Neural Sheaf Diffusion

A Topological Perspective on Heterophily and Oversmoothing in GNNs

Cristian Bodnar
Department of Computer Science
University of Cambridge

Based on joint work with Francesco Di Giovanni, Benjamin Chamberlain, Pietro Liò, Michael M. Bronstein

University of Oxford
March 11, 2022

Motivation

Graph Convolutional Networks

Let $G = (V, E)$ be a graph with node feature matrix $\mathbf{X}$, adjacency matrix $\mathbf{A}$, degree matrix $\mathbf{D}$ and normalised Laplacian Δ_0. Consider the GCN [Kipf & Welling, 2017] equation:

$$\begin{aligned}\mathrm{GCN}(\mathbf{X}, \mathbf{A}) : &= \sigma\Big((\mathbf{D} + \mathbf{I})^{-1/2}(\mathbf{A} + \mathbf{I})(\mathbf{D} + \mathbf{I})^{-1/2}\mathbf{X}\mathbf{W}\Big) = \sigma\Big(\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}\mathbf{X}\mathbf{W}\Big) \\ &= \sigma\Big((\mathbf{I} - \tilde{\Delta}_0)\mathbf{X}\mathbf{W}\Big)\end{aligned} \tag{1}$$

Graph Convolutional Networks

Let $G = (V, E)$ be a graph with node feature matrix $\mathbf{X}$, adjacency matrix $\mathbf{A}$, degree matrix $\mathbf{D}$ and normalised Laplacian Δ_0. Consider the GCN [Kipf & Welling, 2017] equation:

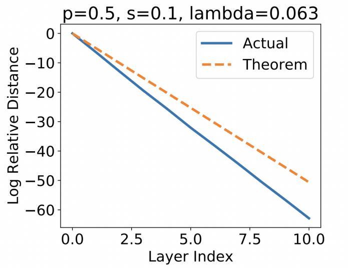
$$\begin{aligned} \mathrm{GCN}(\mathbf{X}, \mathbf{A}) :&= \sigma\Big((\mathbf{D}+\mathbf{I})^{-1/2}(\mathbf{A}+\mathbf{I})(\mathbf{D}+\mathbf{I})^{-1/2}\mathbf{X}\mathbf{W}\Big) = \sigma\Big(\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}\mathbf{X}\mathbf{W}\Big) \\ &= \sigma\Big((\mathbf{I} - \tilde{\Delta}_0)\mathbf{X}\mathbf{W}\Big) \end{aligned} \tag{1}$$

From Equation 1, it is clear the GCN is nothing but a non-linear, parametric and discretised diffusion equation:

$$\dot{\mathbf{X}}(t) = -\tilde{\Delta}_0\mathbf{X}(t) \rightsquigarrow \mathbf{X}(t+1) = \mathbf{X}(t) - \tilde{\Delta}_0\mathbf{X}(t) = (\mathbf{I} - \tilde{\Delta}_0)\mathbf{X}(t) \tag{2}$$

The Oversmoothing Problem

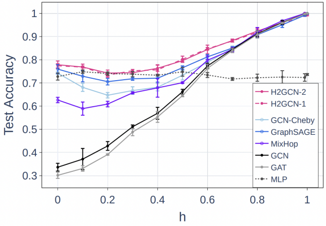
In Deep Graph Neural Networks (GNNs), it has been observed [Oono & Suzuki, 2020, Cai & Wang, 2020] that the features become progressively smoother with increased depth, resulting in a drop in the node classification performance. This is known as the **"oversmoothing problem"** of GNNs.



With more layers, GCN approaches a "smooth" subspace where all the node features are constant [Oono & Suzuki, 2020].

The Heterophily Problem

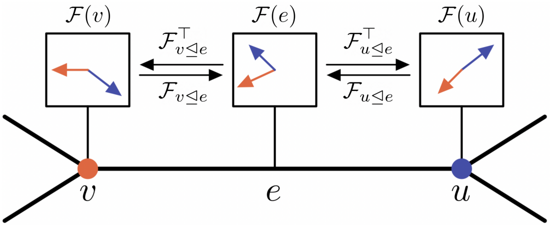
Numerous studies (e.g. [J. Zhu et al., 2020]) remarked that GNNs struggle in heterophilic settings (i.e. graphs where a node tends to be connected to nodes belonging to other classes). We call this the **"heterophily problem"** of GNNs.



The performance of GNNs is strongly correlated to the homophily level of a graph [J. Zhu et al., 2020].

Cellular Sheaves

Cellular Sheaves

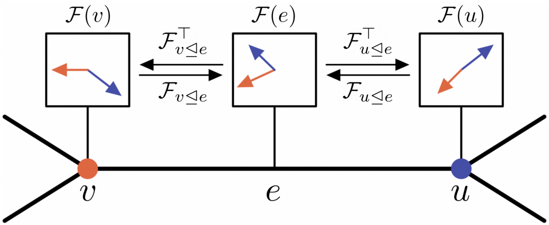


Definition

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

1. A vector space $\mathcal{F}(v)$ for each $v \in V$.

Cellular Sheaves

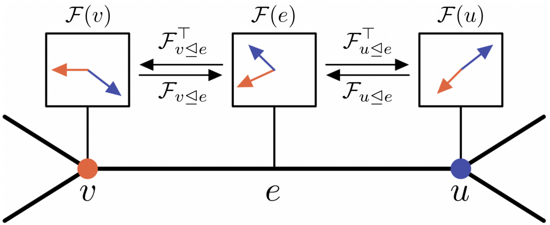


Definition

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

1. A vector space $\mathcal{F}(v)$ for each $v \in V$.
2. A vector space $\mathcal{F}(e)$ for each $e \in E$.

Cellular Sheaves

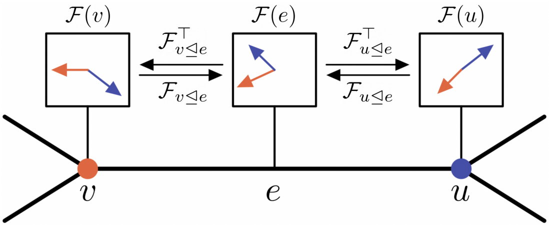


Definition

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

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

Cellular Sheaves



Definition

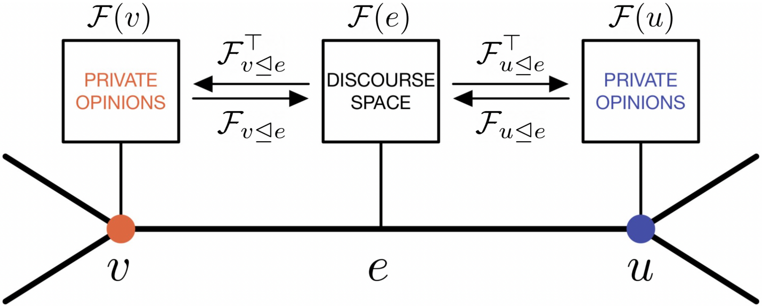
A *cellular sheaf* $(G, \mathcal{F})$ on an undirected graph $G = (V, E)$ consists of:

1. A vector space $\mathcal{F}(v)$ for each $v \in V$.
2. A vector space $\mathcal{F}(e)$ for each $e \in E$.
3. 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* and the linear maps are also known as *restriction maps*.

Opinion Dynamics

From an opinion dynamics perspective [Hansen & Ghrist, 2020], if the nodes represent people, $\mathcal{F}(v)$ represents the space of private opinions of an individual and $\mathcal{F}_{v \trianglelefteq e}$ encodes how this private opinion manifest in a discourse space $\mathcal{F}(e)$.



Cochains and the coboundary maps

We can form a space formed of all the spaces $\mathcal{F}(v)$ for all nodes v. Similarly, we can define a space formed of all the individual $\mathcal{F}(e)$.

Definition

For a sheaf $(\mathcal{F}, G)$ we define the space of 0-cochains $C^0(G;\mathcal{F}) := \bigoplus_{v \in V} \mathcal{F}(v)$ and 1-cochains $C^1(G;\mathcal{F}) := \bigoplus_{e \in E} \mathcal{F}(e)$.

Cochains and the coboundary maps

We can form a space formed of all the spaces $\mathcal{F}(v)$ for all nodes v. Similarly, we can define a space formed of all the individual $\mathcal{F}(e)$.

Definition

For a sheaf $(\mathcal{F}, G)$ we define the space of 0-cochains $C^0(G;\mathcal{F}) := \bigoplus_{v \in V} \mathcal{F}(v)$ and 1-cochains $C^1(G;\mathcal{F}) := \bigoplus_{e \in E} \mathcal{F}(e)$.

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 in the discourse space.

Definition

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

The Sheaf Laplacian

We can use the co-boundary map to define a *sheaf Laplacian* [Hansen & Ghrist, 2019].

Definition

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 $L_{\mathcal{F}} := \delta^\top \delta$ or, equivalently, $L_{\mathcal{F}}(\mathbf{x})_v := \sum_{v,u \trianglelefteq e} \mathcal{F}_{v \trianglelefteq e}^\top (\mathcal{F}_{v \trianglelefteq e} \mathbf{x}_v - \mathcal{F}_{u \trianglelefteq e} \mathbf{x}_u)$

The Sheaf Laplacian

We can use the co-boundary map to define a *sheaf Laplacian* [Hansen & Ghrist, 2019].

Definition

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 $L_{\mathcal{F}} := \delta^\top \delta$ or, equivalently, $L_{\mathcal{F}}(\mathbf{x})_v := \sum_{v,u \trianglelefteq e} \mathcal{F}_{v \trianglelefteq e}^\top (\mathcal{F}_{v \trianglelefteq e} \mathbf{x}_v - \mathcal{F}_{u \trianglelefteq e} \mathbf{x}_u)$

$$L_{\mathcal{F}} := \begin{array}{cc} & \begin{array}{ccc} v & \cdots & u \end{array} \\ \begin{array}{c} v \\ \vdots \\ u \end{array} & \begin{bmatrix} \sum_{v \trianglelefteq e} \mathcal{F}_{v \trianglelefteq e}^\top \mathcal{F}_{v \trianglelefteq e} & \cdots & -\mathcal{F}_{v \trianglelefteq e}^\top \mathcal{F}_{u \trianglelefteq e} \\ \vdots & \ddots & \vdots \\ -\mathcal{F}_{u \trianglelefteq e}^\top \mathcal{F}_{v \trianglelefteq e} & \cdots & \sum_{u \trianglelefteq e} \mathcal{F}_{u \trianglelefteq e}^\top \mathcal{F}_{u \trianglelefteq e} \end{bmatrix} \end{array}$$

The Sheaf Laplacian

We can use the co-boundary map to define a *sheaf Laplacian* [Hansen & Ghrist, 2019].

Definition

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 $L_{\mathcal{F}} := \delta^\top \delta$ or, equivalently, $L_{\mathcal{F}}(\mathbf{x})_v := \sum_{v,u \trianglelefteq e} \mathcal{F}_{v \trianglelefteq e}^\top (\mathcal{F}_{v \trianglelefteq e} \mathbf{x}_v - \mathcal{F}_{u \trianglelefteq e} \mathbf{x}_u)$

$$L_{\mathcal{F}} := \begin{array}{cc} & \begin{array}{ccc} v & \cdots & u \end{array} \\ \begin{array}{c} v \\ \vdots \\ u \end{array} & \begin{bmatrix} \sum_{v \trianglelefteq e} \mathcal{F}_{v \trianglelefteq e}^\top \mathcal{F}_{v \trianglelefteq e} & \cdots & -\mathcal{F}_{v \trianglelefteq e}^\top \mathcal{F}_{u \trianglelefteq e} \\ \vdots & \ddots & \vdots \\ -\mathcal{F}_{u \trianglelefteq e}^\top \mathcal{F}_{v \trianglelefteq e} & \cdots & \sum_{u \trianglelefteq e} \mathcal{F}_{u \trianglelefteq e}^\top \mathcal{F}_{u \trianglelefteq e} \end{bmatrix} \end{array}$$

The normalised Laplacian $\Delta_{\mathcal{F}} := D^{-1/2} L_{\mathcal{F}} D^{-1/2}$, where D is the block-diagonal of $L_{\mathcal{F}}$.

The Sheaf Laplacian

We can use the co-boundary map to define a *sheaf Laplacian* [Hansen & Ghrist, 2019].

Definition

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 $L_{\mathcal{F}} := \delta^\top \delta$ or, equivalently, $L_{\mathcal{F}}(\mathbf{x})_v := \sum_{v,u \trianglelefteq e} \mathcal{F}_{v \trianglelefteq e}^\top (\mathcal{F}_{v \trianglelefteq e} \mathbf{x}_v - \mathcal{F}_{u \trianglelefteq e} \mathbf{x}_u)$

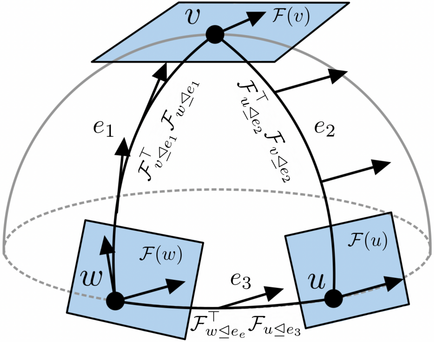
$$L_{\mathcal{F}} := \begin{array}{c} \\ v \\ \vdots \\ u \end{array} \begin{array}{c} \begin{array}{ccc} v & \cdots & u \end{array} \\ \begin{bmatrix} \sum_{v \trianglelefteq e} \mathcal{F}_{v \trianglelefteq e}^\top \mathcal{F}_{v \trianglelefteq e} & \cdots & -\mathcal{F}_{v \trianglelefteq e}^\top \mathcal{F}_{u \trianglelefteq e} \\ \vdots & \ddots & \vdots \\ -\mathcal{F}_{u \trianglelefteq e}^\top \mathcal{F}_{v \trianglelefteq e} & \cdots & \sum_{u \trianglelefteq e} \mathcal{F}_{u \trianglelefteq e}^\top \mathcal{F}_{u \trianglelefteq e} \end{bmatrix} \end{array}$$

The normalised Laplacian $\Delta_{\mathcal{F}} := D^{-1/2} L_{\mathcal{F}} D^{-1/2}$, where D is the block-diagonal of $L_{\mathcal{F}}$.

Assuming all stalks have dimension d and the graph has n nodes, this is an $nd \times nd$ matrix, generalising the $n \times n$ (normalised) graph Laplacian matrix.

Discrete Vector Bundles

The sheaves $(G, \mathcal{F})$ with orthogonal restriction maps (i.e. $\mathcal{F}_{v \trianglelefteq e} \in O(d)$) are called *discrete $O(d)$-bundles* because they are a discrete equivalent of vector bundles.



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.

The Power of Sheaf Diffusion

Sheaf Diffusion

We can represent the d-dimensional vector-features of the nodes as a matrix $\mathbf{X} \in \mathbb{R}^{(nd)\times f}$ with f feature channels, whose columns are vectors in $C^0(G; \mathcal{F})$.

Sheaf Diffusion

We can represent the d-dimensional vector-features of the nodes as a matrix $\mathbf{X} \in \mathbb{R}^{(nd)\times f}$ with f feature channels, whose columns are vectors in $C^0(G;\mathcal{F})$.

We are interested in the asymptotic behaviour of spatially discretised *sheaf diffusion* process governed by the PDE:

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

Sheaf Diffusion

We can represent the d-dimensional vector-features of the nodes as a matrix $\mathbf{X} \in \mathbb{R}^{(nd)\times f}$ with f feature channels, whose columns are vectors in $C^0(G; \mathcal{F})$.

We are interested in the asymptotic behaviour of spatially discretised *sheaf diffusion* process governed by the PDE:

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

Theorem (Hodge Theorem [Hansen & Ghrist, 2020])

In the infinite-time limit, the features converge to the projection of $\mathbf{X}(0)$ *into* $\ker(\Delta_{\mathcal{F}})$, *which is the subspace of signals* $\mathbf{x}$ *such that* $\mathcal{F}_{v\trianglelefteq e}D_v^{-1/2}\mathbf{x}_v = \mathcal{F}_{u\trianglelefteq e}D_u^{-1/2}\mathbf{x}_u$ *for all* $e \in E$.

Sheaf Diffusion

We can represent the d-dimensional vector-features of the nodes as a matrix $\mathbf{X} \in \mathbb{R}^{(nd)\times f}$ with f feature channels, whose columns are vectors in $C^0(G;\mathcal{F})$.

We are interested in the asymptotic behaviour of spatially discretised *sheaf diffusion* process governed by the PDE:

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

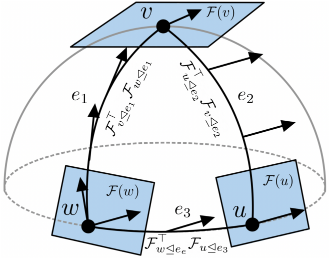
Theorem (Hodge Theorem [Hansen & Ghrist, 2020])

In the infinite-time limit, the features converge to the projection of $\mathbf{X}(0)$ *into* $\ker(\Delta_{\mathcal{F}})$*, which is the subspace of signals* $\mathbf{x}$ *such that* $\mathcal{F}_{v \trianglelefteq e} D_v^{-1/2}\mathbf{x}_v = \mathcal{F}_{u \trianglelefteq e} D_u^{-1/2}\mathbf{x}_u$ *for all* $e \in E$.

Intuition.

All the private opinions converge towards a configuration where everyone agrees with all their neighbours in the discourse space (up to a $D^{-1/2}$ normalisation).

Harmonic Space of the Sheaf Laplacian

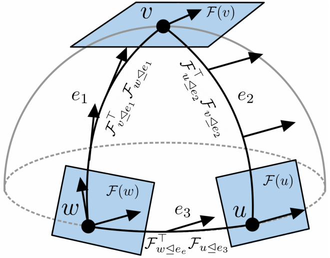


The transport is not path-independent because the vector returns in another position.

Let $(G, \mathcal{F})$ be a discrete $O(d)$-bundle. Given nodes $v, u \in V$ and a path $\gamma_{v \to u} = (v, v_1, \dots, v_\ell, u)$ 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:

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

Harmonic Space of the Sheaf Laplacian



The transport is not path-independent because the vector returns in another position.

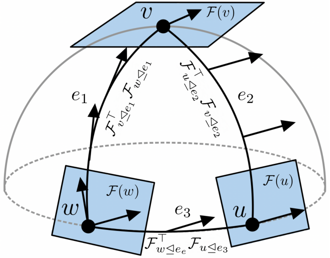
Let $(G, \mathcal{F})$ be a discrete $O(d)$-bundle. Given nodes $v, u \in V$ and a path $\gamma_{v\to u} = (v, v_1, \dots, v_\ell, u)$ from v to u, we consider a notion of **transport** from the stalk $\mathcal{F}(v)$ to the stalk $\mathcal{F}(u)$ via map composition:

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

Main idea

The harmonic space of the Laplacian is characterised by the path-independence of the transport.

Harmonic Space of the Sheaf Laplacian



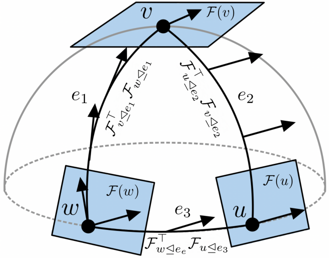
The transport is not path-independent because the vector returns in another position.

Let $\mathcal{F}$ be a discrete $O(d)$ bundle over a connected graph G with n nodes.

Proposition

Let $r := \max_{\gamma_{v\to u}, \gamma'_{v\to u}} ||\mathbf{P}^{\gamma}_{v\to u} - \mathbf{P}^{\gamma'}_{v\to u}||$, *then we have* $\lambda_0^{\mathcal{F}} \leq \frac{r^2}{2}$.

Harmonic Space of the Sheaf Laplacian



The transport is not path-independent because the vector returns in another position.

Let $\mathcal{F}$ be a discrete $O(d)$ bundle over a connected graph G with n nodes.

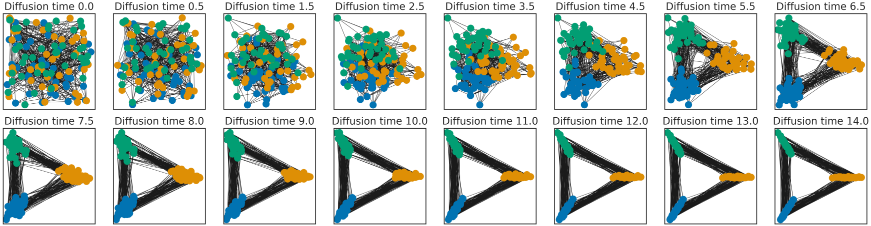
Proposition

Let $r := \max_{\gamma_{v\to u},\gamma'_{v\to u}} ||\mathbf{P}^{\gamma}_{v\to u} - \mathbf{P}^{\gamma'}_{v\to u}||$, *then we have* $\lambda_0^{\mathcal{F}} \leq \frac{r^2}{2}$.

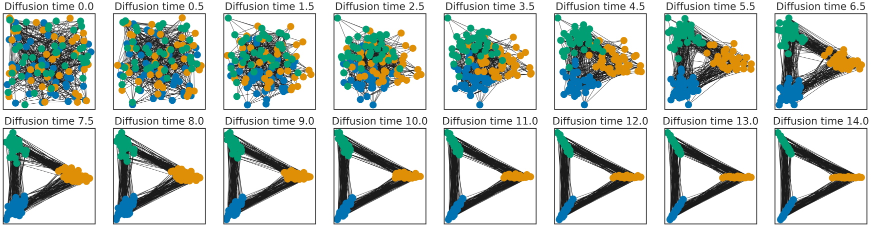
Proposition

Let $||(\mathbf{P}^{\gamma}_{v\to v} - \mathbf{I})\mathbf{x}|| \geq \epsilon||\mathbf{x}||$ *for all cycles* $\gamma_{v\to v}$. *Then* $\lambda_0^{\mathcal{F}} \geq \epsilon^2(2\mathrm{diam}(G)n\,d_{max})^{-1}$.

The Separation Power of Sheaf Diffusion



The Separation Power of Sheaf Diffusion



Definition

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 diffusion time-limit for a dense subset $\mathcal{X}_{\mathcal{F}} \subset \mathbb{R}^{nd \times f}$ of initial conditions.

The Curse of Symmetric Transport

Definition

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

The Curse of Symmetric Transport

Definition

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

Proposition

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 $\mathcal{H}^1_{\mathrm{sym}}$ has linear separation power over $\mathcal{G}$.

The Curse of Symmetric Transport

Definition

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

Proposition

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 $\mathcal{H}^1_{\mathrm{sym}}$ has linear separation power over $\mathcal{G}$.

Proposition

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 $\mathcal{H}^1_{\mathrm{sym}}$ cannot linearly separate any graph in $\mathcal{G}$ for any initial conditions $\mathbf{X}(0) \in \mathbb{R}^{n \times f}$.

Lying is good

Definition

Dropping the equality constraint, $\mathcal{H}^d_* := \{(\mathcal{F}, G) \mid \det(\mathcal{F}_{v \trianglelefteq e}) \neq 0\}$

Lying is good

Definition

Dropping the equality constraint, $\mathcal{H}_*^d := \{(\mathcal{F}, G) \mid \det(\mathcal{F}_{v \trianglelefteq e}) \neq 0\}$

Proposition

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

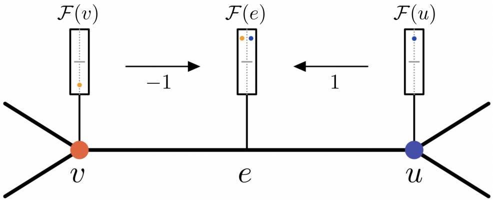
Lying is good

Definition

Dropping the equality constraint, $\mathcal{H}_*^d := \{(\mathcal{F}, G) \mid \det(\mathcal{F}_{v \trianglelefteq e}) \neq 0\}$

Proposition

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



Opposite signs lead to opinion polarisation.

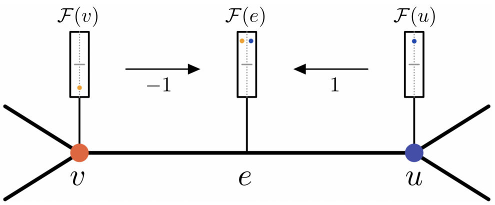
Lying is good

Definition

Dropping the equality constraint, $\mathcal{H}^d_* := \{(\mathcal{F}, G) \mid \det(\mathcal{F}_{v \trianglelefteq e}) \neq 0\}$

Proposition

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



Opposite signs lead to opinion polarisation.

This explains why a recent body of work [Yan et al., 2021, Chien et al., 2021] has found negatively-weighted edges to help in heterophilic settings. If $\mathcal{F}_{v \trianglelefteq e}\mathcal{F}_{u \trianglelefteq e} < 0$, then $\mathcal{F}_{v \trianglelefteq e} \neq \mathcal{F}_{u \trianglelefteq e}$.

The blessing of dimensionality

Even with all this additional flexibility, dimension $d = 1$ still has a major limitation.

Proposition

Let G be a connected graph with $C \geq 3$ classes. $\mathcal{H}^1_*$ cannot separate any $\mathbf{X}(0) \in \mathbb{R}^{n \times f}$.

The blessing of dimensionality

Even with all this additional flexibility, dimension $d = 1$ still has a major limitation.

Proposition

Let G be a connected graph with $C \geq 3$ classes. $\mathcal{H}^1_*$ cannot separate any $\mathbf{X}(0) \in \mathbb{R}^{n \times f}$.

With sufficient stalk dimension (i.e. width), this can be fixed.

Definition

Consider the class of sheaves with diagonal invertible maps and d-dimensional stalks:

$$\mathcal{H}^d_{\mathrm{diag}} := \{(\mathcal{F}, G) \mid \mathcal{F}_{v \trianglelefteq e} = \text{ invertible diagonal matrix}\}$$

The blessing of dimensionality

Even with all this additional flexibility, dimension $d = 1$ still has a major limitation.

Proposition

Let G be a connected graph with $C \geq 3$ classes. $\mathcal{H}^1_*$ cannot separate any $\mathbf{X}(0) \in \mathbb{R}^{n \times f}$.

With sufficient stalk dimension (i.e. width), this can be fixed.

Definition

Consider the class of sheaves with diagonal invertible maps and d-dimensional stalks:

$$\mathcal{H}^d_{\mathrm{diag}} := \{(\mathcal{F}, G) \mid \mathcal{F}_{v \trianglelefteq e} = \text{ invertible diagonal matrix}\}$$

Proposition

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

Beyond diagonal maps

Diagonal restriction maps are extremely simple. Can we do better?

Beyond diagonal maps

Diagonal restriction maps are extremely simple. Can we do better?

Definition

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

Beyond diagonal maps

Diagonal restriction maps are extremely simple. Can we do better?

Definition

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

Theorem

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

Beyond diagonal maps

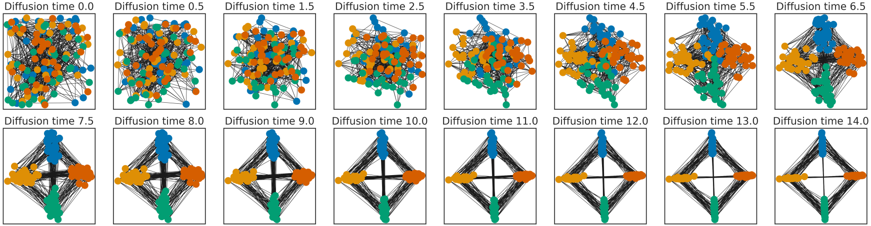
Diagonal restriction maps are extremely simple. Can we do better?

Definition

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

Theorem

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



The Power of Sheaf Diffusion: Review

In summary, we showed that:

1. Sheaf diffusion on a trivial sheaf with symmetric (scalar) restriction maps (as implicitly used in standard graph convolutions) has linear separation power only in certain (homophilic) settings.

The Power of Sheaf Diffusion: Review

In summary, we showed that:

1. Sheaf diffusion on a trivial sheaf with symmetric (scalar) restriction maps (as implicitly used in standard graph convolutions) has linear separation power only in certain (homophilic) settings.
2. Dealing with heterophilic data and oversmoothing requires non-symmetric restriction maps.

The Power of Sheaf Diffusion: Review

In summary, we showed that:

1. Sheaf diffusion on a trivial sheaf with symmetric (scalar) restriction maps (as implicitly used in standard graph convolutions) has linear separation power only in certain (homophilic) settings.
2. Dealing with heterophilic data and oversmoothing requires non-symmetric restriction maps.
3. Higher-dimensional stalks and more complex restriction maps lead to stronger separation power.

The Power of Sheaf Diffusion: Review

In summary, we showed that:

1. Sheaf diffusion on a trivial sheaf with symmetric (scalar) restriction maps (as implicitly used in standard graph convolutions) has linear separation power only in certain (homophilic) settings.
2. Dealing with heterophilic data and oversmoothing requires non-symmetric restriction maps.
3. Higher-dimensional stalks and more complex restriction maps lead to stronger separation power.
4. Node classification can be reduced to finding the right sheaf.

Asymptotics of Sheaf Convolutions

Sheaf Convolutions as Discrete Diffusion

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

$$\mathbf{X}(t+1) = \mathbf{X}(t) - \Delta_{\mathcal{F}}\mathbf{X}(t) = (\mathbf{I}_{nd} - \Delta_{\mathcal{F}})\mathbf{X}(t) \tag{3}$$

Sheaf Convolutions as Discrete Diffusion

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

$$\mathbf{X}(t+1) = \mathbf{X}(t) - \Delta_{\mathcal{F}}\mathbf{X}(t) = (\mathbf{I}_{nd} - \Delta_{\mathcal{F}})\mathbf{X}(t) \tag{3}$$

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 **Sheaf Convolutional Network (SCN)** originally proposed by [Hansen & Gebhart, 2020]:

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

Sheaf Convolutions as Discrete Diffusion

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

$$\mathbf{X}(t+1) = \mathbf{X}(t) - \Delta_{\mathcal{F}}\mathbf{X}(t) = (\mathbf{I}_{nd} - \Delta_{\mathcal{F}})\mathbf{X}(t) \tag{3}$$

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 **Sheaf Convolutional Network (SCN)** originally proposed by [Hansen & Gebhart, 2020]:

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

Question

Can the parametrised version avoid $\ker(\Delta_{\mathcal{F}})$ in the limit? This is important if $\mathcal{F}$ is not perfect. For GCNs, we know that the answer is (usually) no [Oono & Suzuki, 2020, Cai & Wang, 2020].

Tracking the Sheaf Dirichlet Energy

The Sheaf Dirichlet energy tells us how harmonic a signal is with respect to the (normalised) sheaf Laplacian.

Definition

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

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

Tracking the Sheaf Dirichlet Energy

The Sheaf Dirichlet energy tells us how harmonic a signal is with respect to the (normalised) sheaf Laplacian.

Definition

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

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

Idea

We can track this quantity as $\mathbf{x}$ is passed through the layers of an SCN.

Harmonic Convergence in Trivial Bundles

Define by $\mathcal{H}^d_{\mathrm{orth,sym}} := \mathcal{H}^d_{\mathrm{orth}} \cap \mathcal{H}^d_{\mathrm{sym}}$ and let $\lambda_* = \max\left((\lambda_{\min} - 1)^2, (\lambda_{\max} - 1)^2\right)$

Theorem

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

Harmonic Convergence in Trivial Bundles

Define by $\mathcal{H}^d_{\mathrm{orth,sym}} := \mathcal{H}^d_{\mathrm{orth}} \cap \mathcal{H}^d_{\mathrm{sym}}$ and let $\lambda_* = \max\left((\lambda_{\mathrm{min}} - 1)^2, (\lambda_{\mathrm{max}} - 1)^2\right)$

Theorem

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

Importantly, the symmetry condition is necessary.

Proposition

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

Harmonic Convergence in $d = 1$

Theorem

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).

Asymptotics of Sheaf Convolutions: Review

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.

Asymptotics of Sheaf Convolutions: Review

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. Asymmetric relations play an important role in that.

Learning Sheaves

Learning Sheaves

In practice, the ground truth sheaf for a task is unknown, so we aim to learn it from data. We consider the following diffusion-type equation:

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

Learning Sheaves

In practice, the ground truth sheaf for a task is unknown, so we aim to learn it from data. We consider the following diffusion-type equation:

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

Main point

The sheaf evolves over time as a function of the data $(G, \mathcal{F}(t)) = g(G, \mathbf{X}(t); \theta)$.

Learning Sheaves

In practice, the ground truth sheaf for a task is unknown, so we aim to learn it from data. We consider the following diffusion-type equation:

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

Main point

The sheaf evolves over time as a function of the data $(G, \mathcal{F}(t)) = g(G, \mathbf{X}(t); \theta)$.

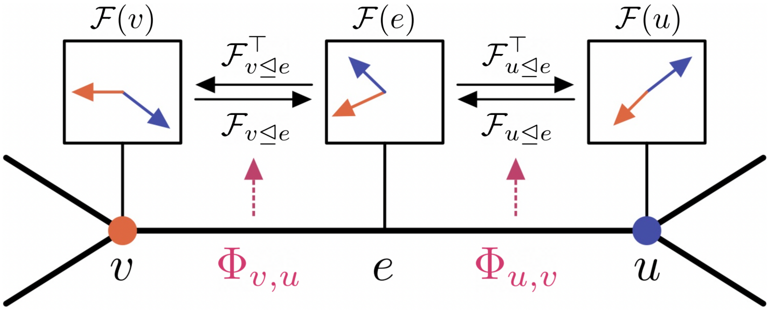
We also consider a discrete version of this equation, with different weights at each layer t.

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

Learning the restriction maps

Each $d \times d$ matrix $\mathcal{F}_{v \trianglelefteq e}$ is learned via a parametric function $\Phi : \mathbb{R}^{d\times 2} \to \mathbb{R}^{d \times d}$:

$$\mathcal{F}_{v \trianglelefteq e := (v,u)} = \Phi(\mathbf{x}_v, \mathbf{x}_u) \tag{7}$$



The restriction maps are learned from data.

Computational Complexity

Let n, m be the number of nodes and edges of the graph and consider that all stalks have dimension d. The computational complexity varies with the type of maps we learn:

- **Diagonal**: Matrix multiplication becomes scalar multiplication, so complexity is $\mathcal{O}(d(n+m))$.

Computational Complexity

Let n, m be the number of nodes and edges of the graph and consider that all stalks have dimension d. The computational complexity varies with the type of maps we learn:

- **Diagonal**: Matrix multiplication becomes scalar multiplication, so complexity is $\mathcal{O}(d(n+m))$.
- **Orthogonal / General**: Because matrix multiplication is needed to construct the transport maps: $\mathcal{O}(d^3(n+m))$. However, at inference time, one can use a pre-computed Laplacian and complexity becomes $\mathcal{O}(d^2(n+m))$.

Computational Complexity

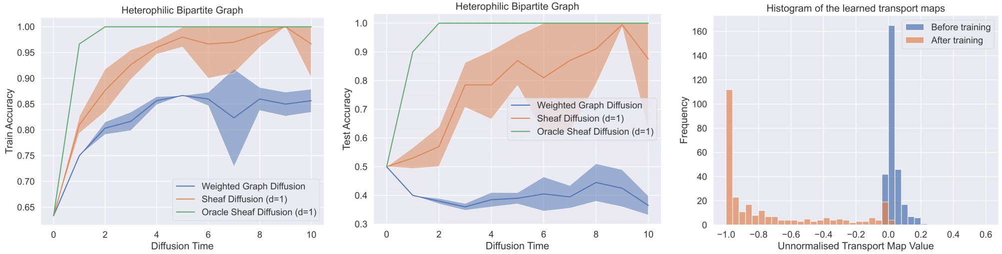
Let n, m be the number of nodes and edges of the graph and consider that all stalks have dimension d. The computational complexity varies with the type of maps we learn:

- **Diagonal**: Matrix multiplication becomes scalar multiplication, so complexity is $\mathcal{O}(d(n+m))$.
- **Orthogonal / General**: Because matrix multiplication is needed to construct the transport maps: $\mathcal{O}(d^3(n+m))$. However, at inference time, one can use a pre-computed Laplacian and complexity becomes $\mathcal{O}(d^2(n+m))$.
- **Other**: One can flexibly adjust between these two extremes by making the transport maps more sparse (e.g. block-diagonal).

Results

Synthetic Experiment: Opinion Polarisation

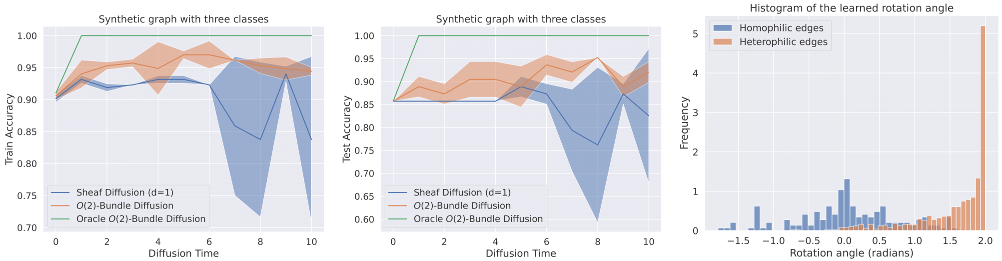
We have a bipartite graph with equally sized partitions that we try to distinguish. $\mathbf{X}(0)$ is not linearly separable. We use a simple sheaf diffusion process with a learned sheaf Laplacian (i.e. no weights and non-linearities)



Training (*Left*) and Testing (*Middle*) accuracy as a function of diffusion time. Learned sheaf Laplacian for $t >> 0$. (*Right*)

Synthetic Experiment in 2D

We have a graph with $C = 3$ classes and homophily level $h = 0.2$. We compare a vector bundle diffusion model with sheaf diffusion with $d = 1$. As before, we use no weights and no activations.



Training (*Left*) and Testing (*Middle*) accuracy as a function of diffusion time. Learned transport rotation angle for $t >> 0$. (*Right*)

Real-World Evaluation

We evaluate on multiple node-classifications tasks with various degrees of homophily [Rozemberczki et al, 2019, Pei et al, 2019].

	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

References

Cristian Bodnar, Francesco Di Giovanni, Benjamin Paul Chamberlain, Pietro Liò, Michael M. Bronstein (2022)
Neural Sheaf Diffusion: A Topological Perspective on Heterophily and Oversmoothing in GNNs
Preprint 2022

Thomas N. Kipf, Max Welling
Semi-Supervised Classification with Graph Convolutional Networks
ICLR 2017

Kenta Oono, Taiji Suzuki (2020)
Graph Neural Networks Exponentially Lose Expressive Power for Node Classification
ICLR 2020

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, Danai Koutra
Beyond Homophily in Graph Neural Networks: Current Limitations and Effective Designs
NeurIPS 2020

References

Jakob Hansen, Robert Ghrist
Opinion Dynamics on Discourse Sheaves
SIAM J. Appl. Math., 81(5), 2033–2060

Jakob Hansen, Robert Ghrist
Toward a Spectral Theory of Cellular Sheaves
Journal of Applied and Computational Topology volume 3, pages 315–358 (2019)

Yujun Yan, Milad Hashemi, Kevin Swersky, Yaoqing Yang, Danai Koutra
Two Sides of the Same Coin: Heterophily and Oversmoothing in Graph Convolutional Neural Networks
Preprint (2021)

Eli Chien, Jianhao Peng, Pan Li, Olgica Milenkovic
Adaptive Universal Generalized PageRank Graph Neural Network
ICLR 2021

References

Chen Cai, Yusu Wang
A Note on Over-Smoothing for Graph Neural Networks
ICML 2020 Graph Representation Learning Workshop

Jakob Hansen, Thomas Gebhart
Sheaf Neural Networks
NeurIPS 2020 Workshop on TDA and Beyond

Benedek Rozemberczki, Carl Allen, Rik Sarkar
Multi-scale Attributed Node Embedding
Journal of Complex Networks

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, Bo Yang
Geom-GCN: Geometric Graph Convolutional Networks
ICLR 2020

Thank you for your attention!

Email: `cb2015@cam.ac.uk`
Twitter: `@crisbodnar`