NEURAL SHEAF DIFFUSION: A TOPOLOGICAL PERSPECTIVE ON HETEROPHILY AND OVERSMOOTHING IN GNNS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cellular sheaves equip graphs with "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 state-of-the-art results in heterophilic settings. Overall, our work provides new connections between GNNs and algebraic topology and would be of interest to both fields.

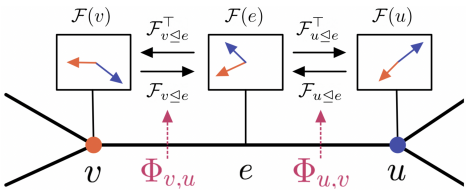


Figure 1: A sheaf $(G, \mathcal{F})$ shown for a single edge of the graph. The *stalks* are isomorphic to $\mathbb{R}^2$. The *restriction maps* $\mathcal{F}_{v \trianglelefteq e}$, $\mathcal{F}_{u \trianglelefteq e}$ and their adjoints move the vector features between these spaces. In practice, we learn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

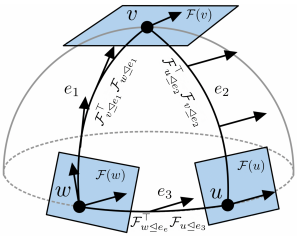


Figure 2: Analogy between parallel transport on a sphere and transport on a discrete vector bundle. A tangent vector is moved from $\mathcal{F}(w) \to \mathcal{F}(v) \to \mathcal{F}(u)$ and back.

1 BACKGROUND

Note. This represents a condensed version of a full paper on sheaves, heterophily and oversmoothing in GNNs. Therefore, the main content covers only some of our essential theoretical and experimental results. For additional introductory material, theory and experiments, please consult the Appendix.

Definition 1. *A cellular sheaf $(G, \mathcal{F})$ (Curry, 2014) 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 $e \in E$, and 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 are called* stalks, *while the linear maps are referred to as* restriction maps *(Figure 1).*

The space formed by all the spaces associated to the nodes of the graph is called the space of 0-*cochains* and is denoted by $C^0(G,\mathcal{F}) := \bigoplus_{v\in V}\mathcal{F}(v)$. Similarly, $C^1(G,\mathcal{F}) := \bigoplus_{e\in E}\mathcal{F}(e)$ – the space of 1-*cochains* – contains the data associated with all the edges of the graph.

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 and similarly for 1-cochains. From an opinion dynamics perspective (Hansen & Ghrist, 2021), $\mathbf{x}_v$ can be thought of as the private opinion of node v, while $\mathcal{F}_{v\trianglelefteq e}$ expressed how that opinion manifests publicly in a discourse space formed by $\mathcal{F}(e)$. It is natural to define a linear *co-boundary map* δ between $C^0(G,\mathcal{F})$ and $C^1(G,\mathcal{F})$, which measures the disagreement between all nodes.

Definition 2. *For some arbitrary choice of orientation for each edge $e = u \to v \in E$, $\delta\colon C^0(G,\mathcal{F}) \to C^1(G,\mathcal{F})$, $\delta(\mathbf{x})_e := \mathcal{F}_{v\trianglelefteq e}\mathbf{x}_v - \mathcal{F}_{u\trianglelefteq e}\mathbf{x}_u$.*

Given a cellular sheaf $(G,\mathcal{F})$, using the co-boundary operator δ, one can define a *Sheaf Laplacian* operator (Hansen & Ghrist, 2019) measuring the average disagreement between all the nodes.

Definition 3. *The sheaf Laplacian of a sheaf $(G,\mathcal{F})$ is a map $L_{\mathcal{F}} : C^0(G,\mathcal{F}) \to C^0(G,\mathcal{F})$ given by $L_{\mathcal{F}} := \delta^\top\delta$, $L_{\mathcal{F}}(\mathbf{x})_v := \sum_{v,u\trianglelefteq e}\mathcal{F}^\top_{v\trianglelefteq e}(\mathcal{F}_{v\trianglelefteq e}\mathbf{x}_v - \mathcal{F}_{u\trianglelefteq e}\mathbf{x}_u)$*

The sheaf Laplacian is a positive semi-definite block matrix. 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 $L_{\mathcal{F}vu} = -\mathcal{F}^\top_{v\trianglelefteq e}\mathcal{F}_{u\trianglelefteq e}$. A normalised version of this Laplacian can also be defined. Let D be the block-diagonal of $L_{\mathcal{F}}$. Then the normalised-sheaf Laplacian $\Delta_{\mathcal{F}} := D^{-1/2}L_{\mathcal{F}}D^{-1/2}$.

For simplicity, we assume from now on that all the stalks have 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 $\mathbb{R}$ and the linear maps to id, the underlying sheaf is trivial, $d = 1$ and one recovers the well-known $n \times n$ (normalised) graph Laplacian matrix. The harmonic space of $L_{\mathcal{F}}$ is characterised by the *global sections* of the sheaf.

Definition 4. *The global sections of a sheaf $H^0(G;\mathcal{F}) := \{\mathbf{x} \in C^0(G;\mathcal{F}) \mid \mathcal{F}_{v\trianglelefteq e}\mathbf{x}_v = \mathcal{F}_{u\trianglelefteq e}\mathbf{x}_u\}$*

This set corresponds to the signals that agree with the restriction maps globally (i.e. along all edges of the graph). The central theorem of discrete Hodge theory says that the space formed by all these 0-cochains is isomorphic to the kernel of the sheaf Laplacian.

The sheaves $(G,\mathcal{F})$ with orthogonal restriction maps (i.e. $\mathcal{F}_{v\trianglelefteq e} \in O(d)$ the Lie group of $d \times d$ orthogonal matrices), will play an important role in our analysis. Such sheaves are called *discrete $O(d)$-bundles* since they can be seen as a discrete equivalent of vector bundles from differential geometry (Tu, 2011). Intuitively, these are objects describing how vector spaces are attached to the points of a manifold. In our case, the role of the manifold is played by the graph. The sheaf Laplacian of a discrete $O(d)$-bundle is equivalent to a *connection Laplacian* (Singer & Wu, 2012), describing how the elements of a vector space are transported via rotations in another neighbouring space. This is analogous to how tangent vectors are transported across a manifold (see Figure 2).

2 THE SEPARATION POWER OF SHEAF DIFFUSION

Preliminaries Let $G = (V,E)$ be a graph and consider that all nodes have features that are d-dim 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-dim vectors. 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 $C^0(G;\mathcal{F})$. Finally, we are interested in the *sheaf diffusion* process governed by the PDE:

$$\mathbf{X}(0) = \mathbf{X}, \quad \dot{\mathbf{X}}(t) = -\Delta_{\mathcal{F}}\mathbf{X}(t) \tag{1}$$

We now analyse the ability of certain classes of sheaf Laplacian operators to linearly separate the features in the limit of the diffusion processes they induce. We consider this a proxy metric for the capacity of certain diffusion processes to avoid oversmoothing.

Definition 5. *A hypothesis class of sheaves with d-dimensional stalks $\mathcal{H}^d$ has linear separation power over a set 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 1 for a dense subset $\mathcal{X}_{\mathcal{F}} \subset \mathbb{R}^{nd\times f}$ of initial conditions.*

The restriction to a set of initial conditions that is dense in the ambient space is necessary because, in the limit, diffusion behaves like a projection in the harmonic space and there will always be degenerate initial conditions that will yield a zero projection.

Definition 6. *Consider the class of sheaves with symmetric and invertible transport maps and d-dimensional stalks:* $\mathcal{H}^d_{\text{sym}} := \{(\mathcal{F}, G) \mid \mathcal{F}_{v \trianglelefteq e} = \mathcal{F}_{u \trianglelefteq e},\ \det(\mathcal{F}_{v \trianglelefteq e}) \neq 0\}$

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 I). Therefore, this hypothesis class is of particular interest since it includes those graph Laplacians typically used by graph convolutional models such as GCN (Kipf & Welling, 2017) and ChebNet (Defferrard et al., 2016).

We first show that this class of sheaf Laplacians can linearly separate the classes in binary classification settings under certain homophily assumptions:

Proposition 7. *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 8. *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 $\mathbf{X}(0) \in \mathbb{R}^{n \times f}$.*

We now consider a larger hypothesis class that also includes non-symmetric relations.

Definition 9. $\mathcal{H}^d := \{(\mathcal{F}, G) \mid \det(\mathcal{F}_{v \trianglelefteq e}) \neq 0\}$

Proposition 10. *Let $\mathcal{G}$ contain all the connected graphs with two classes. Then, $\mathcal{H}^1$ has linear separation power over $\mathcal{G}$.*

These results show that heterophilic settings require an asymmetric transport of the features between neighbouring nodes belonging to different classes. This provides a sheaf-theoretic explanation for why a recent body of work (Yan et al., 2021; Chien et al., 2021; Bo et al., 2021) has found negatively-weighted edges to help in heterophilic settings. From this perspective, negatively-signed edges constrain the product $\mathcal{F}_{v \trianglelefteq e}\mathcal{F}_{u \trianglelefteq e}$ of an implicit underlying sheaf to be negative and hence $\mathcal{F}_{v \trianglelefteq e} \neq \mathcal{F}_{u \trianglelefteq e}$.

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 any $\mathbf{X} \in \mathbb{R}^{n \times f}$.*

This is essentially a consequence of $\dim\big(\ker(\Delta_{\mathcal{F}})\big)$ being at most one in this case. From a GNN perspective, this means that in the infinite depth setting, sufficient *stalk width* (i.e., dimension) is needed in order to solve tasks involving more than two classes.

Definition 12. *Consider the class of sheaves with diagonal invertible maps and d-dimensional stalks* $\mathcal{H}^d_{\text{diag}} := \{(\mathcal{F}, G) \mid \mathcal{F}_{v \trianglelefteq e} =$ *invertible diagonal matrix*$\}$

Proposition 13. *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 proposition illustrates the benefits of using higher-dimensional stalks, while maintaining a simple and computationally convenient class of diagonal restriction maps. By using 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.

Definition 14. *The class of discrete $O(d)$-bundles* $\mathcal{H}^d_{\text{orth}} := \{(\mathcal{F}, G) \mid \mathcal{F}_{v \trianglelefteq e} \in O(d)\}$

We show that for stalks of dimension $d \in \{2, 4\}$, one can classify at least $C = 2d$ classes.

	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-SD	**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)-SD	**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-SD	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	75.69±2.00	87.16±0.37
Cont Diag-SD	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)-SD	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-SD	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

Table 1: Results on node classification datasets sorted by their homophily level. The first section includes discrete GNN models, while the second section includes continuous models. Top three models are coloured by **First**, **Second**, **Third**. Our models are marked **-SD**.

Theorem 15. *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 $\mathcal{G}$.*

This shows that orthogonal maps are able to make more efficient use of the space available to them than diagonal restriction maps.

3 Model & Results

We consider the following diffusion-type equation, which contains the sheaf diffusion equation as a particular case.

$$\dot{\mathbf{X}}(t) = -\sigma\Big(\Delta_{\mathcal{F}(t)}(\mathbf{I}_n \otimes \mathbf{W}_1)\mathbf{X}(t)\mathbf{W}_2\Big), \tag{2}$$

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 $(G, \mathcal{F}(t)) = g(G, \mathbf{X}(t); \theta)$.

We also consider a discrete version of this equation, using a new set of weights at each layer t.

$$\mathbf{X}_{t+1} = \mathbf{X}_t - \sigma\Big(\Delta_{\mathcal{F}(t)}(\mathbf{I} \otimes \mathbf{W}^t_1)\mathbf{X}_t\mathbf{W}^t_2\Big), \tag{3}$$

For both, models we use an initial MLP to compute $\mathbf{X}(0)$ from the raw features and a final linear layer to perform the node classification.

Real-world experiments We test our models on real-world datasets proposed by Rozemberczki et al. (2021); Pei et al. (2020) to evaluate heterophilic learning. These datasets have an edge homophily coefficient h ranging from $h = 0.11$ (very heterophilic) to $h = 0.81$ (very homophilic) and therefore offer a view of how a model performs in both regimes. We evaluate our models on the 10 fixed splits provided by Pei et al. (2020) and report the mean accuracy and standard deviation. Each split contains $48\%/32\%/20\%$ of nodes per class for training, validation and testing, respectively.

Results From the results in Table 1 we see that the discrete version of our models are first in $5/6$ benchmarks with high heterophily ($h < 0.3$) and second-ranked in the other $1/6$. At the same time, the models exhibit strong performance on the homophilic graphs (Cora, Pubmed, Citeseer) being within approximately 1% of the top-performing model. The second part of Table 1 includes continuous models, which our model outperforms on $7/9$ benchmarks with particularly large improvements on Chameleon and Squirrel. Among the discrete diffusion models, the $O(d)$-vector 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 the strong performance of the model learning diagonals maps.

REFERENCES

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.

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.

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

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

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

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

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*, 2021a.

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

Allen Hatcher. *Algebraic topology*. Cambridge Univ. Press, Cambridge, 2000.

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

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

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

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.

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*, pp. 2401–2409. PMLR, 2017.

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

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

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

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

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.

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

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.

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

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

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

Loring W Tu. Manifolds. In *An Introduction to Manifolds*, pp. 47–83. Springer, 2011.

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

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

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.

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

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.

A INTRODUCTION

Graph Neural Networks (GNNs) Sperduti (1994); Goller & Kuchler (1996); Gori et al. (2005); Scarselli et al. (2008); Bruna et al. (2014); Defferrard et al. (2016); Kipf & Welling (2017); Gilmer et al. (2017) 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.

We focus on two main problems often observed in GNNs: their poor performance in heterophilic graphs Zhu et al. (2020) and their oversmoothing behaviour Nt & Maehara (2019); Oono & Suzuki (2020). The former arises from the fact that GNNs are usually 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 deeper GNNs producing features that are too smooth to be useful.

In this work, 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 (Curry, 2014), a subfield of algebraic topology (Hatcher, 2000). 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 2 we analyse how by considering a hierarchy of increasingly general sheaves, starting from a trivial one, a diffusion equation based on the sheaf Laplacian (Hansen & Ghrist, 2019) can solve increasingly more complicated node-classification tasks in the infinite time limit. In this regime, we show that oversmoothing can be avoided by equipping the graph with the right sheaf structure for a task.

In Section C, we study the behaviour of a non-linear, parametric, and discrete version of this process. This results in a *Sheaf Convolutional Network* (Hansen & Gebhart, 2020) that generalises GCN of Kipf & Welling (2017). We prove that this discrete diffusion process does not necessarily converge to the harmonic space of the sheaf Laplacian, even when the norm of the weights is controlled from above, in contrast to GCNs (Cai & Wang, 2020; Oono & Suzuki, 2019). In other words, the weights can steer the diffusion process to a larger extent, which is important when the underlying sheaf is only approximately correct.

All these results are based on the properties of the harmonic space of the sheaf Laplacian. We begin by studying these properties from a spectral perspective in Section B. 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.

Finally, in Section D, we apply our theory for designing simple and practical GNN models. We describe how to construct such models by learning sheaves from data. The resulting sheaf models obtain state-of-the-art results in heterophilic graphs and show strong performance in homophilic ones.

To summarise our contributions: We (1) introduce a novel Graph ML framework based on cellular sheaves; (2) Prove new results about the spectral properties of the sheaf Laplacian that improve our understanding of its harmonic space; (3) Study the linear separation power of sheaf diffusion equations in node classification tasks and demonstrate it can avoid oversmoothing; (4) Prove that a nonlinear sheaf-based GCN-like operator can steer the diffusion process, unlike in the classical case; (5) Show how to learn sheaves from data and construct sheaf diffusion-based GNN models that obtain state-of-the-art results on heterophilic datasets.

B HARMONIC SPACE OF SHEAF LAPLACIANS

In this section we study general properties of the harmonic space $\ker(\Delta_{\mathcal{F}})$ of the normalised sheaf Laplacian. As with the normalised graph Laplacian, the normalised sheaf Laplacian is preferred for most practical purposes because its spectrum is bounded. Since a major role in the analysis below is played by discrete vector bundles, we focus 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, the block diagonal of $\mathcal{L}_{\mathcal{F}}$ has a block structure since $\mathcal{L}_{\mathcal{F}vv} = d_v \mathbf{I}_d$, where d_v is the degree of node v. Accordingly, if a signal $\tilde{\mathbf{x}} \in \ker(L_{\mathcal{F}})$, then the signal $\mathbf{x} : v \mapsto \sqrt{d_v}\mathbf{x}_v \in \ker(\Delta_{\mathcal{F}})$ and similarly for the inverse transformation. In general the harmonic space may be trivial: this happens when the constraints $\mathcal{F}_{v \trianglelefteq e}\mathbf{x}_v = \mathcal{F}_{u \trianglelefteq e}\mathbf{x}_u$ are not compatible with each other. Since the harmonic space will be related to the linear separation power of sheaf diffusion, we investigate below when this space is non-empty and derive to what extent the graph structure affects that. 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 $\mathcal{F}(u)$ via map composition:

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

In general then, transport maps act on node stalks and are constructed by composing single restriction maps (and their transpose) along edges.

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 (we note again that, differently from the classical case, the kernel of $\Delta_{\mathcal{F}}$ may be trivial, so by 'spectral gap' we refer here to the value of the smallest eigenvalue of $\Delta_{\mathcal{F}}$).

Proposition 16. *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 $\lambda^{\mathcal{F}}_0 \leq \frac{r^2}{2}$.*

A consequence of the previous 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*. Next, we address the opposite direction.

Proposition 17. *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 $\mathbf{x}_v \in \ker(\mathbf{P}^{\gamma}_{v \to v} - \mathbf{I})$.*

The previous proposition highlights the interplay between the graph and the sheaf structure. In fact, a simple consequence of this result is that for any cycle-free subset $S \subset V$, we have that any 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 18. *Let $\mathcal{F}$ be a discrete $O(d)$ bundle over a connected graph G with n nodes and let $||(\mathbf{P}^{\gamma}_{v \to v} - \mathbf{I})\mathbf{x}|| \geq \epsilon ||\mathbf{x}||$ 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. We note that

the Cheeger-like inequality presented here is not unique and other types of bounds on $\lambda_0^{\mathcal{F}}$ have been derived by Bandeira et al. (2013).

We conclude this section by further analysing the dimensionality of the harmonic space of discrete $O(d)$-bundles.

Lemma 19. *Let $\mathcal{F}$ be a discrete $O(d)$ bundle over a connected graph G. Then $\dim(H^0) \leq d$ and $\dim(H^0) = d$ iff the transport is path-independent.*

C Asymptotics of Sheaf Convolutions

In Section 2 we analysed the ability of sheaf diffusion to linearly separate the node-classes in the limit. However, when considering a discrete, parametric and non-linear version of this process, it is important to know how much the weights can steer it. This is particularly relevant if the underlying sheaf is only approximately correct for the task to be solved.

The continous diffusion process from Equation 1 has the following Euler discretisation with unit step-size:

$$\mathbf{X}(t+1) = \mathbf{X}(t) - \Delta_{\mathcal{F}}\mathbf{X}(t) = (\mathbf{I}_{nd} - \Delta_{\mathcal{F}})\mathbf{X}(t) \tag{4}$$

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 & Gebhart (2020):

$$\mathbf{Y} = \sigma\Big(\big(\mathbf{I}_{nd} - \Delta_{\mathcal{F}}\big)(\mathbf{I}_n \otimes \mathbf{W}_1)\mathbf{X}\mathbf{W}_2\Big) \in \mathbb{R}^{nd \times f_2}, \tag{5}$$

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.

It is natural to call this model a **Sheaf Convolutional Network** (SCN) since when $\Delta_{\mathcal{F}}$ is the usual normalised graph Laplacian, $\mathbf{W}_1$ becomes a scalar and one recovers GCN of Kipf & Welling (2017). To study this prototypical nonlinear and parametric discrete diffusion model in the limit of an infinite number of layers, we adapt the proof technique of Cai & Wang (2020) to track the *sheaf Dirichlet energy* as the signal is passed through multiple SCN layers.

Definition 20. *The sheaf Dirichlet energy $E_{\mathcal{F}}(\mathbf{x})$ of a signal $\mathbf{x} \in C^0(G, \mathcal{F})$ is defined as:*

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

Similarly, for multiple channels the energy is $\mathrm{trace}(\mathbf{X}^\top \Delta_{\mathcal{F}}\mathbf{X})$. It is easy to see that $\mathbf{x} \in \ker(\Delta_{\mathcal{F}}) \Leftrightarrow E_{\mathcal{F}}(\mathbf{x}) = 0$. Therefore, we can use the energy of a signal as a metric to measure its distance from the harmonic space.

We begin by studying the sheaves for which the features end up asymptotically in $\ker(\Delta_{\mathcal{F}})$. The first such example is for $O(d)$-bundles with symmetric relations. Let $\lambda_* = \max\big((\lambda_{\min} - 1)^2, (\lambda_{\max} - 1)^2\big)$, where $\lambda_{\min}, \lambda_{\max}$ are the smallest and largest non-zero eigenvalues of $\Delta_{\mathcal{F}}$.

Theorem 21. *Let $(\mathcal{F}, G)$ be an $O(d)$-bundle in $\mathcal{H}^d_{\mathrm{orth,sym}}$ and assume σ = ReLU or LeakyReLU. Then $E_{\mathcal{F}}(\mathbf{Y}) \leq \lambda_* \|\mathbf{W}_1\|_2^2 \|\mathbf{W}_2^\top\|_2^2 E_{\mathcal{F}}(\mathbf{X})$.*

In particular, this means that if $\lambda_* \|\mathbf{W}_1\|_2^2 \|\mathbf{W}_2^\top\|_2^2 < 1$, the signal converges exponentially fast to $\ker(\Delta_{\mathcal{F}})$. In some sense, this is not surprising because when we have symmetric relations along all edges (i.e. $\mathcal{F}_{v \leq e} = \mathcal{F}_{u \leq e}$) and hence the conditions in Theorem 21 are satisfied, then the harmonic space $\ker(\Delta_{\mathcal{F}})$ contain the same information as the kernel of the classical normalised Laplacian Δ_0.

Proposition 22. *If $\mathcal{F}$ is an $O(d)$-bundle in $\mathcal{H}^d_{\mathrm{orth,sym}}$, then $\mathbf{x} \in \ker \Delta_{\mathcal{F}}$ if and only if $\mathbf{x}^k \in \ker \Delta_0$ for all $1 \leq k \leq d$.*

Importantly, the symmetry of the relations along all edges is a necessary condition in Theorem 21. As soon as we have an asymmetric transport map, we can find an arbitrarily small linear transformation $\mathbf{W}$ that increases the energy.

Proposition 23. *For any connected graph G and $\varepsilon > 0$, there exist a sheaf $(G, \mathcal{F}) \notin \mathcal{H}^d_{\text{sym}}$, $\mathbf{W}$ with $\|\mathbf{W}\|_2 < \varepsilon$ and feature vector $\mathbf{x}$ such that $E_{\mathcal{F}}((\mathbf{I} \otimes \mathbf{W})\mathbf{x}) > E_{\mathcal{F}}(\mathbf{x})$.*

Beyond $O(d)$-bundles, we also have the following result for general sheaves with stalks of dimension $d = 1$ which generalises that of Cai & Wang (2020); Oono & Suzuki (2019) for GCNs:

Theorem 24. *Let $(\mathcal{F}, G)$ be a sheaf in $\mathcal{H}^1_+$ and assume σ is ReLU or LeakyReLU. Then $E_{\mathcal{F}}(\mathbf{Y}) \leq \lambda_* \|\mathbf{W}_1\|_2^2 \|\mathbf{W}_2^\top\|_2^2 E_{\mathcal{F}}(\mathbf{X})$.*

As before, having positively-signed relations is a necessary condition in the non-linear case to ensure oversmoothing happens. However, in this case, the proof also holds for negatively-signed relations when using a linear model (i.e. when σ is the identity). Due to this result, we note that Propositons 8 and 11 also generalise to GCNs. Any GCN using a weighted graph Laplacian that oversmooths as in Theorem 24 cannot linearly separate more than two classes. Furthermore, such a GCN cannot separate the classes of a bipartitie graph with equally-sized partitions.

We have two main takeaways: (1) Discrete sheaf diffusion models can avoid (in general) the kernel of the diffusion operator even in the deep linear case. (2) The asymmetry of the transport maps play (again) an important role in that.

D SHEAF LEARNING

The advantage of learning a sheaf is that one does not require any sort of embedding of the nodes in an ambient space (as e.g. done in Chamberlain et al. (2021a)). Instead, everything can be learned *locally*. Each $d \times d$ matrix $\mathcal{F}_{v \trianglelefteq e}$ is learned via a parametric function $\Phi : \mathbb{R}^{d \times 2} \to \mathbb{R}^{d \times d}$:

$$\mathcal{F}_{v \trianglelefteq e := (v,u)} = \Phi(\mathbf{x}_v, \mathbf{x}_u) \tag{6}$$

For simplicity, the equation above uses a single feature channel, but in practice, all channels are supplied as input. This function retains the inductive bias of locality specific to GNNs since it only utilises the features of the nodes forming the edge. At the same time, it is important that this function is non-symmetric in order to be able learn asymmetric transport maps along each edge. 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 parametrization 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 do not interact.

Orthogonal In this case, the model effectively learns a discrete vector bundle. Orthogonal matrices provide several advantages: (1) they are able to 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 (Mhammedi et al., 2017).

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 Treating the number of channels and layers as constant, a typical message passing GNN scales with $\mathcal{O}(n + m)$, where n is the number of nodes and m is the number of edges. A sheaf diffusion model with diagonal maps has complexity $\mathcal{O}(d(n + m))$ because the matrix multiplication required to compute the transport maps reduces to a scalar multiplication.

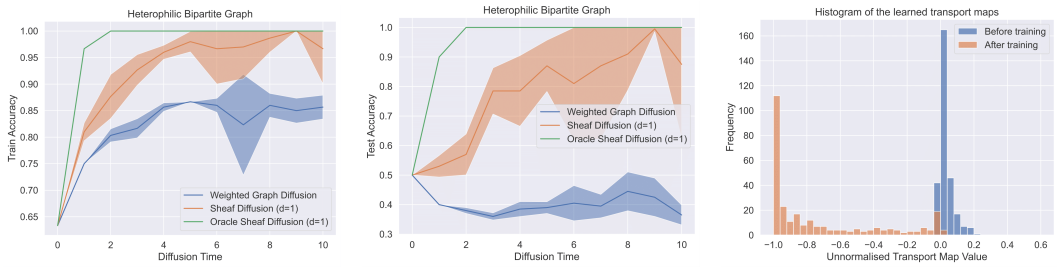


Figure 3: (*Left*) Test accuracy as a function of diffusion time. (*Right*) Histogram of the learned scalar transport maps. The performance of the sheaf diffusion model is superior to that of weighted-graph diffusion. The model correctly learns to invert the features of the two classes via the transport maps.

When learning orthogonal or general $d \times d$ matrices, the complexity becomes $\mathcal{O}(d^3(n+m))$ because matrix multiplication is $\mathcal{O}(d^3)$. We note that additional costs for the latter methods such as that of SVD for computing $D^{-1/2}$ for general matrices and that of parametrising the orthogonal group can be done in $O(d^3)$, so the overall complexity is not affected.

In practice, we use $1 \leq d \leq 5$ and all the operations above benefit from batched GPU computations in PyTorch (Paszke et al., 2019) which effectively results in a constant overhead. For learning orthogonal matrices, we rely on the library Torch Householder (Obukhov, 2021) which provides support for fast transformations with very large batch sizes.

E Additional Experiments

E.1 Synthetic experiments

Opinion polarisation (1D) We first consider a simple synthetic setup given by a connected bipartite graph, where the two partitions form two equally sized classes. We sample the features from two overlapping isotropic Gaussian distributions in order to make the classes linearly non-separable at initialisation time. From Proposition 8 we know that diffusion models using symmetric restriction maps cannot separate the classes in the limit, while a diffusion process using asymmetric maps should be able to.

Therefore, we consider two simple versions of the model from Equation 2, where we set $\mathbf{W}_1 = \mathbf{I}_d$, $\mathbf{W}_2 = \mathbf{I}_f$ and $\sigma = \mathrm{id}$. In the first, the maps $\mathcal{F}_{v \trianglelefteq e = (v,u)}$ are learned by a simple layer of the form $\sigma(\mathbf{w}^\top [\mathbf{x}_v || \mathbf{x}_u])$, where $\sigma = \tanh$. For the second model, we use a similar layer but constraint $\mathcal{F}_{v \trianglelefteq e} = \mathcal{F}_{u \trianglelefteq e} > 0$, which results in a weighted graph Laplacian.

Figure 3 presents the results of this experiment across five seeds. As expected, for diffusion time zero (i.e. no diffusion), we see that a linear classifier cannot separate the classes. Also as expected, the diffusion process using symmetric maps cannot perfectly fit the data. In contrast, with the more general sheaf diffusion, as more diffusion is performed and the signal approaches the harmonic space, the model gets better and the features become linearly separable. In the second 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 Propositions 10 the model learns a negative transport map for all edges.

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 15, 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 $1D$ stalks and another using $2D$ stalks. As we see from Figure 4, 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

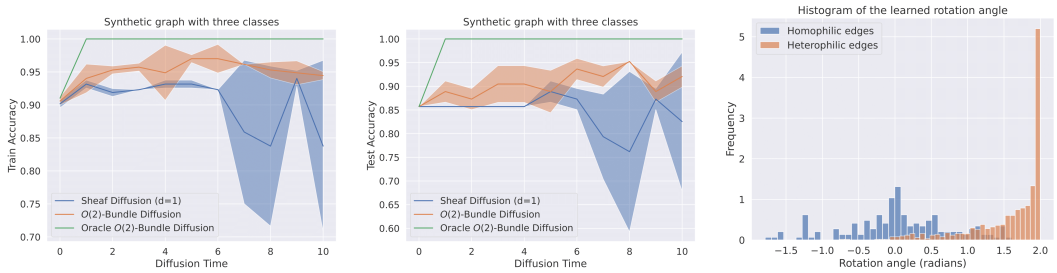


Figure 4: (*Left*) Train accuracy as a function of diffusion time. (*Middle*) Test accuracy as a function of diffusion time. (*Right*) Histogram of the learned rotation angle of the $2D$ transport maps. The performance of the bundle model is superior to that of the one-dimensional sheaf. The transport 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.

	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

Table 2: Ablation study for the dimension of the stalks. Positional encodings improve performance on some of our models.

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 $120° = 360°/C$, 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.

Positional encoding ablation Based on the Remark from the previous section, 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. (2020). In Table 2 we see that positional encodings do indeed 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.

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 6) and $C = 4$ (Figure 5) 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 8. 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 7.

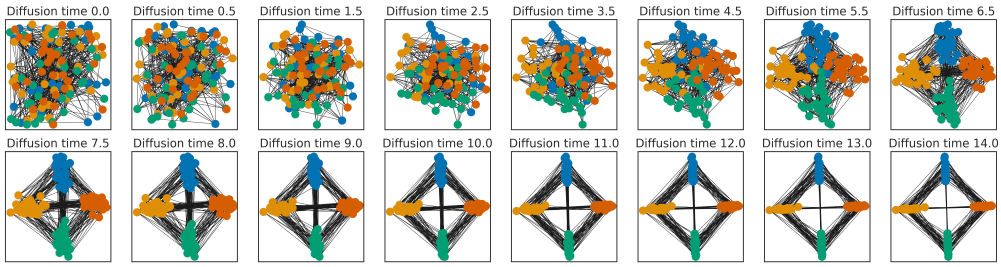


Figure 5: Sheaf diffusion process disentangling the $C = 4$ classes over time. The nodes are coloured by their class.

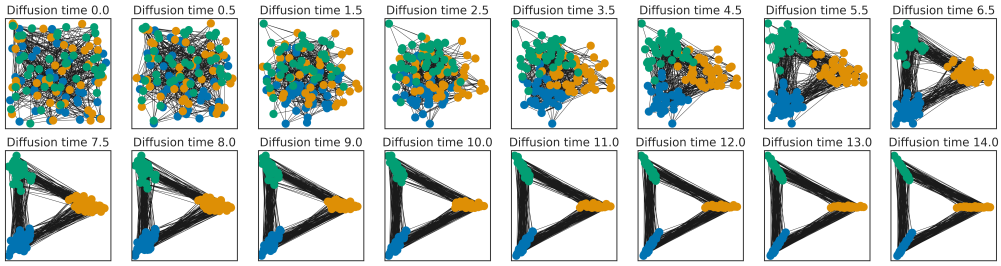


Figure 6: Sheaf diffusion process disentangling the $C = 3$ classes over time. The nodes are coloured by their class.

E.2 Additional details on real-world experiments

Baselines for real-world experiments In Table 1 we evaluate against an ample set of GNN models that can be placed in four categories: (1) classical: GCN (Kipf & Welling, 2017), GAT (Veličković et al., 2018), GraphSAGE (Hamilton et al., 2017); (2) models specifically designed for heterophilic settings: GGCN (Yan et al., 2021), Geom-GCN (Pei et al., 2020), H2GCN (Zhu et al., 2020), GPRGNN (Chien et al., 2021), FAGCN (Bo et al., 2021), MixHop (Abu-El-Haija et al., 2019); (3) models addressing oversmoothing: GCNII (Chen et al., 2020), PairNorm (Zhao & Akoglu, 2020); (4) continuous models: CGNNs (Xhonneux et al., 2020), GRAND (Chamberlain et al., 2021b), and BLEND (Chamberlain et al., 2021a). For (4) we fine-tune and evaluate the models ourselves. The other results are taken from Yan et al. (2021), except for FAGCN and MixHop, which come from Lingam et al. (2021) and Zhu et al. (2020), respectively. All of these were evaluated on the same set of splits as ours.

F 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:

$$\mathbf{X}_{t+1} = (1 + \varepsilon)\mathbf{X}_t - \sigma\Big(\Delta_{\mathcal{F}(t)}(\mathbf{I} \otimes \mathbf{W}^t_1)\mathbf{X}_t\mathbf{W}^t_2\Big). \tag{7}$$

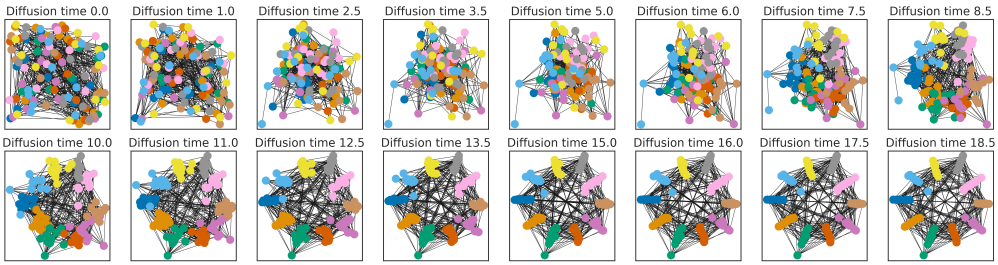


Figure 7: Sheaf diffusion process disentangling the $C = 10$ classes over time. The nodes are coloured by their class.

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.

Learning sheaves and the k-WL test According to our theoretical results, one should learn a pair of non-equal restriction maps along the edges between two classes. Suppose this is done via a local functions $\mathcal{F}_{v \trianglelefteq e := (v,u)} = \phi(\mathbf{x}_v, \mathbf{x}_u)$. Then, $\phi(\mathbf{x}_v, \mathbf{x}_u) \neq \phi(\mathbf{x}_u, \mathbf{x}_v)$ only if $\mathbf{x}_u \neq \mathbf{x}_v$. This leads to the following remark.

Remark. *Let G be a graph with an initial colouring $\mathbf{x}$ such that $\exists (v,u) \in E$ and v and u are not k-WL distinguishable. Then no model upper bounded by k-WL can learn an asymmetric relation along the edge (v,u).*

This observation motivates our decision to evolve the geometry at each layer of the model since at initialisation time many nodes might not be distinguishable. At the same time, it suggests that provably expressive architectures might be able to learn better sheaves.

Hyperparameters for discrete models For the discrete models we searched in the following hyperparameters:

- Hidden channels: $[8, 16, 32]$ for WebKB datasets and $[8, 16, 32, 64]$ for all the other datasets.
- d: $[1, 2, 3, 4, 5]$
- Layers: $[2, 3, 4, 5, 6, 7, 8]$
- Learning rate: 0.02 for the WebKB datasets and 0.01 for all the other datasets.
- Weight decay for the regular model parameters: searched in a log-uniform range over $[-4.5, 11.0]$
- Weight decay for the parameters learning the sheaf: searched in a log-uniform range over $[-4.5, 11.0]$
- Dropout on the inputs: searched uniformly over $[0, 0.9]$.
- Dropout for the other layers: searched uniformly over $[0, 0.9]$.
- Early stopping patience: 100 epochs for the Wiki datasets and 200 for the others.
- Maximum training epochs: 1000 epochs for the Wiki datasets and 500 for the others.

Hyperparameters for continous models For the continous models we searched over the following hyperparameters:

- Hidden channels: $[8, 16, 32, 64]$
- d: $[1, 2, 3, 4, 5]$
- Learning rate: searched in a log-uniform range over $[0.01, 0.1]$
- Weight decay : searched in a log-uniform range over $[-6.9, 13.8]$
- Dropout: searched uniformly in $[0, 0.95]$.

- ODE Integration method: euler.
- Maximum integration time: searched uniformly over $[1.0, 9.0]$.
- Training epochs: 50.

We train all models using the Adam optimiser (Kingma & Ba, 2015). All models use ELU activations.

G Related Work, Discussion and Conclusion

Sheaf Neural Networks Sheaf Neural Networks with a *fixed* sheaf Laplacian were originally introduced by Hansen & Gebhart (2020) in a toy experimental setting. Since then, they have remained completely unexplored. We hope that this paper will fill this lacuna. In contrast to their work, we provide an ample theoretical analysis justifying the use of sheaves in Graph ML. On the practical side, we study for the first time how sheaves can be *learned from data* using neural networks. Finally, this is the first successful application of sheaf-based neural networks in a real-world setting.

Limitations One of the main limitations of our theoretical analysis is that it does not discuss the learnability and generalisation properties of sheaves, only their existence. Nonetheless, this setting was sufficient to produce many valuable insights about heterophily and oversmoothing and a basic understanding of what various types of sheaves can do. Much more theoretical work remains to be done in this direction, and we expect to see further cross-fertilization between ML and algebraic topology research in the future.

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 do not only evolve the features at each layer but also the underlying geometry of the graph. In practice, we demonstrated that this framework can achieve state-of-the-art results in heterophilic settings.

H Harmonic Space Proofs

Proof of Proposition 16. We first note that on a discrete $O(d)$ bundle the degree operator $D_v = d_v \mathbf{I}$ since by orthogonality $\mathcal{F}_{v \trianglelefteq e}^{\top} \mathcal{F}_{v \trianglelefteq e} = \mathbf{I}$. 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 $\tilde{\mathbf{z}}_u \sqrt{d_u} = \mathbf{z}_u$.

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

where we have again used that the maps are orthogonal. Since $(\mathcal{F}_{u \trianglelefteq e}^{\top} \mathcal{F}_{w \trianglelefteq e}) \mathbf{P}^{\gamma}_{v \to w} = \mathbf{P}^{\gamma'}_{v \to u}$ we find that the right hand side can be bound from above by $r^2 ||\tilde{\mathbf{z}}_v||^2$. Therefore, by using Definition 20 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

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

□

Proof of Proposition 17. Assume that $\mathbf{x} \in H^0(G, \mathcal{F})$ and consider $v \in V$ and any cycle based at v denoted by $\gamma_{v \to v} = (v_0 = v, v_1, \ldots, v_L = v)$. We have that

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

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. □

Proof of Proposition 18. If $\epsilon = 0$ there is nothing to prove. Assume that $\epsilon > 0$. By Proposition 17 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:

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

From Definition 20 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

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

□

Proof of Lemma 19. 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 17, 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 $\dim(H^0(G, \mathcal{F})) < d$. □

I PROOFS FOR THE POWER OF SHEAF DIFFUSION

Definition 25. *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 26. *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 & Ghrist (2021).

Lemma 27. *Solutions $\mathbf{X}(t)$ to the diffusion in Equation 1 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 $(G, \mathcal{F})$.

Proof of Proposition 7. 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:

$$\mathbf{a}_v = \begin{cases} \sqrt{d_v + h_v(\alpha - 1)}, & v \in A \\ \sqrt{d_v}, & v \in B \end{cases} \tag{8}$$

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 27. 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 $v \in A, u \in B$. □

Proof of Proposition 8. 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 27, 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 $|A| = |B|$:

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

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. □

Proof of Proposition 10. 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 $\mathbf{X}(0) \in \mathbb{R}^{n \times f}$.

We can construct a sheaf $(\mathcal{F}, G) \in \mathcal{H}^1(G)$ as follows. For all nodes v and edges e, $\mathcal{F}(v) \cong \mathcal{F}(e) \cong \mathbb{R}$. For all $v \in A$, set $\mathcal{F}_{v \trianglelefteq e} = \alpha$. For all $u \in B$, set $\mathcal{F}_{u \trianglelefteq e} = \beta$. Additionally, let $\alpha < 0 < \beta$.

Since the graph is connected, by Lemma 19, the only harmonic eigenvector of $\Delta_{\mathcal{F}}$ is:

$$\mathbf{y}_v = \begin{cases} \beta\sqrt{\sum_{v \trianglelefteq e} \alpha^2} \\ \alpha\sqrt{\sum_{v \trianglelefteq e} \beta^2} \end{cases} = \begin{cases} \beta|\alpha|\sqrt{d_v}, & v \in A \\ \alpha|\beta|\sqrt{d_v}, & v \in B \end{cases} \tag{9}$$

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 $v \in A, u \in B$. □

Definition 28. *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 29. *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 of Proposition 29. 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 (2006)).

Lemma 30. *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.*

Proof of Proposition 11. 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 19, $\dim(H^0) = 1$.

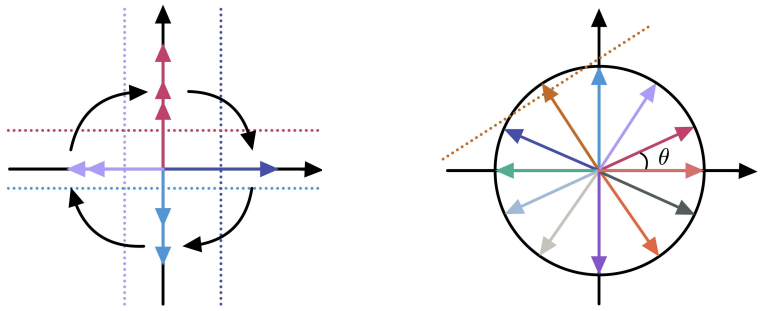
In that case, let $\mathbf{h}$ be the unit-normalised harmonic eigenvector of $\Delta_{\mathcal{F}}$. By Lemma 27, for any node v, its scalar feature in channel $k \leq f$ is given by $x^k_v(\infty) = \langle \mathbf{x}^k(0), \mathbf{h} \rangle \mathbf{h}_v$. Note that we can always find three nodes 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:

$$\mathbf{x}^k_u(\infty) = \langle \mathbf{x}^k(0), \mathbf{h} \rangle \mathbf{h}_u = \alpha\langle \mathbf{x}^k(0), \mathbf{h} \rangle \mathbf{h}_v + (1-\alpha)\langle \mathbf{x}^k(0), \mathbf{h} \rangle \mathbf{h}_w = \alpha\mathbf{x}^k_v(\infty) + (1-\alpha)\mathbf{x}^k_w(\infty). \tag{10}$$

Since this is true for all channels $k < f$, 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 30, the class of v is not linearly separable from the other classes. □

Proof of Proposition 13. 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. □



(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 8: Proof sketch for Lemma 33 and Proposition 34.

Lemma 31. *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 $\{\mathbf{h}^1, \dots, \mathbf{h}^d\}$ form an orthogonal eigenbasis for the harmonic space of $\Delta_{\mathcal{F}}$, where:*

$$\mathbf{h}_v^i = \begin{cases} \mathbf{e}_i \sqrt{d_v^{\mathcal{F}}} \\ \mathbf{P}_{v \to w} \mathbf{e}_i \sqrt{d_v^{\mathcal{F}}} \end{cases} = \begin{cases} \mathbf{e}_i \sqrt{\sum_{v \trianglelefteq e} \alpha_e^2}, & v = v^* \\ \mathbf{P}_{v* \to w} \mathbf{e}_i \sqrt{\sum_{v \trianglelefteq e} \alpha_e^2}, & \textit{otherwise} \end{cases} \tag{11}$$

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

$$E_{\mathcal{F}}(\mathbf{h}_v^i) = \frac{1}{2} \sum_{v,u,e:=(v,u)} \| \frac{1}{\sqrt{d_v^{\mathcal{F}}}} \mathcal{F}_{v \leq e} \mathbf{h}_v - \frac{1}{\sqrt{d_u^{\mathcal{F}}}} \mathcal{F}_{u \trianglelefteq e} \mathbf{h}_u \|_2^2 \tag{12}$$

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

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

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

$$= \frac{1}{2} \sum_{v,u,e:=(v,u)} \| \mathcal{F}_{u \trianglelefteq e} \mathbf{P}_{v^* \to u} \mathbf{e}_i - \mathcal{F}_{u \trianglelefteq e} \mathbf{P}_{v^* \to u} \mathbf{e}_i \|_2^2 = 0 \quad \text{Orthogonality of } \mathcal{F}_{v \trianglelefteq e} \tag{16}$$

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

$$\langle \mathbf{h}_v^i, \mathbf{h}_v^j \rangle = \langle \mathbf{P}_{v* \to w} \mathbf{e}_i \sqrt{d_v^{\mathcal{F}}}, \mathbf{P}_{v* \to w} \mathbf{e}_j \sqrt{d_v^{\mathcal{F}}} \rangle = \sqrt{d_v^{\mathcal{F}}} \sqrt{d_v^{\mathcal{F}}} \langle \mathbf{e}_i, \mathbf{e}_j \rangle = 0 \tag{17}$$

□

Lemma 32. *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 15 in dimension two in the following lemma and then we will look at the general case.

Lemma 33. *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 8a. 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:

$$\mathbf{R}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{R}_2 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \tag{18}$$

and by $\mathcal{C} = \{1, \dots, C\}$ 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 31, 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 8a 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:

$$\begin{aligned}
&\langle \mathbf{P}_{v^* \to u} \sum_i \alpha_i \mathbf{e}_i \sqrt{d_u}, \mathbf{P}_{v^* \to w} \sum_j \alpha_j \mathbf{e}_j \sqrt{d_w} \rangle = \\
&= \sqrt{d_u d_w} \Big[\sum_{i \neq j} \alpha_i \alpha_j \langle \mathbf{P}_{v^* \to u}\mathbf{e}_i, \mathbf{P}_{v^* \to w}\mathbf{e}_j \rangle + \sum_k \alpha_k^2 \langle \mathbf{P}_{v^* \to u}\mathbf{e}_k, \mathbf{P}_{v^* \to w}\mathbf{e}_k \rangle \Big] \\
&= \sqrt{d_u d_w} \Big[\sum_{i<j} \alpha_i \alpha_j \Big(\langle \mathbf{P}_{v^* \to u}\mathbf{e}_i, \mathbf{P}_{v^* \to w}\mathbf{e}_j \rangle + \langle \mathbf{P}_{v^* \to u}\mathbf{e}_j, \mathbf{P}_{v^* \to w}\mathbf{e}_i \rangle \Big) \\
&\quad + \sum_k \alpha_k^2 \langle \mathbf{P}_{v^* \to u}\mathbf{e}_k, \mathbf{P}_{v^* \to w}\mathbf{e}_k \rangle \Big] && (19) \\
&= \sum_k \alpha_k^2 \langle \mathbf{P}_{v^* \to u}\mathbf{e}_k, \mathbf{P}_{v^* \to w}\mathbf{e}_k \rangle && \text{(by Lemma 32)}
\end{aligned}$$

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 8a. □

Proof of Theorem 15. To generalise the proof, 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 19 must be zero for any coefficients α.

Therefore, $\mathcal{P} = \{\mathbf{P}_0, \dots, \mathbf{P}_{d-1}\}$ 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 (Schafer, 2017). 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 33 are a classic 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:

$$\mathbf{R}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{R}_2 = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{R}_3 = \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}, \quad \mathbf{R}_4 = \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

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 33, we can linearly separate up to eight classes. □

Proposition 34. *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 of Propositon 34. 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 8b.

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 19 between the harmonic features of the nodes, we can verify that the angle between any two classes is different from zero. By Lemma 32, 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 $(i-j)\theta$. □

J PROOFS FOR OVERSMOOTHING

Proof of Proposition 22. 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 35. *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 $E_{\mathcal{F}}((\mathbf{I} \otimes \mathbf{W})\mathbf{x}) > 0$.*

Proof. 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 20.

Proposition 36.

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

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. □

Proof of Theorem 21. We first prove a couple of Lemmas before proving the Theorem. The structure of the proof follows that of Cai & Wang (2020), which in turn generalises that of Oono & Suzuki (2019). 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).

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

Lemma 37. *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 $[0, 2]$ (Hansen & Ghrist, 2019, Proposition 5.5). We note that the original proof of Cai & Wang (2020) bounds the expression by $(1 - \lambda_{\min})^2$ instead of λ_*, which appears to be an error. □

Lemma 38. $E_{\mathcal{F}}(\mathbf{X}\mathbf{W}) \leq \|\mathbf{W}^\top\|_2^2 E_{\mathcal{F}}(\mathbf{X})$

Proof. Following the proof of Cai & Wang (2020) we have:

$$\begin{aligned}
E_{\mathcal{F}}(\mathbf{X}\mathbf{W}) &= \mathrm{Tr}(\mathbf{W}^\top \mathbf{X}^\top \Delta_{\mathcal{F}} \mathbf{X}\mathbf{W}) && (31)\\
&= \mathrm{Tr}(\mathbf{X}^\top \Delta_{\mathcal{F}} \mathbf{X}\mathbf{W}\mathbf{W}^\top) && \text{trace cyclic property} \quad (32)\\
&\leq \mathrm{Tr}(\mathbf{X}^\top \Delta_{\mathcal{F}} \mathbf{X}) \|\mathbf{W}\mathbf{W}^\top\|_2 && \text{see Lemma 3.1 in Lee (2008)} \quad (33)\\
&= \mathrm{Tr}(\mathbf{X}^\top \Delta_{\mathcal{F}} \mathbf{X}) \|\mathbf{W}^\top\|_2^2 && (34)
\end{aligned}$$

□

Lemma 39. *For conditions as in Theorem 21,* $E\mathcal{F}\big((\mathbf{I}_n \otimes \mathbf{W})\mathbf{x}\big) \leq \|\mathbf{W}\|_2^2 E(f)$.

Proof. First, we note that for orthogonal matrices, $D_v = \mathbf{I} \sum_{v \trianglelefteq e} \alpha_e^2 = \mathbf{I} d_v$ (Hansen & Ghrist, 2019, Lemma 4.4)

$$E_{\mathcal{F}}\big((\mathbf{I}\otimes\mathbf{W})\mathbf{x}\big) = \frac{1}{2}\sum_{v,w\trianglelefteq e}\|\mathcal{F}_{v\trianglelefteq e}D_v^{-1/2}\mathbf{W}f_v - \mathcal{F}_{w\trianglelefteq e}D_w^{-1/2}\mathbf{W}\mathbf{x}_w\|_2^2 \tag{35}$$

$$= \frac{1}{2}\sum_{v,w\trianglelefteq e}\|\mathcal{F}_e\mathbf{W}\big(d_v^{-1/2}\mathbf{x}_v - d_w^{-1/2}\mathbf{x}_w\big)\|_2^2 \tag{36}$$

$$= \frac{1}{2}\sum_{v,w\trianglelefteq e}\|\mathbf{W}\big(d_v^{-1/2}\mathbf{x}_v - d_w^{-1/2}\mathbf{x}_w\big)\|_2^2 \qquad \mathcal{F}_e \text{ is orthogonal} \tag{37}$$

$$\leq \frac{1}{2}\sum_{v,w\trianglelefteq e}\|\mathbf{W}\|_2^2\|d_v^{-1/2}\mathbf{x}_v - d_w^{-1/2}\mathbf{x}_w\|_2^2 \qquad \text{property of the operator norm} \tag{38}$$

$$= \frac{1}{2}\sum_{v,w\trianglelefteq e}\|\mathbf{W}\|_2^2\|\mathcal{F}_e\big(d_v^{-1/2}\mathbf{x}_v - d_w^{-1/2}\mathbf{x}_w\big)\|_2^2 \qquad \mathcal{F}_e \text{ is orthogonal} \tag{39}$$

$$= \frac{1}{2}\|\mathbf{W}\|_2^2\sum_{v,w\trianglelefteq e}\|\mathcal{F}_e\big(D_v^{-1/2}\mathbf{x}_v - D_w^{-1/2}\mathbf{x}_w\big)\|_2^2 = \|\mathbf{W}\|_2^2 E_{\mathcal{F}}(\mathbf{x}) \tag{40}$$

The proof can also be extended easily extended to vector bundles over weighted graphs (i.e. allowing weighted edges as in Ghrist & Riess (2020)). 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 40. *For conditions as in Theorem 21,* $E_{\mathcal{F}}\big(\sigma(\mathbf{x})\big) \leq E(\mathbf{x})$.

Proof.

$$E\big(\sigma(\mathbf{x})\big) = \frac{1}{2}\sum_{v,w\trianglelefteq e}\|\mathcal{F}_{v\trianglelefteq e}D_v^{-1/2}\sigma(\mathbf{x}_v) - \mathcal{F}_{w\trianglelefteq e}D_w^{-1/2}\sigma(\mathbf{x}_w)\|_2^2 \tag{41}$$

$$= \frac{1}{2}\sum_{v,w\trianglelefteq e}\|\mathcal{F}_e\big(d_v^{-1/2}\sigma(\mathbf{x}_v) - d_w^{-1/2}\sigma(\mathbf{x}_w)\big)\|_2^2 \tag{42}$$

$$= \frac{1}{2}\sum_{v,w\trianglelefteq e}\|d_v^{-1/2}\sigma(\mathbf{x}_v) - d_w^{-1/2}\sigma(\mathbf{x}_w)\|_2^2 \qquad \text{orthogonality of } \mathcal{F}_e \tag{43}$$

$$= \frac{1}{2}\sum_{v,w\trianglelefteq e}\|\sigma\Big(\frac{\mathbf{x}_v}{\sqrt{d_v}}\Big) - \sigma\Big(\frac{\mathbf{x}_w}{\sqrt{d_w}}\Big)\|_2^2 \qquad c\text{ReLU}(x) = \text{ReLU}(cx) \text{ for } c > 0 \tag{44}$$

$$\leq \frac{1}{2}\sum_{v,w\trianglelefteq e}\|\frac{\mathbf{x}_v}{\sqrt{d_v}} - \frac{\mathbf{x}_w}{\sqrt{d_w}}\|_2^2 \qquad \text{Lipschitz continuity of ReLU} \tag{45}$$

$$= \frac{1}{2}\sum_{v,w\trianglelefteq e}\|\mathcal{F}_e\big(d_v^{-1/2}\mathbf{x}_v - d_w^{-1/2}\mathbf{x}_w\big)\|_2^2 \qquad \text{orthogonality of } \mathcal{F}_e \tag{46}$$

$$= E_{\mathcal{F}}(\mathbf{x}) \tag{47}$$

□

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

Proof of Theorem 24. If $d = 1$, then Lemma 39 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 40 holds in this case.

Lemma 41. *For conditions as in Theorem 24,* $E_{\mathcal{F}}\big(\sigma(\mathbf{x})\big) \leq E(\mathbf{x})$.

Proof.

$$
\begin{aligned}
E\big(\sigma(\mathbf{x})\big) &= \frac{1}{2}\sum_{v,w\trianglelefteq e} \|\mathcal{F}_{v\trianglelefteq e} D_v^{-1/2}\sigma(x_v) - \mathcal{F}_{w\trianglelefteq e} D_w^{-1/2}\sigma(x_w)\|_2^2 && (48)\\
&= \frac{1}{2}\sum_{v,w\trianglelefteq e} \||\mathcal{F}_{v\trianglelefteq e}| D_v^{-1/2}\sigma(x_v) - |\mathcal{F}_{w\trianglelefteq e}| D_w^{-1/2}\sigma(x_w)\|_2^2 \quad \mathcal{F}_{v\trianglelefteq e}\mathcal{F}_{w\trianglelefteq e} > 0 && (49)\\
&= \frac{1}{2}\sum_{v,w\trianglelefteq e} \|\sigma\Big(\frac{|\mathcal{F}_{v\trianglelefteq e}| x_v}{\sqrt{d_v}}\Big) - \sigma\Big(\frac{|\mathcal{F}_{w\trianglelefteq e}| x_w}{\sqrt{d_w}}\Big)\|_2^2 \quad c\mathrm{ReLU}(x) = \mathrm{ReLU}(cx) \text{ for } c > 0 && (50)\\
&\leq \frac{1}{2}\sum_{v,w\trianglelefteq e} \|\frac{|\mathcal{F}_{v\trianglelefteq e}| x_v}{\sqrt{d_v}} - \frac{|\mathcal{F}_{w\trianglelefteq e}| x_w}{\sqrt{d_w}}\|_2^2 \quad \text{Lipschitz continuity of ReLU} && (51)\\
&= \frac{1}{2}\sum_{v,w\trianglelefteq e} \|\mathcal{F}_{v\trianglelefteq e} D_v^{-1/2} x_v - \mathcal{F}_{w\trianglelefteq e} D_w^{-1/2} x_w\|_2^2 \quad \mathcal{F}_{v\trianglelefteq e}\mathcal{F}_{w\trianglelefteq e} > 0 && (52)\\
&= E_{\mathcal{F}}(\mathbf{x}) && (53)
\end{aligned}
$$

□

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. □