Neural Sheaf Diffusion: A Topological Perspective on Heterophily and Oversmoothing in GNNs

Cristian Bodnar[*]
University of Cambridge
cb2015@cam.ac.uk

Francesco Di Giovanni[†]
Twitter
fdigiovanni@twitter.com

Benjamin P. Chamberlain
Twitter

Pietro Liò
University of Cambridge

Michael Bronstein
University of Oxford & Twitter

Abstract

Cellular sheaves equip graphs with a "geometrical" structure by assigning vector spaces and linear maps to nodes and edges. Graph Neural Networks (GNNs) implicitly assume a graph with a trivial underlying sheaf. This choice is reflected in the structure of the graph Laplacian operator, the properties of the associated diffusion equation, and the characteristics of the convolutional models that discretise this equation. In this paper, we use cellular sheaf theory to show that the underlying geometry of the graph is deeply linked with the performance of GNNs in heterophilic settings and their oversmoothing behaviour. By considering a hierarchy of increasingly general sheaves, we study how the ability of the sheaf diffusion process to achieve linear separation of the classes in the infinite time limit expands. At the same time, we prove that when the sheaf is non-trivial, discretised parametric diffusion processes have greater control than GNNs over their asymptotic behaviour. On the practical side, we study how sheaves can be learned from data. The resulting sheaf diffusion models have many desirable properties that address the limitations of classical graph diffusion equations (and corresponding GNN models) and obtain competitive results in heterophilic settings. Overall, our work provides new connections between GNNs and algebraic topology and would be of interest to both fields.

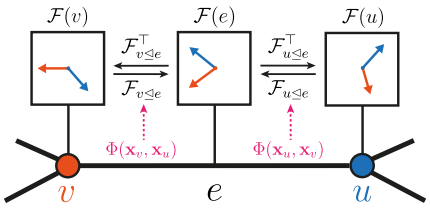


Figure 1: A sheaf $(G, \mathcal{F})$ shown for a single edge of the graph. The *stalks* are isomorphic to $\mathbb{R}^2$. The *restriction maps* $\mathcal{F}_{v \trianglelefteq e}$, $\mathcal{F}_{u \trianglelefteq e}$ and their adjoints move the vector features between these spaces. In practice, we learn the sheaf (i.e. the restrictions maps) from data via a parametric function Φ.

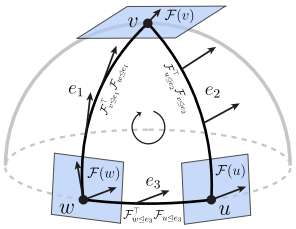


Figure 2: Analogy between parallel transport on a sphere and transport on a discrete vector bundle (cellular sheaf). A tangent vector is moved from $\mathcal{F}(w) \to \mathcal{F}(v) \to \mathcal{F}(u)$ and back. Because the vector returns in a different position, the transport is not path-independent.

[*]Work done as a research intern at Twitter.
[†]Proved the results in Section 3.1.
Our code is available at `https://github.com/twitter-research/neural-sheaf-diffusion`.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

1 Introduction

Graph Neural Networks (GNNs) [11, 19, 26–28, 37, 56, 62] have recently become very popular in the ML community as a model of choice to deal with relational and interaction data due to their multiple successful applications in domains ranging from social science and particle physics to structural biology and drug design. In this work, we focus on two main problems often observed in GNNs: their poor performance in heterophilic graphs [72] and their oversmoothing behaviour [46, 48]. The former arises from the fact that many GNNs are built on the strong assumption of *homophily*, i.e., that nodes tend to connect to other similar nodes. The latter refers to a phenomenon of some deeper GNNs producing features that are too smooth to be useful.

Contributions. We show that these two fundamental problems are linked by a common cause: the underlying "geometry" of the graph (used here in a very loose sense). When this geometry is trivial, as is typically the case, the two phenomena described above emerge. We make these statements precise through the lens of (cellular) sheaf theory [9, 17, 25, 42, 54, 60], a subfield of algebraic topology and geometry. Intuitively, a cellular sheaf associates a vector space to each node and edge of a graph, and a linear map between these spaces for each incident node-edge pair (Figure 1).

In Section 3 we analyse how by considering a hierarchy of increasingly general sheaves, starting from a trivial one, a diffusion equation based on the sheaf Laplacian [32] can solve increasingly more complicated node-classification tasks in the infinite time limit. In this regime, we show that oversmoothing and problems due to heterophily can be avoided by equipping the graph with the right sheaf structure for the task. In Section 4, we study the behaviour of a non-linear, parametric, and discrete version of this process. This results in a *Sheaf Convolutional Network* [30] that generalises GCN of Kipf and Welling [37]. We prove that this discrete diffusion process is more flexible and has greater control over its asymptotic behaviour than GCNs [12, 49].

All these results are based on the properties of the harmonic space of the sheaf Laplacian, which we study from a spectral perspective in Section 3.1. We provide a new Cheeger-type inequality for the spectral gap of the sheaf Laplacian and note that these results might be of independent interest for spectral sheaf theory [32]. Finally, in Section 5, we apply our theory to designing simple and practical GNN models. We describe how to construct Sheaf Neural Networks by learning sheaves from data, thus making these types of models applicable beyond the toy experimental setting where they were originally introduced [30]. The resulting models obtain competitive results both in heterophilic and homophilic graphs.

2 Background

Cellular Sheaves. A *cellular sheaf* [17, 60] over a graph (Figure 1) is a mathematical object associating a vector space to each node and edge in the graph and a map between these spaces for each incident node-edge pair. We define this formally below:

Definition 1. *A cellular sheaf $(G, \mathcal{F})$ on an undirected graph $G = (V, E)$ consists of:*

- *A vector space $\mathcal{F}(v)$ for each $v \in V$.*
- *A vector space $\mathcal{F}(e)$ for each $e \in E$.*
- *A linear map $\mathcal{F}_{v \trianglelefteq e} : \mathcal{F}(v) \to \mathcal{F}(e)$ for each incident $v \trianglelefteq e$ node-edge pair.*

The vector spaces of the nodes and edges are called *stalks*, while the linear maps are referred to as *restriction maps*. The space formed by all the spaces associated with the nodes of the graph is called the space of 0-*cochains* $C^0(G; \mathcal{F}) := \bigoplus_{v \in V} \mathcal{F}(v)$, where $\bigoplus$ denotes the direct sum of vector spaces. For a 0-cochain $\mathbf{x} \in C^0(G; \mathcal{F})$, we use $\mathbf{x}_v$ to refer to the vector in $\mathcal{F}(v)$ of node v. Hansen and Ghrist [33] have constructed a convenient mental model for these objects based on opinion dynamics. In this context, $\mathbf{x}_v$ is the 'private opinion' of node v, while $\mathcal{F}_{v \trianglelefteq e}\mathbf{x}_v$ expresses how that opinion manifests publicly in a 'discourse space' formed by $\mathcal{F}(e)$. A particularly important subspace of $C^0(G; \mathcal{F})$ is the space of *global sections* $H^0(G; \mathcal{F}) := \{\mathbf{x} \in C^0(G; \mathcal{F}) : \mathcal{F}_{v \trianglelefteq e}\mathbf{x}_v = \mathcal{F}_{u \trianglelefteq e}\mathbf{x}_u\}$ containing those private opinions $\mathbf{x}$ for which all neighbours (v, u) agree with each other in the discourse space. Given a cellular sheaf $(G, \mathcal{F})$, we can define a *sheaf Laplacian* operator [32] measuring the aggregated 'disagreement of opinions' at each node:

Definition 2. *The sheaf Laplacian of a sheaf $(G, \mathcal{F})$ is a linear map $L_{\mathcal{F}} : C^0(G, \mathcal{F}) \to C^0(G, \mathcal{F})$ defined node-wise as $L_{\mathcal{F}}(\mathbf{x})_v := \sum_{v, u \trianglelefteq e} \mathcal{F}^{\top}_{v \trianglelefteq e}(\mathcal{F}_{v \trianglelefteq e}\mathbf{x}_v - \mathcal{F}_{u \trianglelefteq e}\mathbf{x}_u)$.*

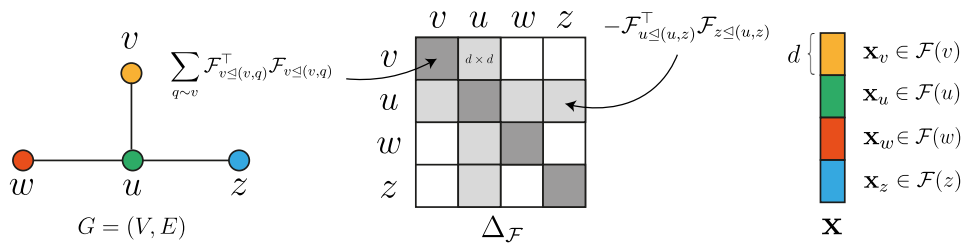


Figure 3: A graph *(left)*, the Laplacian matrix of a sheaf with d-dimensional stalks over the graph *(middle)* , and a 0-cochain $\mathbf{x}$ represented as a block-vector stacking the vectors of all nodes *(right)*.

The sheaf Laplacian is a positive semi-definite block matrix (Figure 3). The diagonal blocks are $L_{\mathcal{F}vv} = \sum_{v \trianglelefteq e} \mathcal{F}^\top_{v \trianglelefteq e} \mathcal{F}_{v \trianglelefteq e}$, while the non-diagonal blocks $L_{\mathcal{F}vu} = -\mathcal{F}^\top_{v \trianglelefteq e} \mathcal{F}_{u \trianglelefteq e}$. Denoting by D the block-diagonal of $L_{\mathcal{F}}$, the normalised sheaf Laplacian is given by $\Delta_{\mathcal{F}} := D^{-1/2} L_{\mathcal{F}} D^{-1/2}$. For simplicity, we assume that all the stalks have a fixed dimension d. In that case, the sheaf Laplacian is a $nd \times nd$ real matrix, where n is the number of nodes of G. When the vector spaces are set to $\mathbb{R}$ (i.e., $d = 1$) and the linear maps to the identity map over $\mathbb{R}$, the underlying sheaf is trivial and one recovers the well-known $n \times n$ graph Laplacian matrix and its normalised version Δ_0. In general, $\Delta_{\mathcal{F}}$ is preferred to $L_{\mathcal{F}}$ for most practical purposes due to its bounded spectrum and, therefore, we focus on the former. A cochain $\mathbf{x}$ is called *harmonic* if $L_{\mathcal{F}}\mathbf{x} = 0$ or, equivalently, if $\mathbf{x} \in \ker(L_{\mathcal{F}})$. This means harmonic cochains are characterised by zero disagreement along all the edges of the graph and it is not difficult to see that, in fact, $H^0(G; \mathcal{F})$ and $\ker(L_{\mathcal{F}})$ are isomorphic as vector spaces [33].

The sheaves with orthogonal maps (i.e. $\mathcal{F}_{v \trianglelefteq e} \in O(d)$ the Lie group of $d \times d$ orthogonal matrices) provide a more geometric interpretation of sheaves and play an important role in our analysis. Such sheaves are called *discrete $O(d)$ bundles* and can be seen as a discrete version of vector bundles [23, 58, 70] from differential geometry [64]. Intuitively, these are objects describing how vector spaces are attached to the points of a manifold. In our discrete case, the role of the manifold is played by the graph and the sheaf Laplacian describes how the elements of a vector space are transported via rotations in another neighbouring vector space similarly to how tangent vectors are moved across a manifold via parallel transport (connection; see Figure 2). Due to this analogy, the sheaf Laplacian on $O(d)$ *bundles* is also referred to as *connection Laplacian* [61].

Heat Diffusion and GCNs. Consider a graph with adjacency matrix $\mathbf{A}$, diagonal degree matrix $\mathbf{D}$, normalised graph Laplacian $\Delta_0 := \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, and an $n \times f$ feature matrix $\mathbf{X}$. We can define the heat diffusion equation and its Euler discretisation with a unit step as follows:

$$\dot{\mathbf{X}}(t) = -\Delta_0 \mathbf{X}(t) \rightsquigarrow \mathbf{X}(t+1) = \mathbf{X}(t) - \Delta_0 \mathbf{X}(t) = (\mathbf{I} - \Delta_0)\mathbf{X}(t). \tag{1}$$

Comparing this with the Graph Convolutional Network [37] model we observe that GCN is an augmented heat diffusion process with an additional $f \times f$ weight matrix $\mathbf{W}$ and a nonlinearity σ:

$$\mathrm{GCN}(\mathbf{X}, \mathbf{A}) := \sigma(\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\mathbf{X}\mathbf{W}) = \sigma((\mathbf{I} - \Delta_0)\mathbf{X}\mathbf{W}). \tag{2}$$

From this perspective, it is not surprising that GCN is particularly affected by heterophily and oversmoothing since heat diffusion makes the features of neighbouring nodes increasingly smooth. In what follows, we consider a much more general and powerful family of (sheaf) diffusion processes leading to more expressive sheaf convolutions.

3 The Expressive Power of Sheaf Diffusion

Preliminaries. Let us now assume G to be a graph with d-dimensional node feature vectors $\mathbf{x}_v \in \mathcal{F}(v)$. The features of all nodes are represented as a single vector $\mathbf{x} \in C^0(G; \mathcal{F})$ stacking all the individual d-dimensional vectors (Figure 3). Additionally, if we allow for f feature channels, everything can be represented as a matrix $\mathbf{X} \in \mathbb{R}^{(nd) \times f}$, whose columns are vectors in $C^0(G; \mathcal{F})$. We are interested in the spatially discretised *sheaf diffusion* process governed by the following PDE:

$$\mathbf{X}(0) = \mathbf{X}, \quad \dot{\mathbf{X}}(t) = -\Delta_{\mathcal{F}} \mathbf{X}(t). \tag{3}$$

It can be shown that in the time limit, each feature channel is projected into $\ker(\Delta_{\mathcal{F}})$ [32]. As described above (up to a $D^{-1/2}$ normalisation), this space contains the signals that agree with the restriction maps of the sheaf along all the edges. Thus, sheaf diffusion can be seen as a 'synchronisation' process over the graph, where all the private opinions converge towards global agreement.

In this section, we investigate the expressive power of this process within the infinite time limit. Because the asymptotic behaviour of sheaf diffusion is determined by the properties of $\ker(\Delta_{\mathcal{F}})$, in Section 3.1 we investigate when this subspace is non-trivial (i.e. it contains more than just the zero vector). In Section 3.2, we use this characterisation of the harmonic space to study what sort of sheaf diffusion processes will asymptotically produce projections into $\ker(\Delta_{\mathcal{F}})$ that can linearly separate the classes for various kinds of graphs and initial conditions. Since diffusion converges exponentially fast, the following results are also relevant for models with finite integration time or layers.

3.1 Harmonic Space of Sheaf Laplacians

A major role in the analysis below is played by discrete vector bundles and we concentrate on this case. We note though, that our results below generalise to the general linear group $\mathcal{F}_{v \trianglelefteq e} \in GL(d)$, the Lie group of $d \times d$ invertible matrices, provided we can also control the norm of the restriction maps from below. Given a discrete $O(d)$-bundle, $\mathcal{F}^{\top}_{v \trianglelefteq e}\mathcal{F}_{v \trianglelefteq e} = \mathbf{I}_d$ and the block diagonal of $L_{\mathcal{F}}$ has a diagonal structure since $L_{\mathcal{F}vv} = d_v \mathbf{I}_d$, where d_v is the degree of node v. Accordingly, if a signal $\tilde{\mathbf{x}} \in \ker(L_{\mathcal{F}})$, then the signal $\mathbf{x} : v \mapsto \sqrt{d_v}\tilde{\mathbf{x}}_v \in \ker(\Delta_{\mathcal{F}})$ and similarly for the inverse transformation.

Key to our analysis is studying *transport* operators induced by the restriction maps of the sheaf. Given nodes $v, u \in V$ and a path $\gamma_{v \to u} = (v, v_1, \ldots, v_\ell, u)$ from v to u, we consider a notion of *transport* from the stalk $\mathcal{F}(v)$ to the stalk $\mathcal{F}(u)$, constructed by composing restriction maps (and their transposes) along the edges:

$$\mathbf{P}^{\gamma}_{v \to u} := (\mathcal{F}^{\top}_{u \trianglelefteq e}\mathcal{F}_{v_\ell \trianglelefteq e}) \ldots (\mathcal{F}^{\top}_{v_1 \trianglelefteq e}\mathcal{F}_{v \trianglelefteq e}) : \mathcal{F}(v) \to \mathcal{F}(u).$$

For general sheaf structures, the graph transport is *path dependent*, meaning that how the vectors are transported across two nodes depends on the path between them (see Figure 2). In fact, we show that this property characterises the *spectral gap* of a sheaf Laplacian, i.e. the smallest eigenvalue of $\Delta_{\mathcal{F}}$.

Proposition 3. *If $\mathcal{F}$ is a discrete $O(d)$ bundle over a connected graph and $r := \max_{\gamma_{v \to u}, \gamma'_{v \to u}} ||\mathbf{P}^{\gamma}_{v \to u} - \mathbf{P}^{\gamma'}_{v \to u}||$, then we have $\lambda^{\mathcal{F}}_0 \leq r^2/2$.*

A consequence of this result is that there is always a non-trivial harmonic space (i.e. $\lambda^{\mathcal{F}}_0 = 0$) if the transport maps generated by an orthogonal sheaf are *path-independent* (i.e. $r = 0$). Next, we address the opposite direction.

Proposition 4. *If $\mathcal{F}$ is a discrete $O(d)$ bundle over a connected graph and $\mathbf{x} \in H^0(G, \mathcal{F})$, then for any cycle γ based at $v \in V$ we have $\mathbf{x}_v \in \ker(\mathbf{P}^{\gamma}_{v \to v} - \mathbf{I})$.*

This proposition highlights the interplay between the graph and the sheaf structure. A simple consequence of this result is that for any cycle-free subset $S \subset V$, we have that any sheaf (or connection-) Laplacian restricted to S always admits a non-trivial harmonic space. A natural question connected to the previous result is whether a Cheeger-like inequality holds in the other direction. This turns out to be the case:

Proposition 5. *Let $\mathcal{F}$ be a discrete $O(d)$ bundle over a connected graph G with n nodes and let $||(\mathbf{P}^{\gamma}_{v \to v} - \mathbf{I})\mathbf{x}_v|| \geq \epsilon ||\mathbf{x}_v||$ for all cycles $\gamma_{v \to v}$. Then $\lambda^{\mathcal{F}}_0 \geq \epsilon^2 (2\mathrm{diam}(G) n\, d_{max})^{-1}$.*

While the bound above is of little use in practice, it shows how the spectral gap of a sheaf Laplacian is indeed related to the deviation of the transport maps from being path-independent, as measured by ϵ. We note that the Cheeger-like inequality presented here is not unique and other types of bounds on $\lambda^{\mathcal{F}}_0$ have been derived [2]. We conclude this section by further analysing the dimensionality of the harmonic space of discrete $O(d)$-bundles:

Lemma 6. *Let $\mathcal{F}$ be a discrete $O(d)$ bundle over a connected graph G. Then $\dim(H^0) \leq d$ and $\dim(H^0) = d$ if and only if the transport is path-independent.*

3.2 The Linear Separation Power of Sheaf Diffusion

In what follows, we use the results above to analyse the ability of certain classes of sheaves to linearly separate the features in the limit of the diffusion processes they induce. We utilise this as a proxy for the capacity of certain diffusion processes to avoid oversmoothing.

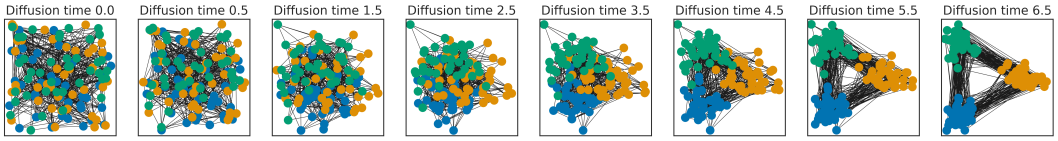


Figure 4: Diffusion process on $O(2)$-bundles progressively separates the classes of the graph.

Definition 7. *A hypothesis class of sheaves with d-dimensional stalks $\mathcal{H}^d$ has* linear separation power *over a family of graphs $\mathcal{G}$ if for any labelled graph $G = (V, E) \in \mathcal{G}$, there is a sheaf $(\mathcal{F}, G) \in \mathcal{H}^d$ that can linearly separate the classes of G in the time limit of Equation 3 for almost all initial conditions.*

Note that the restriction to almost all initial conditions is necessary because, in the limit, diffusion behaves like a projection in the harmonic space and there will always be degenerate initial conditions (e.g. the zero matrix) that will yield a zero projection. We will now show how the choice of the sheaf impacts the behaviour of the diffusion process. For this purpose, we will consider a hierarchy of increasingly general classes of sheaves.

Symmetric invertible. $\mathcal{H}^d_{\text{sym}} := \{(\mathcal{F}, G) : \mathcal{F}_{v \trianglelefteq e} = \mathcal{F}_{u \trianglelefteq e},\ \det(\mathcal{F}_{v \trianglelefteq e}) \neq 0\}$. We note that for $d = 1$, the sheaf Laplacians induced by this class of sheaves coincides with the set of the well-known weighted graph Laplacians with strictly positive weights, which also includes the usual graph Laplacian (see proof in Appendix B). Therefore, this hypothesis class is of particular interest since it includes those graph Laplacians typically used by graph convolutional models such as GCN [37] and ChebNet [19]. We first show that this class of sheaf Laplacians can linearly separate the classes in binary classification settings under certain homophily assumptions:

Proposition 8. *Let $\mathcal{G}$ be the set of connected graphs $G = (V, E)$ with two classes $A, B \subset V$ such that for each $v \in A$, there exists $u \in A$ and an edge $(v, u) \in E$. Then $\mathcal{H}^1_{\text{sym}}$ has linear separation power over $\mathcal{G}$.*

In contrast, under certain heterophilic conditions, this hypothesis class is not powerful enough to linearly separate the two classes no matter what the initial conditions are:

Proposition 9. *Let $\mathcal{G}$ be the set of connected bipartite graphs $G = (A, B, E)$, with partitions A, B forming two classes and $|A| = |B|$. Then $\mathcal{H}^1_{\text{sym}}$ cannot linearly separate any graph in $\mathcal{G}$ for any initial conditions $\mathbf{X}(0) \in \mathbb{R}^{n \times f}$.*

Non-symmetric invertible. $\mathcal{H}^d := \{(\mathcal{F}, G) : \det(\mathcal{F}_{v \trianglelefteq e}) \neq 0\}$. This larger hypothesis class addresses the above limitation by allowing non-symmetric relations:

Proposition 10. *Let $\mathcal{G}$ contain all the connected graphs $G = (V, E)$ with two classes $A, B \subseteq V$. Consider a sheaf $(\mathcal{F}; G) \in \mathcal{H}^1$ with $\mathcal{F}_{v \trianglelefteq e} = -\alpha_e$ if $v \in A$ and $\mathcal{F}_{u \trianglelefteq e} = \alpha_e$ if $u \in B$ with $\alpha_e > 0$ for all $e \in E$. Then the diffusion induced by $(\mathcal{F}; G)$ can linearly separate the classes of G for almost all initial conditions and $\mathcal{H}^1$ has linear separation power over $\mathcal{G}$.*

The type of sheaf above can also be interpreted as a discrete $O(1)$-bundle over a weighted graph with edge weights α_e^2 and transport maps $\mathcal{F}^\top_{v \trianglelefteq e} \mathcal{F}_{u \trianglelefteq e} = -1$ for the inter-class edges and $+1$ for the intra-class edges. In agreement with the results in Section 3.1, it can be checked that this transport is path-independent. Importantly, this result provides a sheaf-theoretic explanation for why a recent body of work [6, 16, 69] has found negatively-weighted edges to help in heterophilic settings.

So far we have only studied the effects of changing the type of sheaves in dimension one. We now consider the effects of adjusting the dimension of the stalks and begin by stating a fundamental limitation of (sheaf) diffusion when $d = 1$.

Proposition 11. *Let G be a connected graph with $C \geq 3$ classes. Then, $\mathcal{H}^1$ cannot linearly separate the classes of G for any initial conditions $\mathbf{X}(0) \in \mathbb{R}^{n \times f}$.*

This is essentially a consequence of $\dim\big(\ker(\Delta_{\mathcal{F}})\big) \leq 1$ in this case, by virtue of Lemma 6. From a GNN perspective, this means that in the infinite depth setting, sufficient *stalk width* (i.e., dimension d) is needed in order to solve tasks involving more than two classes. Note that d is different from the classical notion of feature channels f. As the result above shows, the latter has no effect on the linear separability of the classes in $d = 1$. Next, we will see that the former does.

Diagonal invertible. $\mathcal{H}^d := \{(\mathcal{F}, G) : \text{diagonal } \mathcal{F}_{v \trianglelefteq e}, \det(\mathcal{F}_{v \trianglelefteq e}) \neq 0\}$. The sheaves in this class can be seen as d independent sheaves from $\mathcal{H}^1$ encoded in the d-dimensional diagonals of their restriction maps. This perspective allows us to generalise Proposition 10 to a multi-class setting:

Proposition 12. *Let $\mathcal{G}$ be the set of connected graphs with nodes belonging to $C \geq 3$ classes. Then for $d \geq C$, $\mathcal{H}^d_{\text{diag}}$ has linear separation power over $\mathcal{G}$.*

This result illustrates the benefits of using higher-dimensional stalks while maintaining a simple and computationally convenient class of diagonal restriction maps. Next, with more complex restriction maps, we can show that lower-dimensional stalks can be used to achieve linear separation in the presence of even more classes.

Orthogonal. $\mathcal{H}^d_{\text{orth}} := \{(\mathcal{F}, G) : \mathcal{F}_{v \trianglelefteq e} \in O(d)\}$ is the class of $O(d)$-bundles. Orthogonal maps are able to make more efficient use of the space available to them than diagonal restriction maps:

Proposition 13. *Let $\mathcal{G}$ be the class of connected graphs with $C \leq 2d$ classes. Then, for all $d \in \{2, 4\}$, $\mathcal{H}^d_{\text{orth}}$ has linear separation power over $\mathcal{G}$.*

Figure 4 includes an example diffusion process over an $O(2)$-bundle.

Summary: Different sheaf classes give rise to different behaviours of the diffusion process and consequently, to different separation capabilities. Taken together, these results show that solving any node classification task can be reduced to performing diffusion with the right sheaf.

4 Expressive Power of Sheaf Convolutions

Analogously to how GCN augments heat diffusion, we can construct a **Sheaf Convolutional Network (SCN)** augmenting the sheaf diffusion process. In this section, we analyse the capacity of SCNs to change, *if necessary*, their asymptotic behaviour compared to the base diffusion process. Since the sheaf structure will be ultimately learned from data, this is particularly important for the common setting when the learned sheaf is different from the "ground truth" sheaf for the task to be solved.

The continous diffusion process from Equation 3 has the Euler discretisation with unit step-size $\mathbf{X}(t+1) = \mathbf{X}(t) - \Delta_{\mathcal{F}}\mathbf{X}(t) = (\mathbf{I}_{nd} - \Delta_{\mathcal{F}})\mathbf{X}(t)$. Assuming $\mathbf{X} \in \mathbb{R}^{nd \times f_1}$, we can equip the right side with weight matrices $\mathbf{W}_1 \in \mathbb{R}^{d \times d}$, $\mathbf{W}_2 \in \mathbb{R}^{f_1 \times f_2}$ and a non-linearity σ to arrive at the following model originally proposed by Hansen and Gebhart [30]:

$$\mathbf{Y} = \sigma\Big(\big(\mathbf{I}_{nd} - \Delta_{\mathcal{F}}\big)(\mathbf{I}_n \otimes \mathbf{W}_1)\mathbf{X}\mathbf{W}_2\Big) \in \mathbb{R}^{nd \times f_2}, \tag{4}$$

where f_1, f_2 are the number of input and output feature channels, and $\otimes$ denotes the Kronecker product. Here, $\mathbf{W}_1$ multiplies from the left the vector feature of all the nodes in all channels (i.e. $\mathbf{W}_1\mathbf{x}_v^i$ for all v and channels i), while $\mathbf{W}_2$ multiplies the features from the right and can adjust the number of feature channels, just like in GCNs. As one would expect, when using a trivial sheaf, $\Delta_{\mathcal{F}} = \Delta_0$, $\mathbf{W}_1$ becomes a scalar and one recovers the GCN of Kipf and Welling [37]. To see how SCNs behave compared to their base diffusion process, we investigate how SCN layers affect the *sheaf Dirichlet energy* $E_{\mathcal{F}}(\mathbf{x})$, which sheaf diffusion is known to minimise over time.

Definition 14. $E_{\mathcal{F}}(\mathbf{x}) := \mathbf{x}^\top \Delta_{\mathcal{F}} \mathbf{x} = \frac{1}{2}\sum_{e:=(v,u)} \|\mathcal{F}_{v \trianglelefteq e} D_v^{-1/2}\mathbf{x}_v - \mathcal{F}_{u \trianglelefteq e} D_u^{-1/2}\mathbf{x}_u\|_2^2$

Similarly, for multiple channels the energy is $E_{\mathcal{F}}(\mathbf{X}) := \text{trace}(\mathbf{X}^\top \Delta_{\mathcal{F}} \mathbf{X})$. This is a measure of how close a signal $\mathbf{x}$ is to $\ker(\Delta_{\mathcal{F}})$ and it is easy to see that $\mathbf{x} \in \ker(\Delta_{\mathcal{F}}) \Leftrightarrow E_{\mathcal{F}}(\mathbf{x}) = 0$. We begin by studying the sheaves for which the energy decreases and representations end up asymptotically in $\ker(\Delta_{\mathcal{F}})$. Let $\lambda_* := \max_{i>0}(\lambda_i^{\mathcal{F}} - 1)^2 \leq 1$ and denote by $\mathcal{H}^1_+ := \{(\mathcal{F}, G) \mid \mathcal{F}_{v \trianglelefteq e}\mathcal{F}_{u \trianglelefteq e} > 0\}$.

Theorem 15. *For $(\mathcal{F}, G) \in \mathcal{H}^1_+$ and σ being (Leaky)ReLU, $E_{\mathcal{F}}(\mathbf{Y}) \leq \lambda_* \|\mathbf{W}_1\|_2^2 \|\mathbf{W}_2^\top\|_2^2 E_{\mathcal{F}}(\mathbf{X})$.*

This generalises existent results for GCNs [12, 49] and proves that SCNs using this family of Laplacians, which includes all weighted graph Laplacians, exponentially converge to $\ker(\Delta_{\mathcal{F}})$ if $\lambda_* \|\mathbf{W}_1\|_2^2 \|\mathbf{W}_2^\top\|_2^2 < 1$. In particular, if $E_{\mathcal{F}}(\mathbf{X}) = 0$, then $E_{\mathcal{F}}(\mathbf{Y}) = 0$ and the representations remain trapped inside the kernel no matter what the norm of the weights is. Therefore, in settings as

those described by Propositions 9 and 11, the linear separation capabilities of this class of models are severely limited (see Corollaries 36, 37 in Appendix B).

Finally, the Theorem also extends to bundles with symmetric maps, $\mathcal{H}^d_{\text{orth,sym}} := \mathcal{H}^d_{\text{orth}} \cap \mathcal{H}^d_{\text{sym}}$:

Theorem 16. *If $(\mathcal{F}, G) \in \mathcal{H}^d_{\text{orth,sym}}$ and σ = (Leaky)ReLU, $E_{\mathcal{F}}(\mathbf{Y}) \leq \lambda_* \|\mathbf{W}_1\|_2^2 \|\mathbf{W}_2^\top\|_2^2 E_{\mathcal{F}}(\mathbf{X})$.*

In some sense, this is not surprising because, for this class, $\ker(\Delta_{\mathcal{F}})$ contain the same information as the kernel of the classical normalised graph Laplacian (see Proposition 29 in Appendix C).

More generally, SCNs with sheaves outside $\mathcal{H}^d_{\text{sym}}$, are much more flexible and can easily increase the Dirichlet energy using an arbitrarily small linear transformation $\mathbf{W}_1$:

Proposition 17. *For any connected graph G and $\varepsilon > 0$, there exist a sheaf $(G, \mathcal{F}) \notin \mathcal{H}^d_{\text{sym}}$, $\mathbf{W}_1$ with $\|\mathbf{W}_1\|_2 < \varepsilon$ and feature vector $\mathbf{x}$ such that $E_{\mathcal{F}}((\mathbf{I} \otimes \mathbf{W}_1)\mathbf{x}) > E_{\mathcal{F}}(\mathbf{x})$.*

Importantly, this proves that this family of SCNs can, if necessary, escape the kernel of the Laplacian.

Summary: Not only that sheaf diffusion is more expressive than heat diffusion as shown in Section 3.2, but SCNs are also more expressive than GCNs in the sense that they are generally not constrained to decrease the Dirichlet energy when using low-norm weights. This provides them with greater control than GCNs over their asymptotic behaviour.

5 Neural Sheaf Diffusion and Sheaf Learning

In the previous sections, we discussed the various advantages provided by sheaf diffusion and sheaf convolutions. However, in general, the ground truth sheaf is unknown or unspecified. Therefore, we aim to learn the underlying sheaf from data end-to-end, thus allowing the model to pick the right geometry for solving the task.

Neural Sheaf Diffusion. We propose the diffusion-type model from Equation 5. We note that by setting $\mathbf{W}_1, \mathbf{W}_2$ to identity and $\sigma(\mathbf{x}) = \text{ELU}(\epsilon\mathbf{x})/\epsilon$ with $\epsilon > 0$ small enough or simply $\sigma = \text{id}$, we recover (up to a scaling) the sheaf diffusion equation. Therefore, the model is at least as expressive as sheaf diffusion and benefits from all the positive properties outlined in Section 3.2.

$$\dot{\mathbf{X}}(t) = -\sigma\Big(\Delta_{\mathcal{F}(t)}(\mathbf{I}_n \otimes \mathbf{W}_1)\mathbf{X}(t)\mathbf{W}_2\Big), \tag{5}$$

Crucially, the sheaf Laplacian $\Delta_{\mathcal{F}(t)}$ is that of a sheaf $(G, \mathcal{F}(t))$ that *evolves over time*. More specifically, the evolution of the sheaf structure is described by a learnable function of the data $(G, \mathcal{F}(t)) = g(G, \mathbf{X}(t); \theta)$. This allows the model to make use of the latest available features to manipulate the underlying geometry of the graph and implicitly the behaviour of the diffusion process. Additionally, We use an MLP followed by a reshaping to map the raw features of the dataset to a matrix $\mathbf{X}(0)$ of shape $nd \times f$ and a final linear layer to perform the node classification.

In our experiments, we focus on the time-discretised version of this model from Equation 6, which allows us to use a new set of weights at each layer t while maintaining the nice theoretical properties of the model above.

$$\mathbf{X}_{t+1} = \mathbf{X}_t - \sigma\Big(\Delta_{\mathcal{F}(t)}(\mathbf{I} \otimes \mathbf{W}_1^t)\mathbf{X}_t\mathbf{W}_2^t\Big) \tag{6}$$

We note that this model is different from the SCN model from Equation 4 in two major ways. First, Hansen and Gebhart [30] used a *hand-crafted* sheaf with $d = 1$, constructed in a synthetic setting with full knowledge of the data generating process. In contrast, we *learn* a sheaf, which makes our model applicable to any real-world graph dataset, even in the absence of a sheaf structure. Additionally, motivated by our theoretical results, we make use of the full generality of sheaves by using stalks with $d \geq 1$ and higher-dimensional maps. Second, our model uses a residual parametrisation of the discretised diffusion process, which empirically improves its performance.

Sheaf Learning. The restriction maps are learned using *locally* available information. Each $d \times d$ matrix $\mathcal{F}_{v \trianglelefteq e}$ is learned via a parametric matrix-valued function Φ, with $\mathcal{F}_{v \trianglelefteq e := (v,u)} = \Phi(\mathbf{x}_v, \mathbf{x}_u)$. This function must be non-symmetric to be able to learn asymmetric transport maps along each edge. In practice, we set $\Phi(\mathbf{x}_v, \mathbf{x}_u) = \sigma(\widetilde{\mathbf{W}}[\mathbf{x}_v || \mathbf{x}_u])$ followed by a reshaping of the output, where $\widetilde{\mathbf{W}}$ is a weight matrix. For simplicity, the equations above use a single feature channel, but in practice,

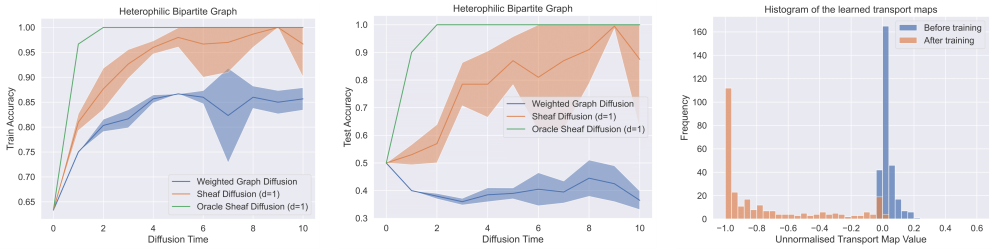


Figure 5: (*Left*) Train and (*Middle*) test accuracy as a function of diffusion time. (*Right*) Histogram of the learned scalar transport maps. The performance of the sheaf diffusion model is superior to that of weighted-graph diffusion and correctly learns to invert the features of the two classes.

all channels are supplied as input. More generally, we can show that if the function Φ has enough capacity and the features are diverse enough, we can learn any sheaf over a graph.

Proposition 18. *Let $G = (V, E)$ be a finite graph with features $\mathbf{X}$. Then, if $(\mathbf{x}_v, \mathbf{x}_u) \neq (\mathbf{x}_w, \mathbf{x}_z)$ for any $(v, u) \neq (w, z) \in E$ and Φ is an MLP with sufficient capacity, Φ can learn any sheaf $(\mathcal{F}; G)$.*

First, this result formally motivates learning a sheaf at each layer since the model can learn to distinguish more nodes after each aggregation step. Second, this suggests that more expressive models (in the Weisfeiler-Lehman sense [7, 8, 44, 68]) could learn a more general family of sheaves. We leave a deeper investigation of these aspects for future work. In what follows, we distinguish between several types of functions Φ depending on the type of matrix they learn.

Diagonal. The main advantage of this parametrisation is that fewer parameters need to be learned per edge and the sheaf Laplacian ends up being a matrix with diagonal blocks, which also results in fewer operations in sparse matrix multiplications. The main disadvantage is that the d dimensions of the stalks interact only via the left $\mathbf{W}_1$ multiplication.

Orthogonal. In this case, the model effectively learns a discrete vector bundle. Orthogonal matrices provide several advantages: (1) they can mix the various dimension of the stalks, (2) the orthogonality constraint prevents overfitting while reducing the number of parameters, (3) they have better understood theoretical properties, and (4) the resulting Laplacians are easier to normalise numerically since the diagonal entries correspond to the degrees of the nodes. In our model, we build orthogonal matrices from a composition of Householder reflections [43].

General. Finally, we consider the most general option of learning arbitrary matrices. The maximal flexibility provided by these maps can be useful, but it also comes with the danger of overfitting. At the same time, the sheaf Laplacian is more challenging to normalise numerically since one has to compute $D^{-1/2}$ for a positive semi-definite matrix D. To perform this at scale, one has to rely on SVD, whose gradients can be infinite if D has repeated eigenvalues. Therefore, this model is more challenging to train.

Computational Complexity. The GCN from Equation 2 has complexity $\mathcal{O}(nc^2 + mc)$, where c is the number of channels and m the number of edges. Assume a sheaf diffusion model with stalk dimension d and f channels such that $d \times f = c$ (i.e. same representation size). Then, when the model uses diagonal maps, the complexity is $\mathcal{O}(nc^2 + mdc)$. When using orthogonal or general matrices, the complexity becomes $\mathcal{O}(n(c^2 + d^3) + m(cd^2 + d^3))$ (see Appendix E.1 for detailed derivations). In practice, we use $1 \leq d \leq 5$, which effectively results in a constant overhead compared to GCN.

6 Experiments

Synthetic experiments. We consider a simple setup given by a connected bipartite graph with equally sized partitions. We sample the features from two overlapping isotropic Gaussian distributions to make the classes linearly non-separable at initialisation time. From Proposition 9 we know that diffusion models using symmetric restriction maps cannot separate the classes in the limit, while a diffusion process using negative transport maps can. Therefore, we use two vanilla sheaf diffusion process by setting $d = 1$, $\mathbf{W}_1 = \mathbf{I}_d$, $\mathbf{W}_2 = \mathbf{I}_f$ and $\sigma = \mathrm{id}$ in Equation 5. In both models, we learn a sheaf at $t = 0$ as a function of $\mathbf{X}(0)$, and we keep the sheaf constant over time. For the first model, we learn a sheaf with general maps $\mathcal{F}_{v \trianglelefteq e} \in \mathbb{R}$. For the second model, we use a similar layer but constraint $\mathcal{F}_{v \trianglelefteq e} = \mathcal{F}_{u \trianglelefteq e}$, obtaining a weighted graph Laplacian.

Table 1: Results on node classification datasets sorted by their homophily level. Top three models are coloured by **First**, **Second**, **Third**. Our models are marked **NSD**.

	Texas	Wisconsin	Film	Squirrel	Chameleon	Cornell	Citeseer	Pubmed	Cora
Hom level	**0.11**	**0.21**	**0.22**	**0.22**	**0.23**	**0.30**	**0.74**	**0.80**	**0.81**
#Nodes	183	251	7,600	5,201	2,277	183	3,327	18,717	2,708
#Edges	295	466	26,752	198,493	31,421	280	4,676	44,327	5,278
#Classes	5	5	5	5	5	5	7	3	6
Diag-NSD	**85.67**±6.95	**88.63**±2.75	**37.79**±1.01	**54.78**±1.81	**68.68**±1.73	**86.49**±7.35	**77.14**±1.85	89.42±0.43	87.14±1.06
O(d)-NSD	**85.95**±5.51	**89.41**±4.74	**37.81**±1.15	**56.34**±1.32	**68.04**±1.58	**84.86**±4.71	76.70±1.57	**89.49**±0.40	86.90±1.13
Gen-NSD	82.97±5.13	**89.21**±3.84	**37.80**±1.22	53.17±1.31	67.93±1.58	**85.68**±6.51	76.32±1.65	89.33±0.35	87.30±1.15
GGCN	**84.86**±4.55	86.86±3.29	37.54±1.56	**55.17**±1.58	**71.14**±1.84	**85.68**±6.63	**77.14**±1.45	89.15±0.37	**87.95**±1.05
H2GCN	**84.86**±7.23	87.65±4.98	35.70±1.00	36.48±1.86	60.11±2.15	82.70±5.28	77.11±1.57	**89.49**±0.38	**87.87**±1.20
GPRGNN	78.38±4.36	82.94±4.21	34.63±1.22	31.61±1.24	46.58±1.71	80.27±8.11	77.13±1.67	87.54±0.38	**87.95**±1.18
FAGCN	82.43±6.89	82.94±7.95	34.87±1.25	42.59±0.79	55.22±3.19	79.19±9.79	N/A	N/A	N/A
MixHop	77.84±7.73	75.88±4.90	32.22±2.34	43.80±1.48	60.50±2.53	73.51±6.34	76.26±1.33	85.31±0.61	87.61±0.85
GCNII	77.57±3.83	80.39±3.40	37.44±1.30	38.47±1.58	63.86±3.04	77.86±3.79	**77.33**±1.48	**90.15**±0.43	**88.37**±1.25
Geom-GCN	66.76±2.72	64.51±3.66	31.59±1.15	38.15±0.92	60.00±2.81	60.54±3.67	**78.02**±1.15	**89.95**±0.47	85.35±1.57
PairNorm	60.27±4.34	48.43±6.14	27.40±1.24	50.44±2.04	62.74±2.82	58.92±3.15	73.59±1.47	87.53±0.44	85.79±1.01
GraphSAGE	82.43±6.14	81.18±5.56	34.23±0.99	41.61±0.74	58.73±1.68	75.95±5.01	76.04±1.30	88.45±0.50	86.90±1.04
GCN	55.14±5.16	51.76±3.06	27.32±1.10	53.43±2.01	64.82±2.24	60.54±5.30	76.50±1.36	88.42±0.50	86.98±1.27
GAT	52.16±6.63	49.41±4.09	27.44±0.89	40.72±1.55	60.26±2.50	61.89±5.05	76.55±1.23	87.30±1.10	86.33±0.48
MLP	80.81±4.75	85.29±3.31	36.53±0.70	28.77±1.56	46.21±2.99	81.89±6.40	74.02±1.90	87.16±0.37	75.69±2.00

Figure 5 presents the results across five seeds. As expected, for diffusion time zero (i.e. no diffusion), we see that a linear classifier cannot separate the classes. At later times, the diffusion process using symmetric maps cannot perfectly fit the data. In contrast, with the more general sheaf diffusion, as time increases and the signal approaches the harmonic space, the model gets better and the features become linearly separable. In the last subfigure, we take a closer look at the sheaf that the model learns in the time limit by plotting a histogram of all the transport (scalar) maps $\mathcal{F}_{v \trianglelefteq e}^{\top} \mathcal{F}_{u \trianglelefteq e}$. In accordance with Proposition 10, the model learns a negative transport map for all edges. This shows that the model manages to avoid oversmoothing (see Appendix F for an experiment with $d > 1$).

Real-world experiments. We test our models on multiple real-world datasets [45, 51, 55, 59, 63] with an edge homophily coefficient h ranging from $h = 0.11$ (very heterophilic) to $h = 0.81$ (very homophilic). Therefore, they offer a view of how a model performs over this entire spectrum. We evaluate our models on the 10 fixed splits provided by Pei et al. [51] and report the mean accuracy and standard deviation. Each split contains $48\%/32\%/20\%$ of nodes per class for training, validation and testing, respectively. As baselines, we use an ample set of GNN models that can be placed in three categories: (1) classical: GCN [37], GAT [65], GraphSAGE [29]; (2) models specifically designed for heterophilic settings: GGCN [69], Geom-GCN [51], H2GCN [72], GPRGNN [16], FAGCN [6], MixHop [1]; (3) models addressing oversmoothing: GCNII [15], PairNorm [71]. All the results are taken from Yan et al. [69], except for FAGCN and MixHop, which come from Lingam et al. [39] and Zhu et al. [72], respectively. All of these were evaluated on the same set of splits as ours. In Appendix F we also include experiments with continuous GNN models.

Results. From Table 1 we see that our models are first in $5/6$ benchmarks with high heterophily ($h < 0.3$) and second-ranked on the remaining one (i.e. Chameleon). At the same time, NSD also shows strong performance on the homophilic graphs by being within approximately 1% of the top model. Overall, NSD models are among the top three models on $8/9$ datasets. The $O(d)$-bundle diffusion model performs best overall confirming the intuition that it can better avoid overfitting, while also transforming the vectors in sufficiently complex ways. We also remark on the strong performance of the model learning diagonals maps, despite the simpler functional form of the Laplacian.

7 Related Work, Discussion, and Conclusion

Sheaf Neural Networks & Sheaf Learning. Sheaf Neural Networks [30] with a *hand-crafted* sheaf Laplacian were originally introduced in a toy experimental setting. Since then, they have remained completely unexplored and we hope that this paper will fill this lacuna. In contrast to [30], we provide an ample theoretical analysis justifying the use of sheaves in Graph ML, study for the first time how sheaves can be *learned from data* using neural networks, and present the first successful application of Sheaf Neural Networks on real-world datasets. Hansen and Ghrist [31] have also considered learning a sheaf Laplacian by framing the problem as the minimization in matrix space of a regularised Dirichlet energy metric. Different from their approach, we learn the sheaf as part of an end-to-end model and using an efficient parametrisation that is independent of the size of the

graph. Inspired by Singer and Wu [61], Barbero et al. [3] have also experimented with inferring a connection Laplacian directly from data at pre-processing time, which yielded promising results. A similar technique was also used to construct a higher-order sheaf Laplacian whose spectrum encodes information about the underlying symmetries of the data [53]. This latter work hints at the powerful data properties that Sheaf Neural Networks could potentially extract from these operators.

Heterophily and Oversmoothing. While good empirical designs jointly addressing these two problems have been proposed before [16, 69], Yan et al. [69] is the only other work connecting the two theoretically. Their analysis [69] is very different in terms of methods and assumptions and, therefore, their results are completely orthogonal. Concretely, the authors analyse the performance of linear SGCs [66] (i.e. GCN without nonlinearities) on random attributed graphs. In contrast, our analysis is not probabilistic, focuses on diffusion PDEs and also extends to GCNs in the non-linear regime. Furthermore, we employ a new set of mathematical tools from cellular sheaf theory, which brings a new language and new tools to analyse these problems. Perhaps the only commonality is that both works find evidence for the benefits of negatively signed edges in GNNs, although with different mathematical motivations. At the same time, other recent works [20, 40] have shown that GCNs with finite layers (typically one) can perform well in heterophilic graphs (including bipartite). This is in no contradiction with our results, which consider an *infinite time/layer* regime (i.e. not finite) and *perfect* linear separation (i.e. a model that cannot fit the data can still achieve high accuracy).

Category Theory and GNNs. From the perspective of category theory [41], cellular sheaves are a *functor* from a *category* describing the incidence structure of the graph to a *category* describing the data living on top of the graph. Informally, this says that the vertices and edges are mapped to some type of data (e.g. vector spaces) and the incidence relations between vertices and edges are mapped to some type of relation between the assigned data (e.g. linear maps between the vector spaces). The generality provided by this perspective could be used to extend the models described in this work to more exotic types of data such as lattices and their associated sheaf Laplacians [24]. At the same time, our work echoes other recent efforts to place GNNs on a categorical foundation [18, 21].

Message Passing Neural Networks. From a message passing perspective [26], the layer from Equation 6 can be seen as a form of GNN-FiLM layer [10, 52]. Each node learns a linear transformation conditioned on the features of the neighbours that is then used to process the neighbouring features. Models of this kind have been recently shown to perform empirically well in heterophilic settings [50]. At the same time, the model bares a superficial resemblance to GAT [65]. For a central node v and a neighbouring node u, GAT learns an attention coefficient a_{vu} as a function of the neighbouring features. In contrast, our model learns a matrix given by the block (v, u) of $\Delta_{\mathcal{F}}$.

Limitations and societal impact. One of the main limitations of our theoretical analysis is that it does not address the generalisation properties of sheaves, but this remains a major impediment for the entire field of deep learning. Nonetheless, our setting was sufficient to produce many valuable insights about heterophily and oversmoothing and a basic understanding of what various types of sheaves can and cannot do. Much more work remains to be done in this direction, and we expect to see further cross-fertilization between ML and algebraic topology in the future. Finally, due to the theoretical nature of this work, we do not foresee any immediate negative societal impacts.

Conclusion. In this work, we used cellular sheaf theory to provide a novel topological perspective on heterophily and oversmoothing in GNNs. We showed that the underlying sheaf structure of the graph is intimately connected with both of these important factors affecting the performance of GNNs. To mitigate this, we proposed a new paradigm for graph representation learning where models not only evolve the features at each layer but also the underlying geometry of the graph. In practice, we demonstrated that this framework achieves competitive results in heterophilic settings.

Acknowledgments and Disclosure of Funding

We are grateful to Iulia Duta, Dobrik Georgiev and Jacob Deasy for valuable comments on an earlier version of this manuscript. CB would also like to thank the Twitter Cortex team for making the research internship a fantastic experience. This research was supported in part by ERC Consolidator grant No. 724228 (LEMAN).

References

[1] Sami Abu-El-Haija, Bryan Perozzi, Amol Kapoor, Hrayr Harutyunyan, Nazanin Alipourfard, Kristina Lerman, Greg Ver Steeg, and Aram Galstyan. Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing. In *The Thirty-sixth International Conference on Machine Learning (ICML)*, 2019. URL http://proceedings.mlr.press/v97/abu-el-haija19a/abu-el-haija19a.pdf.

[2] Afonso S Bandeira, Amit Singer, and Daniel A Spielman. A Cheeger inequality for the graph connection laplacian. *SIAM Journal on Matrix Analysis and Applications*, 34(4):1611–1630, 2013.

[3] Federico Barbero, Cristian Bodnar, Haitz Sáez de Ocáriz Borde, Michael Bronstein, Petar Veličković, and Pietro Liò. Sheaf neural networks with connection laplacians. In *ICML 2022 Workshop on Topology, Algebra, and Geometry in Machine Learning*, 2022.

[4] Lukas Biewald. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

[5] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

[6] Deyu Bo, Xiao Wang, Chuan Shi, and Huawei Shen. Beyond low-frequency information in graph convolutional networks. In *AAAI*. AAAI Press, 2021.

[7] Cristian Bodnar, Fabrizio Frasca, Nina Otter, Yu Guang Wang, Pietro Liò, Guido Montúfar, and Michael Bronstein. Weisfeiler and Lehman Go Cellular: CW Networks. In *NeurIPS*, 2021.

[8] Cristian Bodnar, Fabrizio Frasca, Yuguang Wang, Nina Otter, Guido F Montufar, Pietro Lió, and Michael Bronstein. Weisfeiler and Lehman Go Topological: Message Passing Simplicial Networks. In *ICML*, 2021.

[9] Glen E Bredon. *Sheaf theory*, volume 170. Springer Science & Business Media, 2012.

[10] Marc Brockschmidt. Gnn-film: Graph neural networks with feature-wise linear modulation. In *International Conference on Machine Learning*, pages 1144–1152. PMLR, 2020.

[11] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. In *ICLR*, 2014.

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

[13] Benjamin Paul Chamberlain, James Rowbottom, Davide Eynard, Francesco Di Giovanni, Dong Xiaowen, and Michael M Bronstein. Beltrami flow and neural diffusion on graphs. In *NeurIPS*, 2021.

[14] Benjamin Paul Chamberlain, James Rowbottom, Maria Goronova, Stefan Webb, Emanuele Rossi, and Michael M Bronstein. Grand: Graph neural diffusion. In *ICML*, 2021.

[15] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1725–1735. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20v.html.

[16] Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized pagerank graph neural network. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=n6jl7fLxrP.

[17] Justin Michael Curry. *Sheaves, cosheaves and applications*. University of Pennsylvania, 2014.

[18] Pim de Haan, Taco Cohen, and Max Welling. Natural graph networks. In *NeurIPS*, 2020.

[19] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *NIPS*, 2016.

[20] Lun Du, Xiaozhou Shi, Qiang Fu, Xiaojun Ma, Hengyu Liu, Shi Han, and Dongmei Zhang. Gbk-gnn: Gated bi-kernel graph neural networks for modeling both homophily and heterophily. In *Proceedings of the ACM Web Conference 2022*, pages 1550–1558, 2022.

[21] Andrew Dudzik and Petar Veličković. Graph neural networks are dynamic programmers. *arXiv preprint arXiv:2203.15544*, 2022.

[22] Vijay Prakash Dwivedi, Chaitanya K Joshi, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking graph neural networks. *arXiv:2003.00982*, 2020.

[23] Tingran Gao, Jacek Brodzki, and Sayan Mukherjee. The geometry of synchronization problems and learning group actions. *Discrete & Computational Geometry*, 65(1):150–211, 2021.

[24] Robert Ghrist and Hans Riess. Cellular sheaves of lattices and the tarski laplacian. *arXiv:2007.04099*, 2020.

[25] Robert W Ghrist. *Elementary applied topology*, volume 1. Createspace Seattle, 2014.

[26] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *ICML*, 2017.

[27] Christoph Goller and Andreas Kuchler. Learning task-dependent distributed representations by backpropagation through structure. In *ICNN*, 1996.

[28] Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. In *IJCNN*, 2005.

[29] William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *IEEE Data Engineering Bulletin*, 2017.

[30] Jakob Hansen and Thomas Gebhart. Sheaf neural networks. In *NeurIPS 2020 Workshop on Topological Data Analysis and Beyond*, 2020.

[31] Jakob Hansen and Robert Ghrist. Learning sheaf laplacians from smooth signals. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5446–5450. IEEE, 2019.

[32] Jakob Hansen and Robert Ghrist. Toward a spectral theory of cellular sheaves. *Journal of Applied and Computational Topology*, 3(4):315–358, 2019.

[33] Jakob Hansen and Robert Ghrist. Opinion dynamics on discourse sheaves. *SIAM Journal on Applied Mathematics*, 81(5):2033–2060, 2021.

[34] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

[35] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[36] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[37] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[38] Hyun Deok Lee. On some matrix inequalities. *Korean Journal of Mathematics*, 16(4):565–571, 2008.

[39] Vijay Lingam, Rahul Ragesh, Arun Iyer, and Sundararajan Sellamanickam. Simple truncated svd based model for node classification on heterophilic graphs. *arXiv preprint arXiv:2106.12807*, 2021.

[40] Sitao Luan, Chenqing Hua, Qincheng Lu, Jiaqi Zhu, Mingde Zhao, Shuyuan Zhang, Xiao-Wen Chang, and Doina Precup. Is heterophily a real nightmare for graph neural networks to do node classification? *arXiv preprint arXiv:2109.05641*, 2021.

[41] Saunders Mac Lane. *Categories for the working mathematician*, volume 5. Springer Science & Business Media, 2013.

[42] Saunders MacLane and Ieke Moerdijk. *Sheaves in geometry and logic: A first introduction to topos theory*. Springer Science & Business Media, 2012.

[43] Zakaria Mhammedi, Andrew Hellicar, Ashfaqur Rahman, and James Bailey. Efficient orthogonal parametrisation of recurrent neural networks using householder reflections. In *International Conference on Machine Learning*, pages 2401–2409. PMLR, 2017.

[44] Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and Leman go neural: Higher-order graph neural networks. In *AAAI*, 2019.

[45] Galileo Namata, Ben London, Lise Getoor, Bert Huang, and U Edu. Query-driven active surveying for collective classification. In *10th International Workshop on Mining and Learning with Graphs*, volume 8, page 1, 2012.

[46] H Nt and T Maehara. Revisiting graph neural networks: all we have is low pass filters. *arXiv preprint arXiv:1812.08434v4*, 2019.

[47] Anton Obukhov. Efficient householder transformation in pytorch, 2021. URL `www.github.com/toshas/torch-householder`. Version: 1.0.1, DOI: 10.5281/zenodo.5068733.

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

[49] Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. *arXiv:1905.10947*, 2019.

[50] John Palowitch, Anton Tsitsulin, Brandon Mayer, and Bryan Perozzi. Graphworld: Fake graphs bring real insights for gnns. *arXiv preprint arXiv:2203.00112*, 2022.

[51] Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-gcn: Geometric graph convolutional networks. *arXiv preprint arXiv:2002.05287*, 2020.

[52] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[53] David Pfau, Irina Higgins, Alex Botev, and Sébastien Racanière. Disentangling by subspace diffusion. *Advances in Neural Information Processing Systems*, 33:17403–17415, 2020.

[54] Daniel Rosiak. *Sheaf theory through examples*. MIT Press, 2022.

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

[56] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Trans. Neural Networks*, 20(1):61–80, 2008.

[57] Richard D Schafer. *An introduction to nonassociative algebras*. Courier Dover Publications, 2017.

[58] Luis Scoccola and Jose A Perea. Approximate and discrete euclidean vector bundles. *arXiv preprint arXiv:2104.07563*, 2021.

[59] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–93, 2008.

[60] Allen Dudley Shepard. *A cellular description of the derived category of a stratified space*. PhD thesis, Brown University, 1985.

[61] Amit Singer and H-T Wu. Vector diffusion maps and the connection laplacian. *Communications on pure and applied mathematics*, 65(8):1067–1144, 2012.

[62] Alessandro Sperduti. Encoding labeled graphs by labeling RAAM. In *NIPS*, 1994.

[63] 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.

[64] Loring W Tu. Manifolds. In *An Introduction to Manifolds*, pages 47–83. Springer, 2011.

[65] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

[66] Felix Wu, Tianyi Zhang, Amauri Holanda de Souza Jr, Christopher Fifty, Tao Yu, and Kilian Q Weinberger. Simplifying graph convolutional networks. In *ICML*, 2019.

[67] Louis-Pascal Xhonneux, Meng Qu, and Jian Tang. Continuous graph neural networks. In *ICML*, 2020.

[68] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *ICLR*, 2019.

[69] 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:2102.06462*, 2021.

[70] Fouad El Zein and Jawad Snoussi. Local systems and constructible sheaves. In *Arrangements, local systems and singularities*, pages 111–153. Springer, 2009.

[71] Lingxiao Zhao and Leman Akoglu. Pairnorm: Tackling oversmoothing in gnns. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rkecl1rtwB`.

[72] Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. *Advances in Neural Information Processing Systems*, 33, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Proofs are included in the appendix
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Our code is available at `https://github.com/twitter-research/neural-sheaf-diffusion`.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Section 6 and Appendix E

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Appendix E

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes]

   (b) Did you mention the license of the assets? [Yes] Appendix F

   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The code of our submission.

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Harmonic Space Proofs

Proposition 3. *If $\mathcal{F}$ is a discrete $O(d)$ bundle over a connected graph and $r := \max_{\gamma_{v\to u},\gamma'_{v\to u}} ||\mathbf{P}^{\gamma}_{v\to u} - \mathbf{P}^{\gamma'}_{v\to u}||$, then we have $\lambda_0^{\mathcal{F}} \leq r^2/2$.*

Proof. We first note that on a discrete $O(d)$ bundle the degree operator $D_v = d_v\mathbf{I}$ since by orthogonality $\mathcal{F}^{\top}_{v\trianglelefteq e}\mathcal{F}_{v\trianglelefteq e} = \mathbf{I}$. We can use the Rayleigh quotient to characterize $\lambda_0^{\mathcal{F}}$ as

$$\lambda_0^{\mathcal{F}} = \min_{\mathbf{x}\in\mathbb{R}^{nd}} \frac{\langle \mathbf{x}, \Delta_{\mathcal{F}}\mathbf{x}\rangle}{||\mathbf{x}||^2}.$$

Fix $v \in V$ and choose a minimal path $\gamma_{v\to u}$ for all $u \in V$. For an arbitrary non-zero $\mathbf{z}_v$, consider the signal $\mathbf{z}_u = P^{\gamma}_{v\to u}\mathbf{z}_v$ and we set $\tilde{\mathbf{z}}_u\sqrt{d_u} = \mathbf{z}_u$.

$$||\mathcal{F}_{u\trianglelefteq e}\frac{\mathbf{z}_u}{\sqrt{d_u}} - \mathcal{F}_{w\trianglelefteq e}\frac{\mathbf{z}_w}{\sqrt{d_w}}||^2 = ||\tilde{\mathbf{z}}_u - (\mathcal{F}^{\top}_{u\trianglelefteq e}\mathcal{F}_{w\trianglelefteq e})\tilde{\mathbf{z}}_w||^2 = ||P^{\gamma}_{v\to u}\tilde{\mathbf{z}}_v - (\mathcal{F}^{\top}_{u\trianglelefteq e}\mathcal{F}_{w\trianglelefteq e})\mathbf{P}^{\gamma}_{v\to w}\tilde{\mathbf{z}}_v||^2,$$

where we have again used that the maps are orthogonal. Since $(\mathcal{F}^{\top}_{u\trianglelefteq e}\mathcal{F}_{w\trianglelefteq e})\mathbf{P}^{\gamma}_{v\to w} = \mathbf{P}^{\gamma'}_{v\to u}$ we find that the right hand side can be bound from above by $r^2||\tilde{\mathbf{z}}_v||^2$. Therefore, by using Definition 14 we finally obtain

$$\lambda_0^{\mathcal{F}} = \min_{\mathbf{x}\in R^{nd}} \frac{\langle \mathbf{x}, \Delta_{\mathcal{F}}\mathbf{x}\rangle}{||\mathbf{x}||^2} \leq \frac{\langle \mathbf{z}, \Delta_{\mathcal{F}}\mathbf{z}\rangle}{||\mathbf{z}||^2} = \frac{1}{2}\frac{\sum_{u\sim w}||\mathcal{F}_{u\trianglelefteq e}\frac{\mathbf{z}_u}{\sqrt{d_u}} - \mathcal{F}_{w\trianglelefteq e}\frac{\mathbf{z}_w}{\sqrt{d_w}}||^2}{||\mathbf{z}||^2} \leq \frac{r^2}{2}\frac{\sum_{u\sim w}||\tilde{\mathbf{z}}_v||^2}{||\mathbf{z}||^2}.$$

Since the transport maps are all orthogonal we get

$$||\mathbf{z}||^2 = \sum_u d_u||\mathbf{P}^{\gamma}_{v\to u}\tilde{\mathbf{z}}_v||^2 = \sum_u d_u||\tilde{\mathbf{z}}_v||^2 = \sum_{u\sim w}||\tilde{\mathbf{z}}_v||^2.$$

We conclude that

$$\lambda_0^{\mathcal{F}} \leq \frac{r^2}{2}\frac{\sum_{u\sim w}||\tilde{\mathbf{z}}_v||^2}{||\mathbf{z}||^2} = \frac{r^2}{2}.$$

□

Proposition 4. *If $\mathcal{F}$ is a discrete $O(d)$ bundle over a connected graph and $\mathbf{x} \in H^0(G,\mathcal{F})$, then for any cycle γ based at $v \in V$ we have $\mathbf{x}_v \in \ker(\mathbf{P}^{\gamma}_{v\to v} - \mathbf{I})$.*

Proof. Assume that $\mathbf{x} \in H^0(G,\mathcal{F})$ and consider $v \in V$ and any cycle based at v denoted by $\gamma_{v\to v} = (v_0 = v, v_1, \ldots, v_L = v)$. According to the Hodge Theorem we have that

$$\mathcal{F}_{v_{i+1}\trianglelefteq e}\mathbf{x}_{v_{i+1}} = \mathcal{F}_{v_i\trianglelefteq e}\mathbf{x}_{v_i} \Longrightarrow \mathbf{x}_{v_{i+1}} = (\mathcal{F}^{\top}_{v_{i+1}}\mathcal{F}_{v_i})\mathbf{x}_{v_i} := \rho_{v_i\to v_{i+1}}\mathbf{x}_{v_i}.$$

By composing all the maps we find:

$$\mathbf{x}_v = \rho_{v_{L-1}\to v_L}\cdots\rho_{v_0\to v_1}x_v = \mathbf{P}^{\gamma}_{v\to v}\mathbf{x}_v$$

which completes the proof. □

Proposition 5. *Let $\mathcal{F}$ be a discrete $O(d)$ bundle over a connected graph G with n nodes and let $||(\mathbf{P}^{\gamma}_{v\to v} - \mathbf{I})\mathbf{x}_v|| \geq \epsilon||\mathbf{x}_v||$ for all cycles $\gamma_{v\to v}$. Then $\lambda_0^{\mathcal{F}} \geq \epsilon^2(2\mathrm{diam}(G)n\, d_{max})^{-1}$.*

Proof. If $\epsilon = 0$ there is nothing to prove. Assume that $\epsilon > 0$. By Proposition 4 we derive that the harmonic space is trivial and hence $\lambda_0^{\mathcal{F}} > 0$. Consider a *unit* eigenvector $\mathbf{x} \in \ker(\Delta_{\mathcal{F}} - \lambda_0^{\mathcal{F}} I)$ and let $v \in V$ such that $||\mathbf{x}_v|| \geq ||\mathbf{x}_u||$ for $u \neq v$. There exists a cycle γ based at v such that $\mathbf{P}^{\gamma}_v\mathbf{x}_v \neq \mathbf{x}_v$ for otherwise we could extend $\mathbf{x}_v \neq 0$ to any other node independently of the path choice and hence find a non-trivial harmonic signal. In particular, we can assume this cycle to be non-degenerate, otherwise if there existed a non-trivial degenerate loop contained in γ that does not fix $\mathbf{x}$ we could consider this loop instead of γ for our argument. Let us write this path as $(v_0 = v, v_1, \ldots, v_L = v)$ and consider the rescaled signal $\tilde{\mathbf{x}}_v\sqrt{d_v} = \mathbf{x}_v$. By assumption we have

$$\begin{aligned}
\epsilon||\tilde{\mathbf{x}}_v|| &\leq ||(\mathbf{P}^{\gamma}_{v\to v} - \mathbf{I})\tilde{\mathbf{x}}_v|| = ||(\rho_{v_{L-1}\to v_L}\cdots\rho_{v_0\to v_1} - \mathbf{I})\tilde{\mathbf{x}}_v|| \\
&= ||\mathcal{F}_{v_{L-1}}\rho_{v_{L-2}\to v_{L-1}}\cdots\rho_{v_0\to v_1}\tilde{\mathbf{x}}_v - \mathcal{F}_{v_L=v}\tilde{\mathbf{x}}_v|| \\
&= ||\mathcal{F}_{v_{L-1}}\rho_{v_{L-2}\to v_{L-1}}\cdots\rho_{v_0\to v_1}\tilde{\mathbf{x}}_v - \mathcal{F}_{v_{L-1}}\tilde{\mathbf{x}}_{v_{L-1}} + \mathcal{F}_{v_{L-1}}\tilde{\mathbf{x}}_{v_{L-1}} - \mathcal{F}_{v_L=v}\tilde{\mathbf{x}}_v|| \\
&\leq ||\rho_{v_{L-2}\to v_{L-1}}\cdots\rho_{v_0\to v_1}\tilde{\mathbf{x}}_v - \tilde{\mathbf{x}}_{v_{L-1}}|| + ||\mathcal{F}_{v_{L-1}}\tilde{\mathbf{x}}_{v_{L-1}} - \mathcal{F}_{v_L=v}\tilde{\mathbf{x}}_v||.
\end{aligned}$$

By iterating the approach above we find:

$$\epsilon||\tilde{\mathbf{x}}_v|| \leq \sum_{i=0}^{L}||\mathcal{F}_{v_i}\tilde{\mathbf{x}}_{v_i} - \mathcal{F}_{v_{i+1}}\tilde{\mathbf{x}}_{v_{i+1}}|| \leq \sqrt{L}\left(\sum_{i=0}^{L}||\mathcal{F}_{v_i}\tilde{\mathbf{x}}_{v_i} - \mathcal{F}_{v_{i+1}}\tilde{\mathbf{x}}_{v_{i+1}}||^2\right)^{\frac{1}{2}}$$
$$= \sqrt{L}\left(\sum_{i=0}^{L}||\mathcal{F}_{v_i}\frac{\mathbf{x}_{v_i}}{\sqrt{d_{v_i}}} - \mathcal{F}_{v_{i+1}}\frac{\mathbf{x}_{v_{i+1}}}{\sqrt{d_{v_{i+1}}}}||^2\right)^{\frac{1}{2}}.$$

From Definition 14 we derive that the last term can be bounded from above by $\sqrt{2LE_{\mathcal{F}}(\mathbf{x})} = \sqrt{2L\langle\mathbf{x}, \Delta_{\mathcal{F}}\mathbf{x}\rangle}$. Therefore, we conclude:

$$\epsilon\frac{||\mathbf{x}_v||}{\sqrt{d_v}} \leq \sqrt{2L\langle\mathbf{x}, \Delta_{\mathcal{F}}\mathbf{x}\rangle} = \sqrt{2L\lambda_0^{\mathcal{F}}}||\mathbf{x}|| \leq 2\sqrt{\mathrm{diam}(G)\lambda_0^{\mathcal{F}}}.$$

By construction we get $||\mathbf{x}_v|| \geq 1/\sqrt{n}$, meaning that

$$\lambda_0^{\mathcal{F}} \geq \frac{\epsilon^2}{2\mathrm{diam}(G)}\frac{1}{n\, d_{\text{max}}}.$$

□

Lemma 6. *Let $\mathcal{F}$ be a discrete $O(d)$ bundle over a connected graph G. Then $\dim(H^0) \leq d$ and $\dim(H^0) = d$ if and only if the transport is path-independent.*

Proof. We first note that the argument below extends to weighted $O(d)$-bundles as well. Let $\mathbf{x} \in H^0(G, \mathcal{F})$. According to Proposition 4, given $v, u \in V$, we see that $x_u = \mathbf{P}^{\gamma}_{v\to u}x_v$ for any path $\gamma_{v\to u}$. It means that the harmonic space is uniquely determined by the choice of $\mathbf{x}_v \in \mathcal{F}(v)$. Explicitly, given any cycle γ based at v, we know that $\mathbf{x}_v \in \ker(\mathbf{P}^{\gamma}_{v\to v} - I)$. If the transport is everywhere path-independent, then the kernel coincides with the whole stalk $\mathcal{F}(v)$ and hence we can extend any basis $\{\mathbf{x}_{v_i}\} \in \mathcal{F}(v) \cong \mathbb{R}^d$ to a basis in $H^0(G, \mathcal{F})$ via the transport maps, i.e. $\dim(H^0(G, \mathcal{F})) = d$. If instead there exists a transport map over a cycle $\gamma_{v\to v}$ with non-trivial fixed points, then $\ker(\mathbf{P}^{\gamma}_{v\to v} - I) < \mathcal{F}(v) \cong \mathbb{R}^d$ and hence $\dim(H^0(G, \mathcal{F})) < d$. □

B Proofs for the Power of Sheaf Diffusion

Definition 19. *Let $G = (V, \mathbf{W})$ be a weighted graph, where $\mathbf{W}$ is a matrix with $w_{vu} = w_{uv} \geq 0$ for all $v \neq u \in V$, $w_{vv} = 0$ for all $v \in V$, and (v, u) is an edge if and only if $w_{vu} > 0$.*

The graph Laplacian of a weighted graph is $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is the diagonal matrix of weighted degrees (i.e. $d_v = \sum_u w_{vu}$). Its normalised version is $\widetilde{L} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$.

Proposition 20. *Let G be a graph. The set $\{\Delta_{\mathcal{F}} \mid (G, \mathcal{F}) \in \mathcal{H}^1_{\text{sym}}\}$ is isomorphic to the set of all possible weighted graph Laplacians over G.*

Proof. We prove only one direction. Let $\mathbf{W}$ be a choice of valid weight matrix for the graph G. We can construct a sheaf $(G, \mathcal{F}) \in \mathcal{H}^1_{\text{sym}}$ such that for all edges $v, u \trianglelefteq e$ we have that $\mathcal{F}_{v\trianglelefteq e} = \mathcal{F}_{u\trianglelefteq e} = \pm\sqrt{w_{vu}}$. Then, $\mathcal{L}_{vu} = -w_{vu}$ and $\mathcal{L}_{vv} = \sum_e \|\mathcal{F}_{v\trianglelefteq e}\|^2 = \sum_u w_{vu}$. The equality for the normalised version of the Laplacians follows directly. □

We state the following Lemma without proof based on Theorem 3.1 in Hansen and Ghrist [33].

Lemma 21. *Solutions $\mathbf{X}(t)$ to the diffusion in Equation 3 converge as $t \to \infty$ to the orthogonal projection of $\mathbf{X}(0)$ onto $\ker(\Delta_{\mathcal{F}})$.*

Due to this Lemma, the proofs below rely entirely on the structure of $\ker(\Delta_{\mathcal{F}})$ that one obtains for certain $(G, \mathcal{F})$.

Proposition 8. *Let $\mathcal{G}$ be the set of connected graphs $G = (V, E)$ with two classes $A, B \subset V$ such that for each $v \in A$, there exists $u \in A$ and an edge $(v, u) \in E$. Then $\mathcal{H}^1_{\text{sym}}$ has linear separation power over $\mathcal{G}$.*

Proof. Let $G = (V, E)$ be a graph with two classes $A, B \subset V$ such that for each $v \in A$, there exists $u \in A$ and an edge $(v, u) \in E$. Additionally, let $\mathbf{x}(0)$ be any channel of the feature matrix $\mathbf{X}(0) \in \mathbb{R}^{n \times f}$.

We can construct a sheaf $(\mathcal{F}, G) \in \mathcal{H}^1_{\text{sym}}$ as follows. For all nodes $v \in V$ and edges $e \in E$, $\mathcal{F}(v) \cong \mathcal{F}(e) \cong \mathbb{R}$. For all $v, u \in A$ and edge $(u, v) \in E$, set $\mathcal{F}_{v \trianglelefteq e} = \mathcal{F}_{u \trianglelefteq e} = \sqrt{\alpha} > 0$. Otherwise, set $\mathcal{F}_{v \trianglelefteq e} = 1$.

Denote by h_v the number of neighbours of node v in the same class as v . Note that based on the assumptions, $h_v > 1$ if $v \in A$. Then the only harmonic eigenvector of $\Delta_{\mathcal{F}}$ is:

$$\mathbf{a}_v = \begin{cases} \sqrt{d_v + h_v(\alpha - 1)}, & v \in A \\ \sqrt{d_v}, & v \in B \end{cases} \tag{7}$$

Denote its unit-normalised version $\widetilde{\mathbf{a}} := \frac{\mathbf{a}}{\|\mathbf{a}\|}$. In the limit of the diffusion process, the features converge to $\mathbf{h} = \langle \mathbf{x}(0), \widetilde{\mathbf{a}} \rangle \widetilde{\mathbf{a}}$ by Lemma 21. Assuming, $\mathbf{x}(0) \notin \ker(\Delta_{\mathcal{F}})^{\perp}$, which is nowhere dense in $\mathbb{R}^n$ and, without loss of generality, that $\langle \mathbf{x}(0), \widetilde{\mathbf{a}} \rangle > 0$, for sufficiently large α, $\widetilde{\mathbf{a}}_v \geq \widetilde{\mathbf{a}}_u$ for all $v \in A, u \in B$. □

Proposition 9. *Let $\mathcal{G}$ be the set of connected bipartite graphs $G = (A, B, E)$, with partitions A, B forming two classes and $|A| = |B|$. Then $\mathcal{H}^1_{\text{sym}}$ cannot linearly separate any graph in $\mathcal{G}$ for any initial conditions $\mathbf{X}(0) \in \mathbb{R}^{n \times f}$.*

Proof. Let $G = (A, B, E)$ be a bipartite graph with $|A| = |B|$ and let $\mathbf{x}(0) \in \mathbb{R}^n$ be any channel of the feature matrix $\mathbf{X}(0) \in \mathbb{R}^{n \times f}$.

Consider an arbitrary sheaf $(G, \mathcal{F}) \in \mathcal{H}^1_{\text{sym}}$. Since the graph is connected, the only harmonic eigenvector of $\Delta_{\mathcal{F}}$ is $\mathbf{y} \in \mathbb{R}^n$ with $\mathbf{y}_v = \sqrt{\sum_{v \trianglelefteq e} \|\mathcal{F}_{v \trianglelefteq e}\|^2}$ (i.e. the square root of the weighted degree). Based on Lemma 21, the diffusion process converges in the limit (up to a scaling) to $\langle \mathbf{x}, \mathbf{y} \rangle \mathbf{y}$. For the features to be linearly separable we require that $\langle \mathbf{x}, \mathbf{y} \rangle \neq 0$ and, without loss of generality, for all $v \in A, u \in B$ that $\mathbf{y}_v < \mathbf{y}_u \Leftrightarrow \sum_{v \trianglelefteq e} \|\mathcal{F}_{v \trianglelefteq e}\|^2 < \sum_{u \trianglelefteq e} \|\mathcal{F}_{u \trianglelefteq e}\|^2$.

Suppose for the sake of contradiction there exists a sheaf in $\mathcal{H}^1_{\text{sym}}$ with such a harmonic eigenvector. Then, because $|A| = |B|$:

$$\begin{aligned} \sum_{v \in A} \sum_{v \trianglelefteq e} \|\mathcal{F}_{v \trianglelefteq e}\|^2 < \sum_{u \in B} \sum_{u \trianglelefteq e} \|\mathcal{F}_{u \trianglelefteq e}\|^2 &\Leftrightarrow \sum_{v \in A} \sum_{v \trianglelefteq e} \|\mathcal{F}_{v \trianglelefteq e}\|^2 - \sum_{u \in B} \sum_{u \trianglelefteq e} \|\mathcal{F}_{u \trianglelefteq e}\|^2 < 0 \\ &\Leftrightarrow \sum_{e \in E} \|\mathcal{F}_{v \trianglelefteq e}\|^2 - \|\mathcal{F}_{u \trianglelefteq e}\|^2 < 0 \end{aligned}$$

However, because $(\mathcal{F}, G) \in \mathcal{H}^1_{\text{sym}}$, we have $\mathcal{F}_{v \trianglelefteq e} = \mathcal{F}_{u \trianglelefteq e}$ and the sum above is zero. □

Proposition 10. *Let $\mathcal{G}$ contain all the connected graphs $G = (V, E)$ with two classes $A, B \subseteq V$. Consider a sheaf $(\mathcal{F}; G) \in \mathcal{H}^1$ with $\mathcal{F}_{v \trianglelefteq e} = -\alpha_e$ if $v \in A$ and $\mathcal{F}_{u \trianglelefteq e} = \alpha_e$ if $u \in B$ with $\alpha_e > 0$ for all $e \in E$. Then the diffusion induced by $(\mathcal{F}; G)$ can linearly separate the classes of G for almost all initial conditions and $\mathcal{H}^1$ has linear separation power over $\mathcal{G}$.*

Proof. Let $G = (V, E)$ be a connected graph with two classes $A, B \subset V$. Additionally, let $\mathbf{x}(0)$ be any channel of the feature matrix $\mathbf{X}(0) \in \mathbb{R}^{n \times f}$. Any sheaf of the described type has a single harmonic eigenvector by virtue of Lemma 6 and it has the form:

$$\mathbf{y}_v = \begin{cases} +\sqrt{\sum_{v \trianglelefteq e} \alpha_e}, & v \in A \\ -\sqrt{\sum_{v \trianglelefteq e} \alpha_e}, & v \in B \end{cases} \tag{8}$$

Assume $\mathbf{x}(0) \notin \ker(\Delta_{\mathcal{F}})^{\perp}$, which is nowhere dense in $\mathbb{R}^n$ and, without loss of generality, that $\langle \mathbf{x}(0), \mathbf{y} \rangle > 0$. Then, $\mathbf{y}_v > 0 > \mathbf{y}_u$ for all $v \in A, u \in B$. □

Next, we showed that using signed relations is necessary in $d = 1$ and simply using positive asymmetric relations is not sufficient in this dimension.

Definition 22. *The class of sheaves over G with non-zero maps, one-dimensional stalks, and similarly signed restriction maps* $\mathcal{H}^1_+ := \{(\mathcal{F}, G) \mid \mathcal{F}_{v \trianglelefteq e}\mathcal{F}_{u \trianglelefteq e} > 0\}$

Proposition 23. *Let G be the connected graph with two nodes belonging to two different classes. Then* $\mathcal{H}^1_+$ *cannot linearly separate the two nodes for any initial conditions* $\mathbf{X} \in \mathbb{R}^{2\times f}$.

Proof. Let G be the connected graph with two nodes $V = \{v, u\}$. Then any sheaf $(\mathcal{F}, G) \in \mathcal{H}^1_+(G)$ has restriction maps of the form $\mathcal{F}_{v \trianglelefteq e} = \alpha, \mathcal{F}_{u \trianglelefteq e} = \beta$ and (without loss of generality) $\alpha, \beta > 0$. As before, the only (unormalised) harmonic eigenvector for a sheaf of this form is $\mathbf{y} = (|\alpha|\beta, \alpha|\beta|) = (\alpha\beta, \alpha\beta)$. Since this is a constant vector, the two nodes are not separable in the diffusion limit. □

We state the following result without proof (see Exercise 4.1 in Bishop [5]).

Lemma 24. *Let A and B be two sets of points in* $\mathbb{R}^n$. *If their convex hulls intersect, the two sets of points cannot be linearly separable.*

Proposition 11. *Let G be a connected graph with* $C \geq 3$ *classes. Then,* $\mathcal{H}^1$ *cannot linearly separate the classes of G for any initial conditions* $\mathbf{X}(0) \in \mathbb{R}^{n\times f}$.

Proof. If the sheaf has a trivial global section, then all features converge to zero in the diffusion limit. Suppose $H^0(G, \mathcal{F})$ is non-trivial. Since G is connected and all the restriction maps are invertible, by Lemma 6, $\dim(H^0) = 1$.

In that case, let $\mathbf{h}$ be the unit-normalised harmonic eigenvector of $\Delta_{\mathcal{F}}$. By Lemma 21, for any node v, its scalar feature in channel $k \leq f$ is given by $x^k_v(\infty) = \langle \mathbf{x}^k(0), \mathbf{h}\rangle \mathbf{h}_v$. Note that we can always find three nodes v, u, w belonging to three different classes such that $\mathbf{h}_v \leq \mathbf{h}_u \leq \mathbf{h}_v$. Then, there exists a convex combination $\mathbf{h}_u = \alpha\mathbf{h}_v + (1-\alpha)\mathbf{h}_w$, with $\alpha \in [0, 1]$. Therefore:

$$\mathbf{x}^k_u(\infty) = \langle \mathbf{x}^k(0), \mathbf{h}\rangle \mathbf{h}_u = \alpha\langle \mathbf{x}^k(0), \mathbf{h}\rangle \mathbf{h}_v + (1-\alpha)\langle \mathbf{x}^k(0), \mathbf{h}\rangle \mathbf{h}_w = \alpha\mathbf{x}^k_v(\infty) + (1-\alpha)\mathbf{x}^k_w(\infty). \tag{9}$$

Since this is true for all channels $k \leq f$, it follows that $\mathbf{x}_u(\infty) = \alpha\mathbf{x}_v(\infty) + (1-\alpha)\mathbf{x}_w(\infty)$. Because $\mathbf{x}_u(\infty)$ is in the convex hull of the points belonging to other classes, by Lemma 24, the class of v is not linearly separable from the other classes. □

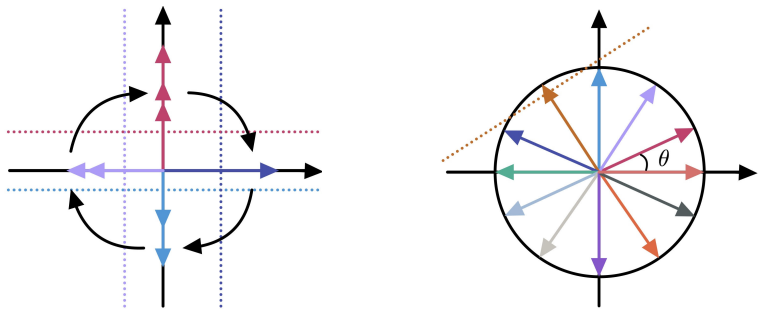
Proposition 12. *Let* $\mathcal{G}$ *be the set of connected graphs with nodes belonging to* $C \geq 3$ *classes. Then for* $d \geq C$, $\mathcal{H}^d_{\text{diag}}$ *has linear separation power over* $\mathcal{G}$.

Proof. Let $G = (V, E)$ be a connected graph with C classes and $(\mathcal{F}, G)$, an arbitrary sheaf in $\mathcal{H}^C(G)$. Because $\mathcal{F}$ has diagonal restriction maps there is no interaction during diffusion between the different dimensions of the stalks. Therefore, the diffusion process can be written as d independent diffusion processes, where the i-th process uses a sheaf $\mathcal{F}^i$ with all stalks isomorphic to $\mathbb{R}$ and $\mathcal{F}^i_{v \trianglelefteq e} = \mathcal{F}_{v \trianglelefteq e}(i, i)$ for all $v \in V$ and incident edges e. Therefore, we can construct d sheaves $\mathcal{F}^i \in \mathcal{H}^1(G)$ with $i < d$ as in Proposition 10, where (in one vs all fashion) the two classes are given by the nodes in class i and the nodes belonging to the other classes.

It remains to restrict that the projection of $\mathbf{x}(0)$ on any of the harmonic eigenvectors of $\Delta_{\mathcal{F}}$ in the standard basis is non-zero. Formally, we require $\mathbf{x}^i(0) \notin \ker(\Delta_{\mathcal{F}^i})^\perp$ for all positive integers $i \leq d$. Since $\ker(\Delta_{\mathcal{F}^i})^\perp$ is nowhere dense in $\mathbb{R}^n$, $\mathbf{x}(0)$ belongs to the direct sum of dense subspaces, which is dense. □

Lemma 25. *Let* $G = (V, E)$ *be a graph and* $(\mathcal{F}, G)$ *a (weighted) orthogonal vector bundle over G with path-independent parallel transport and edge weights* α_e. *Consider an arbitrary node* $v^* \in V$ *and denote by* $\mathbf{e}_i$ *the* i*-th standard basis vector of* $\mathbb{R}^d$. *Then* $\{\mathbf{h}^1, \ldots, \mathbf{h}^d\}$ *form an orthogonal eigenbasis for the harmonic space of* $\Delta_{\mathcal{F}}$, *where:*

$$\mathbf{h}^i_v = \begin{cases} \mathbf{e}_i\sqrt{d^{\mathcal{F}}_v} \\ \mathbf{P}_{v\to w}\mathbf{e}_i\sqrt{d^{\mathcal{F}}_v} \end{cases} = \begin{cases} \mathbf{e}_i\sqrt{\sum_{v \trianglelefteq e}\alpha^2_e}, & v = v^* \\ \mathbf{P}_{v^*\to w}\mathbf{e}_i\sqrt{\sum_{v \trianglelefteq e}\alpha^2_e}, & \textit{otherwise} \end{cases} \tag{10}$$



(a) Aligning the features of each class with the axis of coordinates in a 2D space. Dotted lines indicate linear decision boundaries for each class.

(b) Separating an arbitrary number of classes when the graph is regular. Dotted line shows an example decision boundary for one of the classes.

Figure 6: Proof sketch for Lemma 27 and Proposition 28.

Proof. First, we show that $\mathbf{h}_v^i$ is harmonic.

$$E_{\mathcal{F}}(\mathbf{h}_v^i) = \frac{1}{2} \sum_{v,u,e:=(v,u)} \|\frac{1}{\sqrt{d_v^{\mathcal{F}}}} \mathcal{F}_{v \leq e} \mathbf{h}_v - \frac{1}{\sqrt{d_u^{\mathcal{F}}}} \mathcal{F}_{u \trianglelefteq e} \mathbf{h}_u\|_2^2 \tag{11}$$

$$= \frac{1}{2} \sum_{v,u,e:=(v,u)} \|\mathcal{F}_{v \leq e} \mathbf{P}_{v^* \to v} \mathbf{e}_i - \mathcal{F}_{u \trianglelefteq e} \mathbf{P}_{v^* \to u} \mathbf{e}_i\|_2^2 \tag{12}$$

$$= \frac{1}{2} \sum_{v,u,e:=(v,u)} \|\mathcal{F}_{v \leq e} \mathbf{P}_{u \to v} \mathbf{P}_{v^* \to u} \mathbf{e}_i - \mathcal{F}_{u \trianglelefteq e} \mathbf{P}_{v^* \to u} \mathbf{e}_i\|_2^2 \quad \text{By path independence} \tag{13}$$

$$= \frac{1}{2} \sum_{v,u,e:=(v,u)} \|\mathcal{F}_{v \leq e} \mathcal{F}_{v \trianglelefteq e}^{\top} \mathcal{F}_{u \trianglelefteq e} \mathbf{P}_{v^* \to u} \mathbf{e}_i - \mathcal{F}_{u \trianglelefteq e} \mathbf{P}_{v^* \to u} \mathbf{e}_i\|_2^2 \quad \text{By definition of } \mathbf{P}_{u \to v} \tag{14}$$

$$= \frac{1}{2} \sum_{v,u,e:=(v,u)} \|\mathcal{F}_{u \trianglelefteq e} \mathbf{P}_{v^* \to u} \mathbf{e}_i - \mathcal{F}_{u \trianglelefteq e} \mathbf{P}_{v^* \to u} \mathbf{e}_i\|_2^2 = 0 \quad \text{Orthogonality of } \mathcal{F}_{v \trianglelefteq e} \tag{15}$$

For orthogonality, notice that for any $i, j \leq d$ and $v \in V$, it holds that:

$$\langle \mathbf{h}_v^i, \mathbf{h}_v^j \rangle = \langle \mathbf{P}_{v* \to w} \mathbf{e}_i \sqrt{d_v^{\mathcal{F}}}, \mathbf{P}_{v* \to w} \mathbf{e}_j \sqrt{d_v^{\mathcal{F}}} \rangle = \sqrt{d_v^{\mathcal{F}}} \sqrt{d_v^{\mathcal{F}}} \langle \mathbf{e}_i, \mathbf{e}_j \rangle = 0 \tag{16}$$

□

Lemma 26. *Let $\mathbf{R}_1, \mathbf{R}_2$ be two 2D rotation matrices and $\mathbf{e}_1, \mathbf{e}_2$ the two standard basis vectors of $\mathbb{R}^2$. Then $\langle \mathbf{R}_1 \mathbf{e}_1, \mathbf{R}_2 \mathbf{e}_2 \rangle = -\langle \mathbf{R}_1 \mathbf{e}_2, \mathbf{R}_2 \mathbf{e}_1 \rangle$.*

Proof. The angle between $\mathbf{e}_1$ and $\mathbf{e}_2$ is $\frac{\pi}{2}$. Letting ϕ, θ be the positive rotation angles of the two matrices, the first inner product is equal to $\cos(\pi/2 + (\phi - \theta))$ while the second is $\cos(\pi/2 - (\phi - \theta))$. The result follows from applying the trigonometric identity $\cos(\pi/2 + x) = -\sin x$. □

We first prove Theorem 13 in dimension two in the following lemma and then we will look at the general case.

Lemma 27. *Let $\mathcal{G}$ be the class of connected graphs with $C \leq 4$ classes. Then, $\mathcal{H}^2_{\text{orth}}(G)$ has linear separation power over $\mathcal{G}$.*

Proof. Idea: We can use rotation matrices to align the harmonic features of the classes with the axis of coordinates as in Figure 6a. Then, for each side of each axis, we can find a separating hyperplane separating each class from all the others.

Let G be a connected graph with $C \leq 4$ classes. Denote by $\mathcal{P}$ the following set of rotation matrices together with their signed-flipped counterparts:

$$\mathbf{R}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{R}_2 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \tag{17}$$

and by $\mathcal{C} = \{1, \ldots, C\}$ the set of all class labels. Then, fix a node $v^* \in V$ and construct an injective map $g : C \to \mathcal{P}$ assigning each class label one of the signed basis vectors such that $g(c(v^*)) = \mathbf{R}_1$, where $c(v^*)$ denotes the class of node v^*.

Then, we can construct an sheaf $(G, \mathcal{F}) \in \mathcal{H}^2_{\text{orth}}(G)$ in terms of certain parallel transport maps along each edge, that will depend on $\mathcal{P}$. For all nodes v and edges e, $\mathcal{F}(v) \cong \mathcal{F}(e) \cong \mathbb{R}^2$. For each $u \in V$, we set $\mathbf{P}_{v^* \to u} = g(c(u))$. Then for all $v, u \in V$, set $\mathbf{P}_{v \to u} = \mathbf{P}_{v^* \to v}\mathbf{P}^{-1}_{u \to v^*}$. It is easy to see that the resulting parallel transport is path-independent because it depends purely on the classes of the endpoints of the path.

Based on Lemma 25, the i-th eigenvector of $\Delta_{\mathcal{F}}$ is $\mathbf{h}^i \in \mathbb{R}^{2\times n}$ with $\mathbf{h}^i_u = \mathbf{P}_{v^* \to u}\mathbf{e}_i\sqrt{d_u}$. Now we will show that the projection of $\mathbf{x}(0)$ in this subspace will have a configuration as in Figure 6a up to a rotation.

Let u, w be two nodes belonging to two different classes. Denote by $\alpha_i = \langle \mathbf{x}(0), \mathbf{h}^i \rangle$. Then the inner product between the features of nodes u, w in the limit of the diffusion process is:

$$\begin{aligned}
&\langle \mathbf{P}_{v^* \to u} \sum_i \alpha_i \mathbf{e}_i \sqrt{d_u}, \mathbf{P}_{v^* \to w} \sum_j \alpha_j \mathbf{e}_j \sqrt{d_w} \rangle = \\
&= \sqrt{d_u d_w} \Big[\sum_{i \neq j} \alpha_i \alpha_j \langle \mathbf{P}_{v^* \to u}\mathbf{e}_i, \mathbf{P}_{v^* \to w}\mathbf{e}_j \rangle + \sum_k \alpha_k^2 \langle \mathbf{P}_{v^* \to u}\mathbf{e}_k, \mathbf{P}_{v^* \to w}\mathbf{e}_k \rangle \Big] \\
&= \sqrt{d_u d_w} \Big[\sum_{i<j} \alpha_i \alpha_j \Big(\langle \mathbf{P}_{v^* \to u}\mathbf{e}_i, \mathbf{P}_{v^* \to w}\mathbf{e}_j \rangle + \langle \mathbf{P}_{v^* \to u}\mathbf{e}_j, \mathbf{P}_{v^* \to w}\mathbf{e}_i \rangle \Big) \\
&\quad + \sum_k \alpha_k^2 \langle \mathbf{P}_{v^* \to u}\mathbf{e}_k, \mathbf{P}_{v^* \to w}\mathbf{e}_k \rangle \Big] \qquad (18) \\
&= \sum_k \alpha_k^2 \langle \mathbf{P}_{v^* \to u}\mathbf{e}_k, \mathbf{P}_{v^* \to w}\mathbf{e}_k \rangle \qquad \text{(by Lemma 26)}
\end{aligned}$$

It can be checked that by substituting the transport maps $\mathbf{P}_{v^* \to u}, \mathbf{P}_{v^* \to w}$ with any $\mathbf{R}_a, \mathbf{R}_b$ from $\mathcal{P}$ such that $\mathbf{R}_a \neq \pm\mathbf{R}_b$, the inner product above is zero. Similarly, substituting any $\mathbf{R}_a = -\mathbf{R}_b$, the inner product is $-\sqrt{d_u d_w}\sum_k \alpha_k^2 = -\sqrt{d_u d_w}\|\mathbf{x}(0)\|^2$, which is equal to the product of the norms of the two vectors. Therefore, the diffused features of different classes are positioned at $\frac{\pi}{2}, \pi, \frac{3\pi}{2}$ from each other, as in Figure 6a. □

Proposition 13. *Let $\mathcal{G}$ be the class of connected graphs with $C \leq 2d$ classes. Then, for all $d \in \{2, 4\}$, $\mathcal{H}^d_{\text{orth}}$ has linear separation power over $\mathcal{G}$.*

Proof. To generalise the proof in Lemma 27, we need to find a set $\mathcal{P}$ of size d containing rotation matrices that make the projected features of different classes be pairwise orthogonal for any projection coefficients α. For that, each term in Equation 18 must be zero for any coefficients α.

Therefore, $\mathcal{P} = \{\mathbf{P}_0, \ldots, \mathbf{P}_{d-1}\}$ must satisfy the following requirements:

1. $\mathbf{P}_0 = \mathbf{I} \in \mathcal{P}$, since transport for neighbours in the same class must be the identity. Therefore, $\mathbf{P}_0\mathbf{P}_k = \mathbf{P}_k\mathbf{P}_0 = \mathbf{P}_k$ for all k.
2. Since $\langle \mathbf{P}_0\mathbf{e}_i, \mathbf{P}_k\mathbf{e}_i \rangle = 0$ for all i and $k \neq 0$, it follows that the diagonal elements of $\mathbf{P}_k$ are zero.
3. From $\langle \mathbf{P}_0\mathbf{e}_i, \mathbf{P}_k\mathbf{e}_j \rangle = -\langle \mathbf{P}_0\mathbf{e}_j, \mathbf{P}_k\mathbf{e}_i \rangle$ for all $i \neq j, k \neq 0$ and point (2) it follows that $\mathbf{P}_k^{-1} = \mathbf{P}_k^\top = -\mathbf{P}_k$. Therefore, $\mathbf{P}_k\mathbf{P}_k = -\mathbf{I}$ for all $k \neq 0$.
4. We have $\langle \mathbf{P}_k\mathbf{e}_i, \mathbf{P}_l\mathbf{e}_i \rangle = 0$ for all i and $k \neq l$. Together with (3), it follows that the diagonal elements of $\mathbf{P}_k\mathbf{P}_l$ are zero.

5. We have $\langle \mathbf{P}_k\mathbf{e}_i, \mathbf{P}_l\mathbf{e}_j\rangle = -\langle \mathbf{P}_k\mathbf{e}_j, \mathbf{P}_l\mathbf{e}_i\rangle$ for all $i \neq j$, and $k \neq l$, with $k, l \neq 0$. Together with point (4) it follows that $(\mathbf{P}_k\mathbf{P}_l)^\top = -\mathbf{P}_k\mathbf{P}_l$. Similarly, from point (3) we have that $(\mathbf{P}_k\mathbf{P}_l)^\top = \mathbf{P}_l^\top\mathbf{P}_k^\top = (-\mathbf{P}_l)(-\mathbf{P}_k) = \mathbf{P}_l\mathbf{P}_k$. Therefore, the two matrices are anti-commutative: $\mathbf{P}_k\mathbf{P}_l = -\mathbf{P}_l\mathbf{P}_k$.

We remark that points (1), (3), (5) coincide with the defining algebraic properties of the algebra of complex numbers, quaternions, octonions, sedenions and their generalisations based on the Cayley-Dickson construction [57]. Therefore, the matrices in $\mathcal{P}$ must be a representation of one of these algebras. Firstly, such algebras exist only for d that are powers of two. Secondly, matrix representations for these algebras exist only in dimension two and four. This is because the algebra of octonions and their generalisations, unlike matrix multiplication, is non-associative. As a sanity check, note that the matrices $\mathbf{R}_1, \mathbf{R}_2$ from Lemma 27 are a well-known representation of the unit complex numbers.

We conclude this section by giving out the matrices for $d = 4$, which are the real matrix representations of the four unit quaternions:

$$\mathbf{R}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{R}_2 = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \tag{19}$$

$$\mathbf{R}_3 = \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}, \quad \mathbf{R}_4 = \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

It can be checked that these matrices respect the properties outlined above. Thus, in $d = 4$, we can select the transport maps from the set $\{\pm\mathbf{R}_1, \pm\mathbf{R}_2, \pm\mathbf{R}_3, \pm\mathbf{R}_4\}$ containing eight matrices, which also form a group. Therefore, following the same procedure as in Lemma 27, we can linearly separate up to eight classes. □

Proposition 28. *Let $\mathcal{G}$ be the class of connected regular graphs with a finite number of classes. Then, $\mathcal{H}^2_{\text{orth}}(G)$ has linear separation power over $\mathcal{G}$.*

Proof. Idea: Since the graph is regular, the harmonic features of the nodes will be uniformly scaled and thus positioned on a circle. The aim is to place different classes at different locations on the circle, which would make the classes linearly separable as shown in Figure 6b.

Let G be a regular graph with C classes and define $\theta = \frac{2\pi}{C}$. Denote by $\mathbf{R}_i$ the 2D rotation matrix:

$$\mathbf{R}_i = \begin{bmatrix} \cos(i\theta) & -\sin(i\theta) \\ \sin(i\theta) & \cos(i\theta) \end{bmatrix} \tag{20}$$

Then let $\mathcal{P} = \{\mathbf{R}_i \mid 0 \leq i \leq C-1,\ i \in \mathbb{N}\}$ the set of rotation matrices with an angle multiple of θ. Then we can define a bijection $g : \mathcal{C} \to \mathcal{P}$ and a sheaf $(G, \mathcal{F}) \in \mathcal{H}^2_{\text{orth}}(G)$ as in the proof above. Checking the inner-products from Equation 18 between the harmonic features of the nodes, we can verify that the angle between any two classes is different from zero. By Lemma 26, the cross terms of the inner product vanish:

$$\sum_k \alpha_k^2 \langle \mathbf{R}_i[k], \mathbf{R}_j[k]\rangle = \sum_k \alpha_k^2 \cos((i-j)\theta) = \cos((i-j)\theta)\|\mathbf{x}\|^2 \tag{21}$$

Thus, the angle between classes i, j is $(i-j)\theta$. □

C Energy Flow Proofs

Proposition 29. *If $\mathcal{F}$ is an $O(d)$-bundle in $\mathcal{H}^d_{\text{orth,sym}}$, then $\mathbf{x} \in \ker\Delta_{\mathcal{F}}$ if and only if $\mathbf{x}^k \in \ker\Delta_0$ for all $1 \leq k \leq d$.*

Proof of Proposition 29. Let $\mathbf{x} \in H^0(G, \mathcal{F})$. Then we have

$$\begin{aligned} 0 = E_{\mathcal{F}}(\mathbf{x}) &= \frac{1}{2} \sum_{(v,u)\in E} ||\mathcal{F}_{v \trianglelefteq e} D_v^{-\frac{1}{2}} \mathbf{x}_v - \mathcal{F}_{u \trianglelefteq e} D_u^{-\frac{1}{2}} \mathbf{x}_u||^2 \\ &= \frac{1}{2} \sum_{(v,u)\in E} ||\mathcal{F}_e \left(D_v^{-\frac{1}{2}} \mathbf{x}_v - D_u^{-\frac{1}{2}} \mathbf{x}_u \right)||^2 \\ &= \frac{1}{2} \sum_{(v,u)\in E} ||d_v^{-\frac{1}{2}} \mathbf{x}_v - d_u^{-\frac{1}{2}} \mathbf{x}_u||^2. \end{aligned}$$

The last term vanishes if and only if $\mathbf{x}^k \in \ker \Delta_0$ for each $1 \leq k \leq d$. □

Proposition 17. *For any connected graph G and $\varepsilon > 0$, there exist a sheaf $(G, \mathcal{F}) \notin \mathcal{H}^d_{\text{sym}}$, $\mathbf{W}_1$ with $\|\mathbf{W}_1\|_2 < \varepsilon$ and feature vector $\mathbf{x}$ such that $E_{\mathcal{F}}((\mathbf{I} \otimes \mathbf{W}_1)\mathbf{x}) > E_{\mathcal{F}}(\mathbf{x})$.*

Proof. Let $\mathcal{F}$ be an $O(d)$-bundle over G and $\varepsilon > 0$. Assume that $\mathcal{F}_{v \trianglelefteq e} = \mathcal{F}_{u \trianglelefteq e}$ for each $(u, v) \neq (u_0, v_0)$ and that $\mathcal{F}^\top_{v_0 \trianglelefteq e} \mathcal{F}_{u_0 \trianglelefteq e} - \mathbf{I} := \mathbf{B} \neq 0$ with $\dim(\ker(\mathbf{B})) > 0$. Then there exist a linear map $\mathbf{W} \in \mathbb{R}^{d \times d}$ with $||\mathbf{W}||_2 = \varepsilon$ and $\mathbf{x} \in H^0(G, \mathcal{F})$ such that $E_{\mathcal{F}}((\mathbf{I} \otimes \mathbf{W})\mathbf{x}) > 0$. We sketch the proof. Let $\mathbf{g} \in \ker(\mathbf{B})$. Define then $\mathbf{x} \in C^0(G, \mathcal{F})$ by

$$\mathbf{x}_v = \sqrt{d_v}\mathbf{g}.$$

Then $\mathbf{x} \in H^0(G, \mathcal{F})$. If we now take $\mathbf{W} = \varepsilon \mathbf{P}_{\ker \mathbf{B}^\perp}$ the rescaled orthogonal projection in the orthogonal complement of the kernel of $\mathbf{B}$ we verify the given claim. □

We provide below a proof for the equality in Definition 14.

Proposition 30.

$$\mathbf{x}^\top \Delta_{\mathcal{F}} \mathbf{x} = \frac{1}{2} \sum_{e:=(v,u)} \|\mathcal{F}_{v \trianglelefteq e} D_v^{-1/2} \mathbf{x}_v - \mathcal{F}_{u \trianglelefteq e} D_u^{-1/2} \mathbf{x}_u\|_2^2$$

Proof. We prove the result for the normalised sheaf Laplacian, and other versions can be obtained as particular cases.

$$E(\mathbf{x}) = \mathbf{x}^\top \Delta_{\mathcal{F}} \mathbf{x} = \sum_v \mathbf{x}_v^\top \Delta_{vv} \mathbf{x}_v + \sum_{\substack{w \neq z \\ (w,z)\in E}} \mathbf{x}_w^\top \Delta_{wz} \mathbf{x}_z \tag{22}$$

$$= \sum_{v \trianglelefteq e} \mathbf{x}_v^\top D_v^{-1/2} \mathcal{F}^\top_{v \trianglelefteq e} \mathcal{F}_{v \trianglelefteq e} D_v^{-1/2} \mathbf{x}_v + \sum_{\substack{w < z \\ (w,z)\in E}} \mathbf{x}_w^\top \Delta_{wz} \mathbf{x}_z + \mathbf{x}_z^\top \Delta_{zw} \mathbf{x}_w \tag{23}$$

$$= \frac{1}{2} \sum_{v,w \trianglelefteq e} \Big(\mathbf{x}_v^\top D_v^{-1/2} \mathcal{F}^\top_{v \trianglelefteq e} \mathcal{F}_{v \trianglelefteq e} D_v^{-1/2} \mathbf{x}_v + \mathbf{x}_w^\top D_w^{-1/2} \mathcal{F}^\top_{w \trianglelefteq e} \mathcal{F}_{w \trianglelefteq e} D_w^{-1/2} \mathbf{x}_w \tag{24}$$

$$+ \mathbf{x}_v^\top D_v^{-1/2} \mathcal{F}^\top_{v \trianglelefteq e} \mathcal{F}_{w \trianglelefteq e} D_w^{-1/2} \mathbf{x}_w + \mathbf{x}_w^\top D_w^{-1/2} \mathcal{F}^\top_{w \trianglelefteq e} \mathcal{F}_{v \trianglelefteq e} D_v^{-1/2} \mathbf{x}_v \Big) \tag{25}$$

$$= \frac{1}{2} \sum_{v,w \trianglelefteq e} \mathbf{x}_v^\top D_v^{-1/2} \mathcal{F}^\top_{v \trianglelefteq e} \big(\mathcal{F}_{v \trianglelefteq e} D_v^{-1/2} \mathbf{x}_v - \mathcal{F}_{w \trianglelefteq e} D_w^{-1/2} \mathbf{x}_w \big) \tag{26}$$

$$- \mathbf{x}_w^\top D_w^{-1/2} \mathcal{F}^\top_{w \trianglelefteq e} \big(\mathcal{F}_{v \trianglelefteq e} D_v^{-1/2} \mathbf{x}_v - \mathcal{F}_{w \trianglelefteq e} D_w^{-1/2} \mathbf{x}_w \big) \tag{27}$$

$$= \frac{1}{2} \sum_{v,w \trianglelefteq e} \big(\mathbf{x}_v^\top D_v^{-1/2} \mathcal{F}^\top_{v \trianglelefteq e} - \mathbf{x}_w^\top D_w^{-1/2} \mathcal{F}^\top_{w \trianglelefteq e} \big) \big(\mathcal{F}_{v \trianglelefteq e} D_v^{-1/2} \mathbf{x}_v - \mathcal{F}_{w \trianglelefteq e} D_w^{-1/2} \mathbf{x}_w \big) \tag{28}$$

Note that D_v is symmetric for any node v and so is any $D_v^{-1/2}$. Therefore, the two vectors in the parenthesis are the transpose of each other and the result is their inner product. Thus, we have:

$$E_{\mathcal{F}}(\mathbf{x}) = \frac{1}{2} \sum_{v,w \trianglelefteq e} \|\mathcal{F}_{v \trianglelefteq e} D_v^{-1/2} \mathbf{x}_v - \mathcal{F}_{w \trianglelefteq e} D_w^{-1/2} \mathbf{x}_w\|_2^2 \tag{29}$$

The result follows identically for other types of Laplacian. For the augmented normalized Laplacian, one should simply replace D with $\tilde{D} = D + I$ and for the non-normalised Laplacian, one should simply remove D from the equation. □

Theorem 16. *If $(\mathcal{F}, G) \in \mathcal{H}^d_{\text{orth,sym}}$ and σ = (Leaky)ReLU, $E_{\mathcal{F}}(\mathbf{Y}) \leq \lambda_* \|\mathbf{W}_1\|_2^2 \|\mathbf{W}_2^\top\|_2^2 E_{\mathcal{F}}(\mathbf{X})$.*

Proof. We first prove a couple of Lemmas before proving the Theorem. The structure of the proof follows that of Cai and Wang [12], which in turn generalises that of Oono and Suzuki [49]. The latter proof technique is not directly applicable to our setting because it makes some strong assumptions about the harmonic space of the Laplacian (i.e. that the eigenvectors of the harmonic space have positive entries).

$\lambda_* = \max\big((\lambda_{\min} - 1)^2, (\lambda_{\max} - 1)^2\big)$, where $\lambda_{\min}, \lambda_{\max}$ are the smallest and largest non-zero eigenvalues of $\Delta_{\mathcal{F}}$.

Lemma 31. *For $\mathbf{P} = \mathbf{I} - \Delta_{\mathcal{F}}$, $E(\mathbf{P}\mathbf{x}) < \lambda_* E_{\mathcal{F}}(\mathbf{x})$.*

Proof. We can write $\mathbf{x} = \sum_i c_i \mathbf{h}^i$ as a sum of the eigenvectors $\{\mathbf{h}^i\}$ of $\Delta_{\mathcal{F}}$. Then $\mathbf{x}^\top \Delta_{\mathcal{F}} \mathbf{x} = \sum_i c_i^2 \lambda_i \mathbf{h}^i$, where $\{\lambda_i\}$ are the eigenvalues of $\Delta_{\mathcal{F}}$. Using this for $E_{\mathcal{F}}(\mathbf{P}\mathbf{h})$:

$$E_{\mathcal{F}}(\mathbf{P}\mathbf{x}) = \mathbf{x}^\top \mathbf{P}^\top \Delta \mathbf{P} \mathbf{x} = \mathbf{x}^\top \mathbf{P} \Delta \mathbf{P} f = \sum_i c_i^2 \lambda_i (1 - \lambda_i)^2 \leq \lambda_* \sum_i c_i^2 \lambda_i = \lambda_* E_{\mathcal{F}}(f) \tag{30}$$

The inequality follows from the fact that the eigenvectors of the normalied sheaf Laplacian are in the range $[0, 2]$ [32, Proposition 5.5]. We note that the original proof of Cai and Wang [12] bounds the expression by $(1 - \lambda_{\min})^2$ instead of λ_*, which appears to be an error. □

Lemma 32. $E_{\mathcal{F}}(\mathbf{X}\mathbf{W}) \leq \|\mathbf{W}^\top\|_2^2 E_{\mathcal{F}}(\mathbf{X})$

Proof. Following the proof of Cai and Wang [12] we have:

$$\begin{aligned}
E_{\mathcal{F}}(\mathbf{X}\mathbf{W}) &= \operatorname{Tr}(\mathbf{W}^\top \mathbf{X}^\top \Delta_{\mathcal{F}} \mathbf{X} \mathbf{W}) && (31)\\
&= \operatorname{Tr}(\mathbf{X}^\top \Delta_{\mathcal{F}} \mathbf{X} \mathbf{W} \mathbf{W}^\top) && \text{trace cyclic property} \quad (32)\\
&\leq \operatorname{Tr}(\mathbf{X}^\top \Delta_{\mathcal{F}} \mathbf{X}) \|\mathbf{W}\mathbf{W}^\top\|_2 && \text{see Lemma 3.1 in Lee [38]} \quad (33)\\
&= \operatorname{Tr}(\mathbf{X}^\top \Delta_{\mathcal{F}} \mathbf{X}) \|\mathbf{W}^\top\|_2^2 && (34)
\end{aligned}$$

□

Lemma 33. *For conditions as in Theorem 16, $E\mathcal{F}\big((\mathbf{I}_n \otimes \mathbf{W})\mathbf{x}\big) \leq \|\mathbf{W}\|_2^2 E(f)$.*

Proof. First, we note that for orthogonal matrices, $D_v = \mathbf{I} \sum_{v \trianglelefteq e} \alpha_e^2 = \mathbf{I} d_v$ [32, Lemma 4.4]

$$\begin{aligned}
E_{\mathcal{F}}\big((\mathbf{I} \otimes \mathbf{W})\mathbf{x}\big) &= \frac{1}{2} \sum_{v, w \trianglelefteq e} \|\mathcal{F}_{v \trianglelefteq e} D_v^{-1/2} \mathbf{W} f_v - \mathcal{F}_{w \trianglelefteq e} D_w^{-1/2} \mathbf{W} \mathbf{x}_w\|_2^2 && (35)\\
&= \frac{1}{2} \sum_{v, w \trianglelefteq e} \|\mathcal{F}_e \mathbf{W} \big(d_v^{-1/2} \mathbf{x}_v - d_w^{-1/2} \mathbf{x}_w\big)\|_2^2 && (36)\\
&= \frac{1}{2} \sum_{v, w \trianglelefteq e} \|\mathbf{W} \big(d_v^{-1/2} \mathbf{x}_v - d_w^{-1/2} \mathbf{x}_w\big)\|_2^2 && \mathcal{F}_e \text{ is orthogonal} \quad (37)\\
&\leq \frac{1}{2} \sum_{v, w \trianglelefteq e} \|\mathbf{W}\|_2^2 \|d_v^{-1/2} \mathbf{x}_v - d_w^{-1/2} \mathbf{x}_w\|_2^2 && (38)\\
&= \frac{1}{2} \sum_{v, w \trianglelefteq e} \|\mathbf{W}\|_2^2 \|\mathcal{F}_e \big(d_v^{-1/2} \mathbf{x}_v - d_w^{-1/2} \mathbf{x}_w\big)\|_2^2 && \mathcal{F}_e \text{ is orthogonal} \quad (39)\\
&= \frac{1}{2} \|\mathbf{W}\|_2^2 \sum_{v, w \trianglelefteq e} \|\mathcal{F}_e \big(D_v^{-1/2} \mathbf{x}_v - D_w^{-1/2} \mathbf{x}_w\big)\|_2^2 && (40)\\
&= \|\mathbf{W}\|_2^2 E_{\mathcal{F}}(\mathbf{x}) && (41)
\end{aligned}$$

The proof can also be extended easily extended to vector bundles over weighted graphs (i.e. allowing weighted edges as in Hansen and Ghrist [32]). For the non-normalised Laplacian, the assumption that $\mathcal{F}_e$ is orthogonal can be relaxed to being non-singular and then the upper bound will also depend on the maximum conditioning number over all $\mathcal{F}_e$. □

Lemma 34. *For conditions as in Theorem 16,* $E_{\mathcal{F}}\big(\sigma(\mathbf{x})\big) \leq E(\mathbf{x})$.

Proof.

$$
\begin{aligned}
E\big(\sigma(\mathbf{x})\big) &= \frac{1}{2}\sum_{v,w\trianglelefteq e} \|\mathcal{F}_{v\trianglelefteq e} D_v^{-1/2}\sigma(\mathbf{x}_v) - \mathcal{F}_{w\trianglelefteq e} D_w^{-1/2}\sigma(\mathbf{x}_w)\|_2^2 && (42)\\
&= \frac{1}{2}\sum_{v,w\trianglelefteq e} \|\mathcal{F}_e\big(d_v^{-1/2}\sigma(\mathbf{x}_v) - d_w^{-1/2}\sigma(\mathbf{x}_w)\big)\|_2^2 && (43)\\
&= \frac{1}{2}\sum_{v,w\trianglelefteq e} \|d_v^{-1/2}\sigma(\mathbf{x}_v) - d_w^{-1/2}\sigma(\mathbf{x}_w)\|_2^2 && \text{orthogonality of } \mathcal{F}_e \quad (44)\\
&= \frac{1}{2}\sum_{v,w\trianglelefteq e} \|\sigma\Big(\frac{\mathbf{x}_v}{\sqrt{d_v}}\Big) - \sigma\Big(\frac{\mathbf{x}_w}{\sqrt{d_w}}\Big)\|_2^2 && c\text{ReLU}(x) = \text{ReLU}(cx), c>0 \quad (45)\\
&\leq \frac{1}{2}\sum_{v,w\trianglelefteq e} \|\frac{\mathbf{x}_v}{\sqrt{d_v}} - \frac{\mathbf{x}_w}{\sqrt{d_w}}\|_2^2 && \text{Lipschitz continuity of ReLU} \quad (46)\\
&= \frac{1}{2}\sum_{v,w\trianglelefteq e} \|\mathcal{F}_e\big(d_v^{-1/2}\mathbf{x}_v - d_w^{-1/2}\mathbf{x}_w\big)\|_2^2 && \text{orthogonality of } \mathcal{F}_e \quad (47)\\
&= E_{\mathcal{F}}(\mathbf{x}) && (48)
\end{aligned}
$$

□

Combining these three lemmas for an entire diffusion layer proves the Theorem. □

Theorem 15. *For* $(\mathcal{F}, G) \in \mathcal{H}_+^1$ *and* σ *being (Leaky)ReLU,* $E_{\mathcal{F}}(\mathbf{Y}) \leq \lambda_* \|\mathbf{W}_1\|_2^2 \|\mathbf{W}_2^\top\|_2^2 E_{\mathcal{F}}(\mathbf{X})$.

Proof. If $d = 1$, then Lemma 33 becomes superfluous as $\mathbf{W}_1$ becomes a scalar that can be absorbed into the right-weights. It remains to verify that a version of Lemma 34 holds in this case.

Lemma 35. *For conditions as in Theorem 15,* $E_{\mathcal{F}}\big(\sigma(\mathbf{x})\big) \leq E(\mathbf{x})$.

Proof.

$$
\begin{aligned}
E\big(\sigma(\mathbf{x})\big) &= \frac{1}{2}\sum_{v,w\trianglelefteq e} \|\mathcal{F}_{v\trianglelefteq e} D_v^{-1/2}\sigma(x_v) - \mathcal{F}_{w\trianglelefteq e} D_w^{-1/2}\sigma(x_w)\|_2^2 && (49)\\
&= \frac{1}{2}\sum_{v,w\trianglelefteq e} \||\mathcal{F}_{v\trianglelefteq e}| D_v^{-1/2}\sigma(x_v) - |\mathcal{F}_{w\trianglelefteq e}| D_w^{-1/2}\sigma(x_w)\|_2^2 && \mathcal{F}_{v\trianglelefteq e}\mathcal{F}_{w\trianglelefteq e} > 0 \quad (50)\\
&= \frac{1}{2}\sum_{v,w\trianglelefteq e} \|\sigma\Big(\frac{|\mathcal{F}_{v\trianglelefteq e}|x_v}{\sqrt{d_v}}\Big) - \sigma\Big(\frac{|\mathcal{F}_{w\trianglelefteq e}|x_w}{\sqrt{d_w}}\Big)\|_2^2 && c\sigma(x) = \sigma(cx), c>0 \quad (51)\\
&\leq \frac{1}{2}\sum_{v,w\trianglelefteq e} \|\frac{|\mathcal{F}_{v\trianglelefteq e}|x_v}{\sqrt{d_v}} - \frac{|\mathcal{F}_{w\trianglelefteq e}|x_w}{\sqrt{d_w}}\|_2^2 && \text{ReLU Lipschitz cont.} \quad (52)\\
&= \frac{1}{2}\sum_{v,w\trianglelefteq e} \|\mathcal{F}_{v\trianglelefteq e} D_v^{-1/2}x_v - \mathcal{F}_{w\trianglelefteq e} D_w^{-1/2}x_w\|_2^2 && \mathcal{F}_{v\trianglelefteq e}\mathcal{F}_{w\trianglelefteq e} > 0 \quad (53)\\
&= E_{\mathcal{F}}(\mathbf{x}) && (54)
\end{aligned}
$$

□

We note that if $\mathcal{F}_{v \trianglelefteq e}\mathcal{F}_{w \trianglelefteq e} < 0$ (i.e. the reation is signed), then it is very easy to find counter-examples where ReLU does not work anymore. However, the result still holds in the deep linear case. $\square$

If the features of a SCN/GCN oversmoothing as in Theorem 15 converge to $\ker(\Delta_{\mathcal{F}})$, then the model will no longer be able to linearly separate the classes. This is shown by the following Corollaries.

Corollary 36. *Consider an SCN model f with k layers and a sheaf $(\mathcal{F}; G) \in \mathcal{H}^1_{\text{sym}}$ over a bipartite graph G as in Proposition 9. Then for any finite k, $\mathbf{Y} := f(\mathbf{X})$ is not linearly separable for any input with $E_{\mathcal{F}}(\mathbf{X}) = 0$.*

Proof. By Theorem 15, if $E_{\mathcal{F}}(\mathbf{X}) = 0$, then $E_{\mathcal{F}}(\mathbf{Y}) = 0$ and, therefore, $\mathbf{Y}^i \in \ker(\Delta_{\mathcal{F}})$ for any column i. The proof of Proposition 9 showed that the classes of such a bipartite graph cannot be linearly separated for any such feature matrix $\mathbf{Y}$. $\square$

Corollary 37. *Consider an SCN model f with k layers and a sheaf $(\mathcal{F}; G) \in \mathcal{H}^1_{+}$ over any graph G with more than two classes as in Proposition 11. Then for any finite k, $\mathbf{Y} := f(\mathbf{X})$ is not linearly separable for any input with $E_{\mathcal{F}}(\mathbf{X}) = 0$.*

Proof. By Theorem 15, if $E_{\mathcal{F}}(\mathbf{X}) = 0$, then $E_{\mathcal{F}}(\mathbf{Y}) = 0$ and, therefore, $\mathbf{Y}^i \in \ker(\Delta_{\mathcal{F}})$ for any column i. The proof of Proposition 11 showed that the classes of such a graph cannot be linearly separated for any such feature matrix $\mathbf{Y}$. $\square$

D Sheaf Learning Proofs

Proposition 18. *Let $G = (V, E)$ be a finite graph with features $\mathbf{X}$. Then, if $(\mathbf{x}_v, \mathbf{x}_u) \neq (\mathbf{x}_w, \mathbf{x}_z)$ for any $(v, u) \neq (w, z) \in E$ and Φ is an MLP with sufficient capacity, Φ can learn any sheaf $(\mathcal{F}; G)$.*

Proof. Assume that the node features are k-dimensional and, therefore, the graph feature matrix has shape $\mathbf{X} \in \mathbb{R}^{n \times k}$. Define the finite set $A := \{(\mathbf{x}_v, \mathbf{x}_u) : v \to u \in E\} \subset \mathbb{R}^{2k}$ containing the concatenated features of the nodes for all the oriented edges $v \to u$ of the graph. Then, because each $(\mathbf{x}_v, \mathbf{x}_u)$ is unique, for any dimension d, there exists a (well-defined) function $g : A \to \mathbb{R}^{d \times d}$ sending $(\mathbf{x}_v, \mathbf{x}_u) \mapsto \mathcal{F}_{v \trianglelefteq e = (v,u)}$. We now show that this function can be extended to a smooth function $f : \mathbb{R}^{2k} \to \mathbb{R}^{d \times d}$ and, therefore, it can be approximated by an MLP due to the Universal Approximation Theorem [34, 35].

Let I be an index-set for the elements of A. Then, because A is finite, for any $\mathbf{a}_{i \in I}$, we can find a sufficiently small neighbourhood $U_i \subset \mathbb{R}^{2k}$ such that $\mathbf{a}_i \in U_i$ and $\mathbf{a}_j \notin U_i$ for $j \neq i \in I$. Furthermore, for each $i \in I$, we can find a (smooth) bump function $\psi_i : \mathbb{R}^{2k} \to \mathbb{R}$ such that $\psi_i(\mathbf{a}_i) = 1$ and $\psi_i(\mathbf{a}) = 0$ if $\mathbf{a} \notin U_i$. Then, the function $f(\mathbf{a}) := \sum_{i \in I} g(\mathbf{a}_i)\psi_i(\mathbf{a})$ is smooth and $f|_A = g$. $\square$

E Additional model details and hyperparameters

Hybrid transport maps. Consider the transport maps $-\mathcal{F}^{\top}_{v \trianglelefteq e}\mathcal{F}_{u \trianglelefteq e}$ appearing in the off the diagonal entires of the sheaf Laplacian $L_{\mathcal{F}}$. When learning a sheaf Laplacian there exists the risk that the features are not sufficiently good in the early layers (or in general) and, therefore, it might be useful to consider a hybrid transport map of the form $-\mathcal{F}^{\top}_{v \trianglelefteq e}\mathcal{F}_{u \trianglelefteq e} \bigoplus \mathbf{F}$, where $\bigoplus$ is the direct sum of two matrices and $\mathbf{F}$ represents a fixed (non-learnable map). In particular we consider maps of the form $-\mathcal{F}^{\top}_{v \trianglelefteq e}\mathcal{F}_{u \trianglelefteq e} \bigoplus \mathbf{I}_1 \bigoplus -\mathbf{I}_1$ which essentially appends a diagonal matrix with 1 and -1 on the diagonal to the learned matrix. From a signal processing perspective, these correspond to a low-pass and a high-pass filter that could produce generally useful features. We treat the addition of these fixed parts as an additional hyper-parameter.

Adjusting the activation magnitudes. We note that in practice we find it useful to learn an additional parameter $\varepsilon \in [-1, 1]^d$ (i.e. a vector of size d) in the discrete version of the models:

$$\mathbf{X}_{t+1} = (1 + \varepsilon)\mathbf{X}_t - \sigma\Big(\Delta_{\mathcal{F}(t)}(\mathbf{I} \otimes \mathbf{W}^t_1)\mathbf{X}_t\mathbf{W}^t_2\Big). \tag{55}$$

This allows the model to adjust the relative magnitude of the features in each stalk dimension. This is used across all of our experiments in the discrete models.

Augmented normalised sheaf Laplacian. Similarly to GCN which normalises the Laplacian by the augmented degrees (i.e. $(\mathbf{D}+\mathbf{I}_n)^{-1/2}$, where $\mathbf{D}$ is the usual diagonal matrix of node degrees), we similarly use $(D+\mathbf{I}_{nd})^{-1/2}$ for normalisation to obtain greater numerical stability. This is particularly helpful when learning general sheaves as it increases the numerical stability of SVD.

Table 2: Hyper-parameter ranges for the discrete and continous models.

	Discrete Models	**Continous Models**
Hidden channels	$(8, 16, 32)$ (WebKB) and $(8, 16, 32, 64)$ (others)	$(8, 16, 32, 64)$
Stalk dim d	$1-5$	$1-5$
Layers	$2-8$	N/A
Learning rate	0.02 (WebKB) and 0.01 (others)	Log-uniform $[0.01, 0.1]$
Activation	ELU	ELU
Weight decay (regular parameters)	Log-uniform $[-4.5, 11.0]$	Log-uniform $[-6.9, 13.8]$
Weight decay (sheaf parameters)	Log-uniform $[-4.5, 11.0]$	Log-uniform $[-6.9, 13.8]$
Input dropout	Uniform $[0, 0.9]$	Uniform $[0, 0.9]$
Layer dropout	Uniform $[0, 0.9]$	N/A
Patience (epochs)	100 (Wiki) and 200 (others)	50
Max training epochs	1000 (Wiki) and 500 (others)	50.
Integration time	N/A	Uniform $[1.0, 9.0]$.
Optimiser	Adam [36]	Adam

Hyperparameters and training procedure. We train all models for a fixed maximum number of epochs and perform early stopping when the validation metric has not improved for a pre-specified number of patience epochs. We report the results at the epoch where the best validation metric was obtained for the model configuration with the best validation score among all models. We use the hyperparameter optimisation tools provided by Weights and Biases [4] for this procedure. The complete hyperparameter ranges we optimised over can be found in Table 2. All models were trained and fine-tuned on an Amazon AWS p2.xlarge machine containing 8 NVIDIA K80 GPUs and using a 2.3 GHz (base) and 2.7 GHz (turbo) Intel Xeon E5-2686 v4 Processor.

E.1 Computational Complexity

We can split the computational complexity in the following computational steps:

1. **The linear transformation $\mathbf{X}' = (\mathbf{I} \otimes \mathbf{W}_1^t)\mathbf{X}_t\mathbf{W}_2^t$.** $\mathbf{W}_1$ is a $d \times d$ matrix and $\mathbf{W}_2$ is an $f \times f$ matrix. Therefore, the complexity is $\mathcal{O}\left(n(d^2f + df^2)\right) = \mathcal{O}\left(n(cd + cf)\right) = \mathcal{O}(nc^2)$.
2. **Message Passing.** Since $\Delta_{\mathcal{F}}$ is a sparse matrix, the message passing is implemented as a sparse-dense matrix multiplication $\Delta_{\mathcal{F}}\mathbf{X}'$. When the restriction maps are diagonal, the complexity of this operation is $\mathcal{O}\left(mc\right)$, since the multiplication of each block matrix in $\Delta_{\mathcal{F}}$ and block vector in $\mathbf{X}'$ reduces to a an element-wise vector multiplication. When the restriction maps are non-diagonal, the complexity is $\mathcal{O}\left(mdc\right)$ because each matrix-vector multiplication is $\mathcal{O}(d^2)$ and we need to perform f of them for each node and edge.
3. **Learning the Sheaf.** Assume we learn the restriction maps via $\Phi(\mathbf{x}_v, \mathbf{x}_u) = \sigma(\widetilde{\mathbf{W}}[\mathrm{vec}(\mathbf{X}_v)||\mathrm{vec}(\mathbf{X}_u)])$, where $\mathrm{vec}(\cdot)$ converts the $d \times f$ matrix into a df-sized vector. This operation has to be performed for each incident node-edge pair. Therefore, the complexity is $\mathcal{O}(md^2f) = \mathcal{O}(mcd)$, when learning diagonal maps since $\widetilde{\mathbf{W}}$ is a $d \times 2df$ matrix. When learning a non-diagonal matrix, the number of rows of $\widetilde{\mathbf{W}}$ is $\mathcal{O}(d^2)$ and the complexity becomes $\mathcal{O}(md^3f) = \mathcal{O}(mcd^2)$. Note, however, that in general, the complexity of learning the restriction maps can be significantly reduced to $\mathcal{O}(mc)$ (in the diagonal case) and $\mathcal{O}(m(c + d^2))$ (in the non-diagonal case) by, for instance, using an MLP with constant hidden-size.
4. **Constructing the Laplacian.** To build the Laplacian, we need to perform the matrix-matrix multiplications involved in computing each of the blocks. The complexity of that is $\mathcal{O}(md)$ in the diagonal case and $\mathcal{O}(md^3)$ in the non-diagonal case. Computing the normalisation of the Laplacian is $\mathcal{O}(nd)$ in the diagonal case and $\mathcal{O}(nd^3)$ in the non-diagonal case.

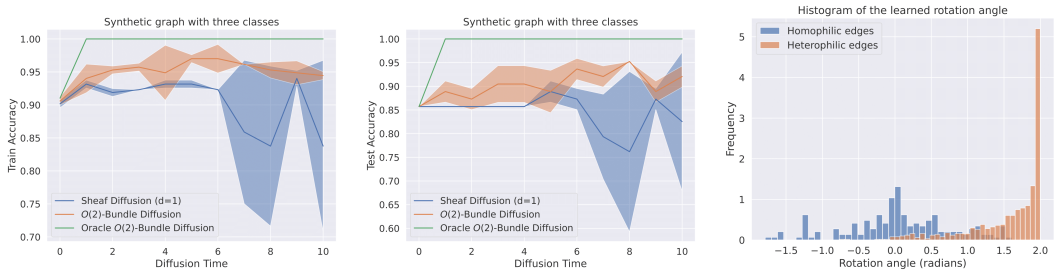


Figure 7: (*Left*) Train accuracy as a function of diffusion time. (*Middle*) Test accuracy as a function of diffusion time. (*Right*) Histogram of the learned rotation angle of the $2D$ transport maps. The performance of the bundle model is superior to that of the one-dimensional sheaf. The transport maps learned by the model are aligned with our expectation: the model learns to rotate more (i.e. to move away) the neighbours belonging to different classes than the neighbours belonging to the same class.

Putting everything together, the final complexity is $\mathcal{O}(nc^2 + mcd)$ in the diagonal case and $\mathcal{O}\left(n(c^2 + d^3) + m(cd^2 + d^3)\right)$ in the non-diagonal case. When learning the sheaf via an MLP with constant hidden-size, the complexity reduces to $\mathcal{O}(nc^2 + mc)$ (same as GCN) and $\mathcal{O}\left(n(c^2 + d^3) + m(c + d^3)\right)$, respectively.

For learning orthogonal matrices, we rely on the library Torch Householder [47] which provides support for fast transformations with large batch sizes.

F Additional Experiments

In this section we provide a series of additional experiments and abaltion studies.

Two dimensional synthetic experiment. In the main text we focused on a synthetic example involving sheaves with one-dimensional stalks. We now consider a graph with three classes and two-dimensional features, with edge homophily level 0.2. We use 80% of the nodes for training and 20% for testing. First, we know that a discrete vector bundle with two-dimensional stalks that can solve the task in the limit exists from Theorem 13, while based on Proposition 11 no sheaf with one-dimensional stalks can perfectly solve the tasks.

Therefore, similarly to the synthetic experiment in the main text, we compare two similar models learning the sheaf from data: one using $1D$ stakls and another using $2D$ stalks. As we see from Figure 7, the discrete vector bundle model has better training and test-time performance than the one-dimensional counterpart. Nonetheless, none of the two models manages to match the perfect performance of the ideal sheaf on this more challenging dataset. From the final subfigure we also see that the model learns to rotate more across the heterophilic edges in order to push away the nodes belonging to other classes. The prevalent angle of this rotation is 2 radians, which is just under $120^\circ = 360^\circ/C$, where $C = 3$ is the number of classes. Thus the model learns to position the three classes at approximately equal arc-lengths from each other for maximum linear separability.

Continous Models. To also understand how the continuous version of our models performs against other PDE-based GNNs we include a category of such SOTA models: CGNNs [67], GRAND [14], and BLEND [13]. Results are included in Table 3. Generally, continuous models do not perform as well as the discrete ones because they are constrained to use the same set of weights for the entire integration time and cannot use dropout. Therefore, the model capacity is difficult to increase without overfitting. Nonetheless, our continuous models generally outperform other state of the art continuous models, which also share the same limitations. Finally, we note that the baselines were fine-tuned over the same hyper-parameter ranges as in Table 2

Positional encoding ablation. Based on Proposition 18 we proceed to analyse the impact of increasing the expressive power of the model by making the nodes more distinguishable. For that, we equip our datasets with additional features consisting of graph Laplacian positional encodings as originally done in Dwivedi et al. [22]. In Table 4 we see that positional encodings do indeed

Table 3: Results on node classification datasets sorted by their homophily level. Top three models are coloured by **First**, **Second**, **Third**. Our models are marked **NSD**.

	Texas	Wisconsin	Film	Squirrel	Chameleon	Cornell	Citeseer	Pubmed	Cora
Hom level	**0.11**	**0.21**	**0.22**	**0.22**	**0.23**	**0.30**	**0.74**	**0.80**	**0.81**
#Nodes	183	251	7,600	5,201	2,277	183	3,327	18,717	2,708
#Edges	295	466	26,752	198,493	31,421	280	4,676	44,327	5,278
#Classes	5	5	5	5	5	5	7	3	6
Cont Diag-NSD	**82.97**±4.37	**86.47**±2.55	**36.85**±1.21	38.17±9.29	**62.06**±3.84	80.00±6.07	76.56±1.19	**89.47**±0.42	86.88±1.21
Cont O(d)-NSD	82.43±5.95	**84.50**±4.34	**36.39**±1.37	**40.40**±2.01	**63.18**±1.69	72.16±10.40	75.19±1.67	89.12±0.30	86.70±1.24
Cont Gen-NSD	**83.78**±6.62	**85.29**±3.31	**37.28**±0.74	**52.57**±2.76	**66.40**±2.28	**84.60**±4.69	**77.54**±1.72	**89.67**±0.40	**87.45**±0.99
BLEND	**83.24**±4.65	84.12±3.56	35.63±0.89	**43.06**±1.39	60.11±2.09	**85.95**±6.82	**76.63**±1.60	**89.24**±0.42	**88.09**±1.22
GRAND	75.68±7.25	79.41±3.64	35.62±1.01	40.05±1.50	54.67±2.54	**82.16**±7.09	76.46±1.77	89.02±0.51	**87.36**±0.96
CGNN	71.35±4.05	74.31±7.26	35.95±0.86	29.24±1.09	46.89±1.66	66.22±7.69	**76.91**±1.81	87.70±0.49	87.10±1.35

improve the performance of the continuous models compared to the numbers reported in the main table. Therefore, we conclude that the interaction between the problem of sheaf learning and that of the expressivity of graph neural networks represents a promising avenue of future research.

Table 4: Ablation study for the dimension of the stalks. Positional encodings improve performance on some of our models.

	Eigenvectors	Texas	Wisconsin	Cornell
Cont Diag-SD	0	82.97 ± 4.37	$\mathbf{86.47 \pm 2.55}$	80.00 ± 6.07
	2	3.51 ± 5.05	85.69 ± 3.73	81.62 ± 8.00
	8	$\mathbf{85.41 \pm 5.82}$	86.28 ± 3.40	$\mathbf{82.16 \pm 5.57}$
	16	82.70 ± 3.86	85.88 ± 2.75	81.08 ± 7.25
Cont $O(d)$-SD	0	82.43 ± 5.95	84.50 ± 4.34	72.16 ± 10.40
	2	84.05 ± 5.85	85.88 ± 4.62	83.51 ± 9.70
	8	$\mathbf{84.87 \pm 4.71}$	$\mathbf{86.86 \pm 3.83}$	$\mathbf{84.05 \pm 5.85}$
	16	83.78 ± 6.16	85.88 ± 2.88	83.51 ± 6.22
Cont Gen-SD	0	$\mathbf{83.78 \pm 6.62}$	85.29 ± 3.31	$\mathbf{84.60 \pm 4.69}$
	2	83.24 ± 4.32	84.12 ± 3.97	81.08 ± 7.35
	8	82.70 ± 5.70	84.71 ± 3.80	83.24 ± 6.82
	16	82.16 ± 6.19	$\mathbf{86.47 \pm 3.09}$	82.16 ± 6.07

Visualising diffusion. To develop a better intuition of the limiting behaviour of sheaf diffusion for node classification tasks we plot the diffusion process using an oracle discrete vector bundle for two graph with $C = 3$ (Figure 9) and $C = 4$ (Figure 8) classes. The diffusion processes converges in the limit to a configuration where the classes are rotated at $\frac{2\pi}{C}$ form each other, just like in the cartoon diagrams of Figure 6. Note that in all cases, the classes are linearly separable in the limit.

We note that this approach generalises to any number of classes, but beyond $C = 4$ it is not guaranteed that they will be linearly separable in $2D$. However, they are still well-separated. We include an example with $C = 10$ classes in Figure 10.

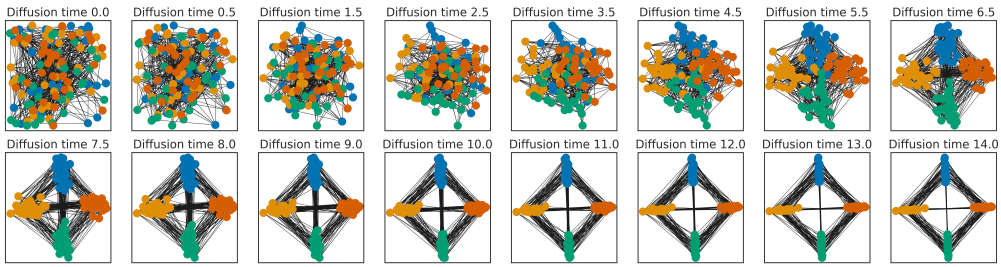


Figure 8: Sheaf diffusion process disentangling the $C = 4$ classes over time. The nodes are coloured by their class.

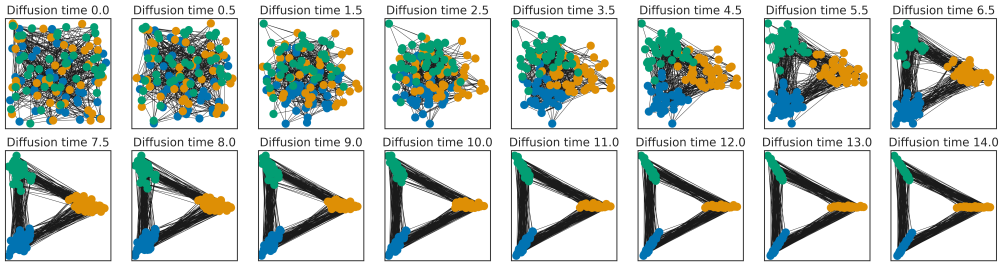


Figure 9: Sheaf diffusion process disentangling the $C = 3$ classes over time. The nodes are coloured by their class.

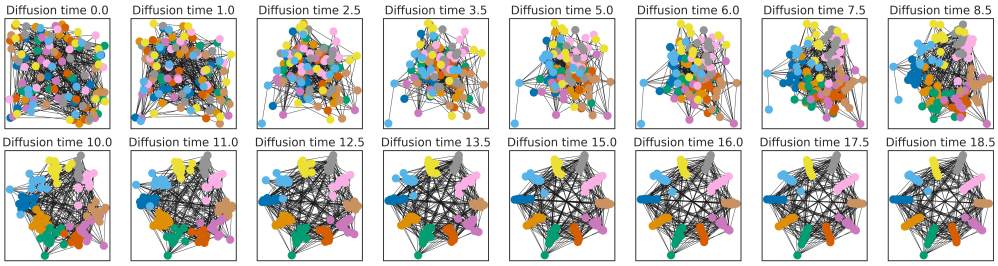


Figure 10: Sheaf diffusion process disentangling the $C = 10$ classes over time. The nodes are coloured by their class.