Graph neural networks through the lens of multi-particle dynamics and gradient flows

Francesco Di Giovanni

Twitter

AMMI

July 22, 2022

Based on *Graph Neural Networks as Gradient Flows*, arXiv:2206.10991, (2022)

Joint work with J. Rowbottom*, B. Chamberlain, T. Markovich, M. Bronstein

Presentation outline

- Graph preliminaries
- Spectral analysis and Dirichlet energy on graphs
- Dynamical systems on graphs
- MPNNs as multi-particle systems and the gradient flow framework (GRAFF)
- Presentation of *Graph Neural Networks as Gradient Flows*

Introduction

Preliminaries on graph operators

- $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ is an *undirected* graph with $|\mathsf{V}| = n$ and $i \sim j$ if $(i, j) \in \mathsf{E}$
- $\mathbf{A}, \mathbf{D}$ are $n \times n$ adjacency and (diagonal) degree matrices
- The *normalized* adjacency is $\bar{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$
- The Laplacian $\boldsymbol{\Delta} = \mathbf{I} - \bar{\mathbf{A}}$ is an operator acting on signals $\mathbf{f} : \mathsf{V} \to \mathbb{R}$ as

$$\boxed{(\boldsymbol{\Delta}\mathbf{f})_i = f_i - \sum_{j \sim i} \frac{f_j}{\sqrt{d_i d_j}}}$$

Preliminaries on graph operators

- $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ is an *undirected* graph with $|\mathsf{V}| = n$ and $i \sim j$ if $(i, j) \in \mathsf{E}$
- $\mathbf{A}, \mathbf{D}$ are $n \times n$ adjacency and (diagonal) degree matrices
- The *normalized* adjacency is $\bar{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$
- The Laplacian $\boldsymbol{\Delta} = \mathbf{I} - \bar{\mathbf{A}}$ is an operator acting on signals $\mathbf{f} : \mathsf{V} \to \mathbb{R}$ as

$$\boxed{(\boldsymbol{\Delta}\mathbf{f})_i = f_i - \sum_{j \sim i} \frac{f_j}{\sqrt{d_i d_j}}}$$

The Laplacian $\boldsymbol{\Delta} \succeq 0 \to$ eigenvalues satisfy $0 = \lambda_0^{\boldsymbol{\Delta}} \leq \ldots \leq \lambda_{n-2}^{\boldsymbol{\Delta}} \leq \rho_{\boldsymbol{\Delta}}$, with $\rho_{\boldsymbol{\Delta}} \leq 2$, and are called (graph) *frequencies*, eigenvectors are denoted by $\{\phi_\ell^{\boldsymbol{\Delta}}\}_{\ell=0}^{n-1}$

Signal on graphs: Dirichlet energy and smoothness

Consider a signal (feature) $\mathbf{f} : \mathsf{V} \to \mathbb{R}$ e.g. temperature of each node

We write $\mathbf{f} = (f_1, \ldots, f_n)^\top \to \mathbf{f} = \sum_\ell c_\ell \phi_\ell^{\mathbf{\Delta}}$

$\mathbf{\Delta}$ can be used to measure smoothness of $\mathbf{f}$: the Dirichlet energy[1] $\mathcal{E}^{\mathrm{Dir}}$ is defined by

$$\boxed{\mathcal{E}^{\mathrm{Dir}}(\mathbf{f}) := \frac{1}{4}\sum_{i\sim j} \Big|\Big| \frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \Big|\Big|^2 = \frac{1}{2}\langle \mathbf{f}, \mathbf{\Delta}\mathbf{f}\rangle = \frac{1}{2}\sum_\ell \lambda_\ell^{\mathbf{\Delta}} c_\ell^2.}$$

$\to$ the frequency components of $\mathbf{f}$ determine the variations of the signal along edges

The quantity $f_i/\sqrt{d_i} - f_j/\sqrt{d_j} := \nabla \mathbf{f}(i,j)$ is the **gradient** of $\mathbf{f}$ along th edge $(i,j) \in \mathsf{E}$

[1] Zhou and Schölkopf (2005)

A rough picture: low-pass vs high-pass filtering

Consider a dynamical process $t \mapsto \mathbf{f}(t) \in \mathbb{R}^n$ starting at $\mathbf{f}_0 \to \mathbf{f}(t) = \sum_\ell c_\ell(t)\boldsymbol{\phi}_\ell^{\boldsymbol{\Delta}}$

A rough picture: low-pass vs high-pass filtering

Consider a dynamical process $t \mapsto \mathbf{f}(t) \in \mathbb{R}^n$ starting at $\mathbf{f}_0 \to \mathbf{f}(t) = \sum_\ell c_\ell(t)\boldsymbol{\phi}_\ell^{\boldsymbol{\Delta}}$

If the high-frequency components $|c_\ell(t)|$, with $\ell >> 0$, decrease with time, then the process acts as '**low-pass** filtering' $\to$ smooths the signal out

A rough picture: low-pass vs high-pass filtering

Consider a dynamical process $t \mapsto \mathbf{f}(t) \in \mathbb{R}^n$ starting at $\mathbf{f}_0 \to \mathbf{f}(t) = \sum_\ell c_\ell(t)\boldsymbol{\phi}_\ell^{\boldsymbol{\Delta}}$

If the high-frequency components $|c_\ell(t)|$, with $\ell >> 0$, decrease with time, then the process acts as '**low-pass** filtering' $\to$ smooths the signal out

If the low-frequency components $|c_\ell(t)|$, with $\ell \sim 0$, decrease with time, then the process acts as '**high-pass** filtering' $\to$ sharpens the signal

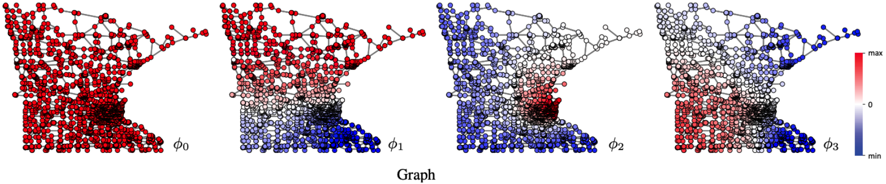


Figure 1: First four Laplacian eigenvectors of Minnesota Road graph. Figure taken from Bronstein et al. (2017)

A prototypical low-pass filtering: the graph heat equation

Consider an input signal $\mathbf{f}_0 : \mathsf{V} \to \mathbb{R}$ and recall that $\mathbf{f} \mapsto \mathcal{E}^{\mathrm{Dir}}(\mathbf{f}) = \frac{1}{2}\langle \mathbf{f}, \boldsymbol{\Delta}\mathbf{f}\rangle$

If we want to *minimize* $\mathcal{E}^{\mathrm{Dir}}$ $\to$ take infinitesimal steps in the direction of steepest descent

$$\text{Heat equation}: \quad \dot{\mathbf{f}}(t) = -\nabla_{\mathbf{f}}\mathcal{E}^{\mathrm{Dir}}(\mathbf{f})(t) = -\boldsymbol{\Delta}\mathbf{f}(t), \quad \mathbf{f}(0) = \mathbf{f}_0.$$

This is a **gradient flow**: $\dot{\mathcal{E}}^{\mathrm{Dir}}(\mathbf{f}(t)) \leq 0$ and $\mathbf{f}(t) \to \mathbf{f}_\infty$ s.t. $\boldsymbol{\Delta}\mathbf{f}_\infty = \mathbf{0}$.

Low-pass dynamics $\to$ 'features become indistinguishable' when $t >> 1$

Multiple channels

Consider $\mathbf{F} : \mathsf{V} \to \mathbb{R}^d$ with matrix representation $\mathbf{F} \in \mathbb{R}^{n \times d} \to \mathcal{E}^{\mathrm{Dir}}$ can be extended as

$$\boxed{\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}) = \frac{1}{4} \sum_{(i,j) \in \mathsf{E}} || \frac{\mathbf{f}_i}{\sqrt{d_i}} - \frac{\mathbf{f}_j}{\sqrt{d_j}} ||^2 = \frac{1}{2} \mathrm{trace}(\mathbf{F}^\top \mathbf{\Delta} \mathbf{F})}$$

The gradient flow of $\mathcal{E}^{\mathrm{Dir}}$ yields heat equation in each feature channel:

$$\boxed{\dot{\mathbf{f}}^r(t) = -\mathbf{\Delta} \mathbf{f}^r(t), \quad 1 \leq r \leq d}$$

The $\otimes$ formalism

We can vectorize a matrix signal $\mathbf{F} \in \mathbb{R}^{n \times d} \to \mathrm{vec}(\mathbf{F}) \in \mathbb{R}^{nd}$

We use the *Kronecker product* $\mathbf{I}_d \otimes \mathbf{\Delta} \in \mathbb{R}^{nd} \times \mathbb{R}^{nd}$ to rewrite $\mathcal{E}^{\mathrm{Dir}}$ as

$$\boxed{\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}) = \frac{1}{2}\langle \mathrm{vec}(\mathbf{F}), (\mathbf{I}_d \otimes \mathbf{\Delta})\mathrm{vec}(\mathbf{F})\rangle}$$

The heat equation can also be rewritten by ‘stacking the columns as’

$$\boxed{\mathrm{vec}(\dot{\mathbf{F}}(t)) = -(\mathbf{I}_d \otimes \mathbf{\Delta})\mathrm{vec}(\mathbf{F}(t))}$$

Upshot: $\otimes$ formalism reduces a *matrix* ODE to a *vector* ODE $\to$ vectorized ODEs are much easier to deal with

A motivating example

How to determine if a dynamical process on a graph is dominated by the low or high frequencies?

A motivating example

How to determine if a dynamical process on a graph is dominated by the low or high frequencies? Use $\mathcal{E}^{\mathrm{Dir}}$ *after* normalization

A motivating example

How to determine if a dynamical process on a graph is dominated by the low or high frequencies? Use $\mathcal{E}^{\mathrm{Dir}}$ *after* normalization

Consider $\dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t) \iff \mathrm{vec}(\dot{\mathbf{F}}(t)) = (\mathbf{I}_d \otimes \bar{\mathbf{A}})\mathrm{vec}(\mathbf{F}(t))$, with $\mathbf{F}(0) = \mathbf{F}_0$

Recall that $\bar{\mathbf{A}} = \mathbf{I} - \boldsymbol{\Delta}$ so we can solve as

$$\boxed{\mathbf{f}^r(t) = e^{\bar{\mathbf{A}}t}\,\mathbf{f}^r(0) = e^{(\mathbf{I}-\boldsymbol{\Delta})t}\,\mathbf{f}^r(0), \quad 1 \leq r \leq d}$$

A motivating example

How to determine if a dynamical process on a graph is dominated by the low or high frequencies? Use $\mathcal{E}^{\mathrm{Dir}}$ *after* normalization

Consider $\dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t) \Longleftrightarrow \mathrm{vec}(\dot{\mathbf{F}}(t)) = (\mathbf{I}_d \otimes \bar{\mathbf{A}})\mathrm{vec}(\mathbf{F}(t))$, with $\mathbf{F}(0) = \mathbf{F}_0$

Recall that $\bar{\mathbf{A}} = \mathbf{I} - \boldsymbol{\Delta}$ so we can solve as

$$\boxed{\mathbf{f}^r(t) = e^{\bar{\mathbf{A}}t}\,\mathbf{f}^r(0) = e^{(\mathbf{I}-\boldsymbol{\Delta})t}\,\mathbf{f}^r(0), \quad 1 \leq r \leq d}$$

Expand each channel in the basis $\{\phi_\ell^{\boldsymbol{\Delta}}\}$ satisfying $\bar{\mathbf{A}}\phi_\ell^{\boldsymbol{\Delta}} = (1-\lambda_\ell^{\boldsymbol{\Delta}})\phi_\ell^{\boldsymbol{\Delta}}$:

$$\boxed{\mathbf{f}^r(t) = \sum_\ell e^{(1-\lambda_\ell^{\boldsymbol{\Delta}})t}\langle \mathbf{f}^r(0), \phi_\ell^{\boldsymbol{\Delta}}\rangle \phi_\ell^{\boldsymbol{\Delta}}}$$

A motivating example

Recall that $\phi_0^{\mathbf{\Delta}}$ is the smoothest eigenvector i.e. $\mathbf{\Delta}\phi_0^{\mathbf{\Delta}} = \mathbf{0}$

The projection along $\phi_0^{\mathbf{\Delta}}$ is the one growing the *fastest*[2] since

$$\boxed{\langle \mathbf{f}^r(t), \phi_0^{\mathbf{\Delta}} \rangle = \mathbf{e}^{(\mathbf{1}-\mathbf{0})\mathbf{t}} \langle \mathbf{f}^r(0), \phi_0^{\mathbf{\Delta}} \rangle}$$

The dynamics are 'dominated' by the low-frequencies: does $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) \to 0$?

[2] Unless $|\langle \mathbf{f}^r(0), \phi_0^{\mathbf{\Delta}} \rangle| = 0$ which is only true in a smaller subspace of $\mathbb{R}^n$

[3] Unless $\langle \mathbf{f}^r(0), \phi_\ell^{\mathbf{\Delta}} \rangle = 0$ for all $\ell > 0$

A motivating example

Recall that $\phi_0^{\boldsymbol{\Delta}}$ is the smoothest eigenvector i.e. $\boldsymbol{\Delta}\phi_0^{\boldsymbol{\Delta}} = \mathbf{0}$

The projection along $\phi_0^{\boldsymbol{\Delta}}$ is the one growing the *fastest*[2] since

$$\boxed{\langle \mathbf{f}^r(t), \phi_0^{\boldsymbol{\Delta}}\rangle = \mathbf{e}^{(\mathbf{1}-\mathbf{0})\mathbf{t}}\langle \mathbf{f}^r(0), \phi_0^{\boldsymbol{\Delta}}\rangle}$$

The dynamics are ‘dominated’ by the low-frequencies: does $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) \to 0$? **No:**[3]

$$\boxed{\mathcal{E}^{\mathrm{Dir}}(\mathbf{f}^r(t)) = \frac{1}{2}\langle \mathbf{f}^r(t), \boldsymbol{\Delta}\mathbf{f}^r(t)\rangle = \sum_{\ell>0} e^{(1-\lambda_\ell^{\boldsymbol{\Delta}})t}(\langle \mathbf{f}^r(0), \phi_\ell^{\boldsymbol{\Delta}}\rangle)^2 \to \infty}$$

[2] Unless $|\langle \mathbf{f}^r(0), \phi_0^{\boldsymbol{\Delta}}\rangle| = 0$ which is only true in a smaller subspace of $\mathbb{R}^n$

[3] Unless $\langle \mathbf{f}^r(0), \phi_\ell^{\boldsymbol{\Delta}}\rangle = 0$ for all $\ell > 0$

A motivating example

Looking at $\mathcal{E}^{\mathrm{Dir}}$ is not enough $\rightarrow$ we should normalize first: in fact we have

$$\boxed{\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)/||\mathbf{F}(t)||) \rightarrow 0, \quad t \rightarrow \infty}$$

and for each channel $1 \leq r \leq d \; \exists \; \mathbf{f}^r_\infty$ s.t.

$$\boxed{\mathbf{f}^r(t)/||\mathbf{f}^r(t)|| \rightarrow \mathbf{f}^r_\infty, \quad \boldsymbol{\Delta}\mathbf{f}^r_\infty = 0}$$

Upshot: Analyse $\mathbf{F}(t)$ via $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)/||\mathbf{F}(t)||)$

Definition

A dynamical system $\dot{\mathbf{F}}(t)$ initialized at $\mathbf{F}(0)$ is *Low-Frequency-Dominant* LFD if $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)/||\mathbf{F}(t)||) \to 0$ for $t \to \infty$.

Low-frequency-dominant: LFD

Definition

A dynamical system $\dot{\mathbf{F}}(t)$ initialized at $\mathbf{F}(0)$ is *Low-Frequency-Dominant* LFD if $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)/||\mathbf{F}(t)||) \to 0$ for $t \to \infty$.

Does it make sense?

Lemma

A dynamical system is LFD *iff for each sequence $t_j \to \infty$ there exist a subsequence $t_{j_k} \to \infty$ and $\mathbf{F}_\infty$ s.t. $\mathbf{F}(t_{j_k})/||\mathbf{F}(t_{j_k})|| \to \mathbf{F}_\infty$ and $\boldsymbol{\Delta}\mathbf{f}^r_\infty = \mathbf{0}$.*

High-frequency-dominant: HFD

Note that $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}) \leq \frac{1}{2}\rho_{\boldsymbol{\Delta}}||\mathbf{F}||^2 \to \mathcal{E}^{\mathrm{Dir}}(\mathbf{F}/||\mathbf{F}||) \leq \frac{1}{2}\rho_{\boldsymbol{\Delta}}$

High-frequency-dominant: HFD

Note that $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}) \leq \frac{1}{2}\rho_{\boldsymbol{\Delta}}||\mathbf{F}||^2 \to \mathcal{E}^{\mathrm{Dir}}(\mathbf{F}/||\mathbf{F}||) \leq \frac{1}{2}\rho_{\boldsymbol{\Delta}}$

Definition

A dynamical system $\dot{\mathbf{F}}(t)$ initialized at $\mathbf{F}(0)$ is *High-Frequency-Dominant* (HFD) if $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)/||\mathbf{F}(t)||) \to \rho_{\boldsymbol{\Delta}}/2$ for $t \to \infty$.

High-frequency-dominant: HFD

Note that $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}) \leq \frac{1}{2}\rho_{\mathbf{\Delta}}||\mathbf{F}||^2 \rightarrow \mathcal{E}^{\mathrm{Dir}}(\mathbf{F}/||\mathbf{F}||) \leq \frac{1}{2}\rho_{\mathbf{\Delta}}$

Definition

A dynamical system $\dot{\mathbf{F}}(t)$ initialized at $\mathbf{F}(0)$ is *High-Frequency-Dominant* (HFD) if $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)/||\mathbf{F}(t)||) \rightarrow \rho_{\mathbf{\Delta}}/2$ for $t \rightarrow \infty$.

Does it make sense?

Lemma

A dynamical system is HFD *iff for each sequence $t_j \rightarrow \infty$ there exist a subsequence $t_{j_k} \rightarrow \infty$ and $\mathbf{F}_\infty$ s.t. $\mathbf{F}(t_{j_k})/||\mathbf{F}(t_{j_k})|| \rightarrow \mathbf{F}_\infty$ and $\mathbf{\Delta}\mathbf{f}^r_\infty = \rho_{\mathbf{\Delta}}\mathbf{f}^r_\infty$.*

The formalism of MPNNs

- Graph $\mathsf{G} = (\mathsf{V}, \mathsf{E})$
- $\mathbf{F}_{\text{input}} \in \mathbb{R}^{n \times p}$ matrix representation of input node features, with rows $\{(\mathbf{f}_i)^\top_{\text{input}}\}_{i=1}^n$
- Encoding map $\psi_{\text{EN}} : \mathbb{R}^p \to \mathbb{R}^{d_0}$
- Update functions $\{\phi^t_{\text{UP}} : \mathbb{R}^{d_t} \to \mathbb{R}^{d_{t+1}}\}$ for $0 \leq t \leq T-1$, with T the *depth*

$$\text{MPNN}: \quad \mathbf{f}_i(t+1) = \phi^t_{\text{UP}}\left(\mathbf{f}_i(t), \{\{\mathbf{f}_j(t) : \; j \sim i\}\}\right), \quad \mathbf{f}_i(0) = \psi_{\text{EN}}((\mathbf{f}_i)_{\text{input}}).$$

Where and why MPNNs struggle?

- *Expressivity*: usually measured via comparison with Weisfeiler-Leman test
- *Low vs High pass*: how do MPNNs perform when we need ‘more than low-pass filters’? $\rightarrow$ related to *over-smoothing* when stacking many layers
- *Over-squashing* $\rightarrow$ are long-range dependencies accounted for? Information flow may be compromised **due to graph topology**

Homophily vs heterophily aka short vs long range interactions

Semi-supervised setting: $\mathsf{V}_{\mathrm{tr}} \subset \mathsf{V}$ labelled $\rightarrow$ predict labels on $\mathsf{V}_{\mathrm{test}}$

Homophily: Neighbours often share labels $\rightarrow$ labels are *smooth* i.e. low-pass is ‘good’

Heterophily: $1 -$ homophily $\rightarrow$ labels are *not* smooth i.e. low-pass is ‘bad’

Homophily vs heterophily aka short vs long range interactions

Semi-supervised setting: $\mathsf{V}_{\mathrm{tr}} \subset \mathsf{V}$ labelled $\rightarrow$ predict labels on $\mathsf{V}_{\mathrm{test}}$

Homophily: Neighbours often share labels $\rightarrow$ labels are *smooth* i.e. low-pass is 'good'

Heterophily: $1-$ homophily $\rightarrow$ labels are *not* smooth i.e. low-pass is 'bad'

Dual perspective: short-range relations vs long-range relations $\rightarrow$ relevant for graph classification and regression tasks on molecules

Graph Convolutional Networks

A layer of **Graph Convolutional Network (GCN)**[4] is defined by:

$$\boxed{\mathbf{F}(t+1) = \mathrm{ReLU}\left(\bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}(t)\right)}$$

$\bar{\mathbf{A}}$ is the message-passing matrix and $\mathbf{W}(t)$ is the 'channel-mixing'

[4] Kipf and Welling (2017)

[5] Nt and Maehara (2019); Oono and Suzuki (2020); Cai and Wang (2020)

Graph Convolutional Networks

A layer of **Graph Convolutional Network (GCN)**[4] is defined by:

$$\boxed{\mathbf{F}(t+1) = \mathrm{ReLU}\left(\bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}(t)\right)}$$

$\bar{\mathbf{A}}$ is the message-passing matrix and $\mathbf{W}(t)$ is the 'channel-mixing'

- Poor performance on heterophilic graphs
- Degradation when increasing depth (over-smoothing): $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) \to 0$ *if singular values of channel-mixing are **sufficiently small***[5]

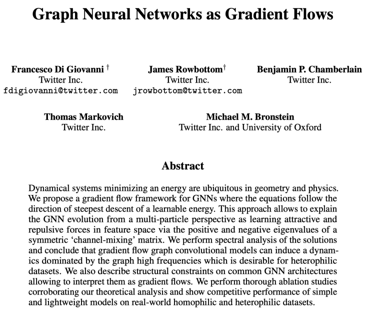
[4] Kipf and Welling (2017)

[5] Nt and Maehara (2019); Oono and Suzuki (2020); Cai and Wang (2020)

A physics-inspired approach

Node features $\rightarrow$ particles in $\mathbb{R}^d$

We propose a gradient flow framework (GRAFF) where MPNNs can be interpreted as multi-particle dynamics that minimize a learnable energy



Graph Neural Networks as Gradient Flows

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

James Rowbottom[†] Twitter Inc. jrowbottom@twitter.com

Benjamin P. Chamberlain Twitter Inc.

Thomas Markovich Twitter Inc.

Michael M. Bronstein Twitter Inc. and University of Oxford

Abstract

Dynamical systems minimizing an energy are ubiquitous in geometry and physics. We propose a gradient flow framework for GNNs where the equations follow the direction of steepest descent of a learnable energy. This approach allows to explain the GNN evolution from a multi-particle perspective as learning attractive and repulsive forces in feature space via the positive and negative eigenvalues of a symmetric 'channel-mixing' matrix. We perform spectral analysis of the solutions and conclude that gradient flow graph convolutional models can induce a dynamics dominated by the graph high frequencies which is desirable for heterophilic datasets. We also describe structural constraints on common GNN architectures allowing to interpret them as gradient flows. We perform thorough ablation studies corroborating our theoretical analysis and show competitive performance of simple and lightweight models on real-world homophilic and heterophilic datasets.

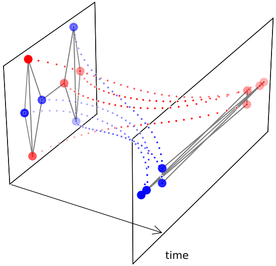


Figure 2: Actual GRAFF dynamics: attractive and repulsive forces lead to a non-smoothing process able to separate labels

Outline of the contributions

- We propose a gradient flow framework for MPNNs where the equations follow the direction of steepest descent of a learnable energy

Outline of the contributions

- We propose a gradient flow framework for MPNNs where the equations follow the direction of steepest descent of a learnable energy
- We show how the channel-mixing $\mathbf{W}$ can learn to induce either LFD or HFD dynamics via its spectrum

Outline of the contributions

- We propose a gradient flow framework for MPNNs where the equations follow the direction of steepest descent of a learnable energy
- We show how the channel-mixing $\mathbf{W}$ can learn to induce either LFD or HFD dynamics via its spectrum
- This allows us to interpret MPNNs as multi-particle dynamics with attractive and repulsive forces generated by positive and negative eigenvalues of $\mathbf{W}$

Outline of the contributions

- We propose a gradient flow framework for MPNNs where the equations follow the direction of steepest descent of a learnable energy
- We show how the channel-mixing $\mathbf{W}$ can learn to induce either LFD or HFD dynamics via its spectrum
- This allows us to interpret MPNNs as multi-particle dynamics with attractive and repulsive forces generated by positive and negative eigenvalues of $\mathbf{W}$
- Show that LFD/HFD dynamics induced by this framework adapt to the underlying homophily/heterophily

Outline of the contributions

- We propose a gradient flow framework for MPNNs where the equations follow the direction of steepest descent of a learnable energy
- We show how the channel-mixing $\mathbf{W}$ can learn to induce either LFD or HFD dynamics via its spectrum
- This allows us to interpret MPNNs as multi-particle dynamics with attractive and repulsive forces generated by positive and negative eigenvalues of $\mathbf{W}$
- Show that LFD/HFD dynamics induced by this framework adapt to the underlying homophily/heterophily

Residual networks as discrete ODEs

A ResNet $\mathbf{F}(t+\tau) = \mathbf{F}(t) + \tau \mathrm{ResNet}(\mathbf{F}(t))$ is the Euler discretization of an ODE[6] (as the step-size $\tau \to 0$)

$$\boxed{\dot{\mathbf{F}}(t) = \mathrm{ResNet}(\mathbf{F}(t))}$$

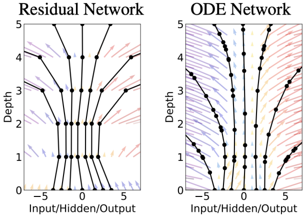


Figure 3: Dynamics of ResNet vs ODE. Figure taken from Chen et al. (2018)

ODE theory $\to$ *analysing and improving ResNets*

[6] Haber and Ruthotto (2018); Chen et al. (2018)

Residual networks as discrete ODEs

A ResNet $\mathbf{F}(t+\tau) = \mathbf{F}(t) + \tau\mathrm{ResNet}(\mathbf{F}(t))$ is the Euler discretization of an ODE[6] (as the step-size $\tau \to 0$)

$$\dot{\mathbf{F}}(t) = \mathrm{ResNet}(\mathbf{F}(t))$$

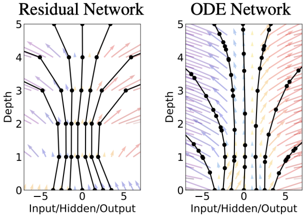


Figure 3: Dynamics of ResNet vs ODE. Figure taken from Chen et al. (2018)

ODE theory $\to$ *analysing and improving ResNets*

What about residual MPNNs?

$$\mathbf{F}(t+\tau) = \mathbf{F}(t) + \tau\mathrm{MPNN}(\mathsf{G}, \mathbf{F}(t)) \to \dot{\mathbf{F}}(t) = \mathrm{MPNN}(\mathsf{G}, \mathbf{F}(t))$$

[6] Haber and Ruthotto (2018); Chen et al. (2018)

Instances of 'continuous' MPNNs

- CGNN[7]: $\dot{\mathbf{F}}(t) = -\mathbf{\Delta}\mathbf{F}(t) + \mathbf{F}(t)\mathbf{W} + \mathbf{F}(0)$
- GRAND[8]: $\dot{\mathbf{F}}(t) = -(\mathbf{I} - \boldsymbol{\mathcal{A}}(\mathbf{F}(t)))\mathbf{F}(t)$, with $\boldsymbol{\mathcal{A}}(\mathbf{F}(t))$ a graph *attention* matrix
- (Linear) PDE-GCN[9]: $\dot{\mathbf{F}}(t) = -\mathbf{\Delta}\mathbf{F}(t)\mathbf{W}(t)^\top\mathbf{W}(t)$
- Second order (wave) equations[10]: $\ddot{\mathbf{F}}(t) = \mathrm{MPNN}(\mathsf{G}, \mathbf{F}(t)) - \gamma\mathbf{F}(t) - \alpha\dot{\mathbf{F}}(t)$

The actual equations are parametric $\rightarrow$ how to choose them?

[7] Xhonneux et al. (2020)
[8] Chamberlain et al. (2021)
[9] Eliasof et al. (2021)
[10] Eliasof et al. (2021), Rusch et al. (2022)

Dynamical systems as gradient flows

Dynamical systems are gradient flows when $\exists\, \mathcal{E} : \mathbb{R}^N \to \mathbb{R}$:

$$\boxed{\dot{\mathbf{F}}(t) = \mathrm{ODE}(\mathbf{F}(t)) = -\nabla_{\mathbf{F}}\mathcal{E}(\mathbf{F}(t)) \implies \dot{\mathcal{E}}(\mathbf{F}(t)) \leq 0.}$$

Gradient flows are easier to analyze and *interpret* since the solution $\mathbf{F}(t)$ is minimizing $\mathcal{E}$

What if we parametrize an energy rather than the MPNN equations?

Dynamical systems as gradient flows

Dynamical systems are gradient flows when $\exists\, \mathcal{E} : \mathbb{R}^N \to \mathbb{R}$:

$$\dot{\mathbf{F}}(t) = \mathrm{ODE}(\mathbf{F}(t)) = -\nabla_{\mathbf{F}}\mathcal{E}(\mathbf{F}(t)) \Longrightarrow \dot{\mathcal{E}}(\mathbf{F}(t)) \leq 0.$$

Gradient flows are easier to analyze and *interpret* since the solution $\mathbf{F}(t)$ is minimizing $\mathcal{E}$

What if we parametrize an energy rather than the MPNN equations?

Goal: Learn $\mathcal{E}_\theta$ **generalizing** $\mathcal{E}^{\mathrm{Dir}}$ $\rightarrow$ *find right notion of smoothness for the problem*

$$\dot{\mathbf{F}}(t) = \mathrm{MPNN}(\mathsf{G}, \mathbf{F}(t)) = -\nabla_{\mathbf{F}}\mathcal{E}_\theta(\mathsf{G}, \mathbf{F}(t))$$

GNNs as Gradient Flows part 1: taking inspiration from harmonic maps

Extending the formalism to graphs

Consider $\mathbf{H} = \mathbf{W}^\top \mathbf{W}$ with $\mathbf{W} \in \mathbb{R}^{d \times d}$ $\rightarrow$ measure smoothness wrt the metric $\mathbf{H}$

Extending the formalism to graphs

Consider $\mathbf{H} = \mathbf{W}^\top \mathbf{W}$ with $\mathbf{W} \in \mathbb{R}^{d \times d}$ $\rightarrow$ measure smoothness wrt the metric $\mathbf{H}$

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}) = \frac{1}{4} \sum_{(i,j) \in \mathsf{E}} ||(\nabla \mathbf{F})_{ij}||^2 \rightarrow \mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}(\mathbf{F}) := \frac{1}{4} \sum_{(i,j) \in \mathsf{E}} ||\mathbf{W}(\nabla \mathbf{F})_{ij}||^2$$

If we minimize $\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}$ *we expect* $||(\nabla \mathbf{F})_{ij}||$ *to shrink 'except' when inside* $\ker(\mathbf{H})$

Generalized harmonic flow on graphs is smoothing

We treat $\mathbf{W}$ as *learnable weights* and study the gradient flow of $\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}$:

$$\boxed{\dot{\mathbf{F}}(t) = -\nabla_{\mathbf{F}}\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}(\mathbf{F}(t)) = -\boldsymbol{\Delta}\mathbf{F}(t)\mathbf{W}^{\top}\mathbf{W}.}$$

[11] Similar to Nt and Maehara (2019); Oono and Suzuki (2020)

[12] This is different from Nt and Maehara (2019); Oono and Suzuki (2020); Cai and Wang (2020)

Generalized harmonic flow on graphs is smoothing

We treat $\mathbf{W}$ as *learnable weights* and study the gradient flow of $\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}$:

$$\dot{\mathbf{F}}(t) = -\nabla_{\mathbf{F}}\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}(\mathbf{F}(t)) = -\boldsymbol{\Delta}\mathbf{F}(t)\mathbf{W}^{\top}\mathbf{W}.$$

Proposition (Informal)

- *No* $\mathbf{W}$ *separates the limit embeddings of nodes with same degree and input features*

[11] Similar to Nt and Maehara (2019); Oono and Suzuki (2020)

[12] This is different from Nt and Maehara (2019); Oono and Suzuki (2020); Cai and Wang (2020)

Generalized harmonic flow on graphs is smoothing

We treat $\mathbf{W}$ as *learnable weights* and study the gradient flow of $\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}$:

$$\dot{\mathbf{F}}(t) = -\nabla_{\mathbf{F}}\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}(\mathbf{F}(t)) = -\boldsymbol{\Delta}\mathbf{F}(t)\mathbf{W}^{\top}\mathbf{W}.$$

Proposition (Informal)

- *No $\mathbf{W}$ separates the limit embeddings of nodes with same degree and input features*
- *If $\mathbf{W}$ has zero kernel, nodes with same degrees converge to the same representation and over-smoothing occurs*[11]

[11] Similar to Nt and Maehara (2019); Oono and Suzuki (2020)

[12] This is different from Nt and Maehara (2019); Oono and Suzuki (2020); Cai and Wang (2020)

Generalized harmonic flow on graphs is smoothing

We treat $\mathbf{W}$ as *learnable weights* and study the gradient flow of $\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}$:

$$\boxed{\dot{\mathbf{F}}(t) = -\nabla_{\mathbf{F}}\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}(\mathbf{F}(t)) = -\boldsymbol{\Delta}\mathbf{F}(t)\mathbf{W}^{\top}\mathbf{W}.}$$

Proposition (Informal)

- *No $\mathbf{W}$ separates the limit embeddings of nodes with same degree and input features*
- *If $\mathbf{W}$ has zero kernel, nodes with same degrees converge to the same representation and over-smoothing occurs*[11]
- *Over-smoothing occurs independently of the spectral radius of $\mathbf{W}$ if its eigenvalues are positive – even for equations which lead to residual MPNNs when discretized*[12]

[11] Similar to Nt and Maehara (2019); Oono and Suzuki (2020)

[12] This is different from Nt and Maehara (2019); Oono and Suzuki (2020); Cai and Wang (2020)

GNNs as Gradient Flows part 2: multi-particle energy approach

A more general energy

We can rewrite $\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}(\mathbf{F}) = \frac{1}{2}\sum_i \langle \mathbf{f}_i, \mathbf{W}^\top \mathbf{W}\mathbf{f}_i\rangle - \frac{1}{2}\sum_{i,j} \bar{a}_{ij}\langle \mathbf{f}_i, \mathbf{W}^\top \mathbf{W}\mathbf{f}_j\rangle$

Replace $\mathbf{W}^\top\mathbf{W}$ with symmetric matrices $\mathbf{\Omega}, \mathbf{W} \in \mathbb{R}^{d\times d} \rightarrow$

$$\mathcal{E}^{\mathrm{tot}}(\mathbf{F}) := \frac{1}{2}\sum_i \langle \mathbf{f}_i, \mathbf{\Omega}\mathbf{f}_i\rangle - \frac{1}{2}\sum_{i,j} \bar{a}_{ij}\langle \mathbf{f}_i, \mathbf{W}\mathbf{f}_j\rangle \equiv \mathcal{E}^{\mathrm{ext}}_{\mathbf{\Omega}}(\mathbf{F}) + \mathcal{E}^{\mathrm{pair}}_{\mathbf{W}}(\mathbf{F})$$

A more general energy

We can rewrite $\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}(\mathbf{F}) = \frac{1}{2}\sum_i \langle \mathbf{f}_i, \mathbf{W}^\top \mathbf{W}\mathbf{f}_i\rangle - \frac{1}{2}\sum_{i,j} \bar{a}_{ij}\langle \mathbf{f}_i, \mathbf{W}^\top \mathbf{W}\mathbf{f}_j\rangle$

Replace $\mathbf{W}^\top\mathbf{W}$ with symmetric matrices $\mathbf{\Omega}, \mathbf{W} \in \mathbb{R}^{d\times d} \rightarrow$

$$\mathcal{E}^{\mathrm{tot}}(\mathbf{F}) := \frac{1}{2}\sum_i \langle \mathbf{f}_i, \mathbf{\Omega}\mathbf{f}_i\rangle - \frac{1}{2}\sum_{i,j} \bar{a}_{ij}\langle \mathbf{f}_i, \mathbf{W}\mathbf{f}_j\rangle \equiv \mathcal{E}^{\mathrm{ext}}_{\mathbf{\Omega}}(\mathbf{F}) + \mathcal{E}^{\mathrm{pair}}_{\mathbf{W}}(\mathbf{F})$$

The gradient flow of $\mathcal{E}^{\mathrm{tot}}$ is

$$\dot{\mathbf{F}}(t) = -\nabla_{\mathbf{F}}\mathcal{E}^{\mathrm{tot}}(\mathbf{F}(t)) = -\mathbf{F}(t)\mathbf{\Omega} + \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}.$$

Attraction vs repulsion

Node-features $\rightarrow$ particles in $\mathbb{R}^d$ with energy $\mathcal{E}^{\text{tot}}$

- $\mathcal{E}^{\text{ext}}_{\mathbf{\Omega}}$ is *independent of the graph topology* $\sim$ **external** field
- $\mathcal{E}^{\text{pair}}_{\mathbf{W}}$ $\sim$ potential energy, with $\mathbf{W}$ defining **pairwise interactions** of adjacent nodes

Attraction vs repulsion

Node-features $\rightarrow$ particles in $\mathbb{R}^d$ with energy $\mathcal{E}^{\mathrm{tot}}$

- $\mathcal{E}^{\mathrm{ext}}_{\mathbf{\Omega}}$ is *independent of the graph topology* $\sim$ **external** field
- $\mathcal{E}^{\mathrm{pair}}_{\mathbf{W}}$ $\sim$ potential energy, with $\mathbf{W}$ defining **pairwise interactions** of adjacent nodes

Decompose $\mathbf{W} = \mathbf{\Theta}_+^\top \mathbf{\Theta}_+ - \mathbf{\Theta}_-^\top \mathbf{\Theta}_-$ into positive and negative eigenvalues

Attraction vs repulsion

$$\mathbf{W} = \mathbf{\Theta}_+^\top \mathbf{\Theta}_+ - \mathbf{\Theta}_-^\top \mathbf{\Theta}_-$$

$$\mathcal{E}^{\mathrm{tot}}(\mathbf{F}) = \frac{1}{2}\sum_i \langle \mathbf{f}_i, (\mathbf{\Omega} - \mathbf{W})\mathbf{f}_i \rangle + \frac{1}{4}\sum_{i,j} ||\mathbf{\Theta}_+(\nabla \mathbf{F})_{ij}||^2 - \frac{1}{4}\sum_{i,j} ||\mathbf{\Theta}_-(\nabla \mathbf{F})_{ij}||^2.$$

The gradient flow minimizes $\mathcal{E}^{\mathrm{tot}} \to \mathbf{W}$ encodes..

- *attraction* via its positive eigenvalues since $||\mathbf{\Theta}_+(\nabla \mathbf{F})_{ij}||^2$ decreases edge-wise
- *repulsion* via its negative eigenvalues since $||\mathbf{\Theta}_-(\nabla \mathbf{F})_{ij}||^2$ increases edge-wise

Spectrum of $\mathbf{W}$ induces LFD or HFD

Consider $\dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W} \iff \mathrm{vec}(\dot{\mathbf{F}}(t)) = (\mathbf{W} \otimes \bar{\mathbf{A}})\mathrm{vec}(\mathbf{F}(t))$

Write the spectrum of $\mathbf{W}$ as $\{\lambda_r^{\mathbf{W}}\}$ with $\lambda_+^{\mathbf{W}} = (\max \lambda_r^{\mathbf{W}})_+$ and $\lambda_-^{\mathbf{W}} = (\min \lambda_r^{\mathbf{W}})_-$

Spectrum of W induces LFD or HFD

Consider $\dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W} \iff \mathrm{vec}(\dot{\mathbf{F}}(t)) = (\mathbf{W} \otimes \bar{\mathbf{A}})\mathrm{vec}(\mathbf{F}(t))$

Write the spectrum of $\mathbf{W}$ as $\{\lambda_r^{\mathbf{W}}\}$ with $\lambda_+^{\mathbf{W}} = (\max \lambda_r^{\mathbf{W}})_+$ and $\lambda_-^{\mathbf{W}} = (\min \lambda_r^{\mathbf{W}})_-$

Proposition (Informal)

If $|\lambda_-^{\mathbf{W}}|(\rho_{\mathbf{\Delta}} - 1) > \lambda_+^{\mathbf{W}}$*, i.e. enough mass is distributed over the negative eigenvalues of the '****channel-mixing****', then graph high frequencies dominate* $\rightarrow$ *what matters is how the spectra of* $\mathbf{\Delta}$ *and* $\mathbf{W}$ *interact*

Spectrum of $\mathbf{W}$ induces LFD or HFD

Consider $\dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W} \iff \mathrm{vec}(\dot{\mathbf{F}}(t)) = (\mathbf{W} \otimes \bar{\mathbf{A}})\mathrm{vec}(\mathbf{F}(t))$

Write the spectrum of $\mathbf{W}$ as $\{\lambda_r^{\mathbf{W}}\}$ with $\lambda_+^{\mathbf{W}} = (\max \lambda_r^{\mathbf{W}})_+$ and $\lambda_-^{\mathbf{W}} = (\min \lambda_r^{\mathbf{W}})_-$

Proposition (Informal)

*If $|\lambda_-^{\mathbf{W}}|(\rho_{\mathbf{\Delta}} - 1) > \lambda_+^{\mathbf{W}}$, i.e. enough mass is distributed over the negative eigenvalues of the '**channel-mixing**', then graph high frequencies dominate $\rightarrow$ what matters is how the spectra of $\mathbf{\Delta}$ and $\mathbf{W}$ interact*

Upshot: *The distribution of positive ($\lambda_+^{\mathbf{W}}$) and negative ($\lambda_-^{\mathbf{W}}$) eigenvalues of $\mathbf{W}$ determine if the dynamics is low/high frequency dominated (L/HFD)*

A comparison with (some) continuous GNN models

Recall the continuous models:

- Linear $\mathrm{PDE}-\mathrm{GCN_D}$: $\dot{\mathbf{F}}_{\mathrm{PDE-GCN_D}}(t) = -\boldsymbol{\Delta}\mathbf{F}(t)\mathbf{K}(t)^\top\mathbf{K}(t)$
- CGNN: $\dot{\mathbf{F}}_{\mathrm{CGNN}}(t) = -\boldsymbol{\Delta}\mathbf{F}(t) + \mathbf{F}(t)\tilde{\boldsymbol{\Omega}} + \mathbf{F}(0)$ with symmetric $\boldsymbol{\Omega}$
- Linear GRAND: $\dot{\mathbf{F}}_{\mathrm{GRAND}}(t) = -\boldsymbol{\Delta}_{\mathrm{RW}}\mathbf{F}(t) = -(\mathbf{I} - \boldsymbol{\mathcal{A}}(\mathbf{F}(0)))\mathbf{F}(t)$

A comparison with (some) continuous GNN models

Recall the continuous models:

- Linear PDE $-$ GCN$_\mathrm{D}$: $\dot{\mathbf{F}}_{\mathrm{PDE-GCN_D}}(t) = -\mathbf{\Delta}\mathbf{F}(t)\mathbf{K}(t)^\top\mathbf{K}(t)$
- CGNN: $\dot{\mathbf{F}}_{\mathrm{CGNN}}(t) = -\mathbf{\Delta}\mathbf{F}(t) + \mathbf{F}(t)\tilde{\mathbf{\Omega}} + \mathbf{F}(0)$ with symmetric $\mathbf{\Omega}$
- Linear GRAND: $\dot{\mathbf{F}}_{\mathrm{GRAND}}(t) = -\mathbf{\Delta}_{\mathrm{RW}}\mathbf{F}(t) = -(\mathbf{I} - \boldsymbol{\mathcal{A}}(\mathbf{F}(0)))\mathbf{F}(t)$

Proposition (Informal)

*The continuous models above are **never*** HFD.

Can graph convolutional models be high-frequency dominated?

Introduce step-size $\tau \leq 1$ and consider gradient flow system

$$\boxed{\mathbf{F}(t+\tau) = \mathbf{F}(t) + \tau \bar{\mathbf{A}} \mathbf{F}(t) \mathbf{W}, \quad \mathbf{W} = \mathbf{W}^{\top},}$$

Let $P_{\mathbf{W}}^{\rho_-}$ be the projection into the eigenspace of $\mathbf{W} \otimes \bar{\mathbf{A}} = \mathbf{W} \otimes (\mathbf{I} - \boldsymbol{\Delta})$ associated with the eigenvalue $\rho_- := |\lambda_-^{\mathbf{W}}|(\rho_{\boldsymbol{\Delta}} - 1)$ and set

$$\boxed{\lambda_+^{\mathbf{W}}(\rho_{\boldsymbol{\Delta}} - 1))^{-1} < |\lambda_-^{\mathbf{W}}| < 2(\tau(2 - \rho_{\boldsymbol{\Delta}}))^{-1}} \tag{1}$$

Can graph convolutional models be high-frequency dominated?

Let m be the *number of layers*

Theorem

If equation 1 holds then there exists $\delta_{\mathrm{HFD}} < \rho_-$ *s.t.*

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(m\tau)) = (1+\tau\rho_-)^{2m}\left(\frac{\rho_{\mathbf{\Delta}}}{2}||P_{\mathbf{W}}^{\rho_-}\mathbf{F}(0)||^2 + \mathcal{O}\left(\left(\frac{1+\tau\delta_{\mathrm{HFD}}}{1+\tau\rho_-}\right)^{2m}\right)\right).$$

The dynamics is HFD *for a.e.* $\mathbf{F}(0)$ *and* $\mathbf{F}(m\tau)/||\mathbf{F}(m\tau)|| \to \mathbf{F}_\infty$ *s.t.* $\mathbf{\Delta}\mathbf{f}_\infty^r = \rho_{\mathbf{\Delta}}\mathbf{f}_\infty^r$.

Can graph convolutional models be high-frequency dominated?

Let m be the *number of layers*

Theorem

If equation 1 holds then there exists $\delta_{\mathrm{HFD}} < \rho_-$ *s.t.*

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(m\tau)) = (1+\tau\rho_-)^{2m}\left(\frac{\rho_{\mathbf{\Delta}}}{2}||P_{\mathbf{W}}^{\rho_-}\mathbf{F}(0)||^2 + \mathcal{O}\left(\left(\frac{1+\tau\delta_{\mathrm{HFD}}}{1+\tau\rho_-}\right)^{2m}\right)\right).$$

The dynamics is HFD *for a.e.* $\mathbf{F}(0)$ *and* $\mathbf{F}(m\tau)/||\mathbf{F}(m\tau)|| \to \mathbf{F}_\infty$ *s.t.* $\mathbf{\Delta}\mathbf{f}^r_\infty = \rho_{\mathbf{\Delta}}\mathbf{f}^r_\infty$.

Conversely, if G *is not bipartite, then for a.e.* $\mathbf{F}(0)$ *the system* $\mathbf{F}(t+\tau) = \tau\bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}$, *with* $\mathbf{W}$ *symmetric, is* LFD *independent of the spectrum of* $\mathbf{W}$.

The role of the residual connection in terms of the spectrum of W

→ linear discrete gradient flows can be HFD due to the negative eigenvalues of $\mathbf{W}$

- Differently from previous results[13], no bound on spectral radius of $\mathbf{W}$ coming from the graph topology as long as $\lambda_+^{\mathbf{W}}$ is small enough
- Without a residual term the dynamics is LFD for a.e. $\mathbf{F}(0)$ *independently* of the sign and magnitude of the eigenvalues of $\mathbf{W}$

[13] Nt and Maehara (2019); Oono and Suzuki (2020); Cai and Wang (2020)

GNNs as Gradient Flows part 4: ablation studies and experiments

Structure of the framework

General ingredients of the framework GRAFF (Gradient Flow Framework)

- *Encoding* block $\psi_{\mathrm{EN}} : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times d}$ is used to process input features $\mathbf{F}_0 \in \mathbb{R}^{n\times p}$

Structure of the framework

General ingredients of the framework GRAFF (Gradient Flow Framework)

- *Encoding* block $\psi_{\mathrm{EN}} : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times d}$ is used to process input features $\mathbf{F}_0 \in \mathbb{R}^{n\times p}$
- *Symmetric* channel-mixing matrices $\mathbf{\Omega}, \mathbf{W} \in \mathbb{R}^{d\times d}$ that are *shared across the layers*

Structure of the framework

General ingredients of the framework GRAFF (Gradient Flow Framework)

- *Encoding* block $\psi_{\mathrm{EN}} : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times d}$ is used to process input features $\mathbf{F}_0 \in \mathbb{R}^{n\times p}$
- *Symmetric* channel-mixing matrices $\mathbf{\Omega}, \mathbf{W} \in \mathbb{R}^{d\times d}$ that are *shared across the layers*
- *Decoding block* $\psi_{\mathrm{DE}} : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times k}$, where k is the number of label classes

Structure of the framework

General ingredients of the framework GRAFF (Gradient Flow Framework)

- *Encoding* block $\psi_{\mathrm{EN}} : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times d}$ is used to process input features $\mathbf{F}_0 \in \mathbb{R}^{n\times p}$
- *Symmetric* channel-mixing matrices $\boldsymbol{\Omega}, \mathbf{W} \in \mathbb{R}^{d\times d}$ that are *shared across the layers*
- *Decoding block* $\psi_{\mathrm{DE}} : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times k}$, where k is the number of label classes

$$\boxed{\mathbf{F}(t+\tau) = \mathbf{F}(t) + \tau\left(-\mathbf{F}(t)\boldsymbol{\Omega} + \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W} + \beta\mathbf{F}(0)\right), \quad \mathbf{F}(0) = \psi_{\mathrm{EN}}(\mathbf{F}_0),}$$

Different choices for $\mathbf{W}$

- *Sum*-variant: $\mathbf{W} = \mathbf{W}' + \mathbf{W}'^{\top} \rightarrow$ 'no-control' on spectrum

[14] Provides justification to Chen et al. (2020)

Different choices for W

- *Sum*-variant: $\mathbf{W} = \mathbf{W}' + \mathbf{W}'^{\top} \rightarrow$ 'no-control' on spectrum
- *(Neg)-Prod*: $\mathbf{W} = \pm\mathbf{W}'^{\top}\mathbf{W}' \rightarrow$ signed eigenvalues

[14] Provides justification to Chen et al. (2020)

Different choices for W

- *Sum*-variant: $\mathbf{W} = \mathbf{W}' + \mathbf{W}'^{\top} \rightarrow$ 'no-control' on spectrum
- *(Neg)-Prod*: $\mathbf{W} = \pm\mathbf{W}'^{\top}\mathbf{W}' \rightarrow$ signed eigenvalues
- $\mathbf{W}$ *diagonally-dominant* (DD): take $\mathbf{W}^0$ *symmetric* with zero diagonal and $\mathbf{w} \in \mathbb{R}^d$ defined by $\mathbf{w}_\alpha = q_\alpha \sum_\beta |\mathbf{W}^0_{\alpha\beta}| + r_\alpha$, and set $\mathbf{W} = \mathrm{diag}(\mathbf{w}) + \mathbf{W}^0 \rightarrow$ by Gershgorin Theorem the model 'can' easily re-distribute mass in the spectrum via q_α, r_α[14].

[14] Provides justification to Chen et al. (2020)

Complexity and number of parameters

GRAFF scales as $\mathcal{O}(|\mathsf{V}|pd + |\mathsf{E}|d)$, where p and d are input feature and hidden dimension

$\rightarrow$ *our model is faster than GCN* with small number of parameters: $pd + d^2 + 3d + dk$

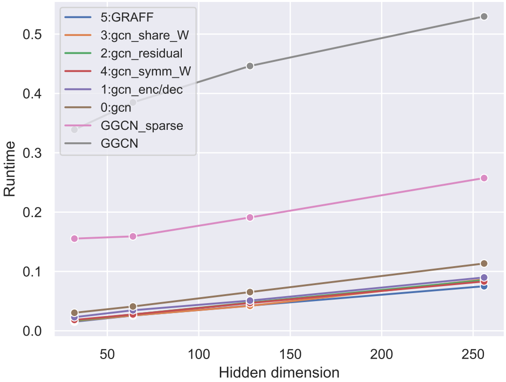


Figure 4: Runtime ablation for inference on Cora dataset

Ablation and synthetic experiments: setting

Recall our claims about role of ‘channel-mixing’ $\mathbf{W}$:

- *Positive eigenvalues of* $\mathbf{W}$ *induce* ***attraction*** *in a residual convolutional model*

Ablation and synthetic experiments: setting

Recall our claims about role of ‘channel-mixing’ $\mathbf{W}$:

- *Positive eigenvalues of* $\mathbf{W}$ *induce* ***attraction*** *in a residual convolutional model*
- *Negative eigenvalues of* $\mathbf{W}$ *induce* ***repulsion*** *in a residual convolutional model*

Ablation and synthetic experiments: setting

Recall our claims about role of ‘channel-mixing’ $\mathbf{W}$:

- *Positive eigenvalues of* $\mathbf{W}$ *induce* ***attraction*** *in a residual convolutional model*
- *Negative eigenvalues of* $\mathbf{W}$ *induce* ***repulsion*** *in a residual convolutional model*
- *A* ***non****-residual convolutional model is always dominated by low-frequencies independent of the spectrum of the* $\mathbf{W}$

Ablation and synthetic experiments: setting

Recall our claims about role of ‘channel-mixing’ $\mathbf{W}$:

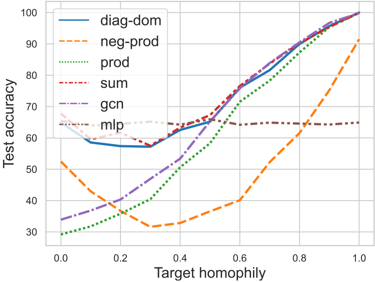
- *Positive eigenvalues of* $\mathbf{W}$ *induce **attraction** in a residual convolutional model*
- *Negative eigenvalues of* $\mathbf{W}$ *induce **repulsion** in a residual convolutional model*
- *A **non**-residual convolutional model is always dominated by low-frequencies independent of the spectrum of the* $\mathbf{W}$

To investigate our claims we use the synthetic Cora dataset of Zhu et al. (2020)

$\rightarrow$ graphs are generated for target levels of homophily via preferential attachment: we expect LFD to be better than HFD with high homophily and vice-versa for low homophily

Ablation and synthetic experiments: part 1

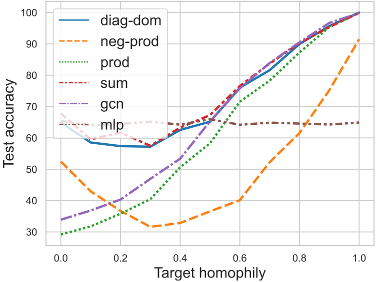
Goal: Explain performance wrt homophily in terms of the spectrum of $\mathbf{W}$



- *Neg-prod* is better than *prod* on low-homophily $\rightarrow$ *confirms* HFD *dynamics*

Ablation and synthetic experiments: part 1

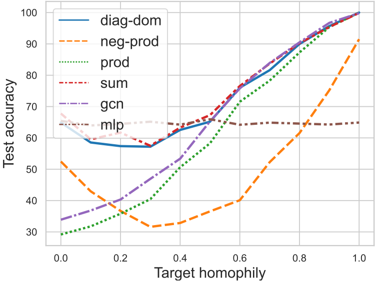
Goal: Explain performance wrt homophily in terms of the spectrum of $\mathbf{W}$



- *Neg-prod* is better than *prod* on low-homophily $\rightarrow$ *confirms* HFD *dynamics*
- *prod* (attraction-only) struggles in low-homophily *even with residual connection*

Ablation and synthetic experiments: part 1

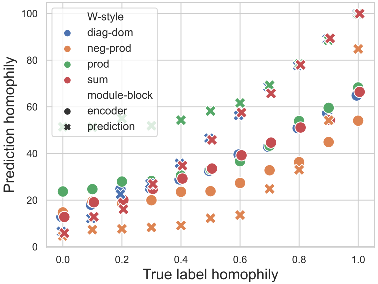
Goal: Explain performance wrt homophily in terms of the spectrum of $\mathbf{W}$



- *Neg-prod* is better than *prod* on low-homophily $\rightarrow$ *confirms* HFD *dynamics*
- *prod* (attraction-only) struggles in low-homophily *even with residual connection*
- 'neutral' variants like *sum* and (DD) are more flexible and outperform GCN confirming that *non*- residual convolutional models are LFD irrespectively of the spectrum of $\mathbf{W}$

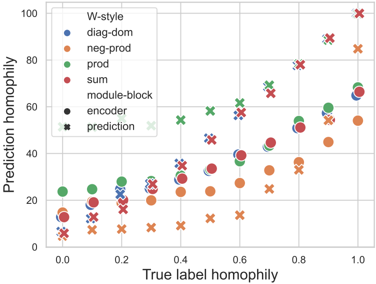
Ablation and synthetic experiments: part 2

Goal: Use homophily to assess if the evolution is *smoothing* $\rightarrow$ compute homophily of the prediction (cross) and compare with that read from the encoding (i.e. *no evolution*)



Ablation and synthetic experiments: part 2

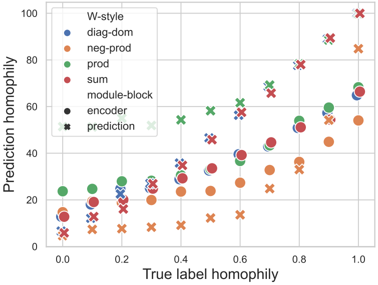
Goal: Use homophily to assess if the evolution is *smoothing* $\rightarrow$ compute homophily of the prediction (cross) and compare with that read from the encoding (i.e. *no evolution*)



- *neg-prod*: homophily decreases after evolution while with *prod* the prediction is smoother than the true homophily

Ablation and synthetic experiments: part 2

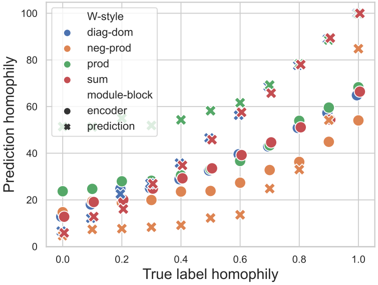
Goal: Use homophily to assess if the evolution is *smoothing* $\rightarrow$ compute homophily of the prediction (cross) and compare with that read from the encoding (i.e. *no evolution*)



- *neg-prod*: homophily decreases after evolution while with *prod* the prediction is smoother than the true homophily
- (DD) and *sum* variants adapt better to the true homophily

Ablation and synthetic experiments: part 2

Goal: Use homophily to assess if the evolution is *smoothing* $\rightarrow$ compute homophily of the prediction (cross) and compare with that read from the encoding (i.e. *no evolution*)



- *neg-prod*: homophily decreases after evolution while with *prod* the prediction is smoother than the true homophily
- (DD) and *sum* variants adapt better to the true homophily
- The encoding compensates when the spectrum of $\mathbf{W}$ has a sign

Conclusions and where to next?

What was the message then?

- Framework where the MPNNs equations minimize a multi-particle learnable energy
- Analysis of the interaction between the spectrum of the graph and the spectrum of the 'channel-mixing' $\rightarrow$ when and why the dynamics is low (high) frequency dominated
- Refined existing asymptotic analysis of MPNNs to account for the role of the spectrum of the channel-mixing
- From a practical perspective, our framework allows for 'educated' choices resulting in a simple, more explainable convolutional model: our results refute the folklore of graph convolutional models being too 'simple' for complex benchmarks.

Limitations and future directions

We restricted to a *constant* bilinear form $\mathbf{W}$, how about non-constant alternatives $\mathbf{W}(\mathbf{F}, t)$ that are *aware* of the features? $\rightarrow$ requirement for local 'heterogeneity' with efficiency

Limitations and future directions

We restricted to a *constant* bilinear form $\mathbf{W}$, how about non-constant alternatives $\mathbf{W}(\mathbf{F}, t)$ that are *aware* of the features? $\rightarrow$ requirement for local 'heterogeneity' with efficiency

What can we say about dynamics that are neither LFD nor HFD?

Limitations and future directions

We restricted to a *constant* bilinear form $\mathbf{W}$, how about non-constant alternatives $\mathbf{W}(\mathbf{F}, t)$ that are *aware* of the features? $\rightarrow$ requirement for local 'heterogeneity' with efficiency

What can we say about dynamics that are neither LFD nor HFD?

The energy formulation points to new models more 'physics' inspired

Thank you!

For more details check out our paper


arXiv:2206.10991v1 [cs.LG] 22 Jun 2022

Graph Neural Networks as Gradient Flows

Francesco Di Giovanni † Twitter Inc. fdigiovanni@twitter.com
James Rowbottom† Twitter Inc. jrowbottom@twitter.com
Benjamin P. Chamberlain Twitter Inc.
Thomas Markovich Twitter Inc.
Michael M. Bronstein Twitter Inc. and University of Oxford

Abstract

Dynamical systems minimizing an energy are ubiquitous in geometry and physics. We propose a gradient flow framework for GNNs where the equations follow the direction of steepest descent of a learnable energy. This approach allows to explain the GNN evolution from a multi-particle perspective as learning attractive and repulsive forces in feature space via the positive and negative eigenvalues of a symmetric 'channel-mixing' matrix. We perform spectral analysis of the solutions and conclude that gradient flow graph convolutional models can induce a dynamics dominated by the graph high frequencies which is desirable for heterophilic datasets. We also describe structural constraints on common GNN architectures allowing to interpret them as gradient flows. We perform thorough ablation studies corroborating our theoretical analysis and show competitive performance of simple and lightweight models on real-world homophilic and heterophilic datasets.

1 Introduction and motivations

Graph neural networks (GNNs) Sperduti [1993], Goller and Kuchler [1996], Gori et al. [2005], Scarselli et al. [2008], Bruna et al. [2014], Defferrard et al. [2016], Kipf and Welling [2017], Battaglia et al. [2016, 2018] and in particular their Message Passing formulation (MPNN) Gilmer et al. [2017], Brandstetter et al. [2022] have become the standard ML tool for dealing with different types of relations and interactions, ranging from social networks to particle physics and drug design. One of the often cited drawbacks of traditional GNN models is their poor 'explainability', making it hard to know why and how they make certain predictions Ying et al. [2019], Yuan et al. [2020], and in which situations they may work and when they would fail. Limitations of GNNs that have attracted attention are over-smoothing Nt and Maehara [2019], Oono and Suzuki [2020], Cai and Wang [2020], over-squashing and bottlenecks Alon and Yahav [2021], Topping et al. [2022], and performance on heterophilic data Pei et al. [2020], Zhu et al. [2020], Chien et al. [2021], Bo et al. [2021], Yan et al. [2021] – where adjacent nodes usally have different labels.

Contributions. We propose a *Gradient Flow Framework* (GRAFF) where the GNN equations follow the direction of steepest descent of a *learnable energy*. Thanks to this framework we can (i) interpret GNNs as a multi-particle dynamics where the learned parameters determine pairwise attractive and repulsive potentials in the feature space. This sheds light on how GNNs can adapt to heterophily and explains their performance and the smoothness of the prediction. (ii) GRAFF leads to residual convolutional models where the *channel-mixing* $\mathbf{W}$ is performed by a shared symmetric bilinear form inducing attraction and repulsion via its positive and negative eigenvalues, respectively. We theoretically investigate the interaction of the graph spectrum with the spectrum of the channel-mixing proving that if there is more mass on the negative eigenvalues of $\mathbf{W}$, then the dynamics

†Equal contribution

@Francesco_dgv, @JRowbottom

References

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. (2017). Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42.

Cai, C. and Wang, Y. (2020). A note on over-smoothing for graph neural networks. *arXiv preprint arXiv:2006.13318*.

Chamberlain, B., Rowbottom, J., Gorinova, M. I., Bronstein, M., Webb, S., and Rossi, E. (2021). Grand: Graph neural diffusion. In *International Conference on Machine Learning*, pages 1407–1418. PMLR.

Chen, M., Wei, Z., Huang, Z., Ding, B., and Li, Y. (2020). Simple and deep graph convolutional networks. In *International Conference on Machine Learning*, pages 1725–1735. PMLR.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018). Neural ordinary differential equations. *Advances in neural information processing systems*, 31.

Eliasof, M., Haber, E., and Treister, E. (2021). Pde-gcn: Novel architectures for graph neural networks motivated by partial differential equations. *Advances in Neural Information Processing Systems*, 34.

Haber, E. and Ruthotto, L. (2018). Stable architectures for deep neural networks. *Inverse Problems*, 34.

Kipf, T. N. and Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17.

Nt, H. and Maehara, T. (2019). Revisiting graph neural networks: All we have is low-pass filters. *arXiv preprint arXiv:1905.09550*.

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

Rusch, T. K., Chamberlain, B. P., Rowbottom, J., Mishra, S., and Bronstein, M. M. (2022). Graph-coupled oscillator networks. In *International Conference on Machine Learning*.

Xhonneux, L.-P., Qu, M., and Tang, J. (2020). Continuous graph neural networks. In *International Conference on Machine Learning*, pages 10432–10441. PMLR.

Zhou, D. and Schölkopf, B. (2005). Regularization on discrete spaces. In *Joint Pattern Recognition Symposium*, pages 361–368. Springer.

Zhu, J., Yan, Y., Zhao, L., Heimann, M., Akoglu, L., and Koutra, D. (2020). Beyond homophily in graph neural networks: Current limitations and effective designs. *Advances in Neural Information Processing Systems*, 33:7793–7804.