Di Giovanni, Benjamin Paul Chamberlain, Pietro Lio, and Michael M Bronstein. Neural sheaf diffusion: A topological perspective on heterophily and over-smoothing in gnns. In *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*, 2022. 6, 8

[40] Gilbert Strang. *Linear algebra and its applications.* Belmont, CA: Thomson, Brooks/Cole, 2006. 6

[41] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021. 6, 8

[42] Oskar Perron. Zur theorie der matrices. *Mathematische Annalen*, 64(2):248–263, 1907. 6

[43] Robert G Gallager. Discrete stochastic processes. *Journal of the Operational Research Society*, 48(1):103–103, 1997. 6

[44] 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. 7

[45] 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. 7

[46] Charles F Van Loan and Nikos Pitsianis. *Approximation with Kronecker products*. Springer, 1993. 7

[47] 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. 7

[48] 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. doi: 10.1162/neco.1989.1.4.541. 7

[49] 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. 7

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

[51] 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. 7

[52] Eran Rosenbluth, Jan Toenshoff, and Martin Grohe. Some might say all you need is sum. *arXiv preprint arXiv:2302.11603*, 2023. 7

[53] 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. 7

[54] 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. 7

[55] Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-gcn: Geometric graph convolutional networks. In *International Conference on Learning Representations*, 2020. 7

[56] Benedek Rozemberczki, Carl Allen, and Rik Sarkar. Multi-scale attributed node embedding. *Journal of Complex Networks*, 9(2):cnab014, 2021. 7

[57] Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 807–816, 2009. 7

[58] Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. Masked label prediction: Unified message passing model for semi-supervised classification. *arXiv preprint arXiv:2009.03509*, 2020. 8

[59] 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. 8

[60] 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. 9, 21

[61] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019. 9

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

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \dots & a_{1q}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{p1}\mathbf{B} & \dots & a_{pq}\mathbf{B} \end{bmatrix} . \tag{14}$$

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

$$\mathrm{vec}(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W}) = (\mathbf{W}^T \otimes \tilde{\mathbf{A}})\,\mathrm{vec}(\mathbf{X}) \tag{15}$$

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

$$\mathrm{vec}(\mathbf{X}) = (\mathbf{U} \otimes \mathbf{V})\mathbf{c} \tag{16}$$

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

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{A}\mathbf{C}) \otimes (\mathbf{B}\mathbf{D})\,. \tag{17}$$

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 $\mathbf{\Delta}$:

$$E(\mathbf{X}) = tr(\mathbf{X}^T\mathbf{\Delta}\mathbf{X}) \tag{18}$$

$$= \mathrm{vec}(\mathbf{X})^T\,\mathrm{vec}(\mathbf{\Delta}\mathbf{X})\,. \tag{19}$$

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

$$E(\mathbf{X}) = \mathbf{c}^T(\mathbf{I}_d \otimes \mathbf{V})^T(\mathbf{I}_d \otimes \mathbf{\Delta})(\mathbf{I}_d \otimes \mathbf{V})\mathbf{c} \tag{20}$$

$$= \mathbf{c}^T(\mathbf{I}_d \otimes \mathbf{V}^T\mathbf{\Delta}\mathbf{V})\mathbf{c}\,. \tag{21}$$

As $\mathbf{\Delta} = \mathbf{V}(\mathbf{I}_n - \mathbf{\Lambda})\mathbf{V}$ 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:

$$E(\mathbf{X}) = \mathbf{c}^T(\mathbf{I}_d \otimes (\mathbf{I}_n - \mathbf{\Lambda}^{\mathbf{\Delta}})\mathbf{c} \tag{22}$$

$$= \sum_{l,r=1}^{n,d} (1 - \lambda_r^{\tilde{\mathbf{A}}})c_{l,r}^2\,. \tag{23}$$

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

$$\begin{aligned} ||\mathbf{X}||_F^2 &= tr(\mathbf{X}^T\mathbf{X}) \\ &= \mathbf{c}^T(\mathbf{U} \otimes \mathbf{V})^T(\mathbf{U} \otimes \mathbf{V})\mathbf{c} \\ &= \sum_{l,r=1}^{n,d} c_{l,r}^2 \end{aligned} \tag{24}$$

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

$$\|\mathbf{A} \otimes \mathbf{B}\|_F = \|\mathbf{A}\|_F \cdot \|\mathbf{B}\|_F \tag{25}$$

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,

$$||\phi(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W})||_F \leq \sigma_1^{\mathbf{W}} \cdot ||\mathbf{X}||_F \,. \tag{26}$$

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

$$\begin{aligned} ||\phi(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W})||_F &= ||\phi(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W}) - \phi(\mathbf{0})||_F && (27)\\ &\leq ||\tilde{\mathbf{A}}\mathbf{X}\mathbf{W} - \mathbf{0}||_F && (28)\\ &= ||\tilde{\mathbf{A}}\mathbf{X}\mathbf{W}||_F \,. && (29) \end{aligned}$$

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

$$\begin{aligned} ||\tilde{\mathbf{A}}\mathbf{X}\mathbf{W}||_F &\leq |\lambda_1^{\tilde{\mathbf{A}}}|\sigma_1^{\mathbf{W}}||\mathbf{X}||_F && (30)\\ &= \sigma_1^{\mathbf{W}}||\mathbf{X}||_F \,, && (31) \end{aligned}$$

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

A.2.1 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 $\mathcal{Q}_i = \mathrm{span}(\mathbf{I}_d \otimes \mathbf{v}_i)$ 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,

$$\begin{aligned} (\mathbf{W} \otimes \tilde{\mathbf{A}})(\mathbf{I}_d \otimes \mathbf{v}_i)\mathbf{c} &= (\mathbf{W}\mathbf{I}_d \otimes \tilde{\mathbf{A}}\mathbf{v}_i)\mathbf{c} \\ &= (\mathbf{I}_d \otimes \mathbf{v}_i)(\mathbf{W} \otimes \lambda_i^{\tilde{\mathbf{A}}}\mathbf{I}_n)\mathbf{c} \\ &= (\mathbf{I}_d \otimes \mathbf{v}_i)\mathbf{c}' \in \mathcal{Q}_i \end{aligned} \tag{32}$$

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

A.2.2 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,

$$\frac{\|\mathbf{T}\mathbf{S}_{(i)}\|_F}{\|\mathbf{T}\mathbf{S}_{(j)}\|_F} = \frac{|\lambda_i^{\tilde{\mathbf{A}}}|}{|\lambda_j^{\tilde{\mathbf{A}}}|} \,. \tag{33}$$

Proof.

$$\begin{aligned}\frac{\|\mathbf{T}\mathbf{S}_{(i)}\|}{\|\mathbf{T}\mathbf{S}_{(j)}\|} &= \frac{\|\left(\mathbf{W}\otimes\tilde{\mathbf{A}}\right)(\mathbf{I}_d\otimes\mathbf{v}_i)\|}{\|\left(\mathbf{W}\otimes\tilde{\mathbf{A}}\right)(\mathbf{I}_d\otimes\mathbf{v}_j)\|}\\ &= \frac{\|\left(\mathbf{W}\otimes\lambda_i^{\tilde{\mathbf{A}}}\mathbf{v}_i\right)\|}{\|\left(\mathbf{W}\otimes\lambda_j^{\tilde{\mathbf{A}}}\mathbf{v}_j\right)\|}\\ &= \frac{|\lambda_i^{\tilde{\mathbf{A}}}|\cdot\|\mathbf{W}\|\cdot\|\mathbf{v}_i\|}{|\lambda_j^{\tilde{\mathbf{A}}}|\cdot\|\mathbf{W}\|\cdot\|\mathbf{v}_j\|}\\ &= \frac{|\lambda_i^{\tilde{\mathbf{A}}}|}{|\lambda_j^{\tilde{\mathbf{A}}}|}\end{aligned} \tag{34}$$

□

A.2.3 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 $\mathbf{v}_1,\dots,\mathbf{v}_n$ and eigenvalues $\lambda_1^{\tilde{\mathbf{A}}},\dots,\lambda_n^{\tilde{\mathbf{A}}}$ 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,

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

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

Proof.

$$\begin{aligned}\lim_{l\to\infty}\frac{\|\mathbf{T}^{(l)}\dots\mathbf{T}^{(1)}\mathbf{S}_{(i)}\|}{\max_p\|\mathbf{T}^{(l)}\dots\mathbf{T}^{(1)}\mathbf{S}_{(p)}\|} &= \lim_{l\to\infty}\frac{||(\mathbf{W}^{(l)}\otimes\tilde{\mathbf{A}})\dots(\mathbf{W}^{(1)}\otimes\tilde{\mathbf{A}})(\mathbf{I}_d\otimes\mathbf{v}_i)||}{\max_p||(\mathbf{W}^{(l)}\otimes\tilde{\mathbf{A}})\dots(\mathbf{W}^{(1)}\otimes\tilde{\mathbf{A}})(\mathbf{I}_d\otimes\mathbf{v}_p)||}\\ &= \lim_{l\to\infty}\frac{||(\mathbf{W}^{(l)}\dots\mathbf{W}^{(1)}\otimes\left(\lambda_i^{\tilde{\mathbf{A}}}\right)^l\mathbf{v}_i)||}{\max_p||(\mathbf{W}^{(l)}\dots\mathbf{W}^{(1)}\otimes\left(\lambda_p^{\tilde{\mathbf{A}}}\right)^l\mathbf{v}_p)||}\\ &= \lim_{l\to\infty}\frac{|\lambda_i^{\tilde{\mathbf{A}}}|^l\cdot||\mathbf{v}_i||}{\max_p||\lambda_p^{\tilde{\mathbf{A}}}|^l|\cdot||\mathbf{v}_p||}\\ &= \begin{cases}1, & \text{if}|\lambda_i^{\tilde{\mathbf{A}}}| = |\lambda_1^{\tilde{\mathbf{A}}}|\\ 0, & \text{otherwise}\end{cases}\end{aligned} \tag{35}$$

□

A.2.4 Proof of Theorem 5.4

Theorem. (Representations converge in probability 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 $\mathbf{v}_1,\dots,\mathbf{v}_n$ and $\mathbf{W}^{(k)}\in\mathbb{R}^{d\times d}$ any matrix. Consider the subspaces $\mathcal{Q}_i = \text{span}(\mathbf{I}_d\otimes\mathbf{v}_i)$ for $i\in[n]$. Then,

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

Proof. Let $\mathbf{\Lambda}$ be the matrix of eigenvalues of $\tilde{\mathbf{A}}$ and the singular value composition of $\mathbf{W}^{(l)}\dots\mathbf{W}^{(1)} = \mathbf{U}\mathbf{\Sigma}\mathbf{N}^T$. We decompose the initial state $\mathbf{X}^{(0)} = (\mathbf{N}\otimes\mathbf{V})\mathbf{c}$ with $\mathbf{c} =$

$(\mathbf{N} \otimes \mathbf{V})^T \mathbf{X}^{(0)}$. We emphasize that the coefficients, therefore, depend on both the initial state $\mathbf{X}^{(0)}$ and the feature transformations $\mathbf{W}^{(l)} \dots \mathbf{W}^{(1)}$. Then,

$$\begin{aligned}
\left\| \frac{\mathbf{X}^{(l)}}{||\mathbf{X}^{(l)}||} - \mathbf{m}^{(l)} \right\| &= \left\| \frac{(\mathbf{W}^{(l)} \otimes \tilde{\mathbf{A}}) \dots (\mathbf{W}^{(1)} \otimes \tilde{\mathbf{A}})(\mathbf{N} \otimes \mathbf{V})\mathbf{c}}{||(\mathbf{W}^{(l)} \otimes \tilde{\mathbf{A}}) \dots (\mathbf{W}^{(1)} \otimes \tilde{\mathbf{A}})(\mathbf{N} \otimes \mathbf{V})\mathbf{c}||} - \mathbf{m}^{(l)} \right\| \\
&= \left\| \frac{(\mathbf{W}^{(l)} \dots \mathbf{W}^{(1)} \otimes \mathbf{V})(\mathbf{N} \otimes \mathbf{\Lambda}^l)\mathbf{c}}{||(\mathbf{W}^{(l)} \dots \mathbf{W}^{(1)} \otimes \mathbf{V})(\mathbf{N} \otimes \mathbf{\Lambda}^l)\mathbf{c}||} - \mathbf{m}^{(l)} \right\| \\
&= \left\| \frac{\left(\mathbf{W}^{(l)} \dots \mathbf{W}^{(1)} \otimes \mathbf{V}\right) \left(\mathbf{N} \otimes \frac{\mathbf{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}}}\right) \mathbf{c}}{|| \left(\mathbf{W}^{(l)} \dots \mathbf{W}^{(1)} \otimes \mathbf{V}\right) \left(\mathbf{N} \otimes \frac{\mathbf{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}^l}}\right) \mathbf{c}||} - \mathbf{m}^{(l)} \right\|
\end{aligned} \tag{36}$$

We simplify the notation and set $\mathbf{P}^{(l)} = \mathbf{W}^{(l)} \dots \mathbf{W}^{(1)}$. We now choose $\mathbf{m}^{(l)} = \frac{\left(\mathbf{P}^{(l)} \otimes \mathbf{V}\right) \left(\mathbf{N} \otimes \frac{\mathbf{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}^l}}\right) \mathbf{c}'}{||\left(\mathbf{P}^{(l)} \otimes \mathbf{V}\right) \left(\mathbf{N} \otimes \frac{\mathbf{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}^l}}\right) \mathbf{c}||}$ 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., $c'_{1,r} = \mathbf{c}_{1,r} \, \forall r$ and $c'_{i,r} = 0 \, \forall i > 1, \forall r$. Continuing,

$$\begin{aligned}
&= \left\| \frac{\left(\frac{\mathbf{P}^{(l)}}{\sigma_1^{\mathbf{P}^{(l)}}} \otimes \mathbf{V}\right)}{|| \left(\frac{\mathbf{P}^{(l)}}{\sigma_1^{\mathbf{P}^{(l)}}} \otimes \mathbf{V}\right) \left(\mathbf{N} \otimes \frac{\mathbf{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}^l}}\right) \mathbf{c}||} \left(\left(\mathbf{N} \otimes \frac{\mathbf{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}^l}}\right) \mathbf{c} - \mathbf{c}' \right) \right\| \\
&\leq \frac{\left\| \frac{\mathbf{P}^{(l)}}{\sigma_1^{\mathbf{P}^{(l)}}} \otimes \mathbf{V} \right\|}{\left\| \left(\frac{\mathbf{P}^{(l)}}{\sigma_1^{\mathbf{P}^{(l)}}} \otimes \mathbf{V}\right) \left(\mathbf{N}_d \otimes \frac{\mathbf{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}^l}}\right) \mathbf{c} \right\|} \left\| \left(\mathbf{N}_d \otimes \frac{\mathbf{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}^l}}\right) \mathbf{c} - \mathbf{c}' \right\| \\
&\leq \frac{\sqrt{nd}}{c_{1,1}} \left(\frac{\lambda_2}{\lambda_1}\right)^l \max_{r,i} c_{r,i} \sqrt{(n-1) \cdot d} < \epsilon \\
&\iff \frac{\sqrt{nd}}{\epsilon} \left(\frac{\lambda_2}{\lambda_1}\right)^l \sqrt{(n-1) \cdot d} < \frac{c_{1,1}}{\max_{r,i} c_{r,i}}
\end{aligned} \tag{37}$$

The left-hand side converges to zero for $l \to \infty$, so the constraint on the coefficients becomes weaker with increased depth. As the volume of the critical solution space around solutions satisfying $c_{1,1} = 0$ reduces to zero, our state converges in probability.

We now provide the details for the upper bounds in the last step separately for each term:

$$\begin{aligned}
\left\| \frac{\mathbf{P}^{(l)}}{\sigma_1^{\mathbf{P}^{(l)}}} \otimes \mathbf{V} \right\| &= \sqrt{tr\left(\left(\frac{\mathbf{P}^{(l)}}{\sigma_1^{\mathbf{P}^{(l)}}} \otimes \mathbf{V}\right)^T \left(\frac{\mathbf{W}^{(l)} \dots \mathbf{W}^{(1)}}{\sigma_1^{\mathbf{P}^{(l)}}} \otimes \mathbf{V}\right) \right)} \\
&= \sqrt{tr\left(\frac{\left(\mathbf{P}^{(l)}\right)^T \mathbf{P}^{(l)}}{\left(\sigma_1^{\mathbf{P}^{(l)}}\right)^2} \otimes \mathbf{I}_n \right)} \\
&= \sqrt{\sum_{l,r} \frac{\sigma_l^2}{\sigma_1^2}} \leq \sqrt{nd}
\end{aligned} \tag{38}$$

Next, we provide a lower bound on the denominator:

$$\begin{aligned}
\left\| \left(\frac{\mathbf{P}^{(l)}}{\sigma_1^{\mathbf{P}^{(l)}}} \otimes \mathbf{V} \right) \left(\mathbf{N}_d \otimes \left(\frac{\mathbf{\Lambda}}{\lambda_1^{\tilde{\mathbf{A}}}} \right)^l \right) \mathbf{c} \right\| &= \sqrt{\left(\left(\frac{\mathbf{P}^{(l)}}{\sigma_1^{\mathbf{P}^{(l)}}} \otimes \mathbf{V} \right) \left(\mathbf{N}_d \otimes \frac{\mathbf{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}^l}} \right) \mathbf{c} \right)^T \left(\frac{\mathbf{P}^{(l)}}{\sigma_1^{\mathbf{P}^{(l)}}} \otimes \mathbf{V} \right) \left(\mathbf{N}_d \otimes \left(\frac{\mathbf{\Lambda}}{\lambda_1^{\tilde{\mathbf{A}}}} \right)^l \right) \mathbf{c}} \\
&= \sqrt{\mathbf{c}^T \left(\mathbf{N} \frac{\left(\mathbf{P}^{(l)}\right)^T \mathbf{P}^{(l)}}{\left(\sigma_1^{\mathbf{P}^{(l)}}\right)^2} \mathbf{N} \otimes \left(\frac{\mathbf{\Lambda}}{\lambda_1^{\tilde{\mathbf{A}}}} \right)^{2l} \right) \mathbf{c}} \\
&= \sqrt{\sum_{l,r} \left(\frac{\sigma_l^{\mathbf{P}^{(l)}}}{\sigma_1^{\mathbf{P}^{(l)}}} \right)^2 \mathbf{c}_{l,r}^2 \left(\frac{\lambda_r}{\lambda_1} \right)^{2l}} \\
&\geq c_{1,1}
\end{aligned} \tag{39}$$

$$\begin{aligned}
\left\| \left(\mathbf{N}_d \otimes \frac{\mathbf{\Lambda}^{\mathbf{A}^l}}{\lambda_1^{\mathbf{A}}} \right) \mathbf{c} - \mathbf{c}' \right\| &= \sqrt{\sum_{r,i} \left(\left(\frac{\lambda_i}{\lambda_1} \right)^l c_{r,i} - \left(\frac{\lambda_i}{\lambda_1} \right)^l c'_{r,i} \right)^2} \\
&\leq \sqrt{\sum_{r=1,i=2}^{d,n} \left(\left(\frac{\lambda_i}{\lambda_1} \right)^l c_{r,i} \right)^2} \\
&\leq \left(\frac{\lambda_2}{\lambda_1} \right)^l \sqrt{\sum_{r=1,i=2}^{n,d} c_{r,i}^2} \\
&\leq \left(\frac{\lambda_2}{\lambda_1} \right)^l \sqrt{(n-1) \cdot d \cdot \max_{r,i} c_{r,i}^2} = \left(\frac{\lambda_2}{\lambda_1} \right)^l \max_{r,i} c_{r,i} \sqrt{(n-1) \cdot d}
\end{aligned} \tag{40}$$

□

A.2.5 Proof of Proposition 5.5

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

$$\operatorname*{plim}_{l \to \infty} E\left(\frac{\mathbf{X}^{(l)}}{\|\mathbf{X}^{(l)}\|_F} \right) = 0 \tag{41}$$

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

Proof. We again use the eigendecomposition $\tilde{\mathbf{A}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, the singular value decomposition $\mathbf{W}^{(1)} \dots \mathbf{W}^{(l)} = \mathbf{P} = \mathbf{U}\mathbf{\Sigma}\mathbf{N}^T$ and the coefficients $\mathbf{c} = (\mathbf{U} \otimes \mathbf{V})^T \operatorname{vec}(\mathbf{X}^{(0)})$ of our initial state.

With this, we can simplify the Dirichlet energy

$$
\begin{aligned}
E\left(\frac{\mathbf{X}^{(l)}}{||\mathbf{X}^{(l)}||}\right) &= \frac{\mathrm{tr}(\mathbf{X}^{(l)}\mathbf{\Delta}\mathbf{X}^{(l)})}{||\mathbf{X}^{(l)}||^2} \\
&= \frac{\mathbf{c}^T(\mathbf{U}^T \otimes \mathbf{V}^T)(\mathbf{W}^{(1)} \otimes \tilde{\mathbf{A}}^T)\dots(\mathbf{W}^{(l)} \otimes \tilde{\mathbf{A}})(\mathbf{I}_d \otimes \tilde{\mathbf{\Delta}})(\mathbf{W}^{(l)^T} \otimes \tilde{\mathbf{A}})(\mathbf{W}^{(1)^T} \otimes \tilde{\mathbf{A}})(\mathbf{U} \otimes \mathbf{V})\mathbf{c}}{||\mathbf{X}^{(l)}||^2} \\
&= \frac{\mathbf{c}^T(\mathbf{U}^T\mathbf{W}^{(1)}\dots\mathbf{W}^{(l)}\mathbf{W}^{(l)^T}\dots\mathbf{W}^{(1)^T}\mathbf{U} \otimes \mathbf{V}^T\tilde{\mathbf{A}}^l\tilde{\mathbf{\Delta}}\tilde{\mathbf{A}}^l\mathbf{V})\mathbf{c}}{||\mathbf{X}^{(l)}||^2} \\
&= \frac{\mathbf{c}^T(\mathbf{\Sigma}^2 \otimes \mathbf{\Lambda}^l\mathbf{V}^T\tilde{\mathbf{\Delta}}\mathbf{V}\mathbf{\Lambda}^l)\mathbf{c}}{\mathbf{c}^T(\mathbf{\Sigma}^2 \otimes \mathbf{\Lambda}^l\mathbf{V}^T\mathbf{V}\mathbf{\Lambda}^l)\mathbf{c}} \\
&= \frac{\mathbf{c}^T(\frac{\mathbf{\Sigma}^2}{\sigma_1^2} \otimes \frac{\mathbf{\Lambda}^l(\mathbf{I}_d-\mathbf{\Lambda})\mathbf{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}}\lambda_1^{\tilde{\mathbf{A}}}})\mathbf{c}}{\mathbf{c}^T(\frac{\mathbf{\Sigma}^2}{\sigma_1^2} \otimes \frac{\mathbf{\Lambda}^l\mathbf{\Lambda}^l}{\lambda_1^{\tilde{\mathbf{A}}}\lambda_1^{\tilde{\mathbf{A}}}})\mathbf{c}} \\
&= \frac{\sum_{p=1,r=1}^{N,d} c_{p,r}^2 \frac{\sigma_r^2}{\sigma_1^2}\frac{\lambda_p^{2l}(1-\lambda_p)}{\lambda_1^{2l}})}{\sum_{p=1,r=1}^{N,d} c_{p,r}^2 \frac{\sigma_r^2}{\sigma_1^2}\frac{\lambda_p^{2l}}{\lambda_1^{2l}})} \,.
\end{aligned}
\tag{42}
$$

For convergence, we now need the last term to be smaller than any given $\epsilon > 0$. Using the property that $(1 - \lambda_1^{\tilde{\mathbf{A}}}) = 0$, we again only require a constraint on the coefficients

$$
\frac{c_{1,1}^2}{\max_{p,r} c_{p,r}^2} > \frac{2 \cdot \lambda_2^{2l}}{\lambda_1^{2l}\epsilon} \,. \tag{43}
$$

Which similarly converges in probability as the right-hand side converges to zero. □

A.3 Proof of Proposition 5.6

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:

$$
\frac{||\mathbf{T}\mathbf{S}_{(i)}||_F}{||\mathbf{T}\mathbf{S}_{(j)}||_F} = \frac{||(\mathbf{W} \otimes \tilde{\mathbf{A}})(\mathbf{I}_d \otimes \mathbf{P}_{(i)})||_F}{||(\mathbf{W} \otimes \tilde{\mathbf{A}})(\mathbf{I}_d \otimes \mathbf{P}_{(j)})||_F} \tag{44}
$$

$$
= \frac{||\mathbf{W}||_F \cdot ||\tilde{\mathbf{A}}\mathbf{P}_{(i)}||_F}{||\mathbf{W}|| \cdot ||\tilde{\mathbf{A}}\mathbf{P}_{(j)}||_F} \tag{45}
$$

$$
= \frac{||\tilde{\mathbf{A}}\mathbf{P}_{(i)}||_F}{||\tilde{\mathbf{A}}\mathbf{P}_{(j)}||_F} \tag{46}
$$

□

A.3.1 Proof of Proposition 5.7

Proposition A.1. *(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,*

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

Proof. The proof is mostly analogous to our proof of Theorem 5.5 but relies on the fact that the product of time-inhomogenous row-stochastic matrices $\lim_{l\to\infty}\Pi_{i=0}^{l}\tilde{\mathbf{A}}^{(i)} = \mathbf{1}(\mathbf{y})^T$ converges to constant columns for some $\mathbf{y} \in \mathbb{R}^n$ as established by Wu et al. [23]. The intuition behind the statement builds on the ergodicity and minimum edge weight ϵ, that 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 $\lim_{l\to\infty}\Pi_{i=0}^{l}\tilde{\mathbf{A}}^{(i)} = \mathbf{V}\mathbf{Q}^{(l)}\mathbf{V}^T = \mathbf{V}\begin{bmatrix} y_1^{(l)} & \dots & y_n^{(l)} \\ q_{21}^{(l)} & \dots & q_{2n}^{(l)} \\ \vdots & \ddots & \vdots \\ q_{n2}^{(l)} & \dots & q_{nn}^{(l)}\end{bmatrix}\mathbf{V}^T$ using the same eigenbasis with each $q_{kp}^{(l)}$ converging to zero with some upper bound Cq^l for some $C > 0$ and $q \in [0,1]$, and each $y_i^{(l)}$ converges to $\mathbf{y}^T\mathbf{v}_i$, as given by Wu et al. [23]. For notational simplicity we define $\mathbf{R}^{(l)} = \Pi_{i=0}^{(l)}\mathbf{W}^{(i)}$ and its singular value decomposition as $\mathbf{R}^{(l)} = \mathbf{U}\mathbf{\Sigma}\mathbf{N}^T$. Decomposing the initial state $\mathrm{vec}(\mathbf{X}^{(0)}) = (\mathbf{N}\otimes\mathbf{V})\mathbf{c}$ using $\mathbf{V}$ and $\mathbf{N}$, the Dirichlet energy then simplifies to

$$\begin{aligned}\lim_{l\to\infty} E\left(\frac{\mathbf{V}\mathbf{Q}^{(l)}\mathbf{V}^T\mathbf{X}^{(0)}\mathbf{R}}{||\mathbf{V}\mathbf{Q}^{(l)}\mathbf{V}^T\mathbf{X}^{(0)}\mathbf{R}||_F}\right) &= \lim_{l\to\infty}\frac{\mathbf{c}^T(\mathbf{\Sigma}^2\otimes\mathbf{V}^T\mathbf{V}\mathbf{Q}^{(l)}\mathbf{V}^T\mathbf{V}(\mathbf{I}_n - \Lambda^{\tilde{\mathbf{A}}})\mathbf{V}^T\mathbf{V}\mathbf{Q}^{(l)}\mathbf{V}^T\mathbf{V})\mathbf{c}}{\mathbf{c}(\mathbf{\Sigma}^2\otimes\mathbf{V}^T\mathbf{V}\mathbf{Q}^{(l)}\mathbf{V}^T\mathbf{V}\mathbf{Q}^{(l)}\mathbf{V}^T\mathbf{V})\mathbf{c}} \\ &= \lim_{l\to\infty}\frac{\mathbf{c}^T(\mathbf{\Sigma}^2\otimes\mathbf{Q}^{(l)^T}(\mathbf{I}_n - \Lambda^{\tilde{\mathbf{A}}})\mathbf{Q}^{(l)})\mathbf{c}}{\mathbf{c}(\mathbf{\Sigma}^2\otimes\mathbf{Q}^{(l)^T}\mathbf{Q}^{(l)})\mathbf{c}}\end{aligned}$$

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:

$$\frac{\mathbf{c}^T(\mathbf{\Sigma}^2\otimes\mathbf{Q}^{(l)^T}(\mathbf{I}_n - \Lambda^{\tilde{\mathbf{A}}})\mathbf{Q}^{(l)})\mathbf{c}}{\mathbf{c}(\mathbf{\Sigma}^2\otimes\mathbf{Q}^{(l)^T}\mathbf{Q}^{(l)})\mathbf{c}} = \frac{\mathbf{c}^T(\frac{\mathbf{\Sigma}^2}{\sigma_1^2}\otimes\mathbf{Q}^{(l)^T}(\mathbf{I}_n - \Lambda^{\tilde{\mathbf{A}}})\mathbf{Q}^{(l)})\mathbf{c}}{\mathbf{c}(\frac{\mathbf{\Sigma}^2}{|\sigma_1^2|}\otimes\mathbf{Q}^{(l)^T}\mathbf{Q}^{(l)})\mathbf{c}} \tag{47}$$

$$\leq \frac{\sum_{p,r} c_{p,r}^2\frac{\sigma_r}{\sigma_1}(1-\lambda_p)}{} \tag{48}$$

To ensure that the last term converges to a value below a given ϵ, we again require

$$\frac{c_{1,1}}{\max_{p,r} c_{p,r}} > \frac{2(d-1)\cdot Cq^l}{\epsilon}\,. \tag{49}$$

The right-hand side converges to zero as per assumption, so the overall convergence happens in probability. □

A.4 Over-correlation

A.4.1 Proof of Theorem 6.1

Theorem. Let $\mathcal{R}_j$ be the space of matrices with rank at most j. Then, for all $\tilde{\mathbf{A}} \in \mathbb{R}^{n\times n}$ and for $\mathbf{W}^{(l)}\dots\mathbf{W}^{(1)}$ in probability with respect to the Lebesgue measure, we have

$$\forall\,\epsilon > 0.\,\exists\,N\in\mathbb{N}.\,\forall\,l > N.\,L^{(n)}\in\mathcal{R}_j\colon \mathbb{P}\left(||X^{(n)} - L^{(n)}|| > \epsilon\right) = 0\,. \tag{50}$$

Proof. We show that $\mathbf{X}^{(l)}$ converges to a sum of a matrix with rank at most j and a matrix that vanishes. We construct subspaces given by the Jordan blocks of the Jordan normal form $\tilde{\mathbf{A}} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$. We merge all subspaces corresponding to an eigenvalue with magnitude $|\lambda_1|$ into $\mathcal{Z} = \mathrm{span}(\mathbf{I}_d \otimes \mathbf{P}_1)$ with $\mathbf{P}_1 \in \mathbb{R}^{n \times j}$ where j is the joint algebraic multiplicity of all eigenvalues with magnitude $|\lambda_1|$. Using these subspaces within Theorem 5.4, $\mathbf{X}^{(l)}$ converges in probability to $\mathcal{Z}$. It is then only left to show that $\mathcal{Z}$ corresponds to a subspace of $\mathcal{R}_j$: For every $\mathbf{h} \in \mathcal{Z}$ with $\mathbf{h} = (\mathbf{I}_d \otimes \mathbf{P}_1)\mathbf{c}$, we have

$$\mathrm{vec}^{-1}(\mathbf{h}) = \mathrm{vec}^{-1}((\mathbf{I}_d \otimes \mathbf{P}_1)\mathbf{c}) = \mathbf{P}_1 \,\mathrm{vec}^{-1}(\mathbf{c}^T)\mathbf{I}_d = \mathbf{P}_1 \,\mathrm{vec}^{-1}(\mathbf{c}^T) \tag{51}$$

which has rank at most j. □

A.5 Sum of Kronecker products

A.6 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 $\mathbf{S} = [\mathbf{e}_1 \otimes \mathbf{s}_1 \quad \ldots \quad \mathbf{e}_d \otimes \mathbf{s}_d] \in \mathbb{R}^{nd \times d}$, there exists an $\mathbf{T} = \sum_1^d (\mathbf{W}_i \otimes \tilde{\mathbf{A}}_i)$ such that for all orthogonal bases $\mathbf{S}' = [\mathbf{e}_1 \otimes \mathbf{s}'_1 \quad \ldots \quad \mathbf{e}_d \otimes \mathbf{s}'_d] \in \mathbb{R}^{nd \times d}$

$$\frac{\|\mathbf{T}\mathbf{S}\|_F}{\|\mathbf{T}\mathbf{S}'\|_F} > \frac{\|\mathbf{S}\|_F}{\|\mathbf{S}'\|_F}\,. \tag{52}$$

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,

$$\frac{\|\mathbf{T}\mathbf{S}\|_F}{\|\mathbf{T}\mathbf{S}'\|_F} \geq \frac{|\lambda_1|}{|\lambda_2|}\frac{\|\mathbf{S}\|_F}{\|\mathbf{S}'\|_F} > \frac{\|\mathbf{S}\|_F}{\|\mathbf{S}'\|_F} \tag{53}$$

□

B Experimental details

B.1 Convergence to a Constant State

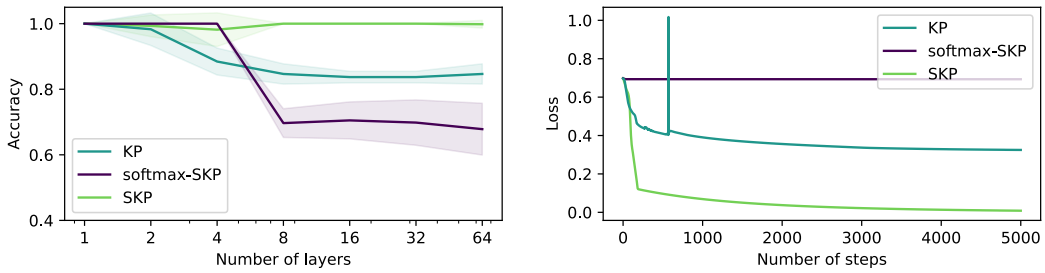
For this experiment, we consider the Cora dataset provided by Pytorch-Geometric [60], consisting of 2708 nodes and 5429. For initialize 128 layers of GAT, GCN, and GraphSAGE using their default initialization and no bias. After each layer, we use a ReLU activation. After each ReLU activation, we track 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

We provide an additional experiment on synthetic data that shows that linearly independent features are not practically achievable, even for non-linear models.

B.2.1 Synthetic Dataset

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



(a) Accuracy: mean and standard deviation. **(b)** Optimization loss for 8 layers.

Figure 5: Accuracies and loss dynamics for the synthetic task.

tasks of this form are trivial when accessing another dimension. Our experimental setting follows this idea: 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 $\mathbf{W}_1^{(k)} \in \mathbb{R}^{6\times 6}$, $\mathbf{W}_2^{(k)} \in \mathbb{R}^{6\times 6}$ 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 $\mathrm{vec}(\mathbf{X}^{(k+1)}) = \phi(\mathbf{T}^{(k)}\,\mathrm{vec}(\mathbf{X}^{(k)})$ with ϕ as the ReLU activation function. After l iterations, we use these node representations for our three classification tasks using the affine transformation

$$\hat{\mathbf{Y}} = \sigma(\mathbf{X}^{(l)}\mathbf{W}_c + \mathbf{b}) \tag{54}$$

with σ as sigmoid activation, a feature transformation $\mathbf{W}_c \in \mathbb{R}^{6\times 3}$ 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 $\tilde{\mathbf{A}}_1^{(1)} = \cdots = \tilde{\mathbf{A}}_1^{(k)}$, $\tilde{\mathbf{A}}_2^{(1)} = \cdots =$

# of layers	1	2	4	8	16	32	64	128
$\mathrm{softmax} - \mathrm{SKP}_{HL}$	100 ± 0	100 ± 0	82 ± 19	71 ± 6	72 ± 6	70 ± 6	67 ± 7	63 ± 8
$\mathrm{softmax} - \mathrm{SKP}_{L}$	100 ± 0	100 ± 0	100 ± 0	70 ± 4	71 ± 6	70 ± 7	68 ± 8	65 ± 8
softmax-SKP	100 ± 0	100 ± 0	100 ± 0	70 ± 4	71 ± 6	70 ± 7	68 ± 8	65 ± 8
KP_{HL}	100 ± 0	100 ± 0	83 ± 0	83 ± 0	83 ± 0	83 ± 0	83 ± 0	82 ± 4
KP_{L}	100 ± 0	100 ± 0	83 ± 0	83 ± 0	83 ± 0	83 ± 0	83 ± 0	82 ± 4
KP	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$\pm$standard deviations accuracies for all considered scenarios. Subscript L denotes the linearized version, subscript HL denotes the version that additionally uses homogeneous aggregation matrices.

$\tilde{\mathbf{A}}_2^{(k)}$. All configurations were run sequentially on a single CPU. The entire runtime was around 30 hours.

Detailed results. We provide additional numerical results in Figure 5 and 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 softmax-SKP. However, the representations are quickly converging to a low-rank state, so both softmax-SKP and KP 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.