Graph neural networks as dynamical systems

Francesco Di Giovanni

Twitter

First Italian School in GDL: July 25–28, Pescara

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

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

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

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

The Laplacian $\boldsymbol{\Delta} \succeq 0 \rightarrow$ 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

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

[1] Zhou and Schölkopf (2005)

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

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

$\to$ 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 $(i,j) \in \mathsf{E}$

[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 $\mathbf{f}_0 \rightarrow \mathbf{f}(t) = \sum_\ell c_\ell(t)\boldsymbol{\phi}_\ell^{\boldsymbol{\Delta}}$

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

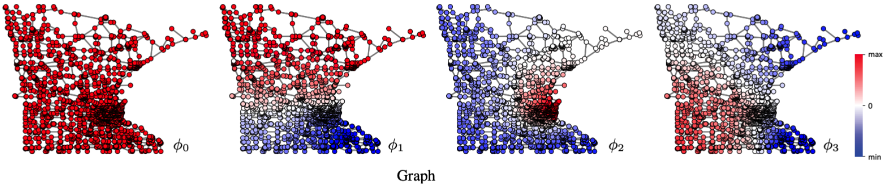


Figure 1: First four Laplacian eigenvectors of Minnesota Road graph. Figure taken from Bronstein et al. (2017)

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. $\boldsymbol{\Delta}\mathbf{f}_\infty = \mathbf{0}$ i.e. $\mathbf{f}_\infty \in \mathrm{span}(\sqrt{d_1}, \ldots, \sqrt{d_n})^\top$

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

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

The gradient flow of $\mathcal{E}^{\mathrm{Dir}}$ yields heat equation in each feature channel[2]:

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

[2] 'Channels' = 'feature components' = 'feature coordinates'

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* $\mathbf{I}_d \otimes \boldsymbol{\Delta} \in \mathbb{R}^{nd} \times \mathbb{R}^{nd}$ to rewrite $\mathcal{E}^{\mathrm{Dir}}$ as

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

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

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

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

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

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

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

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

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

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*[3] since

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

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

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

[4] 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*[3] since

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

The dynamics are ‘dominated’ by the low-frequencies: does $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) \to 0$? **No:**[4]

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

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

[4] 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

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

A motivating example

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

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

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

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

A motivating example

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

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

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

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

Upshot: Analyse $\mathbf{F}(t)$ via $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)/||\mathbf{F}(t)||) \rightarrow$ Rayleigh quotient of $\mathbf{I}_d \otimes \boldsymbol{\Delta}$

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 $t \to \infty$.

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?

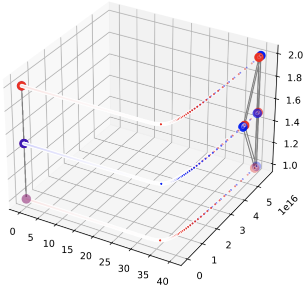
Lemma

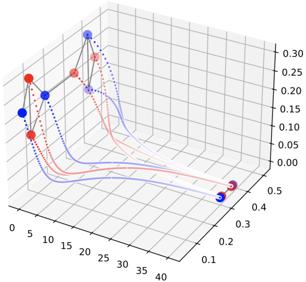
A dynamical system is LFD *iff for each sequence* $t_j \to \infty$ *there exist a subsequence* $t_{j_k} \to \infty$ *and* $\mathbf{F}_\infty$ *s.t.* $\mathbf{F}(t_{j_k})/||\mathbf{F}(t_{j_k})|| \to \mathbf{F}_\infty$ *and* $\boldsymbol{\Delta}\mathbf{f}^r_\infty = \mathbf{0}$.

LFD dynamics: numerical example

A numerical example of LFD dynamics: $T = 4.0$, $\tau = 0.1$

$$\dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{\Lambda}, \quad \mathbf{\Lambda} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

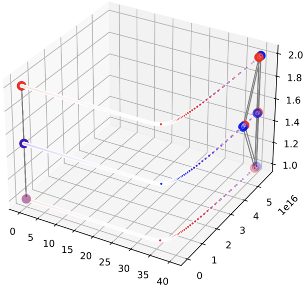




LFD dynamics: numerical example

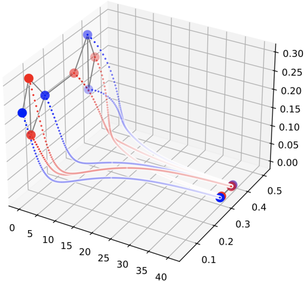
A numerical example of LFD dynamics: $T = 4.0$, $\tau = 0.1$

$$\dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t)\boldsymbol{\Lambda}, \quad \boldsymbol{\Lambda} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$



In both cases the eigenvector $\phi_0^{\boldsymbol{\Delta}}$ dominates the dynamics

- Top: solution becomes unbounded
- Bottom: evolution of $\mathbf{F}(t)/||\mathbf{F}(t)||$
  $\rightarrow$ **convergence to** $\ker(\boldsymbol{\Delta})$ **where**
  we only distinguish nodes based on their degrees



High-frequency-dominant: HFD

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

High-frequency-dominant: HFD

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

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 $t \rightarrow \infty$.

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?

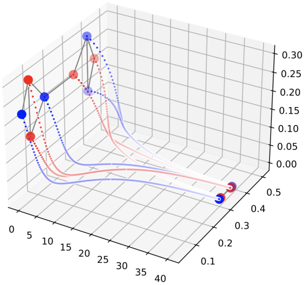
Lemma

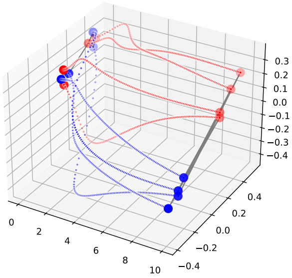
A dynamical system is HFD *iff for each sequence $t_j \rightarrow \infty$ there exist a subsequence $t_{j_k} \rightarrow \infty$ and $\mathbf{F}_\infty$ s.t. $\mathbf{F}(t_{j_k})/||\mathbf{F}(t_{j_k})|| \rightarrow \mathbf{F}_\infty$ and $\mathbf{\Delta}\mathbf{f}^r_\infty = \rho_{\mathbf{\Delta}}\mathbf{f}^r_\infty$.*

Why do we need HFD?

Consider $\dot{\mathbf{F}}(t) = -\bar{\mathbf{A}}\mathbf{F}(t) \rightarrow$ eigenvector $\phi^{\mathbf{\Delta}}_{\rho_{\mathbf{\Delta}}}$ dominates the dynamics

- Evolution of $\mathbf{F}(t)/||\mathbf{F}(t)|| \rightarrow$ **convergence to** $\ker(\rho_{\mathbf{\Delta}}\mathbf{I} - \mathbf{\Delta})$ **where we distinguish nodes based on the largest frequency eigenvector** (right figure)





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)**[5] is defined by:

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

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

[5] Kipf and Welling (2017)

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

Graph Convolutional Networks

A layer of **Graph Convolutional Network (GCN)**[5] is defined by:

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

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

- Poor performance on heterophilic graphs
- Degradation when increasing depth (over-smoothing)[6]

[5] Kipf and Welling (2017)
[6] Nt and Maehara (2019); Oono and Suzuki (2020); Cai and Wang (2020)

Low-pass filters and over-smoothing: review

Theorem (Cai and Wang)

Let $(1-\bar{\lambda})^2 := \max_{\lambda_\ell^{\boldsymbol{\Delta}}}(1-\lambda_\ell^{\boldsymbol{\Delta}})^2$ *and* $s_T = \max_{t \leq T} \operatorname{sing}(\mathbf{W}(t))$. *Then the solution* $\mathbf{F}(T)$ *of GCN satisfies*

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(T)) \leq (s_T(1-\bar{\lambda}))^{2T}\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(0)).$$

Low-pass filters and over-smoothing: review

Theorem (Cai and Wang)

Let $(1-\bar{\lambda})^2 := \max_{\lambda_\ell^{\boldsymbol{\Delta}}}(1-\lambda_\ell^{\boldsymbol{\Delta}})^2$ *and* $s_T = \max_{t \leq T} \mathrm{sing}(\mathbf{W}(t))$. *Then the solution* $\mathbf{F}(T)$ *of GCN satisfies*

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(T)) \leq (s_T(1-\bar{\lambda}))^{2T}\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(0)).$$

- If singular values of $\mathbf{W}(t)$ are controlled in terms of the spectrum of $\boldsymbol{\Delta}$ $\rightarrow$ solution of GCN becomes increasingly smoother

Low-pass filters and over-smoothing: review

Theorem (Cai and Wang)

Let $(1-\bar{\lambda})^2 := \max_{\lambda_\ell^{\boldsymbol{\Delta}}}(1-\lambda_\ell^{\boldsymbol{\Delta}})^2$ *and* $s_T = \max_{t\leq T}\text{sing}(\mathbf{W}(t))$. *Then the solution* $\mathbf{F}(T)$ *of GCN satisfies*

$$\mathcal{E}^{\text{Dir}}(\mathbf{F}(T)) \leq (s_T(1-\bar{\lambda}))^{2T}\mathcal{E}^{\text{Dir}}(\mathbf{F}(0)).$$

- If singular values of $\mathbf{W}(t)$ are controlled in terms of the spectrum of $\boldsymbol{\Delta} \rightarrow$ solution of GCN becomes increasingly smoother
- GCN should succeed with homophily but fail with heterophily

Low-pass filters and over-smoothing: review

Theorem (Cai and Wang)

Let $(1-\bar{\lambda})^2 := \max_{\lambda_\ell^{\boldsymbol{\Delta}}}(1-\lambda_\ell^{\boldsymbol{\Delta}})^2$ *and* $s_T = \max_{t\leq T}\operatorname{sing}(\mathbf{W}(t))$. *Then the solution* $\mathbf{F}(T)$ *of GCN satisfies*

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(T)) \leq (s_T(1-\bar{\lambda}))^{2T}\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(0)).$$

- If singular values of $\mathbf{W}(t)$ are controlled in terms of the spectrum of $\boldsymbol{\Delta} \rightarrow$ solution of GCN becomes increasingly smoother
- GCN should succeed with homophily but fail with heterophily
- If $T >> 1$, we converge to $\ker(\boldsymbol{\Delta})$ i.e. only information to separate nodes is *degree*

Are graph convolutional models doomed?

- What if the singular values of $\mathbf{W}$ are not bounded by $(1-\bar{\lambda})^2$?

Are graph convolutional models doomed?

- What if the singular values of $\mathbf{W}$ are not bounded by $(1 - \bar{\lambda})^2$?
- Can we require more structure on $\mathbf{W}$?

Are graph convolutional models doomed?

- What if the singular values of $\mathbf{W}$ are not bounded by $(1 - \bar{\lambda})^2$?
- Can we require more structure on $\mathbf{W}$?
- What is the interpretation of $\mathbf{W}$?

Are graph convolutional models doomed?

- What if the singular values of $\mathbf{W}$ are not bounded by $(1 - \bar{\lambda})^2$?
- Can we require more structure on $\mathbf{W}$?
- What is the interpretation of $\mathbf{W}$?
- What is the 'minimal requirement' for a graph convolutional framework to be HFD?

A physics-inspired approach

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.

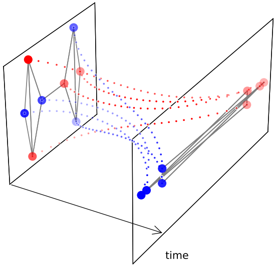


Figure 2: Actual GRAFF dynamics: attractive and repulsive forces lead to a non-smoothing process able to separate labels

Joint w/ J. Rowbottom*, B. Chamberlain, T. Markovich, M. Bronstein (2022)

Outline of the contributions

- We propose a gradient flow framework (GRAFF) 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 (GRAFF) 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 (GRAFF) 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 (GRAFF) 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 $\mathbf{F}(t+\tau) = \mathbf{F}(t) + \tau \text{ResNet}(\mathbf{F}(t))$ is the Euler discretization of an ODE[7] (as the step-size $\tau \to 0$)

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

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

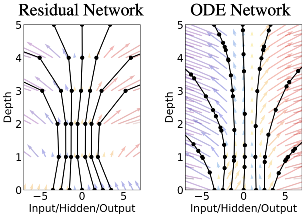


Figure 3: Dynamics of ResNet vs ODE. Figure taken from Chen et al. (2018)

[7] 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[7] (as the step-size $\tau \to 0$)

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

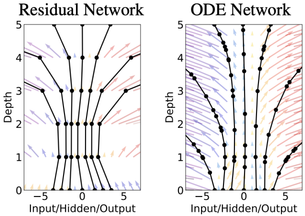


Figure 3: Dynamics of ResNet vs ODE. Figure taken from Chen et al. (2018)

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

What about residual MPNNs?

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

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

Instances of ‘continuous’ MPNNs

The linear GCN[8] system

$$\mathbf{F}(t+1) = \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}(t) \rightarrow \dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}(t) - \mathbf{F}(t)$$

[8] Wu et al. (2019)

Instances of ‘continuous’ MPNNs

The linear GCN[8] system

$$\mathbf{F}(t+1) = \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}(t) \rightarrow \dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}(t) - \mathbf{F}(t)$$

If we use the $\otimes$-formalism: GCN is the unit step-size discretization of

$$\mathrm{vec}(\dot{\mathbf{F}}(t)) = (\mathbf{W}(t)^\top \otimes \bar{\mathbf{A}} - \mathbf{I})\mathrm{vec}(\mathbf{F}(t))$$

$\rightarrow$ we’ll see that the *dampening* term $\mathbf{I}$ is responsible for LFD dynamics

[8] Wu et al. (2019)

Instances of 'continuous' MPNNs

Continuous Graph Neural Network (CGNN)[9]: set $\mathbf{W} = \mathbf{W}^\top \rightarrow$

$$\dot{\mathbf{F}}(t) = -\boldsymbol{\Delta}\mathbf{F}(t) + \mathbf{F}(t)\mathbf{W} + \mathbf{F}(0)$$

[9] Xhonneux et al. (2020)

Instances of 'continuous' MPNNs

Continuous Graph Neural Network (CGNN)[9]: set $\mathbf{W} = \mathbf{W}^\top \rightarrow$

$$\dot{\mathbf{F}}(t) = -\mathbf{\Delta}\mathbf{F}(t) + \mathbf{F}(t)\mathbf{W} + \mathbf{F}(0)$$

- CGNN is a *gradient flow*
- We'll prove that this is **never** HFD
- Source term $\mathbf{F}(0)$ increases expressive power

[9] Xhonneux et al. (2020)

Instances of ‘continuous’ MPNNs

Graph Neural Diffusion (GRAND)[10] is the ‘continuous’ version of GAT[11]

$$\dot{\mathbf{F}}(t) = -(\mathbf{I} - \boldsymbol{\mathcal{A}}(\mathbf{F}(t)))\mathbf{F}(t)$$

- $\boldsymbol{\mathcal{A}}(\mathbf{F}(t))$ is an attention matrix over the edge set
- (Linear) GRAND is a diffusion process with maximum principle $\rightarrow$ *low-pass filter and over-smoothing*

[10] Chamberlain et al. (2021)
[11] Veličković et al. (2018)

Instances of ‘continuous’ MPNNs

PDE-GCN$_D$[12] is a diffusion process given by

$$\dot{\mathbf{F}}(t) = -\boldsymbol{\Delta}\mathbf{F}(t)\mathbf{W}(t)^\top\mathbf{W}(t)$$

$\rightarrow$ We’ll prove that this is a smoothing process and hence **not** suitable for heterophilic graphs

[12] Eliasof et al. (2021)

Second-order variants[13] $\rightarrow$ by design they *prevent over-smoothing*

$$\ddot{\mathbf{F}}(t) = \mathrm{MPNN}(\mathsf{G}, \mathbf{F}(t)) - \gamma\mathbf{F}(t) - \alpha\dot{\mathbf{F}}(t)$$

However, why oscillatory behaviour? Do we need them?

[13] Eliasof et al. (2021); Rusch et al. (2022)

Instances of ‘continuous’ MPNNs

Second-order variants[13] $\rightarrow$ by design they *prevent over-smoothing*

$$\ddot{\mathbf{F}}(t) = \mathrm{MPNN}(\mathsf{G}, \mathbf{F}(t)) - \gamma \mathbf{F}(t) - \alpha \dot{\mathbf{F}}(t)$$

However, why oscillatory behaviour? Do we need them?

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

[13] Eliasof et al. (2021); Rusch et al. (2022)

Instances of 'continuous' MPNNs

Second-order variants[13] $\rightarrow$ by design they *prevent over-smoothing*

$$\ddot{\mathbf{F}}(t) = \mathrm{MPNN}(\mathsf{G}, \mathbf{F}(t)) - \gamma \mathbf{F}(t) - \alpha \dot{\mathbf{F}}(t)$$

However, why oscillatory behaviour? Do we need them?

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

Upshot: Learn an energy rather than the equations!

[13] Eliasof et al. (2021); Rusch et al. (2022)

Dynamical systems as gradient flows

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

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

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?

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** $\mathcal{E}^{\mathrm{Dir}} \to$ *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

Harmonic map flow in continuous space

$f : \mathbb{R}^n \to (\mathbb{R}^d, h)$ smooth with h a constant metric $\to$ The *Dirichlet energy* of f is

$$\mathcal{E}(f,h) = \frac{1}{2}\int_{\mathbb{R}^n} \|\nabla f\|_h^2 \, dx = \frac{1}{2}\sum_{q,r=1}^{d}\sum_{j=1}^{n}\int_{\mathbb{R}^n} h_{qr}\partial_j f^q \partial_j f^r(x) dx$$

$\to$ measures the **smoothness** of f wrt h

[14] Kimmel et al. (1997); Perona and Malik (1990)

Harmonic map flow in continuous space

$f : \mathbb{R}^n \to (\mathbb{R}^d, h)$ smooth with h a constant metric → The *Dirichlet energy* of f is

$$\mathcal{E}(f, h) = \frac{1}{2}\int_{\mathbb{R}^n} \|\nabla f\|_h^2 \, dx = \frac{1}{2}\sum_{q,r=1}^{d}\sum_{j=1}^{n}\int_{\mathbb{R}^n} h_{qr}\partial_j f^q \partial_j f^r(x) dx$$

→ measures the **smoothness** of f wrt h

Eells and Sampson (1964) studied the **gradient flow** of $\mathcal{E}$ given by $\dot{f}(t) = -\nabla_f \mathcal{E}(f(t))$ to find minimizers of $\mathcal{E}$ → extended to manifolds harmonic map flow

For *PDE-based image processing* gradient flows of $\mathcal{E}$ recover the Perona-Malik diffusion[14]

[14] Kimmel et al. (1997); Perona and Malik (1990)

Extending the formalism to graphs

$f : \mathbb{R}^n \to (\mathbb{R}^d, h)$ smooth with h a constant metric $\to$ The *Dirichlet energy* of f is

$$\mathcal{E}(f, h) = \frac{1}{2}\int_{\mathbb{R}^n} \|\nabla f\|_h^2 \, dx = \frac{1}{2}\sum_{q,r=1}^{d}\sum_{j=1}^{n}\int_{\mathbb{R}^n} h_{qr}\partial_j f^q \partial_j f^r(x)dx$$

Extending the formalism to graphs

$f : \mathbb{R}^n \to (\mathbb{R}^d, h)$ smooth with h a constant metric $\to$ The *Dirichlet energy* of f is

$$\mathcal{E}(f,h) = \frac{1}{2}\int_{\mathbb{R}^n} \|\nabla f\|_h^2 \, dx = \frac{1}{2}\sum_{q,r=1}^{d}\sum_{j=1}^{n}\int_{\mathbb{R}^n} h_{qr}\partial_j f^q \partial_j f^r(x)dx$$

$\to$ Replace $\int_{\mathbb{R}^n}$ with $\sum_{i\in\mathsf{V}}$ and $\partial_j|_i$ with $\nabla_{(i,j)\in\mathsf{E}}$:

$$\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}(\mathbf{F}) := \frac{1}{4}\sum_{q,r=1}^{d}\sum_{i\in\mathsf{V}}\sum_{j:(i,j)\in\mathsf{E}} h_{qr}(\nabla \mathbf{f}^q)_{ij}(\nabla \mathbf{f}^r)_{ij} = \frac{1}{4}\sum_{(i,j)\in\mathsf{E}} ||\mathbf{W}(\nabla \mathbf{F})_{ij}||^2.$$

with $\mathbf{H} = \mathbf{W}^\top\mathbf{W}$ with $\mathbf{W} \in \mathbb{R}^{d\times d}$

Extending the formalism to graphs

$f : \mathbb{R}^n \to (\mathbb{R}^d, h)$ smooth with h a constant metric $\to$ The *Dirichlet energy* of f is

$$\mathcal{E}(f, h) = \frac{1}{2}\int_{\mathbb{R}^n} \|\nabla f\|_h^2 \, dx = \frac{1}{2}\sum_{q,r=1}^{d}\sum_{j=1}^{n}\int_{\mathbb{R}^n} h_{qr}\partial_j f^q \partial_j f^r(x)dx$$

$\to$ Replace $\int_{\mathbb{R}^n}$ with $\sum_{i\in\mathsf{V}}$ and $\partial_j|_i$ with $\nabla_{(i,j)\in\mathsf{E}}$:

$$\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}(\mathbf{F}) := \frac{1}{4}\sum_{q,r=1}^{d}\sum_{i\in\mathsf{V}}\sum_{j:(i,j)\in\mathsf{E}} h_{qr}(\nabla \mathbf{f}^q)_{ij}(\nabla \mathbf{f}^r)_{ij} = \frac{1}{4}\sum_{(i,j)\in\mathsf{E}} ||\mathbf{W}(\nabla\mathbf{F})_{ij}||^2.$$

with $\mathbf{H} = \mathbf{W}^\top\mathbf{W}$ with $\mathbf{W} \in \mathbb{R}^{d\times d}$

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 $\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}$:

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

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 (Di G.*, Rowbottom*, et al.)

The dynamics is smoothing. Let $P^{\mathrm{ker}}_{\mathbf{W}}$ *be the projection onto* $\mathrm{ker}(\mathbf{W}^{\top}\mathbf{W})$*, then*

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) \leq e^{-2t\mathrm{gap}(\mathbf{W}^{\top}\mathbf{W})\mathrm{gap}(\boldsymbol{\Delta})}||\mathbf{F}(0)||^2 + \mathcal{E}^{\mathrm{Dir}}((P^{\mathrm{ker}}_{\mathbf{W}} \otimes \mathbf{I}_n)\mathrm{vec}(\mathbf{F}(0))), \quad t \geq 0.$$

$\exists\, \phi_\infty \in \mathbb{R}^d$: *for each* $i \in \mathsf{V}$ *we have* $\mathbf{f}_i(t) \to \sqrt{d_i}\phi_\infty + P^{\mathrm{ker}}_{\mathbf{W}}\mathbf{f}_i(0)$.

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 (Di G.*, Rowbottom*, et al.)

The dynamics is smoothing. Let $P^{\mathrm{ker}}_{\mathbf{W}}$ *be the projection onto* $\ker(\mathbf{W}^\top\mathbf{W})$*, then*

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) \leq e^{-2t\mathrm{gap}(\mathbf{W}^\top\mathbf{W})\mathrm{gap}(\boldsymbol{\Delta})}||\mathbf{F}(0)||^2 + \mathcal{E}^{\mathrm{Dir}}((P^{\mathrm{ker}}_{\mathbf{W}} \otimes \mathbf{I}_n)\mathrm{vec}(\mathbf{F}(0))), \quad t \geq 0.$$

$\exists\, \phi_\infty \in \mathbb{R}^d$: *for each* $i \in \mathsf{V}$ *we have* $\mathbf{f}_i(t) \to \sqrt{d_i}\phi_\infty + P^{\mathrm{ker}}_{\mathbf{W}}\mathbf{f}_i(0)$.

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 (Di G.*, Rowbottom*, et al.)

The dynamics is smoothing. Let $P^{\mathrm{ker}}_{\mathbf{W}}$ *be the projection onto* $\ker(\mathbf{W}^{\top}\mathbf{W})$*, then*

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) \leq e^{-2t\mathrm{gap}(\mathbf{W}^{\top}\mathbf{W})\mathrm{gap}(\boldsymbol{\Delta})}||\mathbf{F}(0)||^2 + \mathcal{E}^{\mathrm{Dir}}((P^{\mathrm{ker}}_{\mathbf{W}} \otimes \mathbf{I}_n)\mathrm{vec}(\mathbf{F}(0))), \quad t \geq 0.$$

$\exists\, \phi_\infty \in \mathbb{R}^d$: *for each* $i \in \mathsf{V}$ *we have* $\mathbf{f}_i(t) \to \sqrt{d_i}\phi_\infty + P^{\mathrm{ker}}_{\mathbf{W}}\mathbf{f}_i(0)$.

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 (Di G.*, Rowbottom*, et al.)

The dynamics is smoothing. Let $P^{\mathrm{ker}}_{\mathbf{W}}$ *be the projection onto* $\ker(\mathbf{W}^{\top}\mathbf{W})$, *then*

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) \leq e^{-2t\mathrm{gap}(\mathbf{W}^{\top}\mathbf{W})\mathrm{gap}(\boldsymbol{\Delta})}||\mathbf{F}(0)||^2 + \mathcal{E}^{\mathrm{Dir}}((P^{\mathrm{ker}}_{\mathbf{W}} \otimes \mathbf{I}_n)\mathrm{vec}(\mathbf{F}(0))), \quad t \geq 0.$$

$\exists\, \boldsymbol{\phi}_{\infty} \in \mathbb{R}^d$: *for each* $i \in \mathsf{V}$ *we have* $\mathbf{f}_i(t) \to \sqrt{d_i}\boldsymbol{\phi}_{\infty} + P^{\mathrm{ker}}_{\mathbf{W}}\mathbf{f}_i(0)$.

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 (Di G.*, Rowbottom*, et al.)

The dynamics is smoothing. Let $P^{\mathrm{ker}}_{\mathbf{W}}$ *be the projection onto* $\ker(\mathbf{W}^{\top}\mathbf{W})$*, then*

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) \leq e^{-2t\mathrm{gap}(\mathbf{W}^{\top}\mathbf{W})\mathrm{gap}(\boldsymbol{\Delta})}||\mathbf{F}(0)||^2 + \mathcal{E}^{\mathrm{Dir}}((P^{\mathrm{ker}}_{\mathbf{W}} \otimes \mathbf{I}_n)\mathrm{vec}(\mathbf{F}(0))), \quad t \geq 0.$$

$\exists\, \phi_\infty \in \mathbb{R}^d$: *for each* $i \in \mathsf{V}$ *we have* $\mathbf{f}_i(t) \to \sqrt{d_i}\phi_\infty + P^{\mathrm{ker}}_{\mathbf{W}}\mathbf{f}_i(0)$.

A few comments on the graph harmonic flow

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

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

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

A few comments on the graph harmonic flow

- 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[15]

[15] Similar to Nt and Maehara (2019); Oono and Suzuki (2020)
[16] This is different from Nt and Maehara (2019); Oono and Suzuki (2020); Cai and Wang (2020)

A few comments on the graph harmonic flow

- 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[15]
- 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[16]

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

[16] 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 $\mathbf{\Omega}, \mathbf{W} \in \mathbb{R}^{d\times d} \to$

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

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 $\mathcal{E}^{\mathrm{tot}}$

- $\mathcal{E}^{\mathrm{ext}}_{\boldsymbol{\Omega}}$ 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

Attraction vs repulsion

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

- $\mathcal{E}^{\text{ext}}_{\boldsymbol{\Omega}}$ 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

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

Attraction vs repulsion

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

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

Attraction vs repulsion

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

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

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

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

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

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

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 $\lambda_-^{\mathbf{W}} = (\min \lambda_r^{\mathbf{W}})_-$

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 $\lambda_-^{\mathbf{W}} = (\min \lambda_r^{\mathbf{W}})_-$

Any eigenvalue of $\mathbf{W} \otimes \bar{\mathbf{A}}$ can be written as $\lambda_r^{\mathbf{W}}\lambda_i^{\bar{\mathbf{A}}} = \lambda_r^{\mathbf{W}}(1 - \lambda_i^{\boldsymbol{\Delta}})$

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 $\lambda_-^{\mathbf{W}} = (\min \lambda_r^{\mathbf{W}})_-$

Any eigenvalue of $\mathbf{W} \otimes \bar{\mathbf{A}}$ can be written as $\lambda_r^{\mathbf{W}}\lambda_i^{\bar{\mathbf{A}}} = \lambda_r^{\mathbf{W}}(1 - \lambda_i^{\boldsymbol{\Delta}})$

Let $P_{\mathbf{W}}^{\rho_-}$ be the projection onto the eigenspace of $\mathbf{W} \otimes \bar{\mathbf{A}}$ associated with $\rho_- := |\lambda_-^{\mathbf{W}}|(\rho_{\boldsymbol{\Delta}} - 1) \to$ Recall that $\rho_{\boldsymbol{\Delta}}$ is the *largest* eigenvalue of $\boldsymbol{\Delta} = \mathbf{I} - \bar{\mathbf{A}}$

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

Proposition (Di G.*, Rowbottom*, et al.)

If $\rho_- > \lambda_+^{\mathbf{W}}$, *then* $\dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}$ *is* HFD *for a.e.* $\mathbf{F}(0)$*: there exists* ϵ_{HFD} *such that* [17]

$$\mathcal{E}^{\text{Dir}}(\mathbf{F}(t)) = e^{2t\rho_-}\left(\frac{\rho_{\mathbf{\Delta}}}{2}||P_{\mathbf{W}}^{\rho_-}\mathbf{F}(0)||^2 + \mathcal{O}(e^{-2t\epsilon_{\text{HFD}}})\right), \quad t \geq 0,$$

and $\mathbf{F}(t)/||\mathbf{F}(t)||$ *converges to* $\mathbf{F}_\infty \in \mathbb{R}^{n\times d}$ *such that* $\mathbf{\Delta}\mathbf{f}_\infty^r = \rho_{\mathbf{\Delta}}\mathbf{f}_\infty^r$*, for* $1 \leq r \leq d$*.*

[17] We have an explicit formula depending on 'spectral gaps' of $\mathbf{\Delta}$ and $\mathbf{W}$

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

Proposition (Di G.*, Rowbottom*, et al.)

If $\rho_{-} > \lambda_{+}^{\mathbf{W}}$, *then* $\dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}$ *is* HFD *for a.e.* $\mathbf{F}(0)$*: there exists* ϵ_{HFD} *such that* [18]

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) = e^{2t\rho_{-}}\left(\frac{\rho_{\boldsymbol{\Delta}}}{2}||P_{\mathbf{W}}^{\rho_{-}}\mathbf{F}(0)||^2 + \mathcal{O}(e^{-2t\epsilon_{\mathrm{HFD}}})\right), \quad t \geq 0,$$

and $\mathbf{F}(t)/||\mathbf{F}(t)||$ *converges to* $\mathbf{F}_{\infty} \in \mathbb{R}^{n\times d}$ *such that* $\boldsymbol{\Delta}\mathbf{f}_{\infty}^{r} = \rho_{\boldsymbol{\Delta}}\mathbf{f}_{\infty}^{r}$, *for* $1 \leq r \leq d$.

[18] We have an explicit formula depending on 'spectral gaps' of $\boldsymbol{\Delta}$ and $\mathbf{W}$

Spectrum of W induces LFD or HFD

Proposition (Di G.*, Rowbottom*, et al.)

If $\rho_- > \lambda_+^{\mathbf{W}}$, *then* $\dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}$ *is* HFD *for a.e.* $\mathbf{F}(0)$*: there exists* ϵ_{HFD} *such that* [19]

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) = e^{2t\rho_-}\left(\frac{\rho_{\mathbf{\Delta}}}{2}||P_{\mathbf{W}}^{\rho_-}\mathbf{F}(0)||^2 + \mathcal{O}(e^{-2t\epsilon_{\mathrm{HFD}}})\right), \quad t \geq 0,$$

and $\mathbf{F}(t)/||\mathbf{F}(t)||$ *converges to* $\mathbf{F}_\infty \in \mathbb{R}^{n\times d}$ *such that* $\mathbf{\Delta}\mathbf{f}_\infty^r = \rho_{\mathbf{\Delta}}\mathbf{f}_\infty^r$, *for* $1 \leq r \leq d$.

[19] We have an explicit formula depending on 'spectral gaps' of $\mathbf{\Delta}$ and $\mathbf{W}$

Spectrum of W induces LFD or HFD

Proposition (Di G.*, Rowbottom*, et al.)

If $\rho_- > \lambda_+^{\mathbf{W}}$, *then* $\dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}$ *is* HFD *for a.e.* $\mathbf{F}(0)$*: there exists* ϵ_{HFD} *such that* [20]

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) = e^{2t\rho_-}\left(\frac{\rho_{\boldsymbol{\Delta}}}{2}||P_{\mathbf{W}}^{\rho_-}\mathbf{F}(0)||^2 + \mathcal{O}(e^{-2t\epsilon_{\mathrm{HFD}}})\right), \quad t \geq 0,$$

and $\mathbf{F}(t)/||\mathbf{F}(t)||$ *converges to* $\mathbf{F}_\infty \in \mathbb{R}^{n\times d}$ *such that* $\boldsymbol{\Delta}\mathbf{f}_\infty^r = \rho_{\boldsymbol{\Delta}}\mathbf{f}_\infty^r$, *for* $1 \leq r \leq d$.

[20] We have an explicit formula depending on ‘spectral gaps’ of $\boldsymbol{\Delta}$ and $\mathbf{W}$

Spectrum of W induces LFD or HFD

Proposition (Di G.*, Rowbottom*, et al.)

If $\rho_- > \lambda_+^{\mathbf{W}}$, *then* $\dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}$ *is* HFD *for a.e.* $\mathbf{F}(0)$*: there exists* ϵ_{HFD} *such that* [21]

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) = e^{2t\rho_-}\left(\frac{\rho_{\mathbf{\Delta}}}{2}||P_{\mathbf{W}}^{\rho_-}\mathbf{F}(0)||^2 + \mathcal{O}(e^{-2t\epsilon_{\mathrm{HFD}}})\right), \quad t \geq 0,$$

and $\mathbf{F}(t)/||\mathbf{F}(t)||$ *converges to* $\mathbf{F}_\infty \in \mathbb{R}^{n\times d}$ *such that* $\mathbf{\Delta}\mathbf{f}_\infty^r = \rho_{\mathbf{\Delta}}\mathbf{f}_\infty^r$, *for* $1 \leq r \leq d$.

[21] We have an explicit formula depending on ‘spectral gaps’ of $\mathbf{\Delta}$ and $\mathbf{W}$

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

Proposition (Di G.*, Rowbottom*, et al.)

If $\rho_- > \lambda_+^{\mathbf{W}}$, *then* $\dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}$ *is* HFD *for a.e.* $\mathbf{F}(0)$*: there exists* ϵ_{HFD} *such that* [22]

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) = e^{2t\rho_-}\left(\frac{\rho_{\boldsymbol{\Delta}}}{2}||P_{\mathbf{W}}^{\rho_-}\mathbf{F}(0)||^2 + \mathcal{O}(e^{-2t\epsilon_{\mathrm{HFD}}})\right), \quad t \geq 0,$$

and $\mathbf{F}(t)/||\mathbf{F}(t)||$ *converges to* $\mathbf{F}_\infty \in \mathbb{R}^{n\times d}$ *such that* $\boldsymbol{\Delta}\mathbf{f}_\infty^r = \rho_{\boldsymbol{\Delta}}\mathbf{f}_\infty^r$, *for* $1 \leq r \leq d$.

[22] We have an explicit formula depending on 'spectral gaps' of $\boldsymbol{\Delta}$ and $\mathbf{W}$

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

Proposition (Di G.*, Rowbottom*, et al.)

If $\rho_- > \lambda_+^{\mathbf{W}}$, *then* $\dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}$ *is* HFD *for a.e.* $\mathbf{F}(0)$*: there exists* ϵ_{HFD} *such that* [22]

$$\mathcal{E}^{\text{Dir}}(\mathbf{F}(t)) = e^{2t\rho_-}\left(\frac{\rho_{\boldsymbol{\Delta}}}{2}||P_{\mathbf{W}}^{\rho_-}\mathbf{F}(0)||^2 + \mathcal{O}(e^{-2t\epsilon_{\text{HFD}}})\right), \quad t \geq 0,$$

and $\mathbf{F}(t)/||\mathbf{F}(t)||$ *converges to* $\mathbf{F}_\infty \in \mathbb{R}^{n\times d}$ *such that* $\boldsymbol{\Delta}\mathbf{f}_\infty^r = \rho_{\boldsymbol{\Delta}}\mathbf{f}_\infty^r$, *for* $1 \leq r \leq d$.

If enough mass is distributed over the negative eigenvalues of the ‘channel-mixing’, graph high frequencies dominate $\rightarrow$ **what matters is how the spectra of $\boldsymbol{\Delta}$ and $\mathbf{W}$ interact**

[22] We have an explicit formula depending on ‘spectral gaps’ of $\boldsymbol{\Delta}$ and $\mathbf{W}$

Source term and a more general family of energies

Equations with a source term may have better expressive power[23]

In our framework: add an extra energy term $\mathcal{E}^{\text{source}}_{\tilde{\mathbf{W}}}(\mathbf{F}) := \beta\langle \mathbf{F}, \mathbf{F}(0)\tilde{\mathbf{W}}\rangle \rightarrow$

$$\dot{\mathbf{F}}(t) = -\mathbf{F}(t)\boldsymbol{\Omega} + \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W} - \beta\mathbf{F}(0)\tilde{\mathbf{W}}.$$

[23] Xhonneux et al. (2020); Chen et al. (2020); Thorpe et al. (2021)

Source term and a more general family of energies

Equations with a source term may have better expressive power[23]

In our framework: add an extra energy term $\mathcal{E}^{\text{source}}_{\tilde{\mathbf{W}}}(\mathbf{F}) := \beta\langle \mathbf{F}, \mathbf{F}(0)\tilde{\mathbf{W}}\rangle \to$

$$\dot{\mathbf{F}}(t) = -\mathbf{F}(t)\boldsymbol{\Omega} + \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W} - \beta\mathbf{F}(0)\tilde{\mathbf{W}}.$$

We can also replace $\bar{\mathbf{A}}$ with $\boldsymbol{\mathcal{A}}$ satisfying $\boldsymbol{\mathcal{A}}_{ij} = 0$ if $(i,j) \notin \mathsf{E} \to$

$$\mathcal{E}^{\text{pair}}_{\boldsymbol{\mathcal{A}},\mathbf{W}}(\mathbf{F}) := -\sum_{(i,j)} \boldsymbol{\mathcal{A}}_{ij}\langle \mathbf{f}_i, \mathbf{W}\mathbf{f}_j\rangle.$$

[23] Xhonneux et al. (2020); Chen et al. (2020); Thorpe et al. (2021)

Comments on non-linear activations

Non-linear function σ can 'activate' the inner products in the energy:

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

[24] Wu et al. (2019); Oono and Suzuki (2020); Chen et al. (2020)

Comments on non-linear activations

Non-linear function σ can 'activate' the inner products in the energy:

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

A few reasons why we keep the gradient flow *linear*

- Perform spectral analysis in closed form[24]
- We have seen no gain in performance when including non-linear activations
- We can 'push the non-linear maps' in either the encoding block or the decoding one

[24] Wu et al. (2019); Oono and Suzuki (2020); Chen et al. (2020)

A comparison with (some) continuous GNN models

Recall the continuous models:

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

A comparison with (some) continuous GNN models

Recall the continuous models:

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

Proposition (Di G.*, Rowbottom*, et al.)

(i) $\mathrm{PDE}-\mathrm{GCN}_D$ *is a smoothing model:* $\dot{\mathcal{E}}^{\mathrm{Dir}}(\mathbf{F}_{\mathrm{PDE-GCN}_D}(t)) \leq 0$.

A comparison with (some) continuous GNN models

Recall the continuous models:

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

Proposition (Di G.*, Rowbottom*, et al.)

(i) $\mathrm{PDE}-\mathrm{GCN}_D$ *is a smoothing model:* $\dot{\mathcal{E}}^{\mathrm{Dir}}(\mathbf{F}_{\mathrm{PDE-GCN}_D}(t)) \leq 0$.

(ii) *For a.e.* $\mathbf{F}(0)$ *it holds:* CGNN *is never* HFD *and if we remove the source term, then* $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}_{\mathrm{CGNN}}(t)/||\mathbf{F}_{\mathrm{CGNN}}(t)||) \leq e^{-\mathrm{gap}(\boldsymbol{\Delta})t}$.

A comparison with (some) continuous GNN models

Recall the continuous models:

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

Proposition (Di G.*, Rowbottom*, et al.)

(i) $\mathrm{PDE}-\mathrm{GCN}_D$ *is a smoothing model:* $\dot{\mathcal{E}}^{\mathrm{Dir}}(\mathbf{F}_{\mathrm{PDE-GCN}_D}(t)) \leq 0$.

(ii) *For a.e.* $\mathbf{F}(0)$ *it holds:* CGNN *is never* HFD *and if we remove the source term, then* $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}_{\mathrm{CGNN}}(t)/||\mathbf{F}_{\mathrm{CGNN}}(t)||) \leq e^{-\mathrm{gap}(\boldsymbol{\Delta})t}$.

(iii) *If* G *is connected,* $\mathbf{F}_{\mathrm{GRAND}}(t) \to \boldsymbol{\mu}$ *as* $t \to \infty$*, with* $\boldsymbol{\mu}^r = \mathrm{mean}(\mathbf{f}^r(0))$, $1 \leq r \leq d$.

GNNs as Gradient Flows part 3: discrete setting

The requirement for symmetry

When classical MPNNs are discretized gradient flows?

The requirement for symmetry

When classical MPNNs are discretized gradient flows?

Let $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n\times n}$ be a symmetric graph vector field $\rightarrow (\boldsymbol{\mathcal{A}})_{ij} = 0,\ (i,j) \notin \mathsf{E}$

Consider a family of linear GNNs with shared weights of the form

$$\mathbf{F}(t+1) = \mathbf{F}(t)\boldsymbol{\Omega} + \boldsymbol{\mathcal{A}}\mathbf{F}(t)\mathbf{W} + \beta\mathbf{F}(0)\tilde{\mathbf{W}}, \quad 0 \leq t \leq T.$$

They are gradient flow of a 'multi-particle' energy iff $\boldsymbol{\Omega}$ and $\mathbf{W}$ are symmetric.

Can graph convolutional models be high-frequency dominated?

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

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

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

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

Can graph convolutional models be high-frequency dominated?

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

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

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

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

Can graph convolutional models be high-frequency dominated?

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

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

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

$$\lambda_+^{\mathbf{W}}(\rho_{\boldsymbol{\Delta}} - 1))^{-1} < |\lambda_-^{\mathbf{W}}| < 2(\tau(2 - \rho_{\boldsymbol{\Delta}}))^{-1} \tag{3}$$

Can graph convolutional models be high-frequency dominated?

Theorem (Di G.*, Rowbottom*, et al.)

If equation 3 holds then there exists $\delta_{\mathrm{HFD}} < \rho_{-}$ *s.t.*

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

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.* $\boldsymbol{\Delta}\mathbf{f}^{r}_{\infty} = \rho_{\boldsymbol{\Delta}}\mathbf{f}^{r}_{\infty}$.

Can graph convolutional models be high-frequency dominated?

Theorem (Di G.*, Rowbottom*, et al.)

If equation 3 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.* $\mathbf{\Delta}\mathbf{f}_\infty^r = \rho_{\mathbf{\Delta}}\mathbf{f}_\infty^r$.

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* $\mathbf{W}$.

HFD **dynamics via negative eigenvalues**

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

- Differently from previous results[25], no bound on spectral radius of $\mathbf{W}$ coming from the graph topology as long as $\lambda_{+}^{\mathbf{W}}$ is small enough

  → Recall that previous over-smoothing results required $\mathbf{W}$ to have *sufficiently small singular values* depending on the spectrum of $\mathbf{\Delta}$

  → If we have symmetry and control the spectrum of $\mathbf{W}$ we can avoid over-smoothing (and in fact be HFD) in terms of positive vs negative eigenvalues of $\mathbf{W}$

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

The role of the residual connection

- Without a residual term the dynamics is LFD for a.e. $\mathbf{F}(0)$ *independently* of the sign and magnitude of the eigenvalues of $\mathbf{W}$

  $\rightarrow$ provides a justification for the residual connection in terms of the spectrum of $\mathbf{W}$

  $\rightarrow$ explains via induced dynamics and spectral analysis the 'expressivity' results in Chen et al. (2020)

Reversing time and sign of the edge weights

Let $\{\lambda_r^{\mathbf{W}}\}$ be the spectrum of $\mathbf{W}$ with orthonormal eigenvectors $\{\phi_r^{\mathbf{W}}\}$ and $\boldsymbol{\Delta} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$

[26] Similar effect as in Bo et al. (2021); Yan et al. (2021)

Reversing time and sign of the edge weights

Let $\{\lambda_r^{\mathbf{W}}\}$ be the spectrum of $\mathbf{W}$ with orthonormal eigenvectors $\{\phi_r^{\mathbf{W}}\}$ and $\boldsymbol{\Delta} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$

Introduce $\mathbf{z}^r(t) : \mathsf{V} \to \mathbb{R}$ defined by $z_i^r(t) = \langle \mathbf{f}_i(t), \phi_r^{\mathbf{W}} \rangle$, then gradient flow becomes:

$$\mathbf{z}^r(t+\tau) = \mathbf{U}(\mathbf{I} + \tau\lambda_r^{\mathbf{W}}(\mathbf{I} - \boldsymbol{\Lambda}))\mathbf{U}^\top \mathbf{z}^r(t) = \mathbf{z}^r(t) + \tau\lambda_r^{\mathbf{W}}\bar{\mathbf{A}}\mathbf{z}^r(t)$$

Along $\phi_r^{\mathbf{W}}$ if $\lambda_r^{\mathbf{W}} < 0$ then the dynamics is equivalent to flipping the sign of the edges [26]

[26] Similar effect as in Bo et al. (2021); Yan et al. (2021)

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 $\psi_{\mathrm{EN}} : \mathbb{R}^{n\times p} \rightarrow \mathbb{R}^{n\times d}$ 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 $\psi_{\text{EN}} : \mathbb{R}^{n \times p} \to \mathbb{R}^{n \times d}$ 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* $\psi_{\text{DE}} : \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times k}$, 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 $\mathbf{\Omega}, \mathbf{W} \in \mathbb{R}^{d\times d}$ 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

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

Different choices for $\mathbf{W}$

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

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

Different choices for $\mathbf{W}$

- *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

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

Different choices for W

- *Sum*-variant: $\mathbf{W} = \mathbf{W}' + \mathbf{W}'^{\top} \to$ 'no-control' on spectrum
- *(Neg)-Prod*: $\mathbf{W} = \pm\mathbf{W}'^{\top}\mathbf{W}' \to$ 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 $\mathbf{W} = \mathrm{diag}(\mathbf{w}) + \mathbf{W}^0 \to$ by Gershgorin Theorem the model 'can' easily re-distribute mass in the spectrum via q_α, r_α[27].

[27] 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: $pd + d^2 + 3d + dk$

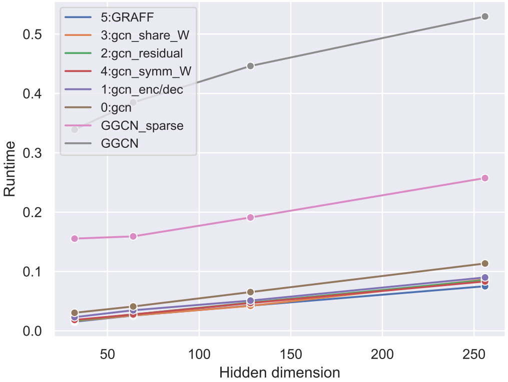


Figure 4: Runtime ablation for inference on Cora dataset

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’ $\mathbf{W}$:

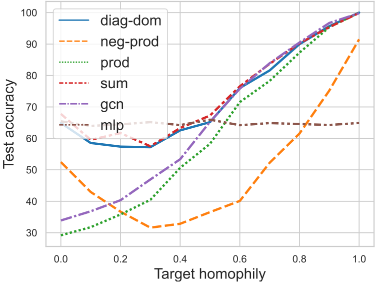
- *Positive eigenvalues of* $\mathbf{W}$ *induce* ***attraction*** *in a residual convolutional model*
- *Negative eigenvalues of* $\mathbf{W}$ *induce* ***repulsion*** *in a residual convolutional model*
- *A* ***non****-residual convolutional model is always dominated by low-frequencies independent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

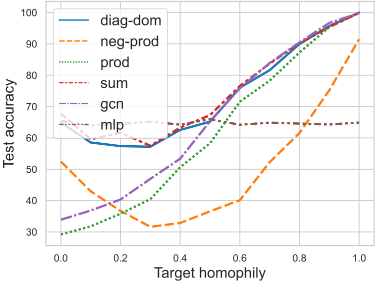
Goal: Explain performance wrt homophily in terms of the spectrum of $\mathbf{W}$



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

Ablation and synthetic experiments: part 1

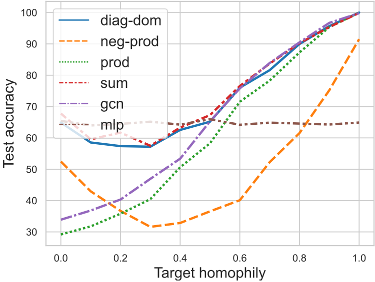
Goal: Explain performance wrt homophily in terms of the spectrum of $\mathbf{W}$



- *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

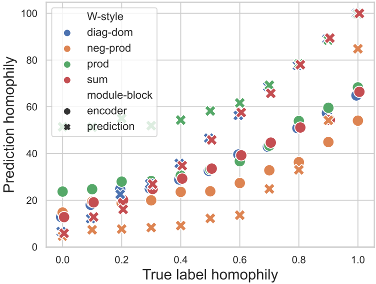
Goal: Explain performance wrt homophily in terms of the spectrum of $\mathbf{W}$



- *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 $\mathbf{W}$

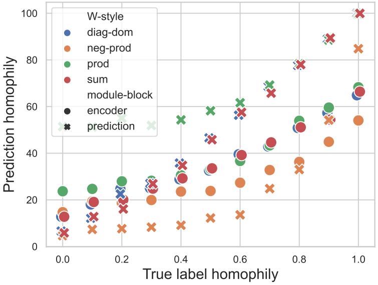
Ablation and synthetic experiments: part 2

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

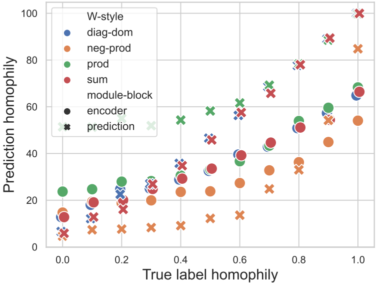
Goal: Use homophily to assess if the evolution is *smoothing* $\to$ compute homophily of the prediction (cross) and compare with that read from the encoding (i.e. *no evolution*)



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

Ablation and synthetic experiments: part 2

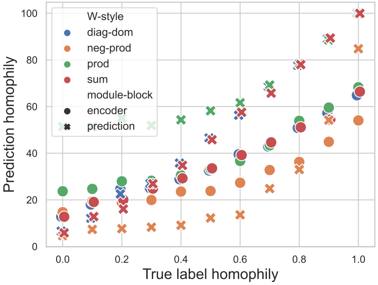
Goal: Use homophily to assess if the evolution is *smoothing* $\rightarrow$ compute homophily of the prediction (cross) and compare with that read from the encoding (i.e. *no evolution*)



- *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

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

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

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 any question/complaint/video-game recommendation do not hesitate to contact me! :-)

fdigiovanni (at) twitter com

@Francesco_dgv

References

Bo, D., Wang, X., Shi, C., and Shen, H. (2021). Beyond low-frequency information in graph convolutional networks. *In AAAI. AAAI Press*.

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.

Eells, J. and Sampson, J. H. (1964). Harmonic mappings of riemannian manifolds. *American journal of mathematics*, 86(1):109–160.

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.

Kimmel, R., Sochen, N., and Malladi, R. (1997). From high energy physics to low level vision. In *International Conference on Scale-Space Theories in Computer Vision*, pages 236–247. Springer.

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*.

Perona, P. and Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *PAMI*, 12(7):629–639.

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*.

Thorpe, M., Nguyen, T. M., Xia, H., Strohmer, T., Bertozzi, A., Osher, S., and Wang, B. (2021). Grand++: Graph neural diffusion with a source term. In *International Conference on Learning Representations*.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph attention networks. In *International Conference on Learning Representations*.

Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., and Weinberger, K. (2019). Simplifying graph convolutional networks. In *International conference on machine learning*, pages 6861–6871. PMLR.

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

Yan, Y., Hashemi, M., Swersky, K., Yang, Y., and Koutra, D. (2021). Two sides of the same coin: Heterophily and oversmoothing in graph convolutional neural networks. *arXiv preprint arXiv:2102.06462*.

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.