SIAM J. MATH. DATA SCI.
Vol. 5, No. 4, pp. 873–898

Understanding Graph Neural Networks with Generalized Geometric Scattering Transforms*

Michael Perlmutter†, Alexander Tong‡, Feng Gao§, Guy Wolf¶, and Matthew Hirn‖

Abstract. The scattering transform is a multilayered wavelet-based architecture that acts as a model of convolutional neural networks. Recently, several works have generalized the scattering transform to graph-structured data. Our work builds on these constructions by introducing windowed and non-windowed geometric scattering transforms for graphs based on two very general classes wavelets, which are in most cases based on asymmetric matrices. We show that these transforms have many of the same theoretical guarantees as their symmetric counterparts. As a result, the proposed construction unifies and extends known theoretical results for many of the existing graph scattering architectures. Therefore, it helps bridge the gap between geometric scattering and other graph neural networks by introducing a large family of networks with provable stability and invariance guarantees. These results lay the groundwork for future deep learning architectures for graph-structured data that have learned filters and also provably have desirable theoretical properties.

Key words. graph neural networks, geometric deep learning, wavelets, scattering transform

MSC codes. 05C62, 05C81, 42C15, 68R10, 68T07

DOI. 10.1137/21M1465056

1. Introduction.

The scattering transform is a wavelet-based model of convolutional neural networks (CNNs), introduced for signals defined on $\mathbb{R}^n$ by Mallat in [18]. Like the front end of a CNN, the scattering transform produces a representation of an inputted signal through an alternating cascade of filter convolutions and pointwise nonlinearities. It primarily differs from CNNs in two respects: (i) It uses predesigned, wavelet filters rather than filters learned through training data, and (ii) it uses the complex modulus $|\cdot|$ as its nonlinear activation

*Received by the editors January 4, 2022; accepted for publication (in revised form) June 6, 2023; published electronically October 25, 2023. The fourth and fifth authors jointly supervised the work.
https://doi.org/10.1137/21M1465056

Funding: This research was partially funded by IVADO (Institut de valorisation des données), grant PRF-2019-3583139727, and FRQNT (Fonds de recherche du Québec, Nature et technologies), grant 299376, Canada CIFAR AI Chair (fourth author); NIH (National Institutes of Health), NIGMS grant R01GM135929 (fifth and fourth authors); and NSF (National Science Foundation), DMS grant 1845856 (fifth author). The content provided here is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

†Boise State University, Department of Mathematics, 1910 University Drive, Boise, ID 83706 USA (mperlmutter@boisestate.edu).

‡Department of Computer Science and Operations Research, Université de Montréal, Montréal, QC H3T 1J4, Canada, and Mila - Quebec AI Institute, Montréal, QC H2S 3H1, Canada (alexander.tong@mila.quebec).

§Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, NY 10032 USA (fg2539@cumc.columbia.edu).

¶Department of Mathematics and Statistics, Université de Montréal, Montréal, QC H3T 1J4, Canada, and Mila - Quebec AI Institute, Montréal, QC H2S 3H1, Canada (guy.wolf@umontreal.ca).

‖Department of Computational Mathematics, Science, and Engineering and Department of Mathematics, Michigan State University, East Lansing, MI 48824 USA (mhirn@msu.edu).

function rather than more common choices, such as the rectified linear unit (ReLU). These differences lead to a network which provably has desirable mathematical properties. In particular, the Euclidean scattering transform is (i) nonexpansive on $\mathbf{L}^2(\mathbb{R}^n)$, (ii) invariant to translations up to a certain scale parameter, and (iii) stable to certain diffeomorphisms. In addition to these theoretical properties, the scattering transform has also been used to achieve very good numerical results in fields such as audio processing [1], medical signal processing [7], computer vision [23], and quantum chemistry [14].

While CNNs have proven tremendously effective for a wide variety of machine learning tasks, they typically assume that inputted data have a Euclidean gridlike structure. However, many data sets of interest, such as social networks, molecules, or surfaces appearing in computer graphics, have an intrinsically non-Euclidean structure and are naturally modeled as graphs or manifolds. This has motivated the rise of geometric deep learning, a field which aims to generalize deep learning methods to non-Euclidean settings. In particular, a number of papers have produced versions of the scattering transform for graphs [11, 12, 13, 32] and manifolds [24]. These constructions provide a model of geometric deep learning architectures, such as graph neural networks (GNNs), in a manner analogous to the way that [18] models CNNs.

In this paper, we construct two new families of wavelet transforms on a graph G from matrices $\mathbf{K}$, which are in most cases asymmetric, and provide a theoretical analysis of both of these wavelet transforms as well as the windowed and nonwindowed scattering transforms constructed from them. Because the matrices $\mathbf{K}$ are in general not symmetric, our wavelet transforms will not be nonexpansive frames on the standard unweighted inner product space. Instead, they will be nonexpansive on a certain weighted inner product space $\mathbf{L}^2(\mathbf{M})$, where $\mathbf{M}$ is an invertible weighting matrix. In important special cases, our matrix $\mathbf{K}$ will be either the lazy random walk matrix $\mathbf{P}$, its transpose $\mathbf{P}^T$, or its symmetric counterpart given by $\mathbf{T} = \mathbf{D}^{-1/2}\mathbf{P}\mathbf{D}^{1/2}$. In these cases, the weighting matrix will depend on the geometry of G.

We will use these wavelets to construct windowed and nonwindowed versions of the scattering transform on G. The windowed scattering transform inputs a signal $\mathbf{x} \in \mathbf{L}^2(\mathbf{M})$ and outputs a sequence of functions which we refer to as the scattering coefficients. We may view the windowed scattering transform as producing a sequence of features for each vertex. Therefore, it is well suited for tasks such as node classification. The nonwindowed scattering transform replaces the low-pass matrix used in the definition of the windowed scattering transform with an averaging operator $\boldsymbol{\mu}$ and instead outputs a sequence of scalar-valued coefficients. In some cases, it can be viewed as the limit of the windowed scattering transform as the scale of the low-pass tends to infinity (evaluated at some fixed coordinate $0 \leq i \leq n-1$). Since the nonwindowed scattering transform produces a single set of coefficients for the entire graph, it is well suited for whole-graph–level tasks, such as graph classification or regression.

1.1. Preliminaries. Let $G = (V, E, W)$ be a weighted, connected graph consisting of vertices V, edges E, and weights W, with n vertices. Without loss of generality, we take the vertices as $V = \{0, \ldots, n-1\}$. If $\mathbf{x} = (\mathbf{x}(0), \ldots, \mathbf{x}(n-1))^T$ is a function defined on V, we will identify $\mathbf{x}$ with the corresponding vector in $\mathbb{R}^n$. Let $\mathbf{A}$ denote the *weighted* adjacency matrix of G, let $\mathbf{d} = (\mathbf{d}(0), \ldots, \mathbf{d}(n-1))^T$ be the weighted degree vector, and let $\mathbf{D} = \operatorname{diag}(\mathbf{d})$. We will let

$$\mathbf{N} := \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$$

be the normalized graph Laplacian, let $0 \leq \omega_0 \leq \omega_1 \leq \ldots \leq \omega_{n-1} \leq 2$ denote the eigenvalues of $\mathbf{N}$, and let $\mathbf{v}_0, \ldots, \mathbf{v}_{n-1}$ be an orthonormal eigenbasis for $\mathbb{R}^n$ (with respect to the standard, unweighted inner product) such that $\mathbf{N}\mathbf{v}_i = \omega_i \mathbf{v}_i$. The matrix $\mathbf{N}$ may be factored as

$$\mathbf{N} = \mathbf{V}\Omega\mathbf{V}^T,$$

where $\Omega = \operatorname{diag}(\omega_0, \ldots, \omega_{n-1})$ and $\mathbf{V}$ is the unitary matrix whose ith column is $\mathbf{v}_i$. One may check that $\omega_0 = 0$ and that we may choose $\mathbf{v}_0 = \mathbf{d}^{1/2}/\|\mathbf{d}^{1/2}\|_2$, where $\mathbf{d}^{1/2}$ is defined componentwise. We note that since G is connected, we have $\omega_1 > 0$.

Our wavelet transforms will be constructed from the matrix $\mathbf{T}_g$ defined by

$$\mathbf{T}_g := \mathbf{V}g(\Omega)\mathbf{V}^T := \mathbf{V}\Lambda_g\mathbf{V}^T, \tag{1.1}$$

where $g : [0,2] \to [0,1]$ is some monotonically decreasing function such that $g(0) = 1$ and $g(2) = 0$ and $\Lambda_g := \operatorname{diag}(g(\omega_0), \ldots, g(\omega_{n-1})) := \operatorname{diag}(\lambda_0, \ldots, \lambda_{n-1})$. We note that by construction, we have $1 = \lambda_0 > \lambda_1 \geq \ldots \geq \lambda_{n-1} \geq 0$. As observed in, e.g., [16], in the case where g is a rational function, the matrix $\mathbf{T}_g$ can be constructed in the spatial domain via functional calculus, and there is no need to explicitly diagonalize $\mathbf{N}$, which may be computationally expensive. Indeed, this approach is used in many popular GNNs [9, 15]. When there is no potential for confusion, we will suppress dependence on g and write $\mathbf{T}$ and Λ in place of $\mathbf{T}_g$ and Λ_g. As our main example, we will set $g(t) = g_\star(t) := 1 - t/2$, yielding

$$\mathbf{T}_{g_\star} = \mathbf{I} - \frac{1}{2}\left(\mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\right) = \frac{1}{2}\left(\mathbf{I} + \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\right). \tag{1.2}$$

In [12], Gama, Ribeiro, and Bruna constructed a graph scattering transform using wavelets which are polynomials in $\mathbf{T}_{g_\star}$, and in [13], Gao, Wolf, and Hirn defined a closely related graph scattering transform from polynomials of the lazy random walk matrix

$$\mathbf{P} := \mathbf{D}^{1/2}\mathbf{T}_{g_\star}\mathbf{D}^{-1/2} = \frac{1}{2}\left(\mathbf{I} + \mathbf{A}\mathbf{D}^{-1}\right).$$

In order to unify and generalize these frameworks, we will let $\mathbf{M}$ be an $n \times n$ invertible matrix and let $\mathbf{K}$ be the matrix defined by

$$\mathbf{K} = \mathbf{K}_{g,\mathbf{M}} := \mathbf{M}^{-1}\mathbf{T}_g\mathbf{M}. \tag{1.3}$$

Note that $\mathbf{K}$ depends on the choice of both g and $\mathbf{M}$ and thus includes a very large family of matrices. However, we shall suppress this dependence in order to avoid cumbersome notation. As important special cases, we note that we may obtain $\mathbf{K} = \mathbf{T}$ by setting $\mathbf{M} = \mathbf{I}$, and we may obtain $\mathbf{P}$ and $\mathbf{P}^T$ by setting $g(t) = g_\star(t)$ and letting $\mathbf{M} = \mathbf{D}^{-1/2}$ and $\mathbf{M} = \mathbf{D}^{1/2}$, respectively. More generally, as we shall see below, the function g controls the eigenvalues of $\mathbf{K}$, while the matrix $\mathbf{M}$ influences the eigenvectors of $\mathbf{K}$. Since the eigenvalues and eigenvectors uniquely determine $\mathbf{K}$, they strongly affect any wavelets derived from $\mathbf{K}$. Thus, the two parameters g and $\mathbf{M}$ allow for a wide range of possible graph wavelet constructions.

In section 2, we will construct two wavelet transforms $\mathcal{W}_J^{(1)}$ and $\mathcal{W}_J^{(2)}$ from functions of $\mathbf{K}$. We note that the matrix $\mathbf{K}$ is not self-adjoint on the standard, unweighted inner product

space (except in the case $\mathbf{K} = \mathbf{T}$). Therefore, we will introduce a weighted inner product space, $\mathbf{L}^2(\mathbf{M})$, of signals defined on V with inner product defined by[1]

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{M}} := \langle \mathbf{M}\mathbf{x}, \mathbf{M}\mathbf{y} \rangle_2, \tag{1.4}$$

where $\langle \cdot, \cdot \rangle_2$ denotes the standard, unweighted inner product on $\mathbb{R}^n$. To better understand this definition, we note that if $\mathbf{M} = \mathbf{D}^{\alpha/2}$, then $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{D}^{\alpha/2}} = \sum_{i=0}^{n-1} \mathbf{x}(i)\mathbf{y}(i)\mathbf{d}(i)^{\alpha}$. Thus, $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{D}^{\alpha/2}}$ is a weighted $\mathbf{L}^2$ inner product with weights depending on the degree of each vertex. We note that the norms $\|\mathbf{x}\|_{\mathbf{M}}^2 := \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{M}}$ and $\|\mathbf{x}\|_2^2 = \langle \mathbf{x}, \mathbf{x} \rangle_2$ are equivalent and that $\frac{1}{\|\mathbf{M}^{-1}\|_2} \|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_{\mathbf{M}} \leq \|\mathbf{M}\|_2 \|\mathbf{x}\|_2$, where for a matrix $\mathbf{B}$, we shall let $\|\mathbf{B}\|_{\mathbf{M}}$ and $\|\mathbf{B}\|_2$ denote its operator norms on $\mathbf{L}^2(\mathbf{M})$ and on the standard, unweighted $\mathbf{L}^2$ space. If $\mathcal{I}$ is a countable indexing set and $\mathcal{X} = \{\mathbf{x}_i\}_{i \in \mathcal{I}}$ is a family of vectors, then we will also let $\|\mathcal{X}\|_{\mathbf{M}}$ denote the $\mathbf{M}\ell^2(\mathbf{L}^2(\mathbf{M}))$ norm of $\mathcal{X}$ and $\|\mathcal{X}\|_2$ denote its norm on the vector-valued $\mathbf{L}^2$ space. If $\Gamma = (\mathbf{B}_i)_{i \in \mathcal{I}}$ is a family of matrices, then we will say that Γ is a frame if there exist $0 < c \leq C < \infty$ such that

$$c\|\mathbf{x}\|_{\mathbf{M}}^2 \leq \|\mathcal{B}\mathbf{x}\|_{\mathbf{M}}^2 := \sum_{i \in \mathcal{I}} \|\mathbf{B}_i \mathbf{x}\|_{\mathbf{M}}^2 \leq C\|\mathbf{x}\|_{\mathbf{M}}^2 \quad \forall \mathbf{x} \in \mathbf{L}^2(\mathbf{M}). \tag{1.5}$$

The following lemma will be useful in studying the frame bounds of the wavelet transforms constructed from $\mathbf{K}$.

Lemma 1.1. *The matrix* $\mathbf{K}$ *is self-adjoint on* $\mathbf{L}^2(\mathbf{M})$.

Proof. Since $\mathbf{T}$ is symmetric, we may use (1.3) and (1.4) to see that

$$\langle \mathbf{K}\mathbf{x}, \mathbf{y} \rangle_{\mathbf{M}} = \langle \mathbf{M}\mathbf{x}, \mathbf{T}\mathbf{M}\mathbf{y} \rangle_2 = \langle \mathbf{M}\mathbf{x}, \mathbf{M}(\mathbf{M}^{-1}\mathbf{T}\mathbf{M})\mathbf{y} \rangle_2 = \langle \mathbf{M}\mathbf{x}, \mathbf{M}\mathbf{K}\mathbf{y} \rangle_2 = \langle \mathbf{x}, \mathbf{K}\mathbf{y} \rangle_{\mathbf{M}}. \qquad \blacksquare$$

It will frequently be useful to consider the eigenvector decompositions of $\mathbf{T}$ and $\mathbf{K}$. By definition, we have $\mathbf{T} = \mathbf{V}\Lambda\mathbf{V}^T$, and therefore $\mathbf{T}\mathbf{v}_i = \lambda_i \mathbf{v}_i$. Since the matrices $\mathbf{T}$ and $\mathbf{K}$ are similar with $\mathbf{K} = \mathbf{M}^{-1}\mathbf{T}\mathbf{M}$, one may use the definition of $\langle \cdot, \cdot \rangle_{\mathbf{M}}$ to verify that the vectors

$$\mathbf{u}_i := \mathbf{M}^{-1}\mathbf{v}_i$$

form an orthonormal eigenbasis for $\mathbf{L}^2(\mathbf{M})$ with $\mathbf{K}\mathbf{u}_i = \lambda_i \mathbf{u}_i$. One may also verify that $\mathbf{w}_i := \mathbf{M}^T \mathbf{v}_i$ is a left eigenvector of $\mathbf{K}$ and $\mathbf{w}_i^T \mathbf{K} = \lambda_i \mathbf{w}_i^T$ for all $0 \leq i \leq n-1$.

In the following section, we will construct wavelets from polynomials of $\mathbf{K}$, where for a polynomial $p(t) = a_k t^k + \ldots + a_1 t + a_0$, we define $p(\mathbf{B}) := a_k \mathbf{B}^k + \ldots + a_1 \mathbf{B} + a_0 \mathbf{I}$. The following lemma uses (1.1) to derive a formula for polynomials of $\mathbf{K}$ and $\mathbf{T}$ and relates the operator norms of polynomials of $\mathbf{K}$ to polynomials of $\mathbf{T}$. It will be useful for studying the wavelet transforms introduced in the following section. For a proof, see section SM1.

Lemma 1.2. *For any polynomial* p, *we have*

$$p(\mathbf{T}) = \mathbf{V}p(\Lambda)\mathbf{V}^T \quad \text{and} \quad p(\mathbf{K}) = \mathbf{M}^{-1}p(\mathbf{T})\mathbf{M} = \mathbf{M}^{-1}\mathbf{V}p(\Lambda)\mathbf{V}^T\mathbf{M}. \tag{1.6}$$

Consequently, for all $\mathbf{x} \in \mathbf{L}^2(\mathbf{M})$, *we have* $\|p(\mathbf{K})\mathbf{x}\|_{\mathbf{M}} = \|p(\mathbf{T})\mathbf{M}\mathbf{x}\|_2$.

[1] To avoid confusion, we note that our definition differs from the notation sometimes used in the literature where $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{M}}$ is defined by $\mathbf{y}^T\mathbf{M}\mathbf{x}$.

In light of Lemma 1.2, for any polynomial p, we may define $p(\mathbf{T})^{1/2}$ and $p(\mathbf{K})^{1/2}$ by

$$p(\mathbf{T})^{1/2} := \mathbf{V}p(\Lambda)^{1/2}\mathbf{V}^T \quad \text{and} \quad p(\mathbf{K})^{1/2} = \mathbf{M}^{-1}\mathbf{V}p(\Lambda)^{1/2}\mathbf{V}^T\mathbf{M}, \tag{1.7}$$

where the square root of the diagaonal matrix $p(\Lambda)$ is defined entrywise. We may readily verify that $p(\mathbf{T})^{1/2}p(\mathbf{T})^{1/2} = p(\mathbf{T})$ and $p(\mathbf{K})^{1/2}p(\mathbf{K})^{1/2} = p(\mathbf{K})$.

1.2. Previous work on graph scattering transforms. Several previous works have introduced different formulations of the graph scattering transform. [12, 11, 32] construct the scattering transform using symmetric wavelets and show their constructions have similar stability and invariance properties to the Euclidean scattering transform. There have also been works empirically demonstrating that the graph scattering transform is effective for tasks such as graph classification [13], vertex classification [27, 20, 21], graph synthesis [31, 4], and combinatorial optimization [19]. However, much of this empirical work has been done using asymmetric wavelets, to which the theoretical guarantees of [12, 11] and [32] do not apply.

In this paper, we will focus on unifying and generalizing the theoretical properties of the different formulations of the graph scattering transform. Analogously to the Euclidean scattering transform, we will show that the windowed graph scattering transform is (i) nonexpansive on $\mathbf{L}^2(\mathbf{M})$, (ii) invariant to permutations of the vertices up to a factor depending on the scale of the low-pass (for certain choices of $\mathbf{K}$), and (iii) stable to graph perturbations. Similarly, we will show that the nonwindowed scattering transform is (i) Lipschitz continuous on $\mathbf{L}^2(\mathbf{M})$, (ii) fully invariant to permutations, and (iii) stable to graph perturbations. Importantly, we note that this is the first work to produce such theoretical guarantees for graph scattering transforms using asymmetric wavelets and is also the first to establish Lipschitz continuity of the nonwindowed graph scattering transform.

In [32], the authors construct a family of wavelet convolutions using the spectral decomposition of the unnormalized graph Laplacian and define a windowed scattering transform as an iterative series of wavelet convolutions and nonlinearities. They then prove results analogous to Theorems 3.2, 3.5, and 3.8 of this paper for their windowed scattering transform. They also introduce a notion of stability to graph perturbations. However, their notion of graph perturbations is significantly different from the one we consider in section 4.

In [12], the authors construct a family of wavelets from polynomials of $\mathbf{T}_g$, in the case where $g(t) = g_\star(t) = 1 - t/2$ and showed that the resulting nonwindowed scattering transform was stable to graph perturbations. This construction was generalized in [11], where the authors introduced a more general class of graph convolutions, constructed from a class of symmetric matrices known as "graph shift operators." The wavelet transform considered in [12] is nearly identical to the $\mathcal{W}_J^{(2)}$ introduced in section 2 in the special case where $g(t) = g_\star(t)$ and $\mathbf{M} = \mathbf{I}$, with the only difference being that our wavelet transform includes a low-pass filter.

In [13], wavelets were constructed from the lazy random walk matrix $\mathbf{P} = \mathbf{D}^{1/2}\mathbf{T}\mathbf{D}^{-1/2}$. These wavelets are essentially the same as the $\mathcal{W}_J^{(2)}$ in the case where $g(t) = g_\star(t)$ and $\mathbf{M} = \mathbf{D}^{-1/2}$, although similarly to [12], the wavelets in [13] do not use a low-pass filter. In all of these previous works, the authors carry out substantial numerical experiments and demonstrate that scattering transforms are effective for a variety of graph deep learning tasks. We also note that Chen, Cheng, and Mallat first introduced a substantially different version of the graph scattering transform in [5, 6] using Haar wavelets. However, the construction and analysis considered there differ substantially from the previously discussed works.

In addition to helping us understand the stability and invariance properties of deep networks on graphs, the scattering transform also helps us investigate the important question, "what sort of filters should be used in a GNN?" Most popular GNNs, such as [15], typically average information over one-hop neighborhoods in order to produce a smooth hidden representation of the vertices, which effectively corresponds to filtering out high-frequency information via a low-pass filter. Scattering, on the other hand, uses multiscale filters that can encode long-range dependencies and effectively capture high-frequency information. In [20, 27], it was shown that this allows for improved numerical performance in situations where high-frequency information is important. Moreover, graph scattering–style networks also been shown to be effective for molecule generation [31] and solving combinatorial optimization problems [19]. In the former case, the large receptive field of the wavelets allows scattering to capture the global structure of the molecule, and in the latter case, the use of wavelets rather than low-pass filters allows one to distinguish a member of the clique from a node which is connected to to most but not all nodes within the clique.

Here we shall focus on unifying and generalizing the theory of several of these previous constructions. Our introduction of the matrix $\mathbf{M}$ allows us to obtain wavelets very similar to either [12] or [13] as special cases. Moreover, the introduction of the tight wavelet frame $\mathcal{W}_J^{(1)}$ allows us to produce a network with provable conservation of energy and nonexpansive properties analogous to [32]. To highlight the generality of our setup, we introduce both windowed and nonwindowed versions of the scattering transform using general (wavelet) frames and provide a detailed theoretical analysis of both.

1.3. Organization, contributions, and summary of main results. In section 2, we will construct two families of graph wavelets $\mathcal{W}_J^{(1)}$ and $\mathcal{W}_J^{(2)}$. In section 3, we will introduce windowed and nonwindowed versions of the graph scattering transform and analyze their continuity and invariance properties. Then in section 4, we will analyze the stability of the networks to perturbations. In section 5, we will discuss the relationship between scattering and other GNNs, and in section 6, we will present numerical experiments before offering a brief conclusion in section 7.

The following is a summary of our main theoretical results. Unless otherwise stated, all results apply to both choices of wavelets, both the windowed and the nonwindowed scattering transform, and to general diffusion matrices $\mathbf{K}$.

- **Section 2:** Propositions 2.1 and 2.2 show that the wavelets $\mathcal{W}_J^{(1)}$ are an isometry and $\mathcal{W}_J^{(2)}$ are a nonexpansive frame on $\mathbf{L}^2(\mathbf{M})$.
- **Section 3:** Proposition 3.1 shows that if $\boldsymbol{\mu} = \mathbf{u}_0$, then the windowed scattering transform converges to the nonwindowed scattering transform as $J \to \infty$. Theorem 3.2 shows that the windowed scattering transform is nonexpansive and that the nonwindowed scattering transform is Lipschitz continuous on $\mathbf{L}^2(\mathbf{M})$. Theorem 3.4 shows that the energy in the mth order scattering coefficients decays exponentially in m, and Theorem 3.5 shows that therefore the scattering transform preserves all of the energy of the input signal if we use the wavelets $\mathcal{W}_J^{(1)}$. Theorems 3.6 and 3.7 show that the windowed scattering transform is equivariant and that the nonwindowed scattering transform is invariant to permutations. Finally, Theorem 3.8 shows that the windowed scattering transform is invariant in the limit at $J \to \infty$ if $\mathbf{K} = \mathbf{P}^T$.

- **Section 4:** Theorems 4.1 and 4.2 provide stability guarantees for $\mathcal{W}^{(1)}$ and $\mathcal{W}^{(2)}$ in the special case where $\mathbf{K} = \mathbf{T}$. Theorem 4.3 then allows these results to be extended to general $\mathbf{K}$. Finally, Theorems 4.7 and 4.11 establish stability of the windowed and nonwindowed scattering transforms.

1.3.1. Contributions. As discussed in section 1.2, there has been a significant amount of work developing, analyzing, and applying different versions of the graph scattering transform. Therefore, in this section, for the sake of clarity, we will highlight several aspects of our paper which are different than these previous works.

- The wavelet family $\mathcal{W}_J^{(2)}$ includes the wavelets of [12] and [13] as special cases, but it also includes many other wavelets, including, in particular, a one-family parameter of wavelets based on diffusion operators of the form $\mathbf{K} = \mathbf{D}^{\alpha}\mathbf{T}\mathbf{D}^{-\alpha}$, where the wavelets from [12] and [13] correspond to $\alpha = 0$ and $\alpha = -.5$. Moreover, the wavelets $\mathcal{W}_J^{(1)}$ are new and are not utilized in any previous version of the graph scattering transform. In our experiments in section 6, we consider scattering transforms which use wavelets $\mathcal{W}^{(\beta)}$, based on diffusion matrices $\mathbf{K} = \mathbf{D}^{-\alpha}\mathbf{T}\mathbf{D}^{\alpha}$, for $\beta = 1, 2$ and $\alpha = -0.5, -0.25, 0, 0.25, 0.5$. We find that the optimal choice of wavelet varies significantly from one data set to another. Therefore, it is our recommendation that practitioners use a validation procedure to select the optimal wavelets for a given task. Moreover, we note that the new wavelets $\mathcal{W}_J^{(1)}$ outperform $\mathcal{W}_J^{(2)}$ on most data sets and that the intermediate values of α, i.e., $\alpha = \pm 0.25$, often deliver superior performance to the values considered in previous work ($\alpha = 0$ or -0.5).
- [12] and [13] considered only nonwindowed versions of the scattering transform, where we consider both windowed and nonwindowed versions. Importantly, unlike its nonwindowed counterpart, the windowed scattering transform can be applied to vertex-level tasks, such as node classification or combinatorial optimization problems. Additionally, utilizing both versions of the scattering transform is crucial to proving that the nonwindowed scattering transform is Lipschitz continuous in Theorem 3.2 that and no analogous result for the nonwindowed scattering transform exists in previous work.
- Our stability result for the nonwindowed scattering transform, Theorem 4.11, considers perturbations both to the graph structure and to the input signal $\mathbf{x}$, whereas the analogous result in [12] only considered perturbations to the graph structure. The proof of this result directly utilizes the Lipschitz continuity of the nonwindowed scattering transform discussed above and therefore would have been nontrivial for the authors of [12] to prove since they did not consider a windowed scattering transform.

2. The graph wavelet transform.

In this section, we will construct two families of graph wavelet transforms based off of the matrix $\mathbf{K} = \mathbf{M}^{-1}\mathbf{T}\mathbf{M}$ introduced in section 1.1. In the following sections, we provide a theoretical analysis of the scattering transforms constructed from each of these wavelet transforms.

Let $J \geq 0$. For $0 \leq j \leq J+1$, let p_j be the polynomial defined by

$$p_j(t) := \begin{cases} 1 - t & \text{if } j = 0, \\ t^{2^{j-1}} - t^{2^j} & \text{if } 1 \leq j \leq J \\ t^{2^J} & \text{if } j = J+1, \end{cases},$$

and let $q_j(t) := p_j(t)^{1/2}$ for $0 \le t \le 1$. We note that by construction,

$$\sum_{j=0}^{J+1} p_j(t) = \sum_{j=0}^{J+1} q_j(t)^2 = 1 \text{ for all } 0 \le t \le 1. \tag{2.1}$$

Given these functions, we define two wavelet transforms by

$$\mathcal{W}_J^{(1)} := \left\{\Psi_j^{(1)}, \Phi_J^{(1)}\right\}_{0\le j\le J}, \quad \Psi_j^{(1)} := q_j(\mathbf{K}), \quad \Phi_J^{(1)} := q_{J+1}(\mathbf{K}) \quad \text{and}$$
$$\mathcal{W}_J^{(2)} := \left\{\Psi_j^{(2)}, \Phi_J^{(2)}\right\}_{0\le j\le J}, \quad \Psi_j^{(2)} := p_j(\mathbf{K}), \quad \Phi_J^{(2)} := p_{J+1}(\mathbf{K}),$$

where $q_j(\mathbf{K})$ is defined as in (1.7). The next two propositions show that $\mathcal{W}_J^{(1)}$ is an isometry and that $\mathcal{W}_J^{(2)}$ is a nonexpansive frame on $\mathbf{L}^2(\mathbf{M})$ (defined in (1.4)). We provide proofs in section SM2.

Proposition 2.1. *$\mathcal{W}_J^{(1)}$ is an isometry from $\mathbf{L}^2(\mathbf{M})$ to $\mathbf{M}\ell^2(\mathbf{L}^2(\mathbf{M}))$. That is,*

$$\|\mathcal{W}_J^{(1)}\mathbf{x}\|_{\mathbf{M}}^2 := \sum_{j=0}^{J} \|\Psi_j^{(1)}\mathbf{x}\|_{\mathbf{M}}^2 + \|\Phi_J^{(1)}\mathbf{x}\|_{\mathbf{M}}^2 = \|\mathbf{x}\|_{\mathbf{M}}^2 \quad \textit{for all } \mathbf{x} \in \mathbf{L}^2(\mathbf{M}).$$

Proposition 2.2. *$\mathcal{W}_J^{(2)}$ is a nonexpansive frame; i.e., there exists a universal constant $c > 0$, which in particular is independent of $\mathbf{M}$, J, or the eignenvalues of $\mathbf{T}$, such that*

$$c\|\mathbf{x}\|_{\mathbf{M}}^2 \le \|\mathcal{W}_J^{(2)}\mathbf{x}\|_{\mathbf{M}}^2 := \sum_{j=0}^{J} \|\Psi_j^{(2)}\mathbf{x}\|_{\mathbf{M}}^2 + \|\Phi_J^{(2)}\mathbf{x}\|_{\mathbf{M}}^2 \le \|\mathbf{x}\|_{\mathbf{M}}^2 \quad \textit{for all } \mathbf{x} \in \mathbf{L}^2(\mathbf{M}).$$

Remark 2.3. If we omit the low-pass operator $\Phi_J^{(2)}$, we can repeat the arguments of Proposition 4.1 of [12] to show that $\{\Psi_j^{(2)}\}_{0\le j\le J}$ is a nonexpansive frame, with a lower bound depending on the geometry of G, when restricted to $\mathbf{x}$ such that $\langle \mathbf{x}, \mathbf{u}_0\rangle_{\mathbf{M}} = 0$. For certain tasks, this may be advantageous. For example, if $\mathbf{K} = \mathbf{P}^T$, we may check that $\mathbf{u}_0$ is a constant vector and that $\Psi_j\mathbf{u}_0 = p_j(\mathbf{K})\mathbf{u}_0 = 0$. Therefore, these restricted wavelets could be used to produce a representation of an input $\mathbf{x}$ which is invariant to the addition of a constant vector.

3. The geometric scattering transform.

In this section, we will construct the scattering transform as a multilayered architecture built off of a frame $\mathcal{W}$, such as the wavelet transforms $\mathcal{W}_J^{(1)}$ and $\mathcal{W}_J^{(2)}$ introduced in section 2. We shall see that the scattering transform is a continuous operator on $\mathbf{L}^2(\mathbf{M})$ whenever $\mathcal{W}$ is nonexpansive. We shall also see that it has desirable conservation of energy bounds when $\mathcal{W} = \mathcal{W}_J^{(1)}$ due to the fact that $\mathcal{W}_J^{(1)}$ is an isometry. On the other hand, we shall see in the following section that the scattering transform has stronger stability guarantees, which are independent of the graph size, when $\mathcal{W} = \mathcal{W}_J^{(2)}$.

3.1. Definitions.

We generalize the wavelet frames $\mathcal{W}_J^{(1)}$ and $\mathcal{W}_J^{(2)}$ defined in section 2. Let $\mathcal{J}$ be any countable indexing set, and let $\mathcal{W} := \{\Psi_j, \Phi\}_{j\in\mathcal{J}}$ be a frame on $\mathbf{L}^2(\mathbf{M})$ with

$$c\|\mathbf{x}\|_{\mathbf{M}}^2 \le \|\mathcal{W}\mathbf{x}\|_{\mathbf{M}}^2 := \sum_{j\in\mathcal{J}} \|\Psi_j\mathbf{x}\|_{\mathbf{M}}^2 + \|\Phi\mathbf{x}\|_{\mathbf{M}}^2 \le C\|\mathbf{x}\|_{\mathbf{M}}^2$$

for some $0 < c \leq C < \infty$. In this paper, we are primarily interested in the case where $\mathcal{J} = \{0, \dots, J\}$ and $\mathcal{W}$ is either $\mathcal{W}_J^{(1)}$ or $\mathcal{W}_J^{(2)}$, meaning that $\Psi_j = \Psi_j^{(i)}$ and $\Phi = \Phi_J^{(i)}$ for $i = 1, 2$, respectively. If a result is specific to these cases, we will write "let $\mathcal{W}$ be either $\mathcal{W}_J^{(1)}$ or $\mathcal{W}_J^{(2)}$," in which case we are implicitly assuming that $\mathcal{J} = \{0, \dots, J\}$. In general, it will be useful to think of the matrices Ψ_j as wavelets and Φ as a low-pass filter, but we emphasize that this specification is not required for many of the results that follow. Indeed, we will define the geometric scattering transform for generic frames in order to highlight the relationship between properties of the scattering transform and properties of the underlying frame.

We let M be the pointwise modulus operator on $M\mathbf{x} := (|\mathbf{x}(0)|, \dots, |\mathbf{x}(n-1)|)$ and let $\mathbf{U}[j]\mathbf{x} := M\Psi_j\mathbf{x}$ for $j \in \mathcal{J}$. We view this transformation $\mathbf{U}[j]$ as a hidden layer of our network and will construct a multilayered architecture by iteratively applying this transformation at different values of j. Formally, for $m > 0$, let $\mathcal{J}^m$ denote the m-fold Cartesian product of $\mathcal{J}$ with itself, and for an index path $\mathbf{p} = (j_1, \dots, j_m) \in \mathcal{J}^m$, let

$$\mathbf{U}[\mathbf{p}]\mathbf{x} := \mathbf{U}[j_m] \dots \mathbf{U}[j_1]\mathbf{x} = M\Psi_{j_m} \cdots M\Psi_{j_1}\mathbf{x}.$$

For $m = 0$, we declare that $\mathcal{J}^0$ is the empty set and interpret $\mathbf{U}[\mathbf{p}_e]\mathbf{x} = \mathbf{x}$ when $\mathbf{p}_e$ is the "empty index." Next, we define the windowed scattering coefficients by

$$\mathbf{S}[\mathbf{p}]\mathbf{x} := \Phi\mathbf{U}[\mathbf{p}]\mathbf{x}.$$

In this definition, the final multiplication by the (low-pass) matrix Φ is interpreted as a local averaging, but one could choose a different type of matrix Φ so long as the frame condition still holds. We will also define nonwindowed scattering coefficients which replace this local averaging by a weighted global averaging. Specifically, we let $\boldsymbol{\mu} \in \mathbf{L}^2(\mathbf{M})$ be a weighting vector and define the nonwindowed scattering coefficients by

$$\overline{\mathbf{S}}[\mathbf{p}]\mathbf{x} := \langle \boldsymbol{\mu}, \mathbf{U}[\mathbf{p}]\mathbf{x} \rangle_{\mathbf{M}}.$$

One natural choice is $\boldsymbol{\mu} = \left(\mathbf{M}^T\mathbf{M}\right)^{-1}\mathbf{1}$, where $\mathbf{1}$ is the vector of all ones. In this case, one may verify that $\overline{\mathbf{S}}[\mathbf{p}]\mathbf{x} = \|\mathbf{U}[\mathbf{p}]\mathbf{x}\|_1$, and we recover a setup similar to [13]. Another natural choice is $\boldsymbol{\mu} = \mathbf{u}_0 = \mathbf{M}^{-1}\mathbf{v}_0$, in which case we recover a setup similar to [12] if we set $\mathbf{M} = \mathbf{I}$.

Given these coefficients, we define an operator $\mathbf{U} : \mathbf{L}^2(\mathbf{M}) \to \mathbf{M}\ell^2(\mathbf{L}^2(\mathbf{M}))$ by

$$\mathbf{U}\mathbf{x} := \{\mathbf{U}[\mathbf{p}]\mathbf{x} : m \geq 0, \mathbf{p} = (j_1, \dots, j_m) \in \mathcal{J}^m\}.$$

Next, we define the windowed and nonwindowed scattering transform $\mathbf{S} : \mathbf{L}^2(\mathbf{M}) \to \mathbf{M}\ell^2(\mathbf{L}^2(\mathbf{M}))$ and $\overline{\mathbf{S}} : \mathbf{L}^2(\mathbf{M}) \to \mathbf{M}\ell^2$ by

$$\mathbf{S}\mathbf{x} := \{\mathbf{S}[\mathbf{p}]\mathbf{x} : m \geq 0, \mathbf{p} \in \mathcal{J}^m\} \quad \text{and} \quad \overline{\mathbf{S}}\mathbf{x} := \{\overline{\mathbf{S}}[\mathbf{p}]\mathbf{x} : m \geq 0, \mathbf{p} \in \mathcal{J}^m\}.$$

When $\mathcal{W}$ is either of the graph wavelet transforms $\mathcal{W}_J^{(1)}$ or $\mathcal{W}_J^{(2)}$ constructed in section 2 and $\mathcal{J} = \{0, \dots, J\}$, we may write $\mathbf{S}_J$ in place of $\mathbf{S}$ if we want to emphasize the dependence on J. Similarly, when we want to emphasize the dependence on $\boldsymbol{\mu}$ we will write $\overline{\mathbf{S}}_{\boldsymbol{\mu}}$ in place of $\overline{\mathbf{S}}$. We will let $\mathbf{S}^\ell$ and $\overline{\mathbf{S}}^\ell$ denote the ℓth layer of the windowed and nonwindowed scattering transforms

$$\mathbf{S}^{(\ell)}\mathbf{x} := \{\mathbf{S}[\mathbf{p}]\mathbf{x} : \mathbf{p} = (j_1, \dots, j_\ell) \in \mathcal{J}^\ell\} \quad \text{and} \quad \overline{\mathbf{S}}^{(\ell)}\mathbf{x} := \{\overline{\mathbf{S}}[\mathbf{p}]\mathbf{x} : \mathbf{p} = (j_1, \dots, j_\ell) \in \mathcal{J}^\ell\}.$$

When $\mathcal{W}$ is either of the wavelet transforms $\mathcal{W}_J^{(1)}$ or $\mathcal{W}_J^{(2)}$ constructed in section 2 and $\boldsymbol{\mu} = \mathbf{u}_0$, we may view the nonwindowed scattering transform as the limit of the windowed scattering transform as $J \to \infty$. Formally, we can prove the following proposition.

Proposition 3.1. *Let $\mathcal{W}$ be either of the wavelet transforms $\mathcal{W}_J^{(1)}$ or $\mathcal{W}_J^{(2)}$ constructed in section 2, and let $\boldsymbol{\mu} = \mathbf{u}_0$. Then for all paths $\mathbf{p} \in \mathcal{J}^m$ and all $\mathbf{x} \in \mathbf{L}^2(\mathbf{M})$,*

$$\lim_{J\to\infty} \|\mathbf{S}_J[\mathbf{p}]\mathbf{x} - (\overline{\mathbf{S}}_{\mathbf{u}_0}[\mathbf{p}]\mathbf{x})\mathbf{u}_0\|_{\mathbf{M}} = 0,$$

where, on the left-hand side, we assume that J is large enough such that $\mathbf{S}_J[\mathbf{p}]$ is well-defined.

For a proof of Proposition 3.1, see section SM3.

3.2. Continuity and conservation of energy properties. The following theorem shows that the windowed scattering transform $\mathbf{S}$ is nonexpansive and that the nonwindowed scattering transform $\overline{\mathbf{S}}$ is Lipschitz continuous whenever the underlying frame $\mathcal{W}$ is nonexpansive.

Theorem 3.2 (nonexpansiveness). *If $\mathcal{W}$ is a frame with $C \leq 1$ in (1.5), then*

$$\|\mathbf{S}\mathbf{x} - \mathbf{S}\mathbf{y}\|_{\mathbf{M}} \leq \|\mathbf{x} - \mathbf{y}\|_{\mathbf{M}} \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbf{L}^2(\mathbf{M}). \tag{3.1}$$

Furthermore, if $\mathcal{W}$ is either of the wavelet transforms $\mathcal{W}_J^{(1)}$ or $\mathcal{W}_J^{(2)}$ constructed in section 2, $\boldsymbol{\mu} = \mathbf{u}_0$, and $\min_i |\mathbf{u}_0(i)| > 0$, then we have

$$\|\overline{\mathbf{S}}_{\mathbf{u}_0}\mathbf{x} - \overline{\mathbf{S}}_{\mathbf{u}_0}\mathbf{y}\|_2 \leq \frac{1}{\sqrt{n}} \frac{\|\mathbf{M}^{-1}\|_2}{\min_i |\mathbf{u}_0(i)|} \|\mathbf{x} - \mathbf{y}\|_{\mathbf{M}}. \tag{3.2}$$

The proof of (3.1) is similar to analogous results in, e.g., [18] and [32]. Equation (3.2) is proved by using Proposition 3.1 to view $\overline{\mathbf{S}}$ as the rescaled limit of $\mathbf{S}_J$ as $J \to \infty$ and applying Fatou's lemma. We note that the scaling factor of $\frac{1}{\sqrt{n}}$ is a consequence of the fact that we have assumed $\mathbf{u}_0$ to have unit norm on $\mathbf{L}^2(\mathbf{M})$. A full proof is provided in section SM5.

Remark 3.3. The proof of (3.2) relies on (3.1), Proposition 3.1, and Fatou's lemma. Therefore, we are not able to establish this inequality for general $\boldsymbol{\mu}$ and $\mathcal{W}$, which do not satisfy the assumptions of Proposition 3.1. However, if Φ is invertible, one can use the relationship $\mathbf{U}\mathbf{x} = \Phi^{-1}\mathbf{S}\mathbf{x}$ to show that $\overline{\mathbf{S}}_{\boldsymbol{\mu}}$ is still Lipschitz continuous since

$$\|\overline{\mathbf{S}}_{\boldsymbol{\mu}}\mathbf{x} - \overline{\mathbf{S}}_{\boldsymbol{\mu}}\mathbf{y}\|_2 \leq \|\boldsymbol{\mu}\|_{\mathbf{M}} \|\Phi^{-1}\|_{\mathbf{M}} \|\mathbf{S}\mathbf{x} - \mathbf{S}\mathbf{y}\|_{\mathbf{M}} \leq \|\boldsymbol{\mu}\|_{\mathbf{M}} \|\Phi^{-1}\|_{\mathbf{M}} \|\mathbf{x} - \mathbf{y}\|_{\mathbf{M}}.$$

The next theorem shows that if $\mathcal{W}$ is either of the wavelet transforms constructed in section 2, then $\mathbf{U}$ experiences rapid energy decay. This implies that it is possible to obtain a good representation of an input signal $\mathbf{x}$ using only a few layers. Our arguments use ideas similar to the proof of Proposition 3.3 in [32], with minor modifications to account for the fact that our wavelet constructions are different. See section SM6 for a complete proof.

Theorem 3.4 (energy decay). *Let $\mathcal{W}$ be either of the wavelet transforms $\mathcal{W}_J^{(1)}$ or $\mathcal{W}_J^{(2)}$ constructed in section 2. Then for all $\mathbf{x} \in \mathbf{L}^2(\mathbf{M})$ and all $m \geq 1$,*

$$\sum_{\mathbf{p}\in\mathcal{J}^{m+1}} \|\mathbf{U}[\mathbf{p}]\mathbf{x}\|_{\mathbf{M}}^2 \leq \left(1 - \frac{\mathbf{d}_{\min}}{\|\mathbf{d}\|_1}\right) \sum_{\mathbf{p}\in\mathcal{J}^m} \|\mathbf{U}[\mathbf{p}]\mathbf{x}\|_{\mathbf{M}}^2. \tag{3.3}$$

Therefore, for all $m \geq 0$,

$$\sum_{\mathbf{p} \in \mathcal{J}^{m+1}} \|\mathbf{U}[\mathbf{p}]\mathbf{x}\|_{\mathbf{M}}^2 \leq \left(1 - \frac{\mathbf{d}_{\min}}{\|\mathbf{d}\|_1}\right)^m \|\mathbf{x}\|_{\mathbf{M}}^2. \tag{3.4}$$

The next theorem shows that if $\mathcal{W} = \mathcal{W}_J^{(1)}$, then the windowed graph scattering transform conserves energy on $\mathbf{L}^2(\mathbf{M})$. Its proof is nearly identical to the proof of Theorem 3.1 in [32]. However, we give a full proof in section SM7 for the sake of completeness.

Theorem 3.5 (energy conservation). *If* $\mathcal{W} = \mathcal{W}_J^{(1)}$, *then*

$$\|\mathbf{S}_J\mathbf{x}\|_{\mathbf{M}} = \|\mathbf{x}\|_{\mathbf{M}} \quad \textit{for all } \mathbf{x} \in \mathbf{L}^2(\mathbf{M}).$$

3.3. Permutation invariance and equivariance. In tasks such as graph classification, two graphs are equivalent if one is a permutation of the other. In this section, we will show that both $\mathbf{U}$ and the windowed graph scattering transform are equivariant with respect to permutations. As a consequence, we will show that the nonwindowed scattering transform is fully permutation invariant and that the windowed scattering transform, under certain assumptions, is permutation invariant up to a factor depending on the scale of the low-pass filter.

Let S_n denote the permutation group on n elements, and, for $\Pi \in S_n$, let $\Pi(G)$ be the graph obtained by permuting the vertices of G. If $G' = \Pi(G)$, we define $\mathbf{M}'$ by $\mathbf{M}' := \Pi\mathbf{M}\Pi^T$. To motivate this definition, we note that if $\mathbf{M}$ is the identity, then $\mathbf{M}'$ is also the identity. Additionally, in the case where $\mathbf{M}$ is the square-root degree matrix $\mathbf{D}^{1/2}$, we note that the square-root degree matrix on G' is given by $(\mathbf{D}')^{1/2} = \Pi\mathbf{D}^{1/2}\Pi^T$, and a similar formula holds when $\mathbf{M} = \mathbf{D}^{-1/2}$. Therefore, in these three cases (which correspond to $\mathbf{K} = \mathbf{T}, \mathbf{P}^T$, and $\mathbf{P}$, respectively, if $g = g_\star$), we may view $\mathbf{M}'$ as the analogue of $\mathbf{M}$ associated to G'. Similarly, we consider the analogues of $\mathcal{W}$ and $\boldsymbol{\mu}$ on G', given by

$$\mathcal{W}' := \Pi\mathcal{W}\Pi^T := \{\Pi\Psi_j\Pi^T, \Pi\Phi\Pi^T\}_{j \in \mathcal{J}} \quad \text{and} \quad \boldsymbol{\mu}' := \Pi\boldsymbol{\mu}. \tag{3.5}$$

We also let $\mathbf{U}', \mathbf{S}'$, and $\overline{\mathbf{S}}'$ denote analogues of $\mathbf{U}, \mathbf{S}$, and $\overline{\mathbf{S}}$ on G' constructed from $\mathcal{W}'$ and $\boldsymbol{\mu}'$.

To further understand the definition of $\mathcal{W}'$, we note that the natural analogue of $\mathbf{T}$ on G' is given by $\mathbf{T}' := \Pi\mathbf{T}\Pi^T$. Therefore, Lemma 1.2 implies that for any polynomial p,

$$\begin{aligned} p\left((\mathbf{M}')^{-1}\mathbf{T}'\mathbf{M}'\right) = (\mathbf{M}')^{-1}p\left(\mathbf{T}'\right)\mathbf{M}' &= \left(\Pi\mathbf{M}\Pi^T\right)^{-1} p\left(\Pi\mathbf{T}\Pi^T\right)\left(\Pi\mathbf{M}\Pi^T\right) \\ &= \Pi\mathbf{M}^{-1}p(\mathbf{T})\mathbf{M}\Pi^T \\ &= \Pi p\left(\mathbf{M}^{-1}\mathbf{T}\mathbf{M}\right)\Pi^T \end{aligned}$$

with a similar formula holding for $q := p^{1/2}$. Thus, if $\mathcal{W}$ is either of the wavelet transforms $\mathcal{W}_J^{(1)}$ or $\mathcal{W}_J^{(2)}$, then $\mathcal{W}'$ is the analogous wavelet transform constructed from $\mathbf{K}' := (\mathbf{M}')^{-1}\mathbf{T}'\mathbf{M}'$. Given our definitions, it is now straightforward to prove the following equivariance theorem.

Theorem 3.6 (equivariance). *Let $\Pi \in S_n$ be a permutation, let $G' = \Pi(G)$, and let $\mathcal{W}'$ be the wavelet transform on G' defined as in (3.5). Then for all $\mathbf{x} \in \mathbf{L}^2(\mathbf{M})$,*

$$\mathbf{U}'\Pi\mathbf{x} = \Pi\mathbf{U}\mathbf{x} \quad \textit{and} \quad \mathbf{S}'\Pi\mathbf{x} = \Pi\mathbf{S}\mathbf{x}.$$

The following result shows that the nonwindowed scattering transform is permuation invariant.

Theorem 3.7 (invariance for the nonwindowed scattering transform). *Let $\Pi \in S_n$ be a permutation, $G' = \Pi(G)$, and let $\mathcal{W}'$ be the wavelet transform on G' defined as in (3.5). Then*

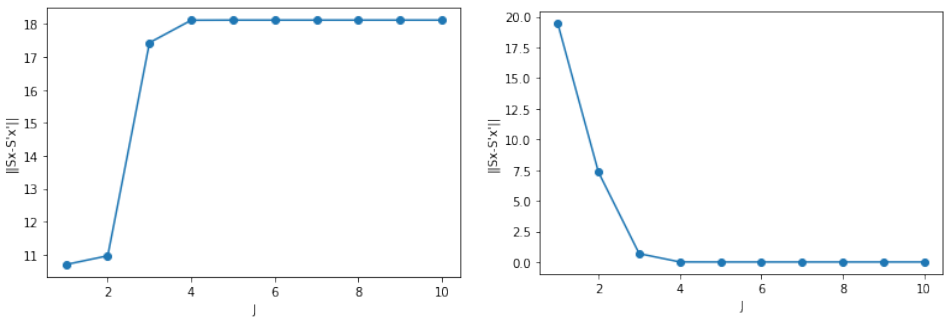
$$\overline{\mathbf{S}}'\Pi\mathbf{x} = \overline{\mathbf{S}}\mathbf{x} \quad \textit{for all } \mathbf{x} \in \mathbf{L}^2(\mathbf{M}).$$

For proofs of Theorems 3.6 and 3.7, see section SM8. Next, we will use Theorem 3.6 to show that if $\mathcal{W}$ is either $\mathcal{W}_J^{(1)}$ or $\mathcal{W}_J^{(2)}$ and $\mathbf{M} = \mathbf{D}^{1/2}$, then the windowed scattering transform is invariant on $\mathbf{L}^2(\mathbf{D}^{1/2})$ up to a factor depending on the scale of the low-pass filter. We note that $0 < \lambda_1 < 1$. Therefore, λ_1^t decays exponentially fast as $t \to \infty$, and so if J is large, the right-hand side of (3.6) below will be nearly zero. We also recall that if our spectral function is given by $g(t) = g_\star(t)$, then this choice of $\mathbf{M}$ will imply that $\mathbf{K} = \mathbf{P}^T$.

Theorem 3.8 (invariance for the windowed scattering transform). *Let $\mathcal{W}$ be either $\mathcal{W}_J^{(1)}$ or $\mathcal{W}_J^{(2)}$, let $\Pi \in S_n$, $G' = \Pi(G)$, and let $\mathcal{W}'$ be the wavelet transform on G' defined as in (3.5). Assume that $\mathbf{M} = \mathbf{D}^{1/2}$, so $\mathbf{K} = \mathbf{P}^T$. Let $t = 2^{J-1}$ if $\mathcal{W} = \mathcal{W}_J^{(1)}$ and $t = 2^J$ if $\mathcal{W} = \mathcal{W}^{(2)}$. Then*

$$\|\mathbf{S}'_J\Pi\mathbf{x} - \mathbf{S}_J\mathbf{x}\|_{\mathbf{D}^{1/2}} \leq \lambda_1^t \|\Pi - \mathbf{I}\|_{\mathbf{D}^{1/2}} \left(1 + \frac{\|\mathbf{d}\|_1}{\mathbf{d}_{\min}}\right)^{1/2} \|\mathbf{x}\|_{\mathbf{D}^{1/2}} \textit{ for all } \mathbf{x} \in \mathbf{L}^2(\mathbf{D}^{1/2}). \tag{3.6}$$

For a proof of Theorem 3.8, see section SM9. We note that while the term on the right-hand side of (3.6) does depend on the permutation Π, one may use the triangle inequlity to bound it uniformly by $\|\Pi - \mathbf{I}\|_{\mathbf{D}^{1/2}} \leq 1 + \sqrt{\frac{\|\mathbf{d}\|_\infty}{\mathbf{d}_{\min}}}$. In particular, this implies that the right-hand side converges to zero as $J \to \infty$. It is also important to note, as illustrated by Figure 1,



(a) $\mathbf{K} = \mathbf{T}$ (b) $\mathbf{K} = \mathbf{P}^T$

Figure 1. *Using $\mathcal{W}_J^{(2)}$, we compare the windowed scattering coefficients $\mathbf{S}_J\mathbf{x}$ of a signal $\mathbf{x}$ on a graph G to the windowed scattering coefficients $\mathbf{S}_J\Pi\mathbf{x}$ of the permuted signal $\Pi\mathbf{x}$ on the graph $G' = \Pi(G)$ for increasing J and with $\mathbf{K}$ equal to both $\mathbf{T}$ and $\mathbf{P}^T$. As we can see, the difference rapidly converges to zero when $\mathbf{K} = \mathbf{P}$ but not when $\mathbf{K} = \mathbf{T}$. Here, we took G to be an Erdős–Reny random graph with 100 vertices and the probability of connecting two vertices with an edge as $p = 0.7$. The entries of $\mathbf{x}$ are i.i.d. standard Gaussian random variables.*

that the windowed scattering transform is not invariant for general $\mathbf{K}$. Indeed, inspecting the proof, we see that the invariance of the windowed scattering transform is a result of the invariance of the low-pass filter Φ_J. For large values of J, this operator essentially projects the signal $\mathbf{x}$ onto the bottom eigenvector $\mathbf{u}_0$. Thus, the invariance of the low-pass filter is a result of $\mathbf{u}_0$ being a constant vector. Since $\mathbf{u}_0 = \mathbf{M}^{-1}\mathbf{d}^{1/2}$, this only occurs when $\mathbf{M} = \mathbf{D}^{1/2}$. We also note that in [32], the scattering transform is constructed from the unnormalized Laplacian, whose bottom eigenvector is always constant, which is key to the invariance result obtained there.

4. Stability to graph perturbations. In this section, let $G = (V, E, W)$ and $\widetilde{G} = (\widetilde{V}, \widetilde{E}, \widetilde{W})$ be weighted, connected graphs with $|V| = |\widetilde{V}| = n$, and let $\mathbf{M}$ and $\widetilde{\mathbf{M}}$ be invertible matrices. Throughout this section, for any object X associated to G or $\mathbf{L}^2(\mathbf{M})$, we will let $\widetilde{X}$ denote the analogous object on $\widetilde{G}$ or $\mathbf{L}^2(\widetilde{\mathbf{M}})$, so, e.g., $\widetilde{\mathbf{d}}$ is the degree vector on $\widetilde{G}$.

Our analysis here is motivated by two problems in the machine learning literature on graphs. The first problem is graph classification, and the second problem is graph alignment. In graph classification problems, one wishes to assign a label to a graph G. In this setting, one requires a representation of G that is invariant to permutations of the vertex indices since this operation does not change the underlying graph. Relatedly, graphs that are "similar" often have similar labels (or even the same label). Thus, the representation should also be stable (or at least continuous) to small perturbations of the graph G. On the other hand, one desires a representation that is able to distinguish between graphs that are "dissimilar," indicating that it should retain as much information about G as possible.

The graph alignment problem involves trying to find the "best" correspondence between the vertices of two graphs G and $\widetilde{G}$ with the same number of vertices. In this setting, one requires a representation of each vertex in the graph, so that the representation of each vertex in G may be compared to the representation of each vertex in $\widetilde{G}$. In this case, permutation invariance is not required and in fact would be a detriment, as a permutation-invariant representation would not be able to align a graph with a permutation of itself. Instead, one requires a permutation-equivariant representation. Additionally, one may want the representation to be locally quasi-invariant in the sense that permutations of small numbers of vertices (in practice located near each other in the graph) will not affect the representation too much. As with the graph classification problem, the representation should be stable to perturbations of the graph and retain as much information about the structure of the graph as possible.

Following [12], as well as [8] and [22], our first measure of the distance between two graphs G and $\widetilde{G}$ will be the "diffusion distances" given by $\text{dist}(\mathbf{T}, \widetilde{\mathbf{T}}) := \|\mathbf{T} - \widetilde{\mathbf{T}}\|_2$. This distance measure it not permutation invariant in the sense that if $\widetilde{G}$ is a permutation of G, then $\|\mathbf{T} - \widetilde{\mathbf{T}}\|_2$ is not zero in general. However, since $\mathbf{T}$ is a diffusion operator, the distance measure provided by $\|\mathbf{T} - \widetilde{\mathbf{T}}\|$ will be stable with respect to relatively small perturbations of the edge weights. Such a distance measure is useful for equivariant models, used in, e.g., graph alignment. For invariant problems, though, such as the graph classification task, if $\widetilde{G}$ is equal to G up to a permutation of its vertex/edge indices, then $\widetilde{G}$ is the same graph as G. As such, the following permutation-invariant graph distance is preferred:

$$\text{dist}_{inva}(G, \widetilde{G}) = \min_{\substack{\Pi \in S_n \\ G' = \Pi(G)}} \|\mathbf{T}' - \widetilde{\mathbf{T}}\|_2 .$$

We will show that the geometric wavelet transform, which is a permutation-equivariant transform, is stable with respect to $\|\mathbf{T} - \widetilde{\mathbf{T}}\|_2$. We will then use this result to prove that the nonwindowed geometric scattering transform, which is a permutation-invariant representation of the graph G, is stable with respect to $\text{dist}_{inva}(G, \widetilde{G})$. Additionally, we show that the windowed geometric scattering transform, which is an equivariant transform, is nevertheless stable with respect to $\text{dist}_{inva}(G, \widetilde{G})$ plus a term that measures the size of the permutation required to align G with $\widetilde{G}$ but that is dampened by the inverse of the scale of the low-pass filter.

In order to carry out this analysis, we introduce additional terms which measure the difference between the matrices $\mathbf{M}$ and $\widetilde{\mathbf{M}}$. Specifically, we let

$$\mathbf{R}_1 := \mathbf{R}_1(\mathbf{M}, \widetilde{\mathbf{M}}) := \mathbf{M}^{-1}\widetilde{\mathbf{M}} \quad \text{and} \quad \mathbf{R}_2 := \mathbf{R}_2(\mathbf{M}, \widetilde{\mathbf{M}}) := \widetilde{\mathbf{M}}\mathbf{M}^{-1}$$

and consider the quantities $\kappa(\mathbf{M}, \widetilde{\mathbf{M}})$ and $R(\mathbf{M}, \widetilde{\mathbf{M}})$ defined by

$$\kappa(\mathbf{M}, \widetilde{\mathbf{M}}) := \max_{i=1,2}\{\max\{\|\mathbf{I} - \mathbf{R}_i\|_2, \|\mathbf{I} - \mathbf{R}_i^{-1}\|_2\}\} \quad \text{and}$$
$$R(\mathbf{M}, \widetilde{\mathbf{M}}) := \max_{i=1,2}\{\max\{\|\mathbf{R}_i\|_2, \|\mathbf{R}_i^{-1}\|_2\}\}.$$

When we choose $\mathbf{K}$ to be $\mathbf{P}$ or $\mathbf{P}^T$, we have $\mathbf{M} = \mathbf{D}^{\pm 1/2}$, which implies that $\mathbf{R}_1 = \mathbf{R}_2 = \text{diag}(\frac{\widetilde{d}_i}{d_i})^{\pm 1/2}$. Therefore, $\kappa(\mathbf{M}, \widetilde{\mathbf{M}})$ measures how different the degree vectors $\mathbf{d}$ and $\widetilde{\mathbf{d}}$ are. More generally, if $\mathbf{M}$ and $\widetilde{\mathbf{M}}$ are diagonal matrices, then $\mathbf{L}^2(\mathbf{M})$ and $\mathbf{L}^2(\widetilde{\mathbf{M}})$ can be viewed as a weighted versions of the standard $\mathbf{L}^2$ space and $\kappa(\mathbf{M}, \widetilde{\mathbf{M}})$ and $R\kappa(\mathbf{M}, \widetilde{\mathbf{M}})$ measure how different these weightings are.

We note that by construction, we have $1 \leq R(\mathbf{M}, \widetilde{\mathbf{M}}) \leq \kappa(\mathbf{M}, \widetilde{\mathbf{M}}) + 1$. Thus, if $\mathbf{M} \approx \widetilde{\mathbf{M}}$, we will have $\kappa(\mathbf{M}, \widetilde{\mathbf{M}}) \approx 0$ and consequently $R(\mathbf{M}, \widetilde{\mathbf{M}}) \approx 1$. We also note that we will have $\kappa(\mathbf{M}, \widetilde{\mathbf{M}}) = 0$ and $R(\mathbf{M}, \widetilde{\mathbf{M}}) = 1$ if either $\mathbf{M} = \mathbf{I}$ (so that $\mathbf{K} = \mathbf{T}$) or if $\mathbf{M} = \mathbf{D}^{\pm 1/2}$ and the graphs G and $\widetilde{G}$ have the same degree vector. The latter situation occurs if, e.g., G and $\widetilde{G}$ are regular graphs of the same degree. Furthermore, we note that if $\mathbf{M}$ is diagonal, then $\mathbf{R}_1 = \mathbf{R}_2$. We only need two separate matrices $\mathbf{R}_1, \mathbf{R}_2$ when $\mathbf{M}$ is not a diagonal matrix. However, in our prototypical examples, we have $\mathbf{M} = \mathbf{I}$ or $\mathbf{M} = \mathbf{D}^{\pm 1/2}$, all of which are diagonal.

4.1. Stability of the wavelet transforms. In this section, we analyze the stability of the wavelets $\mathcal{W}_J^{(1)}$ and $\mathcal{W}_J^{(2)}$ constructed in section 2. Our first two theorems provide stability bounds for $\mathbf{K} = \mathbf{T}$. These results will be extended to general $\mathbf{K}$ by Theorem 4.3.

Theorem 4.1 (stability of $\mathcal{W}_J^{(1)}$ with $\mathbf{K} = \mathbf{T}$). *Suppose that $G = (V, E, W)$ and $\widetilde{G} = (\widetilde{V}, \widetilde{E}, \widetilde{W})$ are weighted, connected graphs with $|V| = |\widetilde{V}| = n$. Let $\lambda_1^* = \max\{\lambda_1, \widetilde{\lambda}_1\}$, and let $\mathbf{M} = \mathbf{I}$ so that $\mathbf{K} = \mathbf{T}$. Let $\mathcal{W}_J^{(1)}$ be the wavelets constructed from $\mathbf{T}$ in section 2, and let $\widetilde{\mathcal{W}}_J^{(1)}$ be the corresponding wavelets on $\widetilde{G}$ constructed from $\widetilde{\mathbf{T}}$. Then there exists a constant $C_{\lambda_1^*}$, depending only on λ_1^*, such that*

$$\|\mathcal{W}_J^{(1)} - \widetilde{\mathcal{W}}_J^{(1)}\|_2^2 \leq C_{\lambda_1^*} 2^J n \|\mathbf{T} - \widetilde{\mathbf{T}}\|_2.$$

The proof of Theorem 4.1 is in section SM10. Our next result provides stability bounds for $\mathcal{W}_J^{(2)}$ when $\mathbf{M} = \mathbf{I}$ (i.e., $\mathbf{K} = \mathbf{T}$). The proof, given in section SM12, is closely modeled after

the proofs of Lemmas 5.1 and 5.2 in [12]. However, due to a small change in the derivation, our result appears in a slightly different form.

Theorem 4.2 (stability of $\mathcal{W}_J^{(2)}$ with $\mathbf{K} = \mathbf{T}$). *Suppose that $G = (V, E, W)$ and $\widetilde{G} = (\widetilde{V}, \widetilde{E}, \widetilde{W})$ are weighted, connected graphs with $|V| = |\widetilde{V}| = n$. Let $\lambda_1^* = \max\{\lambda_1, \widetilde{\lambda}_1\}$, and let $\mathbf{M} = \mathbf{I}$ so that $\mathbf{K} = \mathbf{T}$. Let $\mathcal{W}_J^{(2)}$ be the wavelets constructed from $\mathbf{T}$ in section 2, and let $\widetilde{\mathcal{W}}_J^{(2)}$ be the corresponding wavelets constructed from $\widetilde{\mathbf{T}}$. Then*

$$\left\|\mathcal{W}_J^{(2)} - \widetilde{\mathcal{W}}_J^{(2)}\right\|_2^2 \leq C_{\lambda_1^*}\|\mathbf{T} - \widetilde{\mathbf{T}}\|.$$

Comparing Theorems 4.1 and 4.2, we note that while $\mathcal{W}_J^{(1)}$ has the advantage of being a tight frame, $\mathcal{W}_J^{(2)}$ has the advantage of possessing stronger stability guarantees. Numerical experiments indicate that $\mathcal{W}_J^{(1)}$ is indeed less stable to minor graph perturbations than $\mathcal{W}_J^{(2)}$. In Figure 2, we plot the stability of both $\mathcal{W}_J^{(1)}$ and $\mathcal{W}_J^{(2)}$ on Erdős–Rényi, Watts–Strogatz (small world), and Barabási–Albert (preferential attachment) random graphs. For each random graph G, we obtained a perturbed graph $\widetilde{G}$ by adding mean-zero Guassian noise to each of the edge weights at noise level $\sigma = 0.1$. For all three random graphs, the operator norm of $\mathcal{W}_J^{(1)} - \widetilde{\mathcal{W}}_J^{(1)}$ was significantly larger than $\mathcal{W}_J^{(2)} - \widetilde{\mathcal{W}}_J^{(2)}$. We also note that the bounds in Theorem 4.1 increase exponentially in J, but we do not observe this behavior in practice.

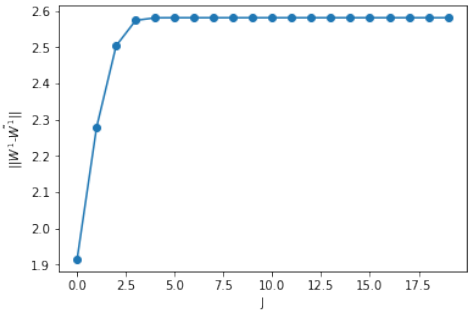
Theorems 4.1 and 4.2 show that the wavelets $\mathcal{W}_J^{(1)}$ and $\mathcal{W}_J^{(2)}$ are stable on $\mathbf{L}^2$ in the special case that $\mathbf{K} = \mathbf{T}$. Our next theorem extends this analysis to general $\mathbf{K}$ and to more general functions of $\mathbf{K}$. In particular, it can be applied to any situation where $\mathcal{W}^{\mathbf{T}} = \{r_j(\mathbf{T})\}_{j\in\mathcal{J}}$ and $\mathcal{W}^{\widetilde{\mathbf{T}}} = \{r_j(\widetilde{\mathbf{T}})\}_{j\in\mathcal{J}}$ form frames on the unweighted $\mathbf{L}^2$ space, where $\mathcal{J}$ is some indexing set and each of the functions r_j is a polynomial or the square root of a polynomial. We note that in the case where $\mathbf{M}$ is close to $\widetilde{\mathbf{M}}$, we have $\kappa(\mathbf{M}, \widetilde{\mathbf{M}}) \approx 0$. Therefore, Theorem 4.3 will imply that $\left\|\mathcal{W}^{\mathbf{K}} - \mathcal{W}^{\widetilde{\mathbf{K}}}\right\|_{\mathbf{M}}^2 \lesssim 6\left\|\mathcal{W}^{\mathbf{T}} - \mathcal{W}^{\widetilde{\mathbf{T}}}\right\|_2^2$. For a proof, see section SM13.

Theorem 4.3 (wavelet stability for general $\mathbf{K}$). *Suppose that $G = (V, E, W)$ and $\widetilde{G} = (\widetilde{V}, \widetilde{E}, \widetilde{W})$ are weighted, connected graphs with $|V| = |\widetilde{V}| = n$, and let $\mathbf{M}$ and $\widetilde{\mathbf{M}}$ be invertible matrices. Let $\mathcal{J}$ be an indexing set, and for $j \in \mathcal{J}$, let $r_j(\cdot)$ be either a polynomial or the square root of a polynomial. Suppose that $\mathcal{W}^{\mathbf{T}} = \{r_j(\mathbf{T})\}_{j\in\mathcal{J}}$ and $\mathcal{W}^{\widetilde{\mathbf{T}}} = \{r_j(\widetilde{\mathbf{T}})\}_{j\in\mathcal{J}}$ are frames on the unweighted $\mathbf{L}^2$ space with $C \leq 1$ in (1.5) for both $\mathcal{W}^{\mathbf{T}}$ and $\mathcal{W}^{\widetilde{\mathbf{T}}}$. Let $\mathbf{K} = \mathbf{M}^{-1}\mathbf{T}\mathbf{M}$, and let $\mathcal{W}^{\mathbf{K}}$ and $\mathcal{W}^{\widetilde{\mathbf{K}}}$ be the frames defined by $\{r_j(\mathbf{K})\}_{j\in\mathcal{J}}$ and $\{r_j(\widetilde{\mathbf{K}})\}_{j\in\mathcal{J}}$. Then*

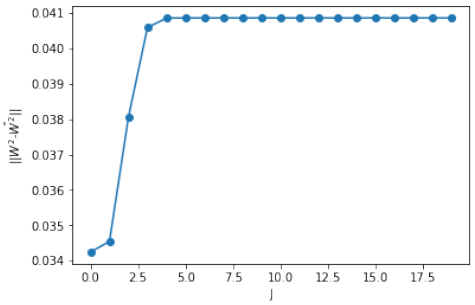
$$\left\|\mathcal{W}^{\mathbf{K}} - \mathcal{W}^{\widetilde{\mathbf{K}}}\right\|_{\mathbf{M}}^2 \leq 6\left(\left\|\mathcal{W}^{\mathbf{T}} - \mathcal{W}^{\widetilde{\mathbf{T}}}\right\|_2^2 + \kappa(\mathbf{M}, \widetilde{\mathbf{M}})^2(\kappa(\mathbf{M}, \widetilde{\mathbf{M}}) + 1)^2\right).$$

The following corollary is an immediate consequence of Theorem 4.3 combined with Theorems 4.1 and 4.2. As noted prior to the statement of Theorem 4.3, in the case where $\mathbf{M}$ is close to $\widetilde{\mathbf{M}}$, the terms involving $\kappa(\mathbf{M}, \widetilde{\mathbf{M}})$ will be small.

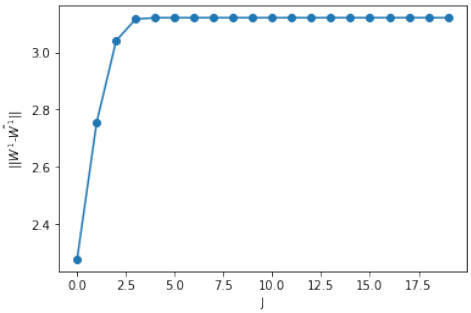
Corollary 4.4 (stability of $\mathcal{W}_J^{(1)}$ and $\mathcal{W}_J^{(2)}$ with general $\mathbf{K}$). *Suppose that $G = (V, E, W)$ and $\widetilde{G} = (\widetilde{V}, \widetilde{E}, \widetilde{W})$ are weighted, connected graphs with $|V| = |\widetilde{V}| = n$. Let $\mathbf{M}$ and $\widetilde{\mathbf{M}}$ be invertible matrices, and let $\lambda_1^* = \max\{\lambda_1, \widetilde{\lambda}_1\}$. Let $\mathcal{W}_J^{(1)}$ and $\mathcal{W}_J^{(2)}$ be the wavelet transforms constructed*



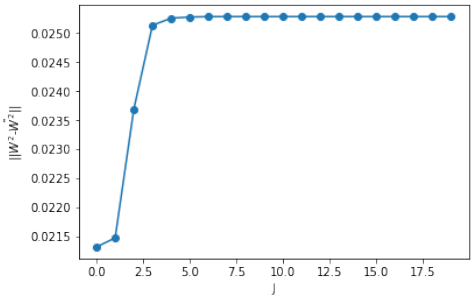
(a) $\mathcal{W}_J^{(1)}$ on Erdős Rényi



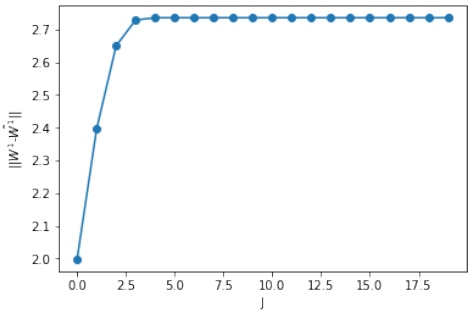
(b) $\mathcal{W}_J^{(2)}$ on Erdős Rényi



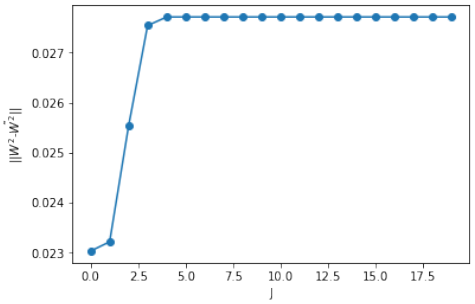
(c) $\mathcal{W}_J^{(1)}$ on Watts Strogatz



(d) $\mathcal{W}_J^{(2)}$ on Watts Strogatz



(e) $\mathcal{W}_J^{(1)}$ on Barabási Albert



(f) $\mathcal{W}_J^{(2)}$ on Barabási Albert

Figure 2. *We plot the operator norm of* $\mathcal{W}_J^{(1)} - \widetilde{W}_J^{(1)}$ *and* $\mathcal{W}_J^{(2)} - \widetilde{W}_J^{(2)}$ *as a function of* J *on Erdős–Rényi* ($n = 100, p = .1$)*, Watts–Strogatz* ($n = 100, k = 20, p = .1$)*, and Barabási–Albert* ($n = 100, m = 10$) *random graphs. The perturbed graphs* $\widetilde{G}$ *were obtained by adding Gaussian noise* ($\sigma = .1$) *to the edge weights of* G*. Observe that for all three random graphs, the wavelets* $\mathcal{W}_J^{(2)}$ *are more stable than* $\mathcal{W}^{(1)}$*. However, it appears that the operator norm of both* $\mathcal{W}_J^{(1)} - \widetilde{W}_J^{(1)}$ *and* $\mathcal{W}^{(2)} - \widetilde{W}_J^{(2)}$ *can be taken to be independent of* J*. In these experiments, we chose* $\mathbf{K} = \mathbf{T}$ *and did not include the low-pass filter* Φ_J *in our wavelets.*

from $\mathbf{K}$ *in section 2, and let* $\widetilde{\mathcal{W}}_J^{(1)}$ *and* $\widetilde{\mathcal{W}}_J^{(2)}$ *be the corresponding wavelet transforms constructed from* $\widetilde{\mathbf{K}}$. *Then*

$$\left\|\mathcal{W}_J^{(1)} - \widetilde{\mathcal{W}}_J^{(1)}\right\|_2^2 \leq C_{\lambda_1^*}\left(2^J n\|\mathbf{T}-\widetilde{\mathbf{T}}\| + \kappa(\mathbf{M},\widetilde{\mathbf{M}})^2(\kappa(\mathbf{M},\widetilde{\mathbf{M}})+1)^2\right) \text{ and}$$
$$\left\|\mathcal{W}_J^{(2)} - \widetilde{\mathcal{W}}_J^{(2)}\right\|_2^2 \leq C_{\lambda_1^*}\left(\|\mathbf{T}-\widetilde{\mathbf{T}}\|_2 + \kappa(\mathbf{M},\widetilde{\mathbf{M}})^2(\kappa(\mathbf{M},\widetilde{\mathbf{M}})+1)^2\right).$$

Remark 4.5. Inspecting Corollary 4.4, we see that our wavelets may become less stable as $\kappa(\mathbf{M},\widetilde{\mathbf{M}})$ increases. In particular, in the case where $\mathbf{M}=\mathbf{D}^\alpha$ and $\widetilde{\mathbf{M}}=\widetilde{\mathbf{D}}^\alpha$, we have

$$\kappa(\mathbf{M},\widetilde{\mathbf{M}}) = \max\left\{\left|1-\left(\frac{d_i}{\tilde{d}_i}\right)^\alpha\right|, \left|1-\left(\frac{\tilde{d}_i}{d_i}\right)^\alpha\right|\right\}.$$

Therefore, values of α closer to zero lead to tighter stability bounds.

One might also wish to replace Corollary 4.4 with an inequality written in terms of $\|\mathbf{K}-\widetilde{\mathbf{K}}\|_{\mathbf{M}}$ rather than $\|\mathbf{T}-\widetilde{\mathbf{T}}\|_2$. This can be done by the following proposition. Recall that if $\mathbf{M}\approx\widetilde{\mathbf{M}}$, then $\kappa(\mathbf{M},\widetilde{\mathbf{M}})\approx 0$ and $R(\mathbf{M},\widetilde{\mathbf{M}})\approx 1$. Therefore, (4.1) implies that $\|\mathbf{T}-\widetilde{\mathbf{T}}\|_2 \lesssim \|\mathbf{K}-\widetilde{\mathbf{K}}\|_{\mathbf{M}}$. For a proof of Proposition 4.6, see section SM14.

Proposition 4.6. *Suppose that* $G=(V,E,W)$ *and* $\widetilde{G}=(\widetilde{V},\widetilde{E},\widetilde{W})$ *are weighted, connected graphs with* $|V|=|\widetilde{V}|=n$. *Then*

$$\|\mathbf{T}-\widetilde{\mathbf{T}}\|_2 \leq \kappa(\mathbf{M},\widetilde{\mathbf{M}})\left(1+R(\mathbf{M},\widetilde{\mathbf{M}})^3\right) + R(\mathbf{M},\widetilde{\mathbf{M}})\|\mathbf{K}-\widetilde{\mathbf{K}}\|_{\mathbf{M}}. \tag{4.1}$$

4.2. Stability of the geometric scattering transform. In this section, we will prove stability bounds for the windowed and nonwindowed scattering transform. We will state these results in terms of the stability bound of the underlying wavelet transform and also the upper frame bound of $\widetilde{\mathcal{W}}$ when considered as on operator on $\mathbf{L}^2(\mathbf{M})$. We do this both to emphasize that the stability of the scattering transform is a consequence of the stability of the underlying frame and so that our result can be applied to scattering transforms built onto of other graph wavelet constructions. We will assume that $G=(V,E,W)$ and $\widetilde{G}=(\widetilde{V},\widetilde{E},\widetilde{W})$ are weighted, connected graphs such that $|V|=|\widetilde{V}|=n$; let $\mathbf{M}$ and $\widetilde{\mathbf{M}}$ be $n\times n$ invertible matrices; and assume that $\mathcal{W}=\{\Psi_j,\Phi\}_{j\in\mathcal{J}}$ and $\widetilde{\mathcal{W}}=\{\widetilde{\Psi}_j,\widetilde{\Phi}\}_{j\in\mathcal{J}}$ are frames on $\mathbf{L}^2(\mathbf{M})$ and $\mathbf{L}^2(\widetilde{\mathbf{M}})$ such that $C\leq 1$ in (1.5). For a family of matrices $\Gamma=(\mathbf{B}_i)_{i\in\mathcal{I}}$, where $\mathcal{I}$ is a countable index set, we define the operator norm of $\Gamma:\mathbf{L}^2(G,\mathbf{M})\to\mathbf{M}\ell^2(\mathbf{L}^2(G,\mathbf{M}))$ as $\|\Gamma\|_{\mathbf{M}}:=\sup_{\|x\|_{\mathbf{M}}=1}\|\Gamma\mathbf{x}\|_{\mathbf{M}}$. If Π is a permutation and $\widetilde{G}'=\Pi(\widetilde{G})$, then we will let $\widetilde{\mathcal{W}}':=\Pi\widetilde{\mathcal{W}}\Pi^T=\{\Pi\widetilde{\Psi}_j\Pi^T,\Pi\widetilde{\Phi}\Pi^T\}_{j\in\mathcal{J}}$ denote the corresponding permuted wavelet frame.

Theorem 4.7 provides stability guarantees for the windowed scattering transform with bounds that $\|\mathcal{W}-\widetilde{\mathcal{W}}\|_{\mathbf{M}}$ and $\|\widetilde{\mathcal{W}}\|_{\mathbf{M}}$. Similarly, Theorem 4.11 guarantees stability for the nonwindowed transform. By Theorems 4.1, 4.2, and 4.3 as well as Proposition 4.10, these results imply that the scattering transforms constructed from $\mathcal{W}_J^{(1)}$ or $\mathcal{W}_J^{(2)}$ are stable in the sense that if G is similar to $\widetilde{G}$ and $\mathbf{M}$ is similar to $\widetilde{\mathbf{M}}$, then the scattering transforms $\mathbf{S}$ and $\widetilde{\mathbf{S}}$ will produce similar representations of an inputted signal $\mathbf{x}$. Many of the ideas in the proof of Theorems 4.7 and 4.11 are similar to those used to prove Theorem 5.3 in [12]. The primary

difference is Lemma SM15.1, which is needed because $\widetilde{\mathcal{W}}$ is not (in general) a nonexpansive frame on $\mathbf{L}^2(\mathbf{M})$, and therefore our results will involve terms related to the operator norm of $\widetilde{\mathcal{W}}$ on $\mathbf{L}^2(\mathbf{M})$, which can be controlled by Proposition 4.10. We also note that both Theorems 4.7 and 4.11 consider pertubations to both the input signal $\mathbf{x}$ and the graph structure, whereas Theorem 5.3 of [12] only considered perturbations to the graph structure.

Theorem 4.7 (stability for the windowed scattering transform). *Let $G = (V, E, W)$ and $\widetilde{G} = (\widetilde{V}, \widetilde{E}, \widetilde{W})$ be weighted, connected graphs with $|V| = |\widetilde{V}| = n$; let $\mathbf{M}$ and $\widetilde{\mathbf{M}}$ be invertible $n \times n$ matrices; and let $\mathcal{J}$ be an indexing set. Let $\mathcal{W} = \{\Psi_j, \Phi\}_{j\in\mathcal{J}}$ and $\widetilde{\mathcal{W}} = \{\widetilde{\Psi}_j, \widetilde{\Phi}\}_{j\in\mathcal{J}}$ be frames on $\mathbf{L}^2(\mathbf{M})$ and $\mathbf{L}^2(\widetilde{\mathbf{M}})$ such that $C \le 1$ in (1.5). Let $\mathbf{S}^\ell$ and $\widetilde{\mathbf{S}}^\ell$ be the ℓth layers of the windowed scattering transforms on G and $\widetilde{G}$. Further assume that $\mathbf{S}^\ell$ is approximately permutation invariant up to a factor of $\mathcal{B}$ in the sense that*

$$\left\|\Pi\mathbf{S}^\ell\mathbf{x} - \mathbf{S}^\ell\mathbf{x}\right\|_{\mathbf{M}} \le \mathcal{B}\|\mathbf{x}\|_{\mathbf{M}} \tag{4.2}$$

for all $\mathbf{x} \in \mathbf{L}^2(\mathbf{M})$ and $\Pi \in S_n$. Then for all $\mathbf{x} \in \mathbf{L}^2(\mathbf{M})$ and $\widetilde{\mathbf{x}} \in \mathbf{L}^2(\widetilde{\mathbf{M}})$,

$$\begin{aligned}\left\|\mathbf{S}^\ell\mathbf{x} - \widetilde{\mathbf{S}}^\ell\widetilde{\mathbf{x}}\right\|_2 \\ \le R(\mathbf{M},\mathbf{I})^2 \inf_{\substack{\Pi\in S_n \\ G'=\Pi(G)}} \left(\mathcal{B}\|\mathbf{x}\|_2 + \|\Pi\mathbf{x} - \widetilde{\mathbf{x}}\|_2 + \sqrt{2}\|\mathcal{W}' - \widetilde{\mathcal{W}}\|_{\mathbf{M}'}\left(\sum_{k=0}^{\ell}\|\widetilde{\mathcal{W}}\|_{\mathbf{M}'}^k\right)\|\widetilde{\mathbf{x}}\|_2\right).\end{aligned} \tag{4.3}$$

Remark 4.8. We can interpret each term in the right-hand side of (4.3). The first term is a direct result of the approximate permutation invariance of the windowed scattering operator. The second term measures the best possible alignment of the input signals $\mathbf{x}$ and $\widetilde{\mathbf{x}}$, whereas the third term measures the best possible alignment of the wavelet operators on G and $\widetilde{G}$. Note that if there is a permutation Π such that $\widetilde{G} = \Pi(G)$ and $\widetilde{\mathbf{x}} = \Pi\mathbf{x}$, then the only nonzero term is $\mathcal{B}\|\mathbf{x}\|_2$. This result is more general than an approximate invariance result, though, since it also characterizes the stability of the windowed scattering transform in terms of the stability of the wavelet operators and the similarity of the input signals $\mathbf{x}$ and $\widetilde{\mathbf{x}}$ (which are often functions of the graphs G and $\widetilde{G}$, respectively).

In order to prove Theorem 4.7, we need the following lemma (proved in section SM15).

Lemma 4.9. *Under the assumptions of Theorem 4.7, we have*

$$\left\|\mathbf{S}^\ell\mathbf{x} - \widetilde{\mathbf{S}}^\ell\mathbf{x}\right\|_{\mathbf{M}} \le \sqrt{2}\|\mathcal{W} - \widetilde{\mathcal{W}}\|_{\mathbf{M}}\left(\sum_{k=0}^{\ell}\|\widetilde{\mathcal{W}}\|_{\mathbf{M}}^k\right)\|\mathbf{x}\|_{\mathbf{M}} \quad \text{for all } \mathbf{x} \in \mathbf{L}^2(\mathbf{M}). \tag{4.4}$$

Proof of Theorem 4.7. Let $\Pi \in S_n$ be a permutation, and let $G' = \Pi(G)$. By Theorem 3.6, we have $\Pi\mathbf{S}^\ell = (\mathbf{S}^\ell)'\Pi$. Therefore, the triangle inequality implies that

$$\left\|\mathbf{S}^\ell\mathbf{x} - \widetilde{\mathbf{S}}^\ell\widetilde{\mathbf{x}}\right\|_2 \le \left\|\mathbf{S}^\ell\mathbf{x} - \Pi\mathbf{S}^\ell\mathbf{x}\|_2 + \|(\mathbf{S}^\ell)'\Pi\mathbf{x} - (\mathbf{S}^\ell)'\widetilde{\mathbf{x}}\|_2 + \|(\mathbf{S}^\ell)'\widetilde{\mathbf{x}} - \widetilde{\mathbf{S}}^\ell\widetilde{\mathbf{x}}\right\|_2. \tag{4.5}$$

The assumption (4.2) implies that

$$\|\mathbf{S}^\ell\mathbf{x} - \Pi\mathbf{S}^\ell\mathbf{x}\|_2 \le R(\mathbf{M},\mathbf{I})\|\mathbf{S}^\ell\mathbf{x} - \Pi\mathbf{S}^\ell\mathbf{x}\|_{\mathbf{M}} \le \mathcal{B}R(\mathbf{M},\mathbf{I})\|\mathbf{x}\|_{\mathbf{M}} \le \mathcal{B}R(\mathbf{M},\mathbf{I})^2\|\mathbf{x}\|_2. \tag{4.6}$$

Similarly, by Theorem 3.2, we have that

$$\|(\mathbf{S}^{\ell})'\Pi\mathbf{x} - (\mathbf{S}^{\ell})'\widetilde{\mathbf{x}}\|_2 \leq R(\mathbf{M}',\mathbf{I})^2\|\Pi\mathbf{x} - \widetilde{\mathbf{x}}\|_2, \tag{4.7}$$

and applying Lemma 4.9 yields

$$\|(\mathbf{S}^{\ell})'\widetilde{\mathbf{x}} - \widetilde{\mathbf{S}}^{\ell}\widetilde{\mathbf{x}}\|_2 \leq \sqrt{2}R(\mathbf{M}',\mathbf{I})^2\|\mathcal{W}' - \widetilde{\mathcal{W}}\|_{\mathbf{M}'}\left(\sum_{k=0}^{\ell}\|\widetilde{\mathcal{W}}\|_{\mathbf{M}'}^k\right)\|\widetilde{\mathbf{x}}\|_2. \tag{4.8}$$

Since $G' = \Pi(G)$, one may check that $R(\mathbf{M},\mathbf{I}) = \max\{\|\mathbf{M}\|_2, \|\mathbf{M}^{-1}\|_2\} = R(\mathbf{M}',\mathbf{I})$. Therefore, combining (4.5) with (4.6), (4.7), and (4.8) yields

$$\left\|\mathbf{S}^{\ell}\mathbf{x} - \widetilde{\mathbf{S}}^{\ell}\widetilde{\mathbf{x}}\right\|_2 \leq R(\mathbf{M},I)^2\left(\mathcal{B}\|\mathbf{x}\|_2 + \|\Pi\mathbf{x} - \widetilde{\mathbf{x}}\|_2 + \sqrt{2}\|\mathcal{W}' - \widetilde{\mathcal{W}}\|_{\mathbf{M}'}\left(\sum_{k=0}^{\ell}\|\widetilde{\mathcal{W}}\|_{\mathbf{M}'}^k\right)\|\widetilde{\mathbf{x}}\|_2\right),$$

and so infimizing over Π completes the proof. ∎

The following proposition shows that if $\mathbf{M}$ is close to $\widetilde{\mathbf{M}}$, then an upper frame bound for $\mathcal{W}$ can be used to produce an upper frame bound for $\widetilde{\mathcal{W}}$ on $\mathbf{L}^2(\mathbf{M})$ provided the wavelet operators satisfy certain conditions. This result yields a more transparent upper bound in the stability results of Theorem 4.7 for the types of wavelet operators covered by the proposition. More specifically, combining Proposition 4.10 with Theorem 4.7 allows one to refine (4.3) to

$$\begin{aligned}&\left\|\mathbf{S}^{\ell}\mathbf{x} - \widetilde{\mathbf{S}}^{\ell}\widetilde{\mathbf{x}}\right\|_2\\ &\quad\leq R(\mathbf{M},\mathbf{I})^2\inf_{\substack{\Pi\in S_n\\ G'=\Pi(G)}}\left(\mathcal{B}\|\mathbf{x}\|_2 + \|\Pi\mathbf{x} - \widetilde{\mathbf{x}}\|_2 + \sqrt{2}\|\mathcal{W}' - \widetilde{\mathcal{W}}\|_{\mathbf{M}'}\left(\sum_{k=0}^{\ell}R(\mathbf{M},\widetilde{\mathbf{M}})^{2k}\right)\|\widetilde{\mathbf{x}}\|_2\right).\end{aligned}$$

We note that this result can be applied to both $\mathcal{W}_J^{(1)}$ and $\mathcal{W}_J^{(2)}$. For a proof, see section SM16.

Proposition 4.10. *Suppose that $G = (V,E,W)$ and $\widetilde{G} = (\widetilde{V},\widetilde{E},\widetilde{W})$ are weighted, connected graphs with $|V| = |\widetilde{V}| = n$. Let $\mathbf{M}$ and $\widetilde{\mathbf{M}}$ be invertible $n\times n$ matrices, and let $\mathbf{K} = \mathbf{M}^{-1}\mathbf{T}\mathbf{M}, \widetilde{\mathbf{K}} = \widetilde{\mathbf{M}}^{-1}\widetilde{\mathbf{T}}\widetilde{\mathbf{M}}$. Let $\mathcal{J}$ be an indexing set, and for $j\in\mathcal{J}$, let $r_j(\cdot)$ be either a polynomial or the square root of a polynomial. Suppose that $\mathcal{W} = \{r_j(\mathbf{K})\}_{j\in\mathcal{J}}$ is a frame on $\mathbf{L}^2(\mathbf{M})$ with $C\leq 1$ in (1.5), and similarly assume that $\widetilde{\mathcal{W}} = \{r_j(\widetilde{\mathbf{K}})\}_{j\in\mathcal{J}}$ is a frame on $\mathbf{L}^2(\widetilde{\mathbf{M}})$ also with $C\leq 1$ in (1.5). Then $\widetilde{\mathcal{W}}$ is a bounded operator on $\mathbf{L}^2(\mathbf{M})$ and*

$$\|\widetilde{\mathcal{W}}\mathbf{x}\|_{\mathbf{M}}^2 = \sum_{j\in\mathcal{J}}\|r_j(\widetilde{\mathbf{K}})\mathbf{x}\|_{\mathbf{M}}^2 \leq R(\mathbf{M},\widetilde{\mathbf{M}})^4\|\mathbf{x}\|_{\mathbf{M}}^2.$$

The following is the analogue of Theorem 4.7 for the nonwindowed scattering transform. For a proof, see section SM17.

Theorem 4.11 (stability for the nonwindowed scattering transform). *Let $G = (V,E,W)$ and $\widetilde{G} = (\widetilde{V},\widetilde{E},\widetilde{W})$ be weighted, connected graphs with $|V| = |\widetilde{V}| = n$; let $\mathbf{M}$ and $\widetilde{\mathbf{M}}$ be invertible $n\times n$ matrices; and let $\mathcal{J}$ be an indexing set. Let $\mathcal{W} = \{\Psi_j,\Phi\}_{j\in\mathcal{J}}$ and $\widetilde{\mathcal{W}} = \{\widetilde{\Psi}_j,\widetilde{\Phi}\}_{j\in\mathcal{J}}$ be*

frames on $\mathbf{L}^2(\mathbf{M})$ *and* $\mathbf{L}^2(\widetilde{\mathbf{M}})$ *such that* $C \leq 1$ *in* (1.5). *Let* $\mathbf{S}^\ell$ *and* $\widetilde{\mathbf{S}}^\ell$ *be the* ℓ*th layers of the windowed scattering transforms on* G *and* $\widetilde{G}$. *Further assume that* $\boldsymbol{\mu} = \mathbf{u}_0$ *and* $\widetilde{\boldsymbol{\mu}} = \widetilde{\mathbf{u}}_0$. *Then*

$$\begin{aligned}(4.9)\qquad \left\|\overline{\mathbf{S}^\ell}\mathbf{x} - \overline{\widetilde{\mathbf{S}}^\ell}\widetilde{\mathbf{x}}\right\|_2^2 \leq 3 \inf_{\Pi \in S_n} \Bigg(& \frac{2R(\mathbf{M},\mathbf{I})^4}{n \min_i |\mathbf{u}_0(i)|^2} \|\Pi\mathbf{x} - \widetilde{\mathbf{x}}\|_2^2 \\ & + 2R(\mathbf{M},\mathbf{I})^2 \|\boldsymbol{\mu}\|_{\mathbf{M}}^2 \left(\sum_{k=0}^{\ell-1} \|\widetilde{W}\|_{\mathbf{M}'}^k \right)^2 \|\mathcal{W}' - \widetilde{\mathcal{W}}\|_{\mathbf{M}'}^2 \|\tilde{\mathbf{x}}\|_2^2 \\ & + R(\mathbf{M}', \widetilde{\mathbf{M}})^2 \|\boldsymbol{\mu}' - \widetilde{\boldsymbol{\mu}}\|_{\mathbf{M}'}^2 \|\widetilde{\mathbf{x}}\|_{\widetilde{\mathbf{M}}}^2 \\ & + R(\widetilde{\mathbf{M}}, \mathbf{I})^4 R(\mathbf{M},\mathbf{I})^2 (1 + \mathbf{R}(\mathbf{M}', \widetilde{\mathbf{M}})) \|\widetilde{\boldsymbol{\mu}}\|_{\widetilde{\mathbf{M}}}^2 \kappa(\mathbf{M}', \widetilde{\mathbf{M}}) \|\widetilde{\mathbf{x}}\|_{\widetilde{\mathbf{M}}} \Bigg). \end{aligned}$$

Remark 4.12. As in Theorem 4.7, we can interpret each term in the right-hand side of (4.9). In fact, the first two terms above are very similar to the second and third terms of (4.3) and have the same interpretation. The third term in (4.9) measures the best possible alignment of the measures $\boldsymbol{\mu}$ and $\widetilde{\boldsymbol{\mu}}$, while the last term measures the best possible alignment of $\mathbf{M}$ and $\widetilde{\mathbf{M}}$ through $\kappa(\mathbf{M}', \widetilde{\mathbf{M}})$. While $\mathbf{x}$, $\mathbf{M}$, and $\boldsymbol{\mu}$ can be chosen independently of G, often they depend on G, and thus it is reasonable to assume that a small perturbation of G, reflected in $\widetilde{G}$, would lead to a small change in these quantities.

Furthermore, when interpreting the right-hand side of the above equation, one should keep in mind that if there exists a permutation Π such that $\mathbf{M}' = \Pi\mathbf{M}\Pi^T$ is close to $\widetilde{\mathbf{M}}$, we will have $\mathbf{R}(\mathbf{M}', \widetilde{\mathbf{M}}) \approx 1$ and $\kappa(\mathbf{M}', \widetilde{\mathbf{M}}) \approx 0$. In our canonical examples, we have either $\mathbf{M} = \mathbf{I}$ or $\mathbf{M} = \mathbf{D}^{\pm 1/2}$, which implies that $R(\mathbf{M},\mathbf{I}) \leq \|\mathbf{d}\|_\infty^{1/2}$. Moreover, if one assumes that $\mathbf{M} = \widetilde{\mathbf{M}} = \mathbf{M}' = \mathbf{I}$ and $\boldsymbol{\mu} = \widetilde{\boldsymbol{\mu}} = \boldsymbol{\mu}' = \frac{1}{\sqrt{n}}\mathbf{1}$ to recover a setup similar to [12], then we have $\kappa(\mathbf{M}', \widetilde{\mathbf{M}}) = 0$ and $\mathbf{K} = \mathbf{T}$, which implies that $\mathbf{u}_0 = \frac{\mathbf{d}^{1/2}}{\|\mathbf{d}^{1/2}\|_2}$, and so the above result simplifies to

$$\left\|\overline{\mathbf{S}^\ell}\mathbf{x} - \overline{\widetilde{\mathbf{S}}^\ell}\widetilde{\mathbf{x}}\right\|_2^2 \leq C_G \inf_{\Pi \in S_n} \left(\frac{\|\mathbf{d}\|_\infty^2 \|\mathbf{d}\|_1}{\min_i |\mathbf{d}(i)|^2} \|\Pi\mathbf{x} - \widetilde{\mathbf{x}}\|_2^2 + \ell^2 \|\mathcal{W}' - \widetilde{\mathcal{W}}\|_{2'}^2 \|\tilde{\mathbf{x}}\|_2^2 \right).$$

Thus, setting $\Pi\mathbf{x} = \widetilde{\mathbf{x}}$, this result essentially includes Corollary 5.4 of [12] as a special case.

5. Scattering view of GNNs.

The main contributions of the theoretical framework established here are twofold. First, it provides an overarching characterization of geometric descriptors captured by scattering features, which have been shown effective both at the node level and at whole-graph–level tasks. Second, it provides a mathematical foundation for studying a wide range of GNNs, analogous to the role played by the Euclidean scattering transform. To further establish this second aspect, we review several prominent GNNs and discuss the relationship between them and the scattering transform.

5.1. Aggregate-transform architectures.

Most GNN architectures can be modeled as alternating between two fundamental types of operations: (i) **Aggregate:** For each signal $\mathbf{x}$, replace $\mathbf{x}(i)$ with a value aggregated from its neighbors, e.g., $\mathbf{x}(i) \leftarrow \sum_{j \in \mathcal{N}_i} w_{ij}\mathbf{x}(j)$, where $\mathcal{N}_i$ a local neighborhood of node i. Different architectures vary in their definition of the $\mathcal{N}_i$ and the weights w_{ij}. However, a common theme is that this operation is applied separately

on each graph signal. (ii) **Transform:** Between aggregation steps, GNNs apply a nodewise transformation on the features of each node, typically in the form of a shallow neural network.

Together, these two operations extend the "convolutions" used in CNNs over images. We note that the typical implementation of such convolutions differs from the traditional mathematical definition of a convolution by considering not only the sliding-window application of a local filter (essentially equivalent to the aggregate step) but also an operation that mixes together resulting filter outputs (essentially equivalent to the transform step). On images, the grid organization of pixels naturally lends itself to learnable filters that can easily translated over the image. However, in graph domains, there is no such natural notion of translation. Therefore, GNNs are limited in the aggregations they can use, and most of them default to employing predetermined aggregations, somewhat analogous to the use of predefined wavelet filters in the original scattering transform. This makes the scattering framework discussed here particularly suitable for providing a solid foundation for not only understanding but also constructing GNNs. Indeed, in the Euclidean case, the scattering transform deviates from CNNs both by removing the channel-mixing (transform) operation and by replacing learned filters with handcrafted ones. Here, however, the only deviation comes from the exclusion of channel-mixing operations in the transform step. Moreover, these channel-mixing operations can be added to the scattering transform in hybrid architectures, such as [20, 21]. To further emphasize this perspective, we will discuss common implementations of the aggregation step in GNNs and how they relate to the scattering framework.

5.2. Message-passing GNNs. As a representative example of message-passing aggregation, we consider the operation used in the Graph Isomorphism Network [30]. Let $\mathbf{X}: V \to \mathbb{R}^m$ be a vector-valued function on the vertices, which may be thought of as either a feature vector for each vertex or an $\mathbf{X}$ as an $n \times m$ feature matrix. There, the hidden state of the network at the ℓth layer is given by

$$\mathbf{X}^{(\ell)}(i) = \mathrm{MLP}^{(\ell)}\left((1+\varepsilon^{(\ell)})\mathbf{X}^{(\ell-1)}(i) + \sum_{j:(i,j)\in E} W(i,j)\mathbf{X}^{(\ell-1)}(j)\right),$$

where $W(i,j)$ are the edge weights of $G = (V, E, W)$ and $\mathrm{MLP}^{(\ell)}$ implements the transform step applied to the vector of aggregated features of each node i. Note that the input into $\mathrm{MLP}^{(\ell)}$ can be written as $(1+\varepsilon^{(\ell)})(\mathbf{H}_{\varepsilon^{(\ell)}}\mathbf{X}^{(\ell-1)})(i)$ with $\mathbf{H}_\varepsilon = \mathbf{I} + (1+\varepsilon)^{-1}\mathbf{A}$. We note that the scaling constant $(1+\varepsilon^{(\ell)})$ can be combined into the MLP. Therefore, we simply consider the aggregation step as applying the filter $\mathbf{H}_\varepsilon$. We note that $\mathbf{H}_\varepsilon$ is quite similar in form to the matrix $\mathbf{T}_{g_\star}$ defined in (1.2). We also note that the decay of $\mathbf{H}_\varepsilon$ is determined by a parameter ε that effectively balances the information retained by node i compared to the information aggregated from its neighbors. Over multiple applications of the filter, the decay parameter controls the propagation speed over the graph, which can essentially be interpreted as controlling the spatial localization of information aggregated by the network in each layer.

5.3. Spectral graph convolutional networks. Spectral networks generalize convolution via the eigendecomposition of the graph Laplacian in a manner analogous to (1.1). The hidden state of the ℓth layer will be an $n \times F_\ell$ feature matrix given by

$$\mathbf{X}^{(\ell)} = \sigma(\mathbf{H}^{(\ell)}\mathbf{X}^{(\ell-1)}\mathbf{\Theta}^{(\ell)}) = \sigma(\mathbf{V}\mathrm{diag}(\mathbf{h}^{(\ell)})\mathbf{V}^T\mathbf{X}^{(\ell-1)}\mathbf{\Theta}^{(\ell)}),$$

where $\sigma(\cdot)$ is a nonlinear activation function (e.g., ReLU, sigmoid, or absolute value in our case) and $\mathbf{h}^{(\ell)}$ is a filter in the graph Fourier domain. Here, the weight matrix $\mathbf{\Theta}^{(\ell)}$ implements the transform step, while the convolutional filter $\mathbf{H}^{(\ell)}$ implements the aggregation step. The primary difference between this network architecture and our model is the transform steps, which are not included in our formulation of the scattering transform. Indeed, in our analysis, we assume that we have a single input feature $\mathbf{x}$. However, we do note that scattering networks that incorporate transform steps were shown to be effective for node classification in [20, 21].

We further note that while some networks propose to parameterize $\mathbf{h}^{(\ell)}$ as a function of Laplacian eigenvalues (e.g., using Chebyshev [9] or Caley [17] polynomials), some of the more popular architectures rely on nearly fully predefined filters. For example, the GCN architecture from [15], which is often used as a representative example of spectral GNNs, essentially implements $\mathbf{H}^{(\ell)}$ as a low-pass filter derived from first-order approximation of a parameterized polynomial. Indeed, up to renormalization (or reparametrization), the filter considered there is given by $\mathbf{H}^{(\ell)} = \mathbf{I} + \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} = 2\mathbf{T}$. Therefore, as alluded to earlier, the relation between the scattering framework and the GCN model can be regarded as stronger than in the traditional Euclidean case since the aggregation step in such GCN architectures (similar to the message-passing case) relies on predetermined filters rather than flexible learned ones.

6. Empirical results. In this section, we empirically study the performance of geometric scattering with different wavelet families $\mathcal{W}_J^{(1)}$ and $\mathcal{W}_J^{(2)}$ and with different choices of $\mathbf{K} = \mathbf{M}^{-1}\mathbf{T}\mathbf{M}$, focusing on the case where $\mathbf{M} = \mathbf{D}^{\alpha}$ for $\alpha \in \{-0.5, -0.25, 0, 0.25, 0.5\}$.[2] Notably, the case where we use the wavelets $\mathcal{W}_J^{(2)}$ and set $\alpha = 0$ or -0.5 corresponds to the settings of [12] and [13], but the other settings do not correspond to wavelets previously used in the geometric scattering literature. We perform graph classification over 11 data sets (see Table 1) and 10 initialization seeds, reporting the mean and standard deviation ($\mu \pm \sigma$) of the test set accuracy (see section SM19 for details). Overall, the optimal setting varies significantly between data sets, as shown in Table 5. For example, on DD, the top-performing model was $\mathcal{W}_J^{(2)}$ with $\alpha = 0.25$. However, this configuration is more than 10 percentage points below the top performer on IMDB-Binary. This highlights the practical importance of having a large family

Table 1

Data set statistics, diameter, nodes, edges, and average clustering coefficient (CC).

	Graphs	Classes	Diameter	Nodes	Edges	CC
COLLAB	5000	3	1.86	74.49	2457.22	0.89
DD	1178	2	19.81	284.32	715.66	0.48
IMDB-B	1000	2	1.86	19.77	96.53	0.95
IMDB-M	1500	3	1.47	13.00	65.94	0.97
MUTAG	188	2	8.22	17.93	19.79	0.00
NCI1	4110	2	13.33	29.87	32.30	0.00
NCI109	4127	2	13.14	29.68	32.13	0.00
PROTEINS	1113	2	11.62	39.06	72.82	0.51
PTC	344	2	7.52	14.29	14.69	0.01
REDDIT-B	2000	2	8.59	429.63	497.75	0.05
REDDIT-5K	4999	5	10.57	508.52	594.87	0.03

[2]Code to reproduce these results can be found at https://github.com/atong01/trainable_symmetry.

Table 2

Test accuracy ($\mu \pm \sigma$) over 11 data sets and 10 seeds of $\mathcal{W}_J^{(1)}$ versus $\mathcal{W}_J^{(2)}$ scattering for different choices of $\alpha \in [-0.5, -0.25, 0, 0.25, 0.5]$, where $\mathbf{M} = \mathbf{D}^\alpha$, corresponding to $\mathbf{K} = \mathbf{D}^{-\alpha}\mathbf{T}\mathbf{D}^\alpha$. In most settings of α, $\mathcal{W}_J^{(1)}$ slightly outperforms $\mathcal{W}_J^{(2)}$. This suggests that $\mathcal{W}_J^{(1)}$-type filters should be explored further.

α	$\mathcal{W}_J^{(1)}$ (exact)	$\mathcal{W}_J^{(2)}$
-0.5	**0.617 ± 0.007**	0.616 ± 0.012
-0.25	**0.640 ± 0.005**	0.626 ± 0.009
0.0	**0.626 ± 0.012**	0.623 ± 0.006
0.25	0.619 ± 0.010	**0.638 ± 0.008**
0.5	**0.626 ± 0.009**	0.616 ± 0.009

Table 3

Examination of the best parameters $\mathcal{W}_J^{(1)}$ versus $\mathcal{W}_J^{(2)}$ and $\alpha \in \{-0.5, -0.25, 0, 0.25, 0.5\}$, showing the number of times each setting was the best performing over 11 data sets and 10 seeds. We find that, on average, $\mathcal{W}_J^{(1)}$ outperforms $\mathcal{W}_J^{(2)}$ (64 to 46) and that $\alpha \in \{-0.25, 0.25, 0.5\}$ outperforms $\alpha \in \{-0.5, 0.0\}$ (28 to 14).

α	$\mathcal{W}_J^{(1)}$	$\mathcal{W}_J^{(2)}$	Total
-0.5	12	2	14
-0.25	12	16	28
0.0	8	5	13
0.25	13	14	27
0.5	19	9	28
Total	64	46	110

Table 4

Test accuracy ($\mu \pm \sigma$) over 11 data sets and 10 seeds of $\mathcal{W}_J^{(1)}$ scattering computation. "Exact" computation of $\mathcal{W}_J^{(1)}$ scattering requires an eigendecomposition of T and therefore $O(n^3)$ time. Computation with the Chebyshev polynomial of order τ on a graph with $|E|$ nonzero edges takes $O(\tau(n + |E|))$ time, which is efficient for sparse graphs with $|E| \approx O(n)$. We do not see a noticeable drop in performance when approximating with $\tau \in \{10, 100\}$ as compared to the exact implementation, suggesting a fast implementation of approximate $\mathcal{W}_J^{(1)}$ scattering using Chebyshev approximation.

τ	10	100	exact
COLLAB	0.692 ± 0.010	0.683 ± 0.012	0.702 ± 0.009
DD	0.685 ± 0.025	0.698 ± 0.024	0.695 ± 0.039
IMDB-BINARY	0.688 ± 0.040	0.666 ± 0.050	0.691 ± 0.031
IMDB-MULTI	0.406 ± 0.046	0.414 ± 0.029	0.398 ± 0.027
MUTAG	0.761 ± 0.074	0.733 ± 0.063	0.688 ± 0.070
NCI1	0.653 ± 0.016	0.659 ± 0.029	0.645 ± 0.030
NCI109	0.668 ± 0.022	0.628 ± 0.013	0.658 ± 0.023
PROTEINS	0.772 ± 0.023	0.799 ± 0.016	0.782 ± 0.022
PTC MR	0.325 ± 0.095	0.333 ± 0.097	0.387 ± 0.089
REDDIT-BINARY	0.805 ± 0.021	0.822 ± 0.021	0.814 ± 0.021
REDDIT-MULTI-5K	0.410 ± 0.013	0.422 ± 0.013	0.417 ± 0.016
Mean	0.624 ± 0.163	0.624 ± 0.160	0.626 ± 0.151

of scattering transforms, which a practitioner can choose from via a validation procedure. When averaged over all data sets, the best choice of α is either -0.25 or 0.25, depending on whether one used $\mathcal{W}_J^{(1)}$ or $\mathcal{W}_J^{(2)}$, as shown in Table 2. We also note that $\mathcal{W}_J^{(1)}$ outperforms

Table 5

Full results, $(\mu \pm \sigma)$ over 10 seeds for 11 data sets with varying α, $\{\mathcal{W}_J^{(1)}, \mathcal{W}_J^{(2)}\}$, and Chebyshev approximation of $\mathcal{W}_J^{(1)}$ with 10 and 100 degree Chebyshev polynomials.

α	model	COLLAB	DD	IMDB-BINARY	IMDB-MULTI	MUTAG	NCI1
-0.5	$\mathcal{W}_J^{(1)}$ $(\tau = 10)$	0.699 ± 0.005	0.654 ± 0.021	**0.740 ± 0.022**	0.384 ± 0.032	0.795 ± 0.064	0.657 ± 0.011
	$\mathcal{W}_J^{(1)}$ $(\tau = 100)$	0.676 ± 0.008	0.721 ± 0.017	0.727 ± 0.021	0.438 ± 0.026	0.725 ± 0.035	0.681 ± 0.022
	$\mathcal{W}_J^{(1)}$ (exact)	0.706 ± 0.006	0.727 ± 0.013	0.698 ± 0.026	0.397 ± 0.023	0.600 ± 0.000	0.622 ± 0.006
	$\mathcal{W}_J^{(2)}$	0.699 ± 0.009	0.660 ± 0.014	0.588 ± 0.012	0.335 ± 0.043	0.775 ± 0.049	0.651 ± 0.006
-0.25	$\mathcal{W}_J^{(1)}(\tau = 10)$	0.684 ± 0.008	0.683 ± 0.013	0.655 ± 0.036	0.438 ± 0.033	0.660 ± 0.070	0.642 ± 0.018
	$\mathcal{W}_J^{(1)}$ $(\tau = 100)$	0.683 ± 0.007	0.694 ± 0.026	0.635 ± 0.051	0.389 ± 0.016	0.700 ± 0.062	0.621 ± 0.014
	$\mathcal{W}_J^{(1)}$ (exact)	0.701 ± 0.010	0.711 ± 0.026	0.692 ± 0.038	0.373 ± 0.014	0.770 ± 0.026	0.650 ± 0.004
	$\mathcal{W}_J^{(2)}$	0.693 ± 0.007	0.674 ± 0.011	0.579 ± 0.017	0.401 ± 0.045	0.740 ± 0.039	0.649 ± 0.009
0.0	$\mathcal{W}_J^{(1)}(\tau = 10)$	0.686 ± 0.010	0.699 ± 0.017	0.683 ± 0.019	**0.444 ± 0.030**	0.750 ± 0.035	0.663 ± 0.022
	$\mathcal{W}_J^{(1)}$ $(\tau = 100)$	0.698 ± 0.005	0.696 ± 0.028	0.695 ± 0.012	0.402 ± 0.014	0.735 ± 0.034	0.637 ± 0.017
	$\mathcal{W}_J^{(1)}$ (exact)	0.690 ± 0.007	0.714 ± 0.015	0.713 ± 0.021	0.421 ± 0.027	0.675 ± 0.063	0.638 ± 0.006
	$\mathcal{W}_J^{(2)}$	0.694 ± 0.007	0.720 ± 0.017	0.584 ± 0.019	0.428 ± 0.016	0.640 ± 0.021	0.667 ± 0.004
0.25	$\mathcal{W}_J^{(1)}$ $(\tau = 10)$	0.701 ± 0.010	0.697 ± 0.023	0.688 ± 0.044	0.423 ± 0.014	0.775 ± 0.026	0.650 ± 0.009
	$\mathcal{W}_J^{(1)}$ $(\tau = 100)$	0.690 ± 0.005	0.686 ± 0.021	0.667 ± 0.012	0.403 ± 0.036	0.725 ± 0.089	0.682 ± 0.006
	$\mathcal{W}_J^{(1)}$ (exact)	0.705 ± 0.005	0.675 ± 0.041	0.684 ± 0.015	0.406 ± 0.016	0.690 ± 0.052	0.617 ± 0.011
	$\mathcal{W}_J^{(2)}$	0.700 ± 0.007	**0.732 ± 0.011**	0.609 ± 0.032	0.427 ± 0.013	**0.845 ± 0.064**	0.658 ± 0.010
0.5	$\mathcal{W}_J^{(1)}$ $(\tau = 10)$	0.690 ± 0.007	0.695 ± 0.021	0.680 ± 0.025	0.346 ± 0.027	0.825 ± 0.035	0.654 ± 0.006
	$\mathcal{W}_J^{(1)}$ $(\tau = 100)$	0.670 ± 0.009	0.695 ± 0.015	0.607 ± 0.008	0.434 ± 0.012	0.780 ± 0.059	0.676 ± 0.008
	$\mathcal{W}_J^{(1)}$ (exact)	**0.708 ± 0.006**	0.651 ± 0.031	0.666 ± 0.034	0.395 ± 0.031	0.705 ± 0.055	**0.696 ± 0.010**
	$\mathcal{W}_J^{(2)}$	0.705 ± 0.010	0.724 ± 0.021	0.608 ± 0.031	0.419 ± 0.045	0.590 ± 0.077	0.659 ± 0.011

α	model	NCI109	PROTEINS	PTC MR	REDDIT-B	REDDIT-M	Mean
-0.5	$\mathcal{W}_J^{(1)}$ $(\tau = 10)$	0.631 ± 0.008	0.772 ± 0.012	0.422 ± 0.106	0.794 ± 0.009	0.403 ± 0.007	0.631 ± 0.016
	$\mathcal{W}_J^{(1)}$ $(\tau = 100)$	0.627 ± 0.009	0.790 ± 0.014	0.292 ± 0.057	0.821 ± 0.021	0.412 ± 0.012	0.627 ± 0.008
	$\mathcal{W}_J^{(1)}$ (exact)	0.624 ± 0.010	0.798 ± 0.013	0.423 ± 0.056	0.789 ± 0.014	0.403 ± 0.007	0.617 ± 0.007
	$\mathcal{W}_J^{(2)}$	0.663 ± 0.010	0.802 ± 0.016	0.366 ± 0.072	0.814 ± 0.011	0.416 ± 0.007	0.616 ± 0.012
-0.25	$\mathcal{W}_J^{(1)}(\tau = 10)$	0.661 ± 0.007	0.790 ± 0.027	0.251 ± 0.044	0.793 ± 0.008	0.406 ± 0.012	0.607 ± 0.008
	$\mathcal{W}_J^{(1)}$ $(\tau = 100)$	0.615 ± 0.015	0.803 ± 0.013	0.286 ± 0.096	0.820 ± 0.023	0.424 ± 0.009	0.609 ± 0.014
	$\mathcal{W}_J^{(1)}$ (exact)	0.676 ± 0.004	0.791 ± 0.009	**0.456 ± 0.020**	0.808 ± 0.017	0.410 ± 0.016	**0.640 ± 0.005**
	$\mathcal{W}_J^{(2)}$	0.680 ± 0.008	0.791 ± 0.038	0.402 ± 0.051	0.841 ± 0.007	**0.435 ± 0.009**	0.626 ± 0.009
0.0	$\mathcal{W}_J^{(1)}(\tau = 10)$	0.675 ± 0.009	0.780 ± 0.024	0.292 ± 0.090	0.791 ± 0.013	0.419 ± 0.006	0.624 ± 0.010
	$\mathcal{W}_J^{(1)}$ $(\tau = 100)$	0.631 ± 0.013	0.797 ± 0.013	0.289 ± 0.069	0.814 ± 0.014	0.430 ± 0.013	0.620 ± 0.009
	$\mathcal{W}_J^{(1)}$ (exact)	0.641 ± 0.004	0.778 ± 0.024	0.384 ± 0.097	0.820 ± 0.009	0.419 ± 0.011	0.626 ± 0.012
	$\mathcal{W}_J^{(2)}$	0.647 ± 0.008	0.783 ± 0.017	0.443 ± 0.054	0.825 ± 0.009	0.421 ± 0.014	0.623 ± 0.006
0.25	$\mathcal{W}_J^{(1)}$ $(\tau = 10)$	**0.690 ± 0.006**	0.759 ± 0.019	0.294 ± 0.018	0.812 ± 0.011	0.412 ± 0.017	0.627 ± 0.003
	$\mathcal{W}_J^{(1)}$ $(\tau = 100)$	0.627 ± 0.005	**0.805 ± 0.007**	0.375 ± 0.119	**0.844 ± 0.014**	0.418 ± 0.012	0.632 ± 0.013
	$\mathcal{W}_J^{(1)}$ (exact)	0.673 ± 0.004	0.787 ± 0.012	0.298 ± 0.102	0.842 ± 0.014	0.433 ± 0.018	0.619 ± 0.010
	$\mathcal{W}_J^{(2)}$	0.654 ± 0.010	0.796 ± 0.019	0.379 ± 0.022	0.822 ± 0.013	0.398 ± 0.010	0.638 ± 0.008
0.5	$\mathcal{W}_J^{(1)}$ $(\tau = 10)$	0.682 ± 0.011	0.758 ± 0.013	0.367 ± 0.083	0.837 ± 0.015	0.412 ± 0.014	0.631 ± 0.009
	$\mathcal{W}_J^{(1)}$ $(\tau = 100)$	0.640 ± 0.007	0.800 ± 0.025	0.422 ± 0.048	0.812 ± 0.017	0.429 ± 0.012	0.633 ± 0.006
	$\mathcal{W}_J^{(1)}$ (exact)	0.679 ± 0.006	0.756 ± 0.020	0.370 ± 0.063	0.813 ± 0.010	0.420 ± 0.008	0.626 ± 0.009
	$\mathcal{W}_J^{(2)}$	0.642 ± 0.005	0.794 ± 0.016	0.453 ± 0.017	0.795 ± 0.019	0.391 ± 0.006	0.616 ± 0.009

$\mathcal{W}_J^{(2)}$ in a majority of cases, as shown in Table 3. Notably, one potential drawback to $\mathcal{W}_J^{(1)}$ is that it requires computing the eigenvectors and eigenvalues of $\mathbf{K}$, which is inefficient for large graphs. However, we show that this difficulty can be overcome via an approximation by Chebyshev polynomials with only a small loss of accuracy, as shown in Table 4.

Specifically, an exact application of the $\mathcal{W}_J^{(1)}$ filters requires an eigendecomposition of the graph, while the $\mathcal{W}_J^{(2)}$ filter can be calculated with a polynomial of the Laplacian.

For large graphs with n nodes and $|E| \leq n^2$ edges, the $\mathcal{W}_J^{(1)}$ filters take $O(n^3)$ time to compute, whereas the $\mathcal{W}_J^{(2)}$ filters take $O(2^J(|E|+n))$ time to compute. This means that $\mathcal{W}_J^{(2)}$ filters are substantially quicker to compute than exact $\mathcal{W}_J^{(1)}$ filters, especially for sparse graphs and $2^J \ll n$. However, we show that the $\mathcal{W}_J^{(1)}$ filter can be approximated with a Chebyshev polynomial approximation of order τ, which has a similar computation time to the $\mathcal{W}_J^{(2)}$ filter with no noticeable performance drop. In Table 4, we compare the performance of the exact $\mathcal{W}_J^{(1)}$ implementation versus a Chebyshev approximation to the filter of order $\tau \in \{10, 100\}$ over 11 data sets. On average, over the data sets, the exact $\mathcal{W}_J^{(1)}$ filter has a test accuracy of 0.626 versus 0.624 for both $\tau = 10$ and $\tau = 100$.

7. Future work. We have introduced a large class of scattering networks with provable guarantees. As alluded to in section 1.2, we believe that our work opens up several new lines of inquiry. One might attempt to learn the optimal choices of the matrix $\mathbf{M}$ and the spectral function g from a parameterized family based on training data, yielding a data-driven architecture with theoretical stability guarantees. Another possible extension would be to consider a construction similar to ours but which uses the spectral decomposition of the unnormalized graph Laplacian rather than the normalized Laplacian. Such a work would generalize [32] in a manner analogous to the way that this work generalizes [12] and [13]. Additionally, one might attempt to incorporate an attention mechanism, such as that used in GAT [29], into the scattering framework for improved numerical performance. Finally, one might wish to study the behavior of the graph scattering transform on data-driven graphs obtained by subsampling a Riemannian manifold $\mathcal{M}$. Such graphs typically arise in high-dimensional data analysis and manifold learning. It can be shown that, under certain conditions, the normalized graph Laplacian converges pointwise [8, 26] or in a spectral sense [2, 3, 10, 25, 28] to the Laplace–Beltrami operator on $\mathcal{M}$ as the number of samples tends to infinity. One might hope to use these results to study the convergence of the graph scattering transforms constructed here to the manifold scattering transform constructed in [24].

REFERENCES

[1] J. Andén and S. Mallat, *Multiscale scattering for audio classification*, in 12th International Society for Music Information Retrieval Conference, 2011, pp. 657–662.

[2] M. Belkin and P. Niyogi, *Convergence of Laplacian eigenmaps*, in Advances in Neural Information Processing Systems, 2007, pp. 129–136.

[3] D. Burago, S. Ivanov, and Y. Kurylev, *A graph discretization of the Laplace-Beltrami operator*, J. Spectr. Theory, 4 (2013), pp. 675–714, https://doi.org/10.4171/JST/83.

[4] E. Castro, A. Benz, A. Tong, G. Wolf, and S. Krishnaswamy, *Uncovering the folding landscape of RNA secondary structure using deep graph embeddings*, in 2020 IEEE International Conference on Big Data, 2020, pp. 4519–4528, https://doi.org/10.1109/BigData50022.2020.9378305.

[5] X. Chen, X. Cheng, and S. Mallat, *Unsupervised deep Haar scattering on graphs*, in Conference on Neural Information Processing Systems, Vol. 27, 2014, pp. 1709–1717.

[6] X. Cheng, X. Chen, and S. Mallat, *Deep Haar scattering networks*, Inf. Inference, 5 (2016), pp. 105–133.

[7] V. Chudacek, R. Talmon, J. Anden, S. Mallat, R. R. Coifman, P. Abry, and M. Doret, *Low dimensional manifold embedding for scattering coefficients of intrapartum fetale heart rate variability*, in 2014 International IEEE Conference in Medicine and Biology, 2014, pp. 6373–6376.

[8] R. R. Coifman and S. Lafon, *Diffusion maps*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 5–30.
[9] M. Defferrard, X. Bresson, and P. Vandergheynst, *Convolutional neural networks on graphs with fast localized spectral filtering*, in Advances in Neural Information Processing Systems, Vol. 29, 2016, pp. 3844–3852.
[10] K. Fujiwara, *Eigenvalues of Laplacians on a closed Riemannian manifold and its nets*, Proc. Amer. Math. Soc., 123 (1995), pp. 2585–2594.
[11] F. Gama, J. Bruna, and A. Ribeiro, *Stability of graph scattering transforms*, in Advances in Neural Information Processing Systems, Vol. 33, 2019, pp. 8038–8048.
[12] F. Gama, A. Ribeiro, and J. Bruna, *Diffusion scattering transforms on graphs*, in International Conference on Learning Representations, 2019.
[13] F. Gao, G. Wolf, and M. Hirn, *Geometric scattering for graph data analysis*, in Proceedings of the 36th International Conference on Machine Learning, Vol. 97, 2019, pp. 2122–2131.
[14] M. Hirn, S. Mallat, and N. Poilvert, *Wavelet scattering regression of quantum chemical energies*, Multiscale Model. Simul., 15 (2017), pp. 827–863.
[15] T. Kipf and M. Welling, *Semi-supervised classification with graph convolutional networks*, in Proceedings of the International Conference on Learning Representations, 2016.
[16] R. Levie, M. Bronstein, W. Huang, L. Bucci, and G. Kutyniok, *Transferability of Spectral Graph Convolutional Neural Networks*, preprint, arXiv:1907.12972v2, 2020.
[17] R. Levie, F. Monti, X. Bresson, and M. M. Bronstein, *Cayleynets: Graph convolutional neural networks with complex rational spectral filters*, IEEE Trans. Signal Process., 67 (2018), pp. 97–109.
[18] S. Mallat, *Group invariant scattering*, Comm. Pure Appl. Math., 65 (2012), pp. 1331–1398.
[19] Y. Min, F. Wenkel, M. Perlmutter, and G. Wolf, *Can Hybrid Geometric Scattering Networks Help Solve the Maximal Clique Problem?*, preprint, https://arxiv.org/abs/2206.01506v2, 2022.
[20] Y. Min, F. Wenkel, and G. Wolf, *Scattering GCN: Overcoming oversmoothness in graph convolutional networks*, in Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 14498–14508.
[21] Y. Min, F. Wenkel, and G. Wolf, *Geometric scattering attention networks*, in Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, 2021, pp. 8518–8522.
[22] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, *Diffusion maps, spectral clustering and reaction coordinates of dynamical systems*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 113–127.
[23] E. Oyallon and S. Mallat, *Deep roto-translation scattering for object classification*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2865–2873.
[24] M. Perlmutter, F. Gao, G. Wolf, and M. Hirn, *Geometric scattering networks on compact Riemannian manifolds*, in Mathematical and Scientific Machine Learning Conference, 2020, pp. 570–604.
[25] Z. Shi, *Convergence of Laplacian Spectra from Random Samples*, preprint, arXiv:1507.00151, 2015.
[26] A. Singer, *From graph to manifold Laplacian: The convergence rate*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 128–134.
[27] A. Tong, F. Wenkel, K. Macdonald, S. Krishnaswamy, and G. Wolf, *Data-driven learning of geometric scattering modules for GNNs*, in 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing, 2021, pp. 1–6, https://doi.org/10.1109/MLSP52302.2021.9596169.
[28] N. Trillos, M. Gerlach, M. Hein, and D. Slepcev, *Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace-Beltrami operator*, Found. Comput. Math., 20 (2018), pp. 827–887, https://doi.org/10.1007/s10208-019-09436-w.
[29] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, *Graph attention networks*, in International Conference on Learning Representations, 2018.
[30] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, *How powerful are graph neural networks?*, in International Conference on Learning Representations, 2019.
[31] D. Zou and G. Lerman, *Encoding robust representation for graph generation*, in International Joint Conference on Neural Networks, 2019, pp. 1–9.
[32] D. Zou and G. Lerman, *Graph convolutional neural networks via scattering*, Appl. Comput. Harmon. Anal., 49 (2019), pp. 1046–1074.