Generalized energy and gradient flow via graph framelets

Andi Han∗ Dai Shi† Zhiqi Shao∗ Junbin Gao∗

Abstract

In this work, we provide a theoretical understanding of the framelet-based graph neural networks through the perspective of energy gradient flow. By viewing the framelet-based models as discretized gradient flows of some energy, we show it can induce both low-frequency and high-frequency-dominated dynamics, via the separate weight matrices for different frequency components. This substantiates its good empirical performance on both homophilic and heterophilic graphs. We then propose a generalized energy via framelet decomposition and show its gradient flow leads to a novel graph neural network, which includes many existing models as special cases. We then explain how the proposed model generally leads to more flexible dynamics, thus potentially enhancing the representation power of graph neural networks.

1 Introduction

Graph neural networks (GNNs) [20, 22, 28, 42, 46] have become the primary tool for representation learning over graph-structured data, such as social networks [9], citation networks [28], molecules [18], traffic networks [14], among others. There generally exists two types of GNN models, i.e., spatial and spectral based models. Spatial GNNs, including MPNN [21], GAT [42], GIN [49], usually propagate information in the neighbourhood and update node representations via a weighted average of the neighbours. Spectral GNNs, including ChebyNet [15], GCN [28], BernNet [26], performs filtering on the spectral domain provided by graph Fourier transform (where the orthonormal system is given by the eigenvectors of graph Laplacian). Wavelet-based graph representation learning [17, 23, 31, 43, 48, 51, 52, 53, 58], a class of spectral methods, provide a multi-resolution analysis of graph signals and thus often lead to better signal representations by capturing information at different scales. In particular, graph framelets [17, 53], a type of tight wavelet frame, further allows separate modelling of low-pass and high-pass signal information, and has been considered to define graph convolution as in [8, 51, 53]. It has been shown that the graph framelet convolution allows more flexible control of approximate (via low-pass filters) and detailed information (via high-pass filters) with great robustness and efficiency, achieving the state-of-the-art results on multiple graph learning tasks [8, 50, 51, 55, 56, 58].

Along with the developments of more advanced models, theoretical understanding of both the power and pitfalls of graph neural networks has called for great attention. Currently, there exists many known limitations of GNNs, including over-smoothing [30], over-squashing [1, 41], limited expressive power [33, 49] and poor performance on heterophilic graphs [37, 57]. In an attempt

∗University of Sydney (andi.han@sydney.edu.au, zsha2911@uni.sydney.edu.au, junbin.gao@sydney.edu.au)
†Western Sydney University (dai.shi@sydney.edu.au)
to better resolve the above issues, many studies have tried to understand GNNs through various frameworks, such as dynamical systems and differential equations \cite{6 16 36 40 45}, Riemannian geometry \cite{34, 41}, algebraic topology \cite{4, 24}. Nevertheless, existing analyses are limited to either spatial GNNs or spectral GNNs based on the Fourier transform. In fact, it has been empirically observed that wavelet/framelet-based models tend to alleviate the previously known issues, such as mitigating the issue of over-smoothing \cite{43} and achieving good performance on heterophilic graphs \cite{8}. Despite the success of wavelet-based models, comparatively little is known on how the multi-scale and frequency-separation properties provided by graph wavelets/framelets potentially avoid the aforementioned pitfalls and enhance the learning capacity of GNNs.

In this paper, we particularly focus on the graph framelet-based models and analyze the model behaviors from the perspective of energy gradient flows. In physics, the evolution of particles is often modelled as differential equations that minimize an energy, known as the gradient flows. The energy functional and its gradient flow provide a characterization of particles’ states and movements, which are essential to understand the dynamics. Recently, such idea has been adapted to graph neural networks \cite{16}. By modelling GNNs as (discretized) gradient flows, one can study the limiting behaviors of GNNs, which are closely related to the notions of over-smoothing, over-separating and heterophilic graphs. Particularly, \cite{16} characterizes the dynamics of GNNs in terms of low-frequency or high-frequency dominance or both. It is known that low-frequency-dominant (LFD) models tend to perform well on homophilic graphs (where neighbouring nodes are likely to come from the same cluster) \cite{35, 46, 29, 16} while in contrast, high-frequency-dominant (HFD) models tend to perform well on heterophilic graphs \cite{3, 16}. In this regard, models that can be both LFD and HFD are usually preferred.

To the best of our knowledge, this is the first work that provides a theoretical understanding of multi-scale graph neural networks and justification for its good empirical performance. Specifically, we highlight our main contributions as follows.

1. We show the framelet convolutions, either spatial-based \cite{8} or spectral-based \cite{51} can be viewed as a discretized gradient flow of some energy. From this point of view, we prove that the framelet-based GNNs can induce both LFD and HFD dynamics, thus theoretically explaining its effectiveness on heterophilic graphs and its capability to potentially avoid over-smoothing, as often empirically observed.

2. We then define a generalized energy via framelet decomposition. We show such energy includes the energy proposed in \cite{16}, which itself is a generalization of the graph Dirichlet energy. We then propose a GNN model, namely gradient flow based framelet graph convolution (GradF-UFG), as discretization of the proposed generalized energy.

3. We show the proposed GradF-UFG includes the framelet-based GNNs as special cases. We also connect the proposed model with the recently introduced energy enhanced framelet convolution and analyze its behaviors. We explain how the proposed model provides more flexibility compared to existing works.

**Organization.** The rest of the paper is organized as follows. Section 2 first summarizes the related works on continuous GNNs and its connections to dynamical systems and differential equations. We also provide an overview on framelet-based graph representation learning. In section 3 we review the preliminary knowledge on graphs, graph framelets and framelet convolutions. We also introduce the notions of graph Dirichlet energy, gradient flow along with the definitions of LFD and HFD
dynamics. Section 4 starts with identifying a connection between the spatial framelet convolution with Dirichlet energy and also characterizes its asymptotic behaviors from the lens of gradient flow. In section 5, we define the framelet generalized energy and propose a GNN model based on its gradient flow, where we also connect to existing works. Finally, in section 6, we perform similar analysis for the spectral framelet convolution.

2 Related works

Dynamical systems, differential equations and GNNs. Neural ODE [12] is introduced as a continuous version of ResNet [25]. Since then, many work has studied its counterparts on graphs and proposed continuous GNNs, such as [2, 38, 47]. Another parallel research direction aims to understand GNNs from continuous dynamical systems and differential equations. GCN [28], one of the most popular GNN models, can be viewed as a discrete Markov process [36] and its linear version is verified as the discretized (isotropic) graph heat diffusion [15], which minimizes the graph Dirichlet energy. Similarly, GAT [42] is related to an anisotropic heat diffusion on graphs as shown in [6]. Furthermore, many recent studies introduce GNNs that are inspired by physical systems and diffusion PDEs, including heat diffusion with a source term [40], non-Euclidean Beltrami flow [7], wave diffusion equation [19], networks of coupled oscillators [39], nonlinear anisotropic diffusion [11]. The recent work [16] provides a framework for analyzing continuous formulation of GNN models as gradient flows of some energy. In particular, the work proposes a general energy where its gradient flow leads to many existing GNN models. Under such framework, [16] verifies that many models can only lead to LFD dynamics, including GCN [28], GRAND [6], CGNN [47] and PDE-GCN [19].

Framelet-based graph learning. The work [17] provides an efficient way to compute graph framelet transforms via Chebyshev polynomial approximation, which allows practical applications such as semi-supervised clustering. Graph spectral framelet convolution has been proposed in [51, 53], which is shown to empirically enhance the graph neural networks with its robustness and multi-scale properties. Later, graph framelets have been integrated for graph signal denoising [55], robust graph embedding [54], dynamic graph [56] and directed graph learning [58], exhibiting great improvement in model performance. More recently, [8] proposes a spatial graph framelet convolution and shows its close connection to the Dirichlet energy. The paper also introduces a perturbed Dirichlet energy in order to mitigate the over-smoothing issue.

3 Preliminaries

Graphs and graph convolution. A graph $G = (V_G, E_G)$ of $n$ nodes (with self-loops) can be represented by graph adjacency matrix $A \in \mathbb{R}^{n \times n}$. In this work, we assume the graph is undirected and unweighted, i.e., $A$ is symmetric with $a_{ij} = 1$ if $(i, j) \in E_G$ and 0 otherwise. In addition, we consider the symmetric normalized adjacency matrix as $\tilde{A} = D^{-1/2} A D^{-1/2}$ where $D$ is the diagonal degree matrix, with the $i$-th diagonal entry given by $d_i = \sum_j a_{ij}$, the degree of node $i$. The normalized graph Laplacian is given by $\tilde{L} = I_n - \tilde{A}$. We use $\rho_L$ to denote the largest eigenvalue (also called the highest frequency) of $\tilde{L}$. From the spectral graph theory [13], $\rho_L \leq 2$ and the equality holds if and only if there exists a connected component of the graph $G$ that is bipartite.
Graph convolution network (GCN) \cite{KipfW17} defines the layer-wise propagation rule via the normalized adjacency matrix as

$$H(\ell + 1) = \sigma(\hat{A}H(\ell)W^\ell),$$  \hspace{1cm} (1)$$

where $H(\ell)$ denotes the feature matrix at layer $\ell$ with $H(0) = X \in \mathbb{R}^{n \times c}$, the input features (also called input signals), and $W^\ell$ is the learnable feature transformation. It can be shown that GCN corresponds to a localized filter via graph Fourier transform, i.e., $h(\ell + 1) = U^T(I_n - \Lambda)Uh(\ell)$ where $U, \Lambda$ are from the eigendecomposition $L = U^T\Lambda U$ and $Uh$ is known as the Fourier transform of a graph signal $h \in \mathbb{R}^n$. In this paper, we let $\{(\lambda_i, u_i)\}_{i=1}^n$ be the set of eigen-pairs of $\hat{L}$ where $u_i$ are the row vectors of $U$.

**Graph framelets and framelet convolution.** Graph (undecimated) framelets \cite{Estrada12} are defined via a filter bank $\eta = \{a; b^{(1)}, ..., b^{(L)}\}$ and its induced (complex-valued) scaling functions $\Psi = \{\alpha; \beta^{(1)}, ..., \beta^{(L)}\}$ where $L$ is the number of high-pass filters. Particularly, it satisfies that $\hat{a}(2\xi) = \hat{a}(\xi)\hat{a}(\xi)$ and $\hat{b}^{(r)}(2\xi) = \hat{b}^{(r)}(\xi)\hat{a}(\xi)$ for all $\xi \in \mathbb{R}, r = 1, ..., L$ where $\hat{a}, \beta^{(r)}$ denote the Fourier transform of $a, \beta^{(r)}$ and $\hat{a}, b^{(r)}$ denote the Fourier series of $a, b^{(r)}$ respectively. The graph framelets are defined by $\varphi_{j,p}(v) = \sum_{i=1}^n \hat{a}(\lambda_i/2^j)u_i(p)u_i(v)$ and $\psi_{j,p}^r(v) = \sum_{i=1}^n \hat{b}^{(r)}(\lambda_i/2^j)u_i(p)u_i(v)$ for $r = 1, ..., L$ and for scale level $j = 1, ..., J$. We use $u_i(v)$ to represent the eigenvector $u_i$ at node $v$. $\varphi_{j,p}$ and $\psi_{j,p}^r$ are known as the low-pass framelets and high-pass framelets at node $p$.

The framelet coefficients of a graph signal $h$ are given by $v_0 = \{(\varphi_{0,p}, h)\}_{p \in \mathcal{V}_0}$, $w_{j} = \{(\psi_{j,p}, h)\}_{p \in \mathcal{V}_j}$. For a multi-channel signal $H \in \mathbb{R}^{n \times c}$, we can compactly write its framelet coefficients as

$$V_0 = U^T\hat{a}\left(\frac{\Lambda}{2}\right)UH, \quad W_{j} = U^T\hat{\beta}^{(r)}\left(\frac{\Lambda}{2^{J+1}}\right)UH, \quad \forall j = 0, ..., J, r = 1, ..., L$$

where $\hat{a}, \hat{\beta}^{(r)}$ applies elementwise to the diagonal of $\Lambda$, and $V_0, W_j$ are respectively the low-pass and high-pass coefficients. Define the framelet transform matrices $W_{0,j}, W_{r,j}$ such that $V_0 = W_{0,j}H = U^T\Lambda_{0,j}UH$, and $W_j = W_{r,j}H = U^T\Lambda_{r,j}UH$ for $r = 1, ..., L, j = 1, ..., J$, where $\Lambda_{r,j}$ is a diagonal matrix with entries $(\Lambda_{0,j})_{ii} = \hat{\alpha}(\lambda_i/2^j)$ and $(\Lambda_{r,j})_{ii} = \hat{\beta}^{(r)}(\lambda_i/2^{J+1})$. By the tightness of the framelet transform, we have $\Lambda_{0,j}^2 + \sum_{r,j} \Lambda_{r,j}^2 = I_n$ and the framelet decomposition and reconstruction are invertible, i.e., $W_{0,j}^TW_{0,j}H + \sum_{r,j} W_{r,j}^TW_{r,j}H = H$. This property allows unique decomposition of any graph signal onto the spectral framelet domain. Refer to \cite{Estrada12,Shuman13} for more detailed discussions.

The **spectral graph framelet convolution** is proposed in \cite{Estrada12}, which is similar to the graph (Fourier) convolution by applying a filter on the spectral domain before reconstructing the signal. The layer-wise propagation rule is given by $H(\ell + 1) = \sigma(W_{0,\ell}^T \text{diag}(\theta_{0,\ell})W_{0,\ell}H(\ell)W^\ell + \sum_{r,j} W_{r,j}^T \text{diag}(\theta_{r,j})W_{r,j}H(\ell)W^\ell)$, where $\theta_{r,j} \in \mathbb{R}^n$ is a learnable filter coefficient and $W^\ell$ is a shared weight matrix across all $r,j$ for layer $\ell$. Rather than performing spectral filtering as in \cite{Estrada12}, the **spatial graph framelet convolution** performs a spatial message passing over the spectral framelet domain \cite{Estrada12} as $H(\ell + 1) = W_{0,j}^T \sigma(\hat{A}W_{0,j}H(\ell)W^\ell) + \sum_{r,j} W_{r,j}^T \sigma(\hat{A}W_{r,j}H(\ell)W^\ell)$.

In this paper, our subsequent analysis focuses on the spatial framelet convolution (or just framelet convolution) due to its connection to the Dirichlet energy as shown in Section 4. Nevertheless, we also obtain similar conclusions for spectral framelet convolution in Section 6.

\*With a slight abuse of notations, we use $\top$ to also represent conjugate transpose if the filters are complex-valued.
Dirichlet energy and gradient flow. The graph Dirichlet energy provides a measure of smoothness of a graph signal $\mathbf{H} \in \mathbb{R}^{n \times c}$, which is defined as $E(\mathbf{H}) = \frac{1}{2} \sum_{i,j} a_{ij} \| \mathbf{h}_i / \sqrt{\mathcal{A}_i} - \mathbf{h}_j / \sqrt{\mathcal{A}_j} \|^2 = \frac{1}{2} \text{tr} (\mathbf{H}^T \hat{\mathbf{L}} \mathbf{H})$, where $\mathbf{h}_i \in \mathbb{R}^c$ is the signal vector of node $i$. The gradient flow of the Dirichlet energy yields the so-called graph heat equation \cite{13} as $\dot{\mathbf{H}}(t) = - \nabla E(\mathbf{H}(t)) = - \hat{\mathbf{L}} \mathbf{H}(t)$. Its Euler discretization leads to the propagation of linear GCN models \cite{46,45}. The process is called Laplacian smoothing \cite{32} and it converges to the kernel of $\hat{\mathbf{L}}$, i.e., $\text{ker}(\hat{\mathbf{L}})$ as $t \to \infty$, resulting in non-separation of nodes with same degrees, known as the over-smoothing issue.

A recent work \cite{16} also shows that even with nonlinear activation and weight matrix as in classic GCN \cite{1}, the process is still dominated by the low frequency, thus eventually converging to the kernel of $\hat{\mathbf{L}}$, for almost every initialization. To quantify such behavior, \cite{16} considers a general dynamics as $\dot{\mathbf{H}}(t) = \text{GNN}_\theta(\mathbf{H}(t), t)$, with $\text{GNN}_\theta(\cdot)$ as an arbitrary graph neural network function, and also characterizes its behavior by low/high-frequency-dominance (L/HFD).

**Definition 1** \cite{16}. $\mathbf{H}(t) = \text{GNN}_\theta(\mathbf{H}(t), t)$ is Low-Frequency-Dominant (LFD) if and only if $E(\mathbf{H}(t)/\|\mathbf{H}(t)\|) \to 0$ as $t \to \infty$, and is High-Frequency-Dominant (HFD) if $E(\mathbf{H}(t)/\|\mathbf{H}(t)\|) \to \rho_L/2$ as $t \to \infty$.

**Lemma 1** \cite{16}. A GNN model is LFD (resp. HFD) if and only if for each $t_j \to \infty$, there exists a subsequence indexed by $t_{jk} \to \infty$ and $\mathbf{H}_\infty$ such that $\mathbf{H}(t_{jk})/\|\mathbf{H}(t_{jk})\| \to \mathbf{H}_\infty$ and $\hat{\mathbf{L}} \mathbf{H}_\infty = 0$ (resp. $\hat{\mathbf{L}} \mathbf{H}_\infty = \rho_L \mathbf{H}_\infty$).

**Remark 1** (LFD, HFD, homophilic and heterophilic). The characterization of LFD and HFD has direct consequences on model performance for homophilic and heterophilic graph datasets. As shown in Lemma 1 a GNN model is LFD if and only if there exists a subsequence of $\mathbf{H}(t)$ that converges to $\text{ker}(\hat{\mathbf{L}})$ (thus smoothing) and is HFD if and only if there exists a subsequence that converges to the eigenvector associated with the largest frequency of $\hat{\mathbf{L}}$ (thus separating). In the case of homophilic graph, where neighboring nodes tend to share the same label, low-frequency information is more important while in contrast for heterophilic graph, high-frequency information is more important. Thus, ideally, a model should be flexible enough to accommodate both LFD and HFD dynamics.

4 Dirichlet energy, gradient flow and framelet convolution

This section studies the following linearized spatial framelet convolution proposed in \cite{8} and its connection to energy and gradient flow.

$$\mathbf{H}(\ell + 1) = \mathbf{W}_{0,J}^T \hat{\mathbf{A}} \mathbf{W}_{0,J} \mathbf{H}(\ell) \mathbf{W}_{0,J} + \sum_{r,j} \mathbf{W}_{r,j}^T \hat{\mathbf{A}} \mathbf{W}_{r,j} \mathbf{H}(\ell) \mathbf{W}_{r,j}. \quad (2)$$

The consideration of the linear dynamics as in \cite{2} is driven by the definition of an energy, which is often a quadratic form \cite{16}. Further, it has been shown in \cite{14,16} that removing nonlinearity often does not severely harm the prediction performance. Nevertheless, we investigate the effect of nonlinearity on the gradient flow in Section 5.3. Furthermore, we require the weight matrices $\mathbf{W}_{r,j} = \mathbf{W}_{r,j}$ to be fixed across all layers and to be symmetric (in order to ensure that it leads to a gradient flow of some energy as discussed in Section 5). The symmetric assumption is not strict, given the universal approximation results in \cite{27} and has also been considered in \cite{16}.

We start with a result showing that the total Dirichlet energy for all framelet coefficients (i.e., $\mathbf{W}_{r,j} \mathbf{H}$) is identical to the Dirichlet energy on the spatial domain.
Proposition 1 (Energy conservation via framelet decomposition [8]). For a graph signal $H$, let $E_{0,J}(H) = \frac{1}{2}tr((W_{0,J}H)\hat{L}W_{0,J}H)$ and similarly $E_{r,J}(H) = \frac{1}{2}tr((W_{r,J}H)\hat{L}W_{r,J}H)$ for $r = 1, \ldots, L, j = 1, \ldots, J$. Then we have $E(H) = E_{0,J}(H) + \sum_{r_J} E_{r,J}(H)$.

Proposition 1 immediately shows that the gradient flow of the Dirichlet energy can be equivalently expressed as $\dot{H}(t) = -\nabla E(H(t)) = -\left((W_{0,J}\hat{L}W_{0,J}H(t) + \sum_{r,J} W_{r,J}\hat{L}W_{r,J}H(t))\right)$ where its discretization (with a stepsize $\tau = 1$) leads to

$$H(\ell + 1) = W_{0,J}\hat{A}W_{0,J}H(\ell) + \sum_{r,J} W_{r,J}\hat{A}W_{r,J}H(\ell)$$

due to $W_{0,J}W_{0,J} + \sum_{r,J} W_{r,J}W_{r,J} = I_n$. Comparing the above dynamics to the framelet convolution in (2), we see the only difference is in the weight matrices $W_{r,J}$. In fact, if the weight matrices $W_{r,J}$ are the same across all $r, j$, the dynamics would be equivalent to that of GCN, thus inducing a LFD dynamics for almost every $H(0)$ [16, Theorem 4.3]. To see this, let $W_{0,J} = W_{r,J} = W$. Then (2) becomes

$$H(\ell + 1) = (W_{0,J}\hat{A}W_{0,J} + \sum_{r,J} W_{r,J}\hat{A}W_{r,J})H(\ell)W$$

$$= \left(U^T A^2_{0,J}(I_n - A)U + \sum_{r,J} U^T A^2_{r,J}(I_n - A)U\right)H(\ell)W$$

$$= U^T(I_n - A)UH(t)W = \hat{A}H(\ell)W$$

where we use the fact that $A^2_{0,J} + \sum_{r,J} A^2_{r,J} = I_n$. However, we show in Theorem 1 that with different $W_{r,J}$, the framelet convolution is able to induce a HFD dynamics.

For the purpose of analysis, we focus on the Haar-type filter [17], which is commonly considered for framelet convolution, such as [8, 51, 55, 56]. We remark that similar analysis can be performed for other filters. The filter bank of Haar filter is given by a low-pass filter $\hat{a}(\xi) = \cos(\xi/2)$, and a high-pass filter $\hat{b}(\xi) = \sin(\xi/2)$. We focus on the cases when the scale is $J = 1$ and $J = 2$, which are sufficient for practical applications [8, 51]. When $J = 1$, we can explicitly write the induced scaling functions as $\hat{a}(\Lambda/2) = \cos(\Lambda/8)$ and $\hat{b}(\Lambda/2) = \sin(\Lambda/8)$. When $J = 2$, the scaling functions are given in [8, Proposition 6] as $\hat{a}(\Lambda/2) = \cos^2(\Lambda/8)\cos(\Lambda/16)$, $\hat{b}(\Lambda/2) = \sin(\Lambda/8)\cos(\Lambda/16)$, $\hat{b}(\Lambda/4) = \sin(\Lambda/16)$.

Theorem 1. The (spatial) graph framelet convolution [2] with Haar-type filter can induce both LFD and HFD dynamics. Specifically, let $W_{0,J} = I_c, W_{r,J} = \lambda^W I_c$ for $r = 1, \ldots, L, j = 1, \ldots, J$. Then the framelet convolution can be HFD when $|\lambda^W| > 1$ and is sufficiently large. On the other hand, when $|\lambda^W| \leq 1$, the framelet convolution can only be LFD.

Proof. Consider the linearized framelet convolution (3) with a stepsize $\tau$ as

$$H(t + \tau) = \tau W_{0,J}\hat{A}W_{0,J}H(t)W_{0,J} + \tau \sum_{r,J} W_{r,J}\hat{A}W_{r,J}H(t)W_{r,J},$$

for some symmetric matrices $W_{0,J}, W_{r,J} \in \mathbb{R}^{c \times c}$ where $c$ is the number of channels (feature dimension).

We consider the Haar-type filter with one high pass ($r = 1$) as in [17]. The Haar filter with one high pass has filter bank given by $\hat{a}(\xi) = \cos(\xi/2), \hat{b}(\xi) = \sin(\xi/2)$. We start the analysis with a single scale $J = 1$ and then extend the analysis to two scales $J = 2$, as commonly used for framelet convolution in [8, 51].
Haar filter with one scale. When $J = 1$, we have

$$W_{0,1} = U^T A_{0,1} U = U^T \cos(A/8) U, \quad W_{1,1} = U^T A_{1,1} U = U^T \sin(A/8) U,$$

Now consider the vectorization of (3) as

$$\text{vec}(H(m\tau)) = \left(\tau W_{0,1} \otimes W_{0,1} \hat{A} W_{0,1} + \tau W_{1,1} \otimes W_{1,1} \hat{A} W_{1,1}\right)^m \text{vec}(H(0)), \quad (4)$$

where $\otimes$ represents the Kronecker product. To simplify the analysis, we assume $W_{0,1} = I_c$, $W_{1,1} = 1^W I_c$ without loss of generality. From standard results on Kronecker product, we can rewrite (4) as

$$\text{vec}(H(m\tau)) = \tau^m \sum_{k,i} \left(\left(\lambda_i^{A_{0,1}} + \lambda_i^W (\lambda_i^{A_{1,1}})^2\right)(1 - \lambda_i)\right)^m c_{k,i}(0) e_k \otimes u_i, \quad (5)$$

where we recall that $\{(\lambda_i, u_i)\}$ are the eigen-pairs of $\hat{L}$ and $c_{k,i}(0) = \langle \text{vec}(H(0)), e_k \otimes u_i \rangle$. We also denote $\lambda_i^{A_{r,j}} = (A_{r,j})_{ii}$ for all $(r,j)$. We show when $|\lambda_i^W| > 1$, the dynamics (5) can be HFD. \(^1\)

When $\lambda_i = 0$, we have $\left|\left(\lambda_i^{A_{0,1}} + \lambda_i^W (\lambda_i^{A_{1,1}})^2\right)(1 - \lambda_i)\right| = 1.$

When $\lambda_i \neq 0$, we have

$$\left|\left(\lambda_i^{A_{0,1}} + \lambda_i^W (\lambda_i^{A_{1,1}})^2\right)(1 - \lambda_i)\right| = \left|\cos^2(\lambda_i/8) + \lambda_i^W \sin^2(\lambda_i/8)\right|(1 - \lambda_i)$$

First given that $\lambda_i \in [0, \rho_L]$, $\rho_L \leq 2$, we can see when $|\lambda_i^W| \leq 1$, $\cos^2(\lambda_i/8) + \lambda_i^W \sin^2(\lambda_i/8) \in (0, 1)$ and is non-increasing in $\lambda_i$. Hence $\left|\left(\lambda_i^{A_{0,1}} + \lambda_i^W (\lambda_i^{A_{1,1}})^2\right)(1 - \lambda_i)\right|$ takes its maximum at $\lambda_i = 0$ regardless of the value of $\lambda_i^W$. This suggests the dominant frequency is 0, thus inducing a LFD dynamics.

On the other hand, when $|\lambda_i^W| > 1$ and is sufficiently large (with respect to the $\rho_L$), then we can show the function $\left|\left(\lambda_i^{A_{0,1}} + \lambda_i^W (\lambda_i^{A_{1,1}})^2\right)(1 - \lambda_i)\right|$ takes its unique maximum at frequency $\rho_L$. \(^2\) Particularly, denote $\delta_{\text{HFD}} := (\cos^2(\rho_L/4) + \lambda_i^W \sin^2(\rho_L/4))(\rho_L - 1)$ and suppose $|\lambda_i^W| > 1$ sufficiently large such that for all $i$ where $\lambda_i \neq \rho_L$, $\left|\left(\lambda_i^{A_{0,1}} + \lambda_i^W (\lambda_i^{A_{1,1}})^2\right)(1 - \lambda_i)\right| < \delta_{\text{HFD}}$ holds. Then we can show the dynamics is HFD where the dominant frequency is $\rho_L$.

More precisely, let $\delta := \max\{\lambda_i \neq \rho_L\} \left|\left(\lambda_i^{A_{0,1}} + \lambda_i^W (\lambda_i^{A_{1,1}})^2\right)(1 - \lambda_i)\right|$. We also denote $P_\rho = \sum_k (e_k \otimes u_{\rho}) (e_k \otimes u_{\rho})^\top$ where $u_{\rho}$ is the eigenvector of $\hat{L}$ associated with eigenvalue $\rho_L$ (assuming the eigenvalue $\rho_L$ is simple). Then we can decompose (5) as

$$\text{vec}(H(m\tau)) = \tau^m \sum_{k,i} \delta_{\text{HFD}} \sum_{k,i: \lambda_i \neq \rho_L} c_{k,i}(0) e_k \otimes u_{\rho}$$

\(^3\)We note that there may exist other settings where the dynamics is HFD. Here we only provide one case where this happens.

\(^3\)We remark that when $\rho_L \to 1$, for sufficiently large $|\lambda_i^W| > 1$, it can happen that the maximum is attained at frequency $\lambda \in (0, 1)$. In this case, the dynamics is neither LFD nor HFD.
where $\delta < \delta_{\text{HFD}}$. By normalizing the results, we obtain \( \frac{\text{vec}(H(m,r))}{\|\text{vec}(H(m,r))\|} \rightarrow \frac{P_{\rho}(\text{vec}(H(0)))}{\|P_{\rho}(\text{vec}(H(0)))\|} \), as $m \rightarrow \infty$, where the latter is a unit vector $h_{\infty}$ satisfying $(I_c \otimes \hat{L})h_{\infty} = \rho_L h_{\infty}$. This suggests, the dynamics is HFD according to Definition 1 and Lemma 1.

**Haar filter with two scales.** When $J = 2$, we have

\[
W_{0,2} = U^\top \cos(\Lambda/8)\cos(\Lambda/16)U, \quad W_{1,1} = U^\top \sin(\Lambda/8)\cos(\Lambda/16)U, \quad W_{1,2} = U^\top \sin(\Lambda/16)U.
\]

Following the same strategy, we assume $W_{0,2} = I_c$ and $W_{1,1} = W_{1,2} = \lambda^W I_c$. Then the vectorization of the dynamics writes

\[
\text{vec}(H(m,r)) = \tau^m \sum_{k,i} \left( (\lambda^A_{0,2})^2 + \lambda^W ((\lambda^A_{1,1})^2 + (\lambda^A_{1,2})^2) \right)^m (1 - \lambda_i)^m c_{k,i}(0)e_k \otimes u_i.
\]

It can be similarly verified that when $|\lambda^W| < 1$, we have $(\lambda^A_{0,2})^2 + \lambda^W ((\lambda^A_{1,1})^2 + (\lambda^A_{1,2})^2) \in (0,1]$ and is non-increasing in $\lambda_i$. Thus in this case, the dynamics can only be LFD for the same reason. On the other hand, for $|\lambda^W| > 1$ sufficiently large, the dominant frequency can be $\rho_L$ and the dynamics is HFD given the definitions of the scaling functions above. Following the same step as above, we can show the dynamics is HFD.

**Remark 2.** Importantly from Theorem 1, we conclude that given $W_{0,J} = I_c$, $W_{r,j} = \lambda^W I_c$, the framelet convolution can be HFD when $|\lambda^W| > 1$ and is sufficiently large, i.e., when the eigenvalue magnitude of the high-pass weight matrices is sufficiently larger than that of the low-pass. We also show that when $|\lambda^W| < 1$, the dynamics can only be LFD. This conclusion is irrespective of whether there exists any negative spectrum in the weight matrices (i.e., in this case, $\lambda^W$ can be negative or positive). In contrast, without framelet decomposition, in order to induce a HFD dynamics, \[16\] requires a sufficiently large negative spectrum of the weight matrix. This shows with framelet decomposition, the parameters are given more flexibility in adapting to the right frequency.

Based on Theorem 1, we shall expect framelet convolution to perform better than GCN on both homophilic and heterophilic graphs, which aligns with empirical observations \[8\].

## 5 Generalized framelet energy and its gradient flow

This section proposes a generalized energy via framelet decomposition. We first recall the general energy $\mathcal{E}(H)$ proposed in \[16\], which is designed to induce both LFD and HFD dynamics. That is, $\mathcal{E}(H) = \frac{1}{2} \sum_i \langle h_i, \Omega h_i \rangle - \frac{1}{2} \sum_{i,j} a_{ij} \langle h_i, Wf_j \rangle$, for some symmetric matrices $\Omega, W$, and can be written alternatively as

\[
\mathcal{E}(H) = \frac{1}{2} \text{tr}(H^\top \Omega) - \frac{1}{2} \text{tr}(H^\top \hat{A}HW) = \frac{1}{2} \langle \text{vec}(H), (\Omega \otimes I_n - W \otimes \hat{A})\text{vec}(H) \rangle
\]  

(6)

It is worth noticing that the generalized energy includes the Dirichlet energy as a special case. By setting $\Omega = W = I_c$, we recover the Dirichlet energy. It has been shown in \[16\] that the positive spectrum of $W$ in \[6\] encodes attractive interactions via graph gradients while the negative spectrum encodes repulsive interactions. Thus, by adjusting the spectrum of $W$ provides a control for frequency reduction and amplification.
Now we propose a framelet generalized energy as follows. Let $E_{r,j}(H) = \frac{1}{2} \text{tr}(W_{r,j}H^T \Omega_{r,j} W_{r,j} H - \sum_{i=1}^n \Lambda_{0,j} (I_n - \Lambda) U H W H^T)$ for all $(r, j) \in I$. Then the total framelet generalized energy is given by

$$E^\text{tot}(H) = E_{0,j}(H) + \sum_{r,j} E_{r,j}(H) = \frac{1}{2} \text{tr}(H^T W_{0,j}^T \Omega_{0,j} W_{0,j} H W) - \frac{1}{2} \text{tr}(H^T W_{0,j}^T \Omega_{0,j} W_{0,j} H W) + \sum_{r,j} \left( \frac{1}{2} \text{tr}(H^T W_{r,j}^T \Omega_{r,j} W_{r,j} H W) - \frac{1}{2} \text{tr}(H^T W_{r,j}^T \Omega_{r,j} W_{r,j} H W) \right)$$

$$= \frac{1}{2} \text{tr}(H^T \Omega H) - \frac{1}{2} \text{tr}(U^T A_{0,j}^2 (I_n - \Lambda) U H W H^T) - \frac{1}{2} \sum_{r,j} \text{tr}(U^T A_{r,j}^2 (I_n - \Lambda) U H W H^T)$$

$$= \frac{1}{2} \text{tr}(H^T \Omega H) - \frac{1}{2} \text{tr}(H^T \Lambda H) - \frac{1}{2} \sum_{r,j} \text{tr}(H^T \Lambda H) = E(H),$$

where we use the fact that $A_{0,j}^2 + \sum_{r,j} A_{r,j}^2 = I_n$. \hfill \qed

5.1 Proposed gradient flow of the generalized energy

We next propose a GNN model, named gradient flow based undecimated framelet convolution (GradF-UFG), which is given by the discretization of the gradient flow of the proposed framelet energy $E^\text{tot}$, i.e. $\dot{H}(t) = -\nabla E^\text{tot}(H(t))$ as

$$H(t + \tau) = H(t) - \tau \sum_{(r,j) \in I} \left( W_{r,j}^T \Omega_{r,j} W_{r,j} H(t) - W_{r,j}^T \hat{\Lambda} W_{r,j} H(t) W_{r,j} \right)$$

with $H(0) = \text{MLP}_\theta(X)$, where $\text{MLP}_\theta(\cdot)$ is an encoder and $X$ is the input feature matrix. The encoder model can be used to alter the dimension of the input features and also include nonlinearity to increase the expressiveness. In Section 5.3, we explore an alternative way to include nonlinearity and study its influence on the energy.

Next we show in the below proposition that the framelet convolution (2) is in fact a special instance of the proposed gradient flow discretization, GradF-UFG.
Proposition 3 (Framelet convolution as gradient flow of the proposed generalized energy $\mathcal{E}^{\text{tot}}$). When $\Omega_{0,j} = \Omega_{r,j} = I_c$, then framelet convolution (2) corresponds to the proposed GradF-UFG with a stepsize $\tau = 1$.

Proof. The proof directly follows from the fact that $W_{0,r}^T W_{0,J} + \sum_{r,j} W_{r,j}^T W_{r,j} = I_n$. 

A multi-particle perspective on the proposed energy and its gradient flow. We can follow [16] by providing a multi-particle interpretation for the proposed generalized framelet energy $\mathcal{E}^{\text{tot}}$ as well as its gradient flow in (7). We may view each node as a particle in $\mathbb{R}^c$ with energy $\mathcal{E}^{\text{tot}}$. Let $\Theta_{r,j}, \Theta_{r,j}^+ \in \mathbb{R}^{c \times c}$ be defined such that $W_{r,j} = (\Theta_{r,j}^+)^\top \Theta_{r,j}^+ - (\Theta_{r,j}^-)^\top \Theta_{r,j}^-$. Then the energy $\mathcal{E}^{\text{tot}}$ can be rewritten as

$$
\mathcal{E}^{\text{tot}}(H) = \sum_{(r,j) \in I} \left( \frac{1}{2} \sum_{i=1}^n \langle W_{r,j} h_i, (\Omega_{r,j} - W_{r,j}) W_{r,j} h_i \rangle 
+ \frac{1}{4} \sum_{i,i' = 1}^n \| \Theta_{r,j}^+ \nabla W_{r,j} H_{ii'} \|^2 - \frac{1}{4} \sum_{i,i' = 1}^n \| \Theta_{r,j}^- \nabla W_{r,j} H_{ii'} \|^2 \right),
$$

where $\nabla W_{r,j} H_{ii'} = W_{r,j} (h_i / \sqrt{d_i} - h_{i'} / \sqrt{d_{i'}})$ is the graph gradient along edge $(i, i') \in E_G$ in terms of the framelet coefficients. The first term in (8) corresponds to an external energy, independent of the particle interaction. The latter two terms in (8) encode the attraction and repulsion respectively along the graph gradients projected onto the framelet domain.

The gradient flow of such energy (7) thus minimizes the divergence between the projected gradients through the positive spectrum of $W_{r,j}$ and maximizes the divergence through the negative spectrum. It should be noticed that when $\Omega_{r,j} = \Omega, W_{r,j} = W$, $\mathcal{E}^{\text{tot}}$ reduces to the energy $\mathcal{E}$ without framelet decomposition in [16], as shown in Proposition 2. In this case, let $W = (\Theta^+)^\top \Theta^+ - (\Theta^-)^\top \Theta^-$. Then minimizing $\| \Theta^+ (\nabla H)_{ij} \|^2$ introduces a smoothing effect and maximizing $\| \Theta^- (\nabla H)_{ij} \|^2$ induces a separating effect to the system. This suggests a sufficiently large negative spectrum (compared to the positive spectrum) is able to yield a HFD dynamics.

In contrast, we have shown in Theorem 1 that with framelet decomposition, all $W_{r,j}$ can be positive definite for a HFD system as long as sufficient weights are put on the high-pass components. In addition, defining energy on different framelets separately in (8) allows effects other than smoothing or separating (e.g. via band-pass filters), which can lead to a more flexible dynamics.

5.2 Discussions and comparisons with framelet enhanced energy

A recent work [8] introduces the following framelet perturbed energy $E^\epsilon(H)$, which provably enhances the Dirichlet energy for the same input signal.

$$
E^\epsilon(H) = \frac{1}{2} \text{tr}((W_{0,j} H)^\top (\hat{L} + \epsilon I_n) W_{0,j} H) + \frac{1}{2} \sum_{r,j} \text{tr}((W_{r,j} H)^\top (\hat{L} - \epsilon I_n) W_{r,j} H),
$$

for some chosen $\epsilon > 0$. Specifically, $E^\epsilon(H)$ slightly decreases the energy for the low-pass while increases the energy for the high-passes. The motivation for such definition of energy is that for the Haar-type filter, one can show $E^\epsilon(H) > E(H)$ for any $\epsilon > 0$ due to the energy gap between the
low-pass and high-pass filters in this case. However, we highlight that this strategy only applies to Haar-type filter and can be difficult to generalize to arbitrary framelet filters.

Now we first analyze the gradient flow of such energy without weight matrices as

\[
\dot{H}(t) = -\nabla E^\epsilon(H(t)) = -W_{0,J}^T(\hat{L} + \epsilon I_n)W_{0,J}H(t) - \sum_{r,j} W_{r,J}^T(\hat{L} - \epsilon I_n)W_{r,J}H(t)
\]

\[
= -\dot{L}H(t) - \epsilon(W_{0,J}^T W_{0,J} - \sum_{r,j} W_{r,J}^T W_{r,J})H(t)
\]

\[
= -U^T(\Lambda + \epsilon \Delta_{\text{gap}})UH(t),
\]

where we let \(W_{0,J}^T W_{0,J} - \sum_{r,j} W_{r,J}^T W_{r,J} = U^T \Delta_{\text{gap}} U\). Particularly for the case of Haar-type filter (with two scales), [8] Proposition 7 shows \(\Delta_{\text{gap}} \geq 0\) is a diagonal matrix with nonnegative diagonal entries. This is the primary reason for enhancement compared to the Dirichlet energy. Hence for the following analysis, we also focus on such choice of filter, i.e., \(\alpha(\Lambda/2) = \cos^2(\Lambda/8) \cos(\Lambda/16),\ \beta(\Lambda/2) = \sin(\Lambda/8) \cos(\Lambda/16),\ \hat{\beta}(\Lambda/4) = \sin(\Lambda/16)\).

We can write the solution of (9) in vectorized form as

\[
\text{vec}(H(t)) = \sum_{k,i} e^{-((\lambda_i + \epsilon \delta_i)t)c_{k,i}(0)}e_k \otimes u_i, \quad c_{k,i}(0) = \langle \text{vec}(H(0)), e_k \otimes u_i \rangle
\]

where we denote \(\delta_i = (\Delta_{\text{gap}})_i = \cos^2(\lambda_i/8) \cos(\lambda_i/16) - \sin^2(\lambda_i/8) \cos^2(\lambda_i/16) - \sin^2(\lambda_i/16)\) as the energy gap at frequency \(\lambda_i\), and \(\{e_k\}_{k=1}^c\) are the standard basis vectors of \(\mathbb{R}^c\). Immediately from the solution (10), we see the dynamics is LFD regardless of the choice of \(\epsilon\). This claim is emphasized in the following proposition.

**Proposition 4.** For any \(\epsilon > 0\), the solution \(H(t)\) from the gradient flow of \(E^\epsilon(H)\), i.e., \(\dot{H}(t) = -\nabla E^\epsilon(H(t))\) is LFD with \(E(H(t)) \to 0\) as \(t \to \infty\).

**Proof.** The Dirichlet energy of \(H(t)\) is given by

\[
E(H(t)) = \frac{1}{2} \langle \text{vec}(H(t)), (I_c \otimes \hat{L}) \text{vec}(H(t)) \rangle = \frac{1}{2} \sum_{k,i} e^{-2t(\lambda_i + \epsilon \delta_i)}c_{k,i}(0)^2 \lambda_i
\]

\[
\leq \frac{\rho L}{2} e^{-2t \vartheta} \|H(0)\|^2
\]

where we denote \(\vartheta := \min_{\lambda_i \in (0, \rho L]} \lambda_i + \epsilon \delta_i > 0\). Hence as \(t \to \infty\), we see \(E(H(t)) \to 0\). \(\square\)

**Remark 3.** Proposition [4] shows that despite enhancing the energy at each discrete timestamp, i.e., \(E^\epsilon(H(t)) \geq E(H(t))\), in the limit \((t \to \infty)\), the dynamics still converges to \(\ker(\hat{L})\), resulting in over-smoothing. Nevertheless, the convergence rate to the kernel is indeed slower once we inject the perturbation \(\epsilon \Delta_{\text{gap}} > 0\).

In order for the dynamics to be HFD, separate weight matrices are needed for each filter, as done in [8]. Specifically, the model proposed in [8], namely energy enhanced undecimated framelet graph convolution (EE-UFG), is given by

\[
H(\ell + 1) = W_{0,J}^T \sigma\left((\hat{A} - \epsilon I_n)W_{0,J}H(\ell)W_{0,J}^\ell + \sum_{r,j} W_{r,J}^T \sigma((\hat{A} + \epsilon I_n)W_{r,J}H(\ell)W_{r,J}^\ell)\right)
\]

(11)
We show in the next proposition that EE-UFG also can be written as the discretization of gradient flow of the proposed generalized framelet energy $E_{\text{tot}}(H)$, thus a special case of the proposed GradF-UFG.

**Proposition 5** (EE-UFG as gradient flow of the proposed generalized energy $E_{\text{tot}}$). When $\Omega_{0,J} = I_c + \epsilon W_{0,J}$, $\Omega_{r,j} = I_c - \epsilon W_{r,j}$, for $r = 1, \ldots, L$, $j = 1, \ldots, J$, then EE-UFG (without activation function) corresponds to the GradF-UFG with a stepsize $\tau = 1$.

**Proof.** The proof is similar to that of Proposition 3 and hence omitted. \hfill \square

From Proposition 5, we see with specific choices of the weight matrices, one can recover the evolution of EE-UFG. We can then follow the similar analysis in Section 4 to show EE-UFG can induce both LFD and HFD dynamics. Nevertheless, we highlight that it is not the perturbation that steers the dynamics from over-smoothing (as shown in Proposition 4), but the separate weight matrices for low-frequency and high-frequency components of the signal.

In fact, we can further compare the framelet generalized energy with and without perturbation. Let $E_{\text{frame}}$ denote the energy $E_{\text{tot}}$ with $\Omega_{0,J} = \Omega_{r,j} = I_c$ that leads to the framelet convolution (as in Proposition 3), and let $E'_{\text{frame}}$ denotes its perturbed energy as in Proposition 5. Then we have

$$E'_{\text{frame}}(H) = E_{\text{frame}}(H) + \frac{\epsilon}{2} \sum_{i=1}^n \left( \langle W_{0,J}h_i, W_{0,J}W_{0,J}h_i \rangle - \sum_{r,j} \langle W_{r,j}h_i, W_{r,j}W_{r,j}h_i \rangle \right). \tag{12}$$

When all the weight matrices are $W_{r,j} = I_c$ as in the case for Dirichlet energy, the perturbation indeed provides an enhancement in the energy. Nonetheless, for the general case in (12), whether the perturbation increases or decreases the energy depends on the relative spectrum of the weight matrices for different framelets.

We finally remark that, as opposed to setting a fixed $\epsilon$ as in [8], the proposed energy allows more flexible choices of the weights that adapt to different datasets.

### 5.3 The influence of the activation functions

So far, we have only considered the linearized dynamics of gradient flow or the nonlinearity is decoupled from the dynamics via an encoder model (as in (7)). Here we study an activated GradF-UFG model, where nonlinearity is added at each timestamp. Although the dynamics no longer qualifies as a gradient flow, we show the analysis is not affected. Particularly, we consider an activated gradient flow similar as in [16]:

$$\dot{H}(t) = \sigma(-\nabla E_{\text{tot}}(H(t))) = \sigma\left(-\sum_{(r,j) \in I} (W_{r,j}^T W_{r,j}H(t)\Omega_{r,j} - W_{r,j}^T \hat{A} W_{r,j}H(t) W_{r,j})\right) \tag{13}$$

Then we show in the following theorem that for common activation functions, the conclusions from the paper remains the same.

**Theorem 2.** For any activation function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ satisfying $x\sigma(x) \geq 0$ for all $x \in \mathbb{R}$, then suppose $H(t)$ solves (13), we have $E_{\text{tot}}(H(t))$ is decreasing in $t$. For its discretization, $H(t + \tau) = H(t) + \tau \sigma(-\nabla E_{\text{tot}}(H(t)))$, we have $E'_{\text{tot}}(H(t)) \leq E_{\text{tot}}(H(t)) + C_m \|H(t + \tau) - H(t)\|^2$, where $C_m$ is the largest eigenvalue magnitude of $\frac{1}{2} \sum_{(r,j) \in I} (\Omega_{r,j} \otimes W_{r,j}^T W_{r,j} - W_{r,j} \otimes W_{r,j}^T \hat{A} W_{r,j})$.\hfill \square
We may consider a even more general energy via graph framelets by including a source term $\theta$. We have thus far focused on spatial framelet convolution \cite{8}. For the spectral framelet convolution, \[ E(\theta) = \frac{1}{2} \sum_{(r,j) \in I} (\Omega_{r,j} \otimes W_{r,j}^T W_{r,j} - W_{r,j} \otimes W_{r,j}^T \tilde{A} W_{r,j}) \vec{H}(t) \] for all $\theta \in \mathbb{R}^n$. Particularly, we define $\theta = 1$, which is symmetric.

The above results suggest that for common activation functions satisfying $x \sigma(x) \geq 0$, such as ReLU and tanh, the activated gradient flow still minimizes the energy.

5.4 A even more general framelet energy with source term

We may consider a even more general energy via graph framelets by including a source term $E(0)$. The source term has been shown to avoid over-smoothing and increase the expressive power \cite{10, 20, 40}. Particularly, we define

\[ E_r(j) = \frac{1}{2} \sum_{(r,j) \in I} (\Omega_{r,j} \otimes W_{r,j}^T W_{r,j} - W_{r,j} \otimes W_{r,j}^T \tilde{A} W_{r,j}) \] for all $(r, j) \in I$. The corresponding GradF-UFG model is given by its discretized gradient flow as

\[ \text{GradF-UFG : } H(t+\tau) = H(t) - \tau \sum_{(r,j) \in I} \left( W_{r,j}^T W_{r,j} H(t) \Omega_{r,j} - W_{r,j}^T \tilde{A} W_{r,j} H(t) W_{r,j} - \beta W_{r,j}^T H(0) \tilde{W}_{r,j} \right) \]

where $H(0) = \text{MLP}_0(X)$.

6 A gradient flow perspective on the spectral framelet convolution

We have thus far focused on spatial framelet convolution \cite{8}. For the spectral framelet convolution \cite{51}, one can similarly characterize its behaviors as an energy gradient flow. Recall the (linearized) spectral framelet convolution (with fixed weight) is given by

\[ H(\ell + 1) = W_{\ell,j}^T \text{diag}(\theta_{\ell,j}) W_{\ell,j} H(\ell) W + \sum_{r,j} W_{r,j}^T \text{diag}(\theta_{r,j}) W_{r,j} H(\ell) W, \quad (14) \]

where $\theta_{r,j} \in \mathbb{R}^n$ is the filter coefficients for framelets at $(r, j)$. In particular, the energy that leads to the evolution in (14) is given by

\[ \mathcal{E}^\text{tot}(H) = \frac{1}{2} \sum_{(r,j) \in I} \text{tr} \left( (W_{r,j}^T W_{r,j} - (W_{r,j}^T \text{diag}(\theta_{r,j}) W_{r,j} H W.) \right). \quad (15) \]
We next show, similar to spatial framelet convolution, the spectral convolution in (14) can lead to both LFD and HFD dynamics.

**Theorem 3.** The spectral graph framelet convolution (14) with Haar-type filter can induce both LFD and HFD dynamics. Specifically, let $\theta_{0,J} = 1_n$ and $\theta_{r,j} = \theta 1_n$ for $r = 1, \ldots, L, J = 1, \ldots, J$ where $1_n$ is a vector of all 1s. Suppose $\theta \geq 0$. Then when $\theta \in [0, 1)$, the spectral framelet convolution is LFD and when $\theta > 1$, the spectral framelet convolution is HFD.

**Proof.** We here focus on the Haar filter with $J = 1$ scale. For $J = 2$, the proof is exactly the same. We can write the discretized gradient flow of (15) with a stepsize $\tau$ as

$$\text{vec}(H(m\tau)) = \tau^m \left( W \otimes (W_{0,1}^T W_{0,1} + \theta W_{1,1}^T W_{1,1}) \right)^m \text{vec}(H(0))$$

$$= \tau^m \sum_{k,i} \left( \lambda^W_k \left( \cos^2(\lambda_i/8) + \theta \sin^2(\lambda_i/8) \right) \right)^m c_{k,i}(0) \phi^W_k \otimes u_i,$$

where we let $\{(\lambda^W_k, \phi^W_k)\}_{k=1}^c$ as the eigen-pairs of $W$ and denote $c_{k,i}(0) = \langle \text{vec}(H(0)), \phi^W_k \otimes u_i \rangle$. Now we see

$$|\lambda^W_k \left( \cos^2(\lambda_i/8) + \theta \sin^2(\lambda_i/8) \right)| \leq \Delta^W \left( \cos^2(\lambda_i/8) + \theta \sin^2(\lambda_i/8) \right),$$

where we let $\Delta^W : = \max_k |\lambda^W_k|$. It is easy to see when $\theta > 1$, $\cos^2(\lambda_i/8) + \theta \sin^2(\lambda_i/8)$ is (monotonically) increasing in $\lambda_i \in [0, \rho_L]$ and its maximum is achieved at $\lambda_i = \rho_L$. On the other hand, when $\theta \in [0, 1)$, the function is (monotonically) decreasing and the maximum is achieved at $\lambda_i = 0$. The proof is complete by following the same steps as for Theorem 1.

**Remark 4.** From Theorem 3, we can draw a similar conclusion as for the spatial framelet convolution. That is, when the weights on the high-pass framelets are higher than that of the low-pass, the spectral framelet convolution is HFD. Otherwise, the model is LFD. This shows an equivalence between the spectral and spatial framelet convolution from the perspective of gradient flow.

### 7 Concluding remarks

In this paper, we have theoretically analyzed the behaviors of framelet-based GNNs via the framework of gradient flows and energy minimization. We have shown the framelet convolutions, either spatial or spectral, can accommodate both low-frequency-dominated (LFD) and high-frequency-dominated (HFD) dynamics, unlike many existing works that can only be LFD. This corroborates the good empirical performance of framelet-based models on heterophilic graphs and on mitigating the over-smoothing issue. The main underlying reason is that the framelet decomposition allows separate weight matrices on different frequencies. More specifically, when the weights on high-frequency components are relatively higher than the low-pass components, the dynamics can be HFD. In addition, we also propose a framelet generalized energy where its gradient flow includes many existing models as special cases. We have explained from various aspects why the proposed model generally leads to more flexible dynamics.

It should be noted that for simplicity of analysis, we have only focused on the Haar-based filter and when the weight matrices share the same eigen-space. This suffices for the scope of this work. It would be interesting to explore necessary conditions (if available) for framelet-based models to be LFD/HFD. We hope this work would motivate more theoretical understanding on the capability of multi-scale spectral GNNs via alternative frameworks.
References

[1] Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In International Conference on Learning Representations, 2020.

[2] Pedro HC Avelar, Anderson R Tavares, Marco Gori, and Luis C Lamb. Discrete and continuous deep residual learning over graphs. arXiv:1911.09554, 2019.

[3] Deyu Bo, Xiao Wang, Chuan Shi, and Huawei Shen. Beyond low-frequency information in graph convolutional networks. In AAAI Conference on Artificial Intelligence, volume 35, pages 3950–3957, 2021.

[4] Cristian Bodnar, Francesco Di Giovanni, Benjamin Paul Chamberlain, Pietro Liò, and Michael M Bronstein. Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in gnns. arXiv:2202.04579, 2022.

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

[6] Ben Chamberlain, James Rowbottom, Maria I Gorinova, Michael Bronstein, Stefan Webb, and Emanuele Rossi. Grand: Graph neural diffusion. In International Conference on Machine Learning, pages 1407–1418. PMLR, 2021.

[7] Benjamin Chamberlain, James Rowbottom, Davide Eynard, Francesco Di Giovanni, Xiaowen Dong, and Michael Bronstein. Beltrami flow and neural diffusion on graphs. Advances in Neural Information Processing Systems, 34:1594–1609, 2021.

[8] Jialin Chen, Yuelin Wang, Cristian Bodnar, Pietro Liò, and Yu Guang Wang. Dirichlet energy enhancement of graph neural networks by framelet augmentation. github, 2022.

[9] Jie Chen, Tengfei Ma, and Cao Xiao. FastGCN: Fast learning with graph convolutional networks via importance sampling. In International Conference on Learning Representations, 2018.

[10] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In International Conference on Machine Learning, pages 1725–1735. PMLR, 2020.

[11] Qi Chen, Yifei Wang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Optimization-induced graph implicit nonlinear diffusion. In International Conference on Machine Learning, pages 3648–3661. PMLR, 2022.

[12] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. Advances in Neural Information Processing Systems, 31, 2018.

[13] Fan RK Chung. Spectral graph theory, volume 92. American Mathematical Soc., 1997.

[14] Zhiyong Cui, Kristian Henrickson, Ruimin Ke, and Yinhai Wang. Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting. IEEE Transactions on Intelligent Transportation Systems, 21(11):4883–4894, 2019.
[15] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in Neural Information Processing Systems*, 29, 2016.

[16] Francesco Di Giovanni, James Rowbottom, Benjamin P Chamberlain, Thomas Markovich, and Michael M Bronstein. Graph neural networks as gradient flows. *arXiv:2206.10991*, 2022.

[17] Bin Dong. Sparse representation on graphs by tight wavelet frames and applications. *Applied and Computational Harmonic Analysis*, 42(3):452–479, 2017.

[18] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in Neural Information Processing Systems*, 28, 2015.

[19] Moshe Eliasof, Eldad Haber, and Eran Treister. PDE-GCN: Novel architectures for graph neural networks motivated by partial differential equations. *Advances in Neural Information Processing Systems*, 34:3836–3849, 2021.

[20] Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. In *International Conference on Learning Representations*, 2018.

[21] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pages 1263–1272. PMLR, 2017.

[22] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, 30, 2017.

[23] David K Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011.

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

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[26] Mingguo He, Zhewei Wei, Hongteng Xu, et al. Bernnet: Learning arbitrary graph spectral filters via bernstein approximation. *Advances in Neural Information Processing Systems*, 34:14239–14251, 2021.

[27] Shell Xu Hu, Sergey Zagoruyko, and Nikos Komodakis. Exploring weight symmetry in deep neural networks. *Computer Vision and Image Understanding*, 187:102786, 2019.

[28] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv:1609.02907*, 2016.

[29] Johannes Klicpera, Stefan Weßenberger, and Stephan Günnemann. Diffusion improves graph learning. In *International Conference on Neural Information Processing Systems*, pages 13366–13378, 2019.
[30] Guohao Li, Matthias Muller, Ali Thabet, and Bernard Ghanem. Deepgcns: Can gcns go as deep as cnns? In *International Conference on Computer Vision*, pages 9267–9276, 2019.

[31] Ming Li, Zheng Ma, Yu Guang Wang, and Xiaosheng Zhuang. Fast Haar transforms for graph neural networks. *Neural Networks*, 128:188–198, 2020.

[32] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *AAAI Conference on Artificial Intelligence*, 2018.

[33] Haggai Maron, Heli Ben-Hamu, Hadar Serviansky, and Yaron Lipman. Provably powerful graph networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[34] Chien-Chun Ni, Yu-Yao Lin, Feng Luo, and Jie Gao. Community detection on networks with Ricci flow. *Scientific reports*, 9(1):1–12, 2019.

[35] Hoang Nt and Takanori Maehara. Revisiting graph neural networks: All we have is low-pass filters. *arXiv:1905.09550*, 2019.

[36] Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2019.

[37] Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-GCN: Geometric graph convolutional networks. In *International Conference on Learning Representations*, 2019.

[38] Michael Poli, Stefano Massaroli, Junyoung Park, Atsushi Yamashita, Hajime Asama, and Jinkyoo Park. Graph neural ordinary differential equations. *arXiv:1911.07532*, 2019.

[39] T Konstantin Rusch, Ben Chamberlain, James Rowbottom, Siddhartha Mishra, and Michael Bronstein. Graph-coupled oscillator networks. In *International Conference on Machine Learning*, pages 18888–18909. PMLR, 2022.

[40] Matthew Thorpe, Tan Minh Nguyen, Hedi Xia, Thomas Strohmer, Andrea Bertozzi, Stanley Osher, and Bao Wang. GRAND++: Graph neural diffusion with a source term. In *International Conference on Learning Representations*, 2021.

[41] Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. In *International Conference on Learning Representations*, 2021.

[42] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[43] Jingyi Wang and Zhidong Deng. A deep graph wavelet convolutional neural network for semi-supervised node classification. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[44] Xiyuan Wang and Muhan Zhang. How powerful are spectral graph neural networks. In *International Conference on Machine Learning*, 2022.
[45] Yifei Wang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Dissecting the diffusion process in linear graph convolutional networks. *Advances in Neural Information Processing Systems*, 34:5758–5769, 2021.

[46] Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In *International Conference on Machine Learning*, pages 6861–6871. PMLR, 2019.

[47] Louis-Pascal Xhonneux, Meng Qu, and Jian Tang. Continuous graph neural networks. In *International Conference on Machine Learning*, pages 10432–10441. PMLR, 2020.

[48] Bingbing Xu, Huawei Shen, Qi Cao, Yunqi Qiu, and Xueqi Cheng. Graph wavelet neural network. In *International Conference on Learning Representations*, 2018.

[49] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2018.

[50] Mengxi Yang, Xuebin Zheng, Jie Yin, and Junbin Gao. Quasi-framelets: Another improvement to graph neural networks. *arXiv:2201.04728*, 2022.

[51] Xuebin Zheng, Bingxin Zhou, Junbin Gao, Yuguang Wang, Pietro Lió, Ming Li, and Guido Montufar. How framelets enhance graph neural networks. In *International Conference on Machine Learning*, pages 12761–12771. PMLR, 2021.

[52] Xuebin Zheng, Bingxin Zhou, Meng Li, Yu Guang Wang, and Junbin Gao. MathNet: Haar-like wavelet multiresolution-analysis for graph representation and learning. *arXiv:2007.11202*, 2020.

[53] Xuebin Zheng, Bingxin Zhou, Yu Guang Wang, and Xiaosheng Zhuang. Decimated framelet system on graphs and fast g-framelet transforms. *Journal of Machine Learning Research*, 23:18–1, 2022.

[54] Bingxin Zhou, Yuanhong Jiang, Yu Guang Wang, Jingwei Liang, Junbin Gao, Shirui Pan, and Xiaoqun Zhang. Graph neural network for local corruption recovery. *arXiv:2202.04936*, 2022.

[55] Bingxin Zhou, Ruikun Li, Xuebin Zheng, Yu Guang Wang, and Junbin Gao. Graph denoising with framelet regularizer. *arXiv:2111.03264*, 2021.

[56] Bingxin Zhou, Xinliang Liu, Yuelhua Liu, Yunying Huang, Pietro Lio, and YuGuang Wang. Spectral transform forms scalable transformer. *arXiv:2111.07602*, 2021.

[57] Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. *Advances in Neural Information Processing Systems*, 33:7793–7804, 2020.

[58] Chunya Zou, Andi Han, Lequan Lin, and Junbin Gao. A simple yet effective SVD-GCN for directed graphs. *arXiv:2205.09335*, 2022.