Journal of Machine Learning Research 22 (2021) 1-59 Submitted 3/20; Revised 6/21; Published 11/21

Transferability of Spectral Graph Convolutional Neural Networks

Ron Levie LEVIE@MATH.LMU.DE
Department of Mathematics
Ludwig-Maximilians-Universität München
80333 München, Germany

Wei Huang WEI.HUANG@USI.CH
Institute of Computational Science
Università della Svizzera italiana
6900 Lugano, Switzerland

Lorenzo Bucci LORENZO.BUCCI@USI.CH
Institute of Computational Science
Università della Svizzera italiana
6900 Lugano, Switzerland

Michael Bronstein M.BRONSTEIN@IMPERIAL.AC.UK
Department of Computing
Imperial College London
London SW7 2BU, United Kingdom

Gitta Kutyniok KUTYNIOK@MATH.LMU.DE
Department of Mathematics
Ludwig-Maximilians-Universität München
80333 München, Germany

Editor: Stefanie Jegelka

Abstract

This paper focuses on spectral graph convolutional neural networks (ConvNets), where filters are defined as elementwise multiplication in the frequency domain of a graph. In machine learning settings where the data set consists of signals defined on many different graphs, the trained ConvNet should generalize to signals on graphs unseen in the training set. It is thus important to transfer ConvNets between graphs. Transferability, which is a certain type of generalization capability, can be loosely defined as follows: if two graphs describe the same phenomenon, then a single filter or ConvNet should have similar repercussions on both graphs. This paper aims at debunking the common misconception that spectral filters are not transferable. We show that if two graphs discretize the same "continuous" space, then a spectral filter or ConvNet has approximately the same repercussion on both graphs. Our analysis is more permissive than the standard analysis. Transferability is typically described as the robustness of the filter to small graph perturbations and re-indexing of the vertices. Our analysis accounts also for large graph perturbations. We prove transferability between graphs that can have completely different dimensions and topologies, only requiring that both graphs discretize the same underlying space in some generic sense.

Keywords: graph convolutional neural network, spectral method, generalization, transferability, stability

1. Introduction

The success of convolutional neural networks (ConvNets) on Euclidean domains ignited an interest in recent years in extending these methods to graph structured data. In a standard ConvNet, the network receives as input a signal defined over a Euclidean rectangle, and at each layer applies a set of convolutions/filters on the outputs of the previous layer, a non linear activation function, and, optionally, pooling. A graph ConvNet has the same architecture, with the only difference that now signals are defined over the vertices of graph domains, and not Euclidean rectangles. Graph structured data is ubiquitous in a range of applications, and can represent 3D shapes, molecules, social networks, point clouds, and citation networks to name a few.

In a machine learning setting, the general architecture of the ConvNet is fixed, but the specific filters to use in each layer are free parameters. In training, the filter coefficients are optimized to minimize some loss function. In some situations, both the graph and the signal defined on the graph are variables in the input space of the ConvNet. Namely, the data set consists of many different graphs, and many different signals on these graphs. We call such a scenario a **multi-graph setting**. In multi-graph settings, if two graphs represent the same underlying phenomenon, and the two signals given on the two graphs are similar in some sense, the output of the ConvNet on both signals should be similar as well. This property is typically termed transferability, and is an essential requirement if we wish the ConvNet to generalize well on the test set, which in general consists of graphs unseen in the training set. In fact, transferability can be seen as a special type of generalization capability. Analyzing and proving transferability is the focus of this paper.

1.1 Convolutional Neural Networks

A classical 1D convolution neural network, as described above, can be written explicitly as follows. We call each application of filters, followed by the activation function and pooling a **layer**. We consider discrete input signals $\mathbf{f} \in \mathbb{R}^{d_1}$, seen as the samples of a continuous signal $f : \mathbb{R} \to \mathbb{R}$ at d_1 sample points. In each Layer $l = 1, \ldots, L$ there are $K_l \in \mathbb{N}$ signal channels. The convolution-operators/filters of the ConvNet map the signal channels of each Layer $l-1$ to the signal channels of Layer l. Moreover, as the layers increase, we consider coarser discrete signals. Namely, signals of Layer l consist of d_l samples, where $d_1 \geq d_2 \geq \ldots \geq d_L$. Consider the affine-linear filters

$$\{g^l_{k'k} \mid {\scriptstyle k=1\ldots K_{l-1},\ k'=1\ldots K_l}\}$$

of Layer $l-1$, and the matrix $A^l = \{a^l_{k'k}\}_{k'k} \in \mathbb{R}^{K_l \times K_{l-1}}$ that mixes the $K_{l-1} \times K_l$ resulting output signals to the K_l channels of Layer l. Note that each $g^l_{k'k}$ denotes a convolution operator plus constant. Denote the signals at Layer l by $\{\mathbf{f}^l_{k'}\}_{k'=1}^{K_l}$. The ConvNet maps Layer $l-1$ to Layer l by

$$\{\mathbf{f}^l_{k'}\}_{k'=1}^{K_l} = Q^l\Big(\rho\Big\{\sum_{k=1}^{K_{l-1}} a^l_{k'k}\ g^l_{k'k}(\mathbf{f}^{l-1}_k)\Big\}_{k'=1}^{K_l}\Big),$$

where $\rho : \mathbb{R} \to \mathbb{R}$, called the **activation function**, operates pointwise on vectors, and the **pooling operator** $Q^l : \mathbb{R}^{d_{l-1}} \to \mathbb{R}^{d_l}$ sub-samples signals from $\mathbb{R}^{d_{l-1}}$ to $\mathbb{R}^{d_l}$. A typical choice for ρ is the ReLU function $\rho(x) = \max\{0, x\}$. The output of the ConvNet are the signals $\{\mathbf{f}^L_{k'}\}_{k'=1}^{K_L}$ at Layer L.

When generalizing this architecture to graphs, there is a need to extend the convolution, activation function, and pooling to graph structured data. Here, graph signals are mappings that assign to each vertex of a graph a value. The activation function operates pointwise on signals, and generalizes trivially to graph signals. For pooling, graph signals are subsampled to signals over coarsened graphs, typically via the Graclus algorithm (Dhillon et al., 2004) (see also (Defferrard et al., 2016, Subsection 2.2)). Next, we explain how filters are generalized to graphs.

1.2 Convolution Operators on Graphs

There are generally two approaches to defining convolution operators on graphs, both generalizing the standard convolution on Euclidean domains (Bronstein et al., 2017; Wu et al., 2020). Spatial approaches generalize the idea of a sliding window to graphs. Here, the main challenge is to define a way to translate a filter kernel along the vertices of the graph, or to aggregate feature information from the neighbors of each node. Some popular examples of spatial methods are (Gori et al., 2005; Scarselli et al., 2009; Monti et al., 2017). Spectral methods are inspired by the convolution theorem in Euclidean domains, that states that convolution in the spatial domain is equivalent to pointwise multiplication in the frequency domain. The challenge here is to define the frequency domain and the Fourier transform of graphs. The basic idea is to define the graph Laplacian, or some other graph operator that we interpreted as a shift operator, and to use its eigenvalues as frequencies and its eigenvectors as the corresponding pure harmonics (Ortega et al., 2018). Decomposing a graph signal to its pure harmonic coefficients is by definition the graph Fourier transform, and filters are defined by multiplying the different frequency components by different values, see Subsection 2.1 for more details. For some examples of spectral methods we refer to (Bruna et al., 2013; Defferrard et al., 2016; Levie et al., 2019b; Gama et al., 2018). Additional references for both methods can be found in (Wu et al., 2020).

One typical motivation for favoring spatial methods is the claim that spectral methods are not transferable, and thus do not generalize well on graphs unseen in the training set. The goal in this paper is to debunk this misconception, and to show that state-of-the-art spectral graph filtering methods are transferable. This paper does not argue against spatial methods, but shows the potential of spectral approaches to cope with data sets having varying graphs. We would like to encourage researches to reconsider spectral methods in such situations. Interestingly, Bianchi et al. (2021) obtained state-of-the-art results using spectral graph filters on variable graphs, without any modification to compensate for the "non-transferability".

1.3 Stability of Spectral Methods

A necessary condition of any reasonable definition of transferability is stability. Namely, given a filter, if the topology of a graph is perturbed, then the filter on the perturbed graph is close to the filter on the un-perturbed graph. Without stability it is not even possible to transfer a filter from a graph to another very close graph, and thus stability is necessary for transferability. Previous work studied the behavior of graph filters with respect to variations in the graph. Segarra et al. (2017) provided numerical results on the robustness of polynomial graph filters to additive Gaussian perturbations of the eigenvectors of the graph Laplacian. Since the eigendecomposition is not stable to perturbations in the topology of the graph, this result does not prove robustness to such perturbations. Isufi et al. (2017b) showed that the expected graph filter under random edge losses is equal to the accurate output. However, Isufi et al. (2017b) did not bound the error in the output in terms of the error in the graph topology. Gama et al. (2019b) studied the stability with respect to diffusion distance of diffusion scattering transforms on graphs, a graph version of the popular scattering transforms, which are pre-defined Euclidean domain ConvNets (Bruna and Mallat, 2013). Zou and Lerman (2020) also studied stability of graph scattering transforms, in terms of perturbations in the Laplacian eigenvectors and vertex permutations. Recently, Gama et al. (2020) studied stability properties of spectral graph filters of a fixed number of vertices. However, in (Gama et al., 2020, Theorems 2 and 3) the assumption that the relative error matrix is normal and is close to a scaled identity matrix is restrictive, and not satisfied in the generic case. In particular, only perturbations which are approximately a multiplication of all of the edge weights by the same scalar are considered in these theorems. A similar restriction is implicit in the analysis of Gama et al. (2019a), which studied stability of graph scattering transforms. Kostrikov et al. (2018) analyzed the stability of a special type of ConvNet on triangle meshes, where filtering is pre-defined via propagating information from vertices to faces and back using the Dirac operator. The error of the ConvNet between two polygon meshes discretizing the same surface was bounded, assuming the two meshes consist of the same number of vertices. This approach to stability is reminiscent of our approach, but in our analysis we do not assume that the two graphs consist of the same number of vertices. Moreover, we consider general spectral graph ConvNets.

1.4 Our Contribution

In the following we summarize our contribution.

1.4.1 Theoretical Settings of Transferability

We prove in this paper the stability of graph spectral filters to general perturbations in the topology. In fact, we present a more permissive framework of transferability, allowing to compare graphs of incompatible sizes and topologies. We consider spectral filters as they are, and do not enhance them with any computational machinery for transferring filters. Thus, one of the main conceptual challenges is to find a way to compare two different graphs, with incompatible graph structures, from a theoretical stance. To accommodate the comparison of incompatible graphs, our approach resorts to non-graph theoretical considerations, assuming that graphs are observed from some underlying non-graph spaces. In our approach, graphs are regarded as discretizations of underlying corresponding “continuous” metric spaces. This makes sense, since a weighted graph can be interpreted as a set of points (vertices) and a decreasing function of their distances (edge weights). We can actually relax the assumption that the “continuous space” is metric, and consider more general topological

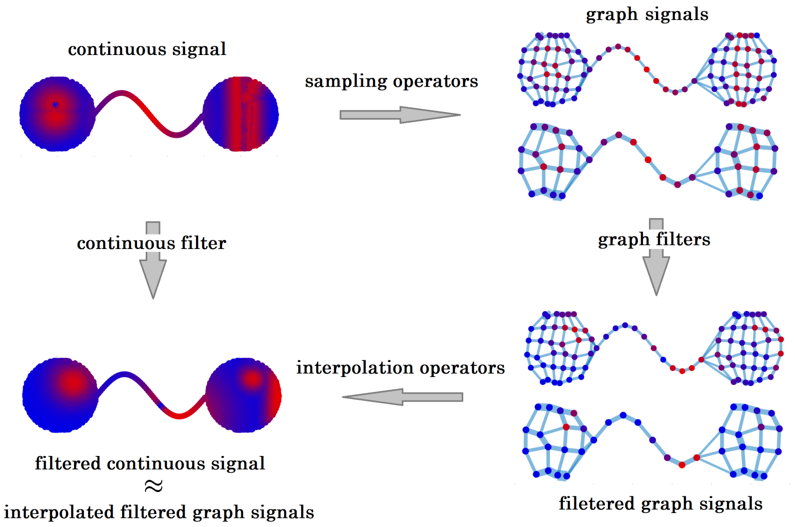

Figure 1: Diagram of the approximation procedure, illustrating how a fixed filter/ConvNet operates on a "continuous" topological space and two graphs discretizing it. Top left: a continuous signal on the topological space. Top right: the sampling of the continuous signal to the two graphs that discretize the topological space. Bottom right: the filter applied on both graph signals. Bottom left: the filter applied on the continuous topological space signal is approximated by the interpolation of either of the two filtered graph signals. As a result, the interpolations of the two filtered graph signals are approximately identical.

spaces[1]. Two graphs are comparable, or represent the same phenomenon, if both discretize the same space. This approach allows us to prove transferability under small perturbations of the adjacency matrix, but more generally, allows us to prove transferability between graphs with incompatible sizes.

More generally, we consider graph sampled from general measure spaces[2], where the **sampling operator** is a linear mapping that takes a signal on the measure space and returns a signal on the graph. We consider a corresponding **interpolation operator**, a linear mapping that takes a signal on the graph and returns a signal on the measure space. This setting is general, and can be used to describe graphs sampled from topological spaces at sample points, graphs coarsened to smaller graphs via, e.g., the Graclus algorithm (Dhillon et al., 2004), and graph perturbations, as discussed in Subsection 3.2.

[1] A topological space is a generalization of a metric space, where distances are no longer defined, but continuity is defined. In metric spaces $\mathcal{M}$, continuity of functions $f : \mathcal{M} \to \mathbb{R}$ is defined via an "ϵ—δ" formulation: f is continuous at $x \in \mathcal{M}$, if for every open interval $B_{\epsilon,f(x)} = \big(f(x)-\epsilon, f(x)+\epsilon\big)$ about $f(x)$, the inverse set $\{y \in \mathcal{M} \mid f(y) \in B_{\epsilon,f(x)}\}$ contains some open ball $B_{\delta,x} = \{y \in \mathcal{M} \mid \mathrm{dist}(y,x) < \delta\}$ about x. Topological spaces generalize the "ϵ—δ" notion of continuity by directly specifying which sets are open, without defining a notion of distance.

[2] A measure space is informally a space in which it is possible to compute the volume of a rich collection of subsets. Using the notion of volume, it is then possible to define integration of functions defined on the measure space, and thus the root mean square error between functions is well defined.

The way to compare two graphs is to consider their embeddings to the "continuous" space they both discretize. For intuition, consider the special case where the "continuous" space is a manifold. Any manifold can be discretized to a graph/polygon-mesh in many different ways, resulting in different graph topologies. A filter designed/learned on one polygon-mesh should have approximately the same repercussion on a different polygon-mesh discretizing the same manifold. The informal term "repercussion" means "the effect that a network/filter has on data." Choosing a rigorous definition for this term is a mathematical modeling challenge that we address as follows. To compare the filter on the two graphs, we consider a generic signal defined on the continuous space, and sampled to both graphs. After applying the graph filter on the sampled signal on both graphs, we interpolate the results back to two continuous signals. In our analysis we show that these two interpolated continuous signals are approximately equal (see Figure 1 for illustration of this procedure).

For the case of graphs sampled from topological spaces, we develop a digital signal processing (DSP) framework akin to the classical Nyquist—Shannon approach, where now analog domains are topological spaces, and digital domains are graphs.

1.4.2 The Basic Assumption of Graphs Discretizing Topological Spaces

In the DSP setting of transferability, the assumption that graphs are discretizations of topological spaces is an ansatz, and it is important to clarify the philosophy behind this choice. One of the fundamental challenges in studying transferability is to determine to which graph changes a network should be sensitive/discriminative and to which changes the network should generalize, or be transferable. The later changes are sometimes termed nuisances in the machine learning jargon, since the network should be designed/trained to ignore them. A network should not be transferable to all graph changes, since then the network cannot be used to discriminate between different types of graphs. On the other hand, the network should be transferable between different graphs that represent the same underlying phenomenon, even if these two graphs are not close to each other in standard measures of graph distance. The ansatz that two graphs represent the same phenomenon if both discretize the same topological space, gives us a theoretical starting point: we know to which graph changes the network should be transferable, so the problem of transferability can be formulated mathematically. What we show is that spectral graph ConvNets always generalize between graphs discretizing the same topological space, regardless of the specific form of their filters. Namely, this type of generalization is built-in to spectral graph ConvNets, and requires no training.

The validity of this ansatz from a modeling stance is justifiable to different extents, depending on the situation. As noted above, it is natural to think of graphs as discretizations of metric spaces. Certainly, this is the case for geometric data sets like meshes, or 3D solids like molecules. There is also evidence that real life networks, like World Wide Web, social networks, protein interaction networks, and biological cellular networks, have underlying geometric structures. For example, in (Song et al., 2005) it was shown that such networks are self similar, in the sense that the coarsened version of the network has the same probability distribution of links as the fine network. Hence, a network and its coarsened version both represent the same underlying phenomenon. It is thus desirable for graph ConvNets to have the same effect on both the original and the coarsened graph in some sense. Follow up works showed that networks can be seen as sampled from a latent underlying geometric space, e.g., a hyperbolic space (Krioukov et al., 2010), or a circle (Serrano et al., 2008). For a comprehensive survey on the underlying geometry of networks we refer the reader to (Boguñá et al., 2021).

One might even stretch the interpretation further, and consider examples like citation networks[3], which seem non-geometric. The idea is to view citation networks as discretizations of some hypothetical underlying metric space. This metric space is the continuous limit of citation networks, where the number of papers tends to infinity. Intuitively, in the limit there is a continuum of papers, and the distance between papers models the probability for the two papers to be linked by a citation. Namely, the distance decreases to zero as the probability increases to one. We do not attempt to study or characterize this hypothetical continuous citation network, but only postulate its existence as a metric space. In practice, the computations in training and applying filters do not use any knowledge of the underlying continuous metric space. Its existence is used only for approximation theoretic analysis.

Other notions of graphs approximating continuous latent spaces are possible. For example, in graphon analysis, simple graphs approximate graphons if the homomorphism densities of the graph and of the graphon are close (Borgs et al., 2008). In this paper we focus however on the sampling approach, leaving the graphon approach for future research.

1.4.3 Concept-based and Principle Transferability

Graph ConvNets can manage transferability in different ways. First, when a graph ConvNet is shown a multi-graph training set, it can learn “concepts” that promote transferability. Let us call this approach **concept-based transferability**. Second, it may be the case that transferability is a mathematical law: a built-in capability of certain types of graph ConvNets, independent of their specific filters, which requires no training. This approach, that we call **principle transferability**, is the focus of this paper.

We believe that the success of spectral graph ConvNets in multi-graph settings relies on both types of transferability. We call the accumulative effect of concept-based transferability and principle transferability **total transferability**. In this paper we prove theoretically that spectral graph ConvNets have principle transferability. We moreover demonstrate principle transferability by concocting experiments that isolate principle transferability from concept-based transferability. This is done by zero shot learning: training the network on one single graph, which prevents it from learning concepts for dealing with varying graphs, and testing the resulting network on other graphs. The performance of such a network on the new graphs only partially degrades, illustrating the effect size of principle transferability in total transferability. Moreover, in our isolated principle transferability experiment, spectral methods outperform spatial methods, which indicates that spectral methods have competitive transferability capabilities.

[3] A citation network is a graph, where each node represents a paper. Two nodes are connected by an edge if there is a citation between the papers. A graph signal is constructed by mapping the content of each paper to a vector representing this content.

1.4.4 Overview of Our Transferability Results

In the following we give a high-level overview of our results.

The transferability inequality. In the transferability theory there is always an original space with an original Laplacian, from which we sample a graph and a graph Laplacian. As explained in Subsection 1.4.1, the original space may be a "continuous" measure space or a discrete graph. Let us call the original space the **continuous space**, and the original Laplacian the **continuous Laplacian**. A **transferability error** is the error between the continuous object and the discretized object. In Section 3 we introduce the **transferability inequality** (Theorem 4), a generic inequality that bounds the **transferability error of filters** in terms of the **transferability error of Laplacians** and the error entailed by sampling-interpolating, called the **consistency error**. Informally, the transferability inequality reads

$$transferability\ of\ filter \quad \leq \quad transferability\ of\ Laplacian \quad + \quad consistency\ error.$$

The transferability inequality asserts that if sampling and interpolation is chosen well, in the sense that sampling a continuous signal and then interpolating it results in a small error, and if the graph Laplacian approximates the continuous Laplacian, then also any graph spectral filter approximates the corresponding filter on the continuous space.

Sufficient conditions for transferability. The transferability inequality states that the transferability of a filter is small if the transferability of the Laplacian and the consistency error are small. In Section 4 we introduce general conditions under which the transferability of the Laplacian and the consistency error are small.

Transferability of graph spectral ConvNets. In Subsection 4.2 we extend the transferability results of filters to transferability of spectral ConvNets. We prove the transferability of graph spectral ConvNets under the assumption of small transferability error of the Laplacian and small consistency error in each coarsened version of the graph in the network (Theorem 16 and Corollary 17). This implies that graph spectral ConvNets are appropriate in multi-graph settings. We support this claim both with basic experiments and by recalling other papers that demonstrate transferability of spectral methods in practice.

Transferability of graphs sampled from topological spaces. In Section 5 we prove that the sufficient conditions for transferability are satisfied for graphs discretizing topological spaces via sampling. To this end, we develop a digital signal processing (DSP) framework akin to the classical Nyquist—Shannon approach, where now analog domains are topological spaces, and digital domains are graphs. Graphs are sampled from topological spaces by evaluation at sample points. We prove that graph Laplacians approximate topological space Laplacians in case the sample points satisfy some quadrature assumptions, namely, if certain integrals over the topological space can be approximated by sums over the sample points.

Transferability of graphs randomly sampled from topological spaces. An important question that arises from the transferability inequality is if it is reasonable to assume that the right hand side of the transferability inequality is small. Another question is if the assumptions of the DSP setting of transferability are reasonable. The answer to these questions depends on the situation. A universal mathematical analysis is not possible, since the answer depends on how the graph data set was constructed, how graphs were sampled and from what model, and how the graph Laplacians were chosen. To give a mathematical solution to this question, in Subsection 5.4 we consider a controlled setting of the data acquisition step. We prove that the quadrature assumptions of our DSP framework are satisfied in high probability in case the sample points of the discrete graphs are drawn randomly from the corresponding topological space (Theorem 32). In this scenario, spectral ConvNets are transferable in high probability.

Main message. The concept that spectral graph ConvNets are not appropriate in situations where the data consists of many different graphs and many different signals on these graphs is a misconception. Graph spectral ConvNets are transferable both in practice and theory. If your data consists of many graphs, among other methods, you should consider spectral graph ConvNets.

All proofs are given in the appendix. We wish to remark that some preliminary results on stability of spectral convolutions of graphs of a fixed size were reported in (Levie et al., 2019a).

2. Theoretical Framework of Graph Spectral Methods

In this section we recall the theory of graph spectral methods. We show that state-of-the-art graph spectral methods are based on a **functional calculus** implementation of convolution operators, and explain the misconception of non-transferability of spectral graph filters. We last show how to use graph spectral methods for directed graphs.

2.1 Spectral Convolution Operators

Consider an undirected weighted graph $\mathcal{G} = \{\mathcal{E}, \mathcal{V}, \mathbf{W}\}$, with vertices $\mathcal{V} = \{1, \ldots, N\}$, edges $\mathcal{E} \subset \mathcal{V}^2$, and adjacency matrix $\mathbf{W}$. The adjacency matrix $\mathbf{W} = (w_{n,m})_{n,m=1}^N$ is symmetric and represents the weights of the edges, where $w_{n,m}$ is nonzero only if vertex n is connected to vertex m by an edge. Consider the degree matrix $\mathbf{D}$, defined as the diagonal matrix with entries $d_{n,n} = \sum_{m=1}^N w_{n,m}$.

The frequency domain of a graph is determined by choosing a shift operator, namely a self-adjoint operator $\boldsymbol{\Delta}$ that respects the connectivity of the graph. As a prototypical example, we consider the unnormalized Laplacian $\boldsymbol{\Delta} = \mathbf{D} - \mathbf{W}$, which depends linearly on $\mathbf{W}$. Other examples of common shift operators are the normalized Laplacian $\boldsymbol{\Delta}_{\mathrm{n}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, and the adjacency matrix itself. In this paper we call a generic self-adjoint shift operator **Laplacian**, and denote it by $\boldsymbol{\Delta}$. Denote the eigenvalues of $\boldsymbol{\Delta}$ by $\{\lambda_n\}_{n=1}^N$, and the eigenvectors by $\{\phi_n : V \to \mathbb{C}\}_{n=1}^N$. The Fourier transform of a graph signal $f : V \to \mathbb{C}$ is given by the vector of frequency intensities

$$\mathcal{F}f = (\langle f, \phi_n \rangle)_{n=1}^N,$$

where $\langle u, v \rangle$ is an inner product in $\mathbb{C}^N$, e.g., the standard dot product. The inverse Fourier transform of the vector $(v_n)_{n=1}^N$ is given by

$$\mathcal{F}^*(v_n)_{n=1}^N = \sum_{n=1}^N v_n \phi_n.$$

Since $\{\phi_n\}_{n=1}^N$ is an orthonormal basis, $\mathcal{F}^*$ is the inverse of $\mathcal{F}$. A spectral graph filter $\mathbf{G}$ based on the coefficients $(g_n)_{n=1}^N$ is defined by

$$\mathbf{G}f = \sum_{n=1}^{N} g_n \left\langle f, \phi_n \right\rangle \phi_n. \tag{1}$$

Any spectral filter defined by (1) is **permutation equivariant**, namely, does not depend on the indexing of the vertices. Re-indexing the vertices in the input results in the same re-indexing of vertices in the output.

Spectral filters implemented by (1) have two disadvantages. First, as shown in Subsection 2.3, they are not transferable. Second, they entail high computational complexity. Formula (1) requires the computation of the eigendecomposition of the Laplacian $\mathbf{\Delta}$, which is computationally demanding and can be unstable when the number of vertices N is large. Moreover, there is no general "graph FFT" algorithm for computing the Fourier transform of a signal $f \in L^2(V)$, and (1) requires computing the frequency components $\langle f, \phi_n \rangle$ and their summation directly.

2.2 Functional Calculus Implementation of Spectral Convolution Operators

To overcome the above two limitations, state-of-the-art methods, like (Defferrard et al., 2016; Isufi et al., 2017a; Levie et al., 2019b; Gama et al., 2018), are implemented via **functional calculus**. Functional calculus is the theory of applying functions $g : \mathbb{C} \to \mathbb{C}$ on normal operators in Hilbert spaces $\mathcal{H}$. In the special case of a self-adjoint or unitary operator $\mathbf{T}$ in the space $\mathcal{H}$, with a discrete spectrum, $g(\mathbf{T})$ is defined by

$$g(\mathbf{T})f = \sum_{n} g(\lambda_n) \left\langle f, \phi_n \right\rangle \phi_n, \tag{2}$$

for any vector f in the Hilbert space, where $\{\lambda_n, \phi_n\}$ is the eigendecomposition of the operator $\mathbf{T}$. The operator $g(\mathbf{T})$ is normal for general $g : \mathbb{C} \to \mathbb{C}$, self-adjoint for $g : \mathbb{C} \to \mathbb{R}$, and unitary for $g : \mathbb{C} \to e^{i\mathbb{R}}$ (where $e^{i\mathbb{R}}$ is the unit complex circle).

Definition (2) is canonical in the following sense. In the special case where

$$g(\lambda) = \frac{\sum_{l=0}^{L} c_l \lambda^l}{\sum_{l=0}^{L} d_l \lambda^l}$$

is a rational function, $g(\mathbf{T})$ can be defined in two ways. First, by (2), and second by compositions, linear combinations, and inversions, as

$$g(\mathbf{T}) = \Big(\sum_{l=0}^{L} c_l \mathbf{T}^l \Big) \Big(\sum_{l=0}^{L} d_l \mathbf{T}^l \Big)^{-1} \tag{3}$$

It can be shown that (2) and (3) are equivalent.

Moreover, definition (2) is also canonical in regard to non-rational functions. Loosely speaking, if a polynomial p approximates the function g, then the operator $p(\mathbf{T})$ approximates the operator $g(\mathbf{T})$. This is formulated as follows. Consider the space $PW(\lambda_M)$ of vectors f comprising finite eigenbasis expansions

$$f = \sum_{n=0}^{M} b_n \phi_n,$$

for a fixed M. If a sequence of polynomials $\{g_k\}_k$ converges to a continuous function g in the sense

$$\lim_{k\to\infty} \sup_{\lambda \leq |\lambda_M|} |g(\lambda) - g_k(\lambda)| = 0,$$

then also

$$\lim_{k\to\infty} \|g(\mathbf{T}) - g_k(\mathbf{T})\| = 0, \tag{4}$$

where the operator norm in (4) is defined by

$$\|g(\mathbf{T}) - g_k(\mathbf{T})\| := \sup_{0\neq f \in PW(\lambda_M)} \frac{\|g(\mathbf{T})f - g_k(\mathbf{T})f\|}{\|f\|}.$$

When filters are defined via (2) with polynomial or rational function g, implementing spectral filters via (3) overcomes the limitation of definition (1). By relying on the spatial operations of compositions, linear combinations, and inversions, the computation of a spectral filter is carried out entirely in the spatial domain, without ever resorting to spectral computations. Thus, no eigendecomposition and Fourier transforms are ever computed. The inversions in $g(\mathbf{T})f$ involve solving systems of linear equations, which can be computed directly if N is small, or by some iterative approximation method for large N. Methods like (Defferrard et al., 2016; Kipf and Welling, 2017; Ortega et al., 2018; Gama et al., 2018) use polynomial filters, and (Isufi et al., 2017a; Levie et al., 2019b; Bianchi et al., 2021) use rational function filters. We term spectral methods based on functional calculus **functional calculus filters**.

2.3 The Misconception of Non-transferability of Spectral Graph Filters

The non-transferability claim is formulated based on the sensitivity of the Laplacian eigendecomposition to small perturbations in $\mathbf{W}$, or equivalently in $\mathbf{\Delta}$. Namely, a small perturbation of $\mathbf{\Delta}$ can result in a large perturbation of the eigendecomposition $\{\lambda_n, \phi_n\}_{n=1}^{N}$, which results in a large change in the filter defined via (1). This claim was stated in (Bronstein et al., 2017) only for spectral filters implemented via (1), for which it is true. However, later papers misinterpreted this claim and applied it to functional calculus filters. This misconception can be found in prominent surveys (Wu et al., 2020), as well as research papers, e.g., (Fey et al., 2018; Maron et al., 2018; Te et al., 2018; Cai et al., 2019; Chen et al., 2019; Bi et al., 2019) (the list if far from exhaustive). The instability argument does not prove non-transferability, since state-of-the-art spectral methods do not explicitly use the eigenvectors, and do not parametrize the filter coefficients g_n via the index n of the eigenvalues. Instead, state-of-the-art methods are based on functional calculus, and define the filter coefficients using a function $g : \mathbb{R} \to \mathbb{C}$, as $g(\lambda_n)$. The parametrization of the filter coefficients by g is indifferent to the specifics of how the spectrum is indexed, and instead represents an overall response in the frequency domain, where the **value** of each frequency determines its response, and not its index. When functional calculus filters are defined by (2), a small perturbation of $\mathbf{\Delta}$ that results in a perturbation of λ_n, also results in a perturbation of the coefficients $g(\lambda_n)$. It turns out that the perturbation in $g(\lambda_n)$ implicitly compensates for the instability of the eigendecomposition, and functional calculus spectral filters are stable. This is seen by using the transferability inequality in a graph perturbation setting (see Subsection 3.2).

As a toy example, consider the graph Laplacian on a graph with three nodes, defined via its eigendecomposition, where $\lambda_1 = 1$ has a 2D eigenspace, spanned by the eigenvectors ϕ_1, ϕ_2, and $\lambda_2 = 2$ has a 1D eigenspace spanned by ϕ_3. Implementation (1) with $g_1 \neq g_2$ is not even uniquely defined by $\mathbf{\Delta}$, as the basis $\{\phi_1, \phi_2\}$ of the eigenspace of λ_1 is not uniquely defined by $\mathbf{\Delta}$. On the other hand, the functional calculus implementation (2) imposes that the frequency response is one constant for the whole eigenspace of λ_1, and the non-uniqueness problem is avoided. More generally, in (Levie et al., 2019a) the stability of functional calculus filters was proved.

2.4 Spectral Graph Filters on Directed Graphs

In Appendix A we explain how functional calculus applies as-is to non-normal matrices, even though the theory is defined only for normal operators. As a result, spectral filters can be defined on directed graphs represented by non-symmetric adjacency matrices.

There is an inner product structure in $\mathbb{C}^N$ under which general diagonalizable matrices can be seen as normal operators. Given an $N \times N$ diagonalizable matrix $\mathbf{A}$ with eigenvectors $\{\boldsymbol{\gamma}_k\}_{k=1}^N$, consider the matrix $\mathbf{\Gamma}$ comprising the eigenvectors as columns. Define the inner product

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{v}^{\mathrm{H}} \mathbf{B} \mathbf{u}, \tag{5}$$

where $\mathbf{B} = \mathbf{\Gamma}^{-\mathrm{H}} \mathbf{\Gamma}^{-1}$ is symmetric, $\mathbf{u}$ and $\mathbf{v}$ are given as column vectors, and for a matrix $\mathbf{C} = (c_{m,k})_{n,m} \in \mathbb{C}^{N \times N}$, the Hermitian transpose $\mathbf{C}^{\mathrm{H}}$ is the matrix consisting of entries $c^{\mathrm{H}}_{m,k} = \overline{c_{k,m}}$. Under the inner product (5), $\mathbf{A}$ is normal. Consider an operator A represented by the matrix $\mathbf{A}$. The adjoint A^* of an operator A is defined to be the unique operator such that

$$\forall \mathbf{u}, \mathbf{v} \in \mathbb{C}^d, \quad \langle A\mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{u}, A^* \mathbf{v} \rangle .$$

The matrix representation of the adjoint A^* is given by

$$\mathbf{A}^* = \mathbf{B}^{-1} \mathbf{A}^{\mathrm{H}} \mathbf{B}. \tag{6}$$

Thus, an operator is self-adjoint if $\mathbf{B}^{-1} \mathbf{A}^{\mathrm{H}} \mathbf{B} = \mathbf{A}$, and unitary if $\mathbf{B}^{-1} \mathbf{A}^{\mathrm{H}} \mathbf{B} = \mathbf{A}^{-1}$.

The above results are proved in Appendix A.

3. The Transferability Inequality

In this section we derive the **transferability inequality**, a generic inequality that bounds the **transferability error of filters** by the **transferability error of Laplacians** plus the error entailed by sampling-interpolating, called the **consistency error**.

3.1 The General Setting of Transferability

For a graph discretizing a "continuous" topological space, as described in Subsections 1.4.1 and 1.4.2, the transferability error between the graph and the topological space is defined as follows. Given a generic signal in the topological space, on the one hand, the signal is sampled to the graph, the discrete filter is applied on the sampled signal, and the filtered signal is interpolated back to the topological space. On the other hand, the filter is applied on the signal directly in the topological space. The error between these two output signals is called the transferability error of the filter. For two graphs with small transferability error between each of the graphs and the topological space, the transferability error of the filter between the two graphs is also small by the triangle inequality. We thus focus on transferability between graphs and topological spaces.

Topological space signals are functions that assign to every point in the topological space a value. The error between pairs of signals is defined as the root mean square error (RMSE). To define RMSE in this abstract setting we must be able to integrate over the topological space, and thus we always assume that the topological space comes with some notions of volume, namely a Borel measure[4].

Instead of focusing on graphs discretizing topological spaces via sampling, we consider a more general setting. In the general setting we study transferability between two domains, $\mathcal{M}$ and the finite domain G. The domains $\mathcal{M}$ and G are assumed to be measure spaces, and we consider the two spaces of signals[5] $L^2(\mathcal{M})$ and $L^2(G)$. We assume that the spaces $L^2(\mathcal{M})$ and $L^2(G)$ are separable, namely, there exist orthonormal bases of $L^2(\mathcal{M})$ and $L^2(G)$. Since filtering is seen as a procedure of increasing certain frequencies, and decreasing others, we need a notion of oscillation of signals in the spaces $L^2(\mathcal{M})$ and $L^2(G)$. For that, we endow the signal spaces with additional structure. In each of the signal spaces $L^2(\mathcal{M})$ and $L^2(G)$ we consider a special normal linear operators (typically self-adjoint) that we call the **Laplacian** of the space. For $L^2(\mathcal{M})$ we denote the Laplacian by $\mathcal{L}$, and for $L^2(G)$ we denote the Laplacian by $\boldsymbol{\Delta}$. We suppose that $\mathcal{L}$ and $\boldsymbol{\Delta}$ have discrete spectra in the following sense.

Definition 1 *Consider the normal operator T with spectrum consisting only of eigenvalues, and denote the eigendecomposition of T by $\{\lambda_j, P_j\}_{j=1}^{\infty}$, with eigenvalues λ_j and projections P_j upon the corresponding eigenspaces W_j. We say that T has **discrete spectrum** if in each bounded disc in $\mathbb{C}$ there are finitely many eigenvalues of T, and the eigenspace of each eigenvalue is finite-dimensional. We consider the eigenvalues in increasing order of $|\lambda_j|$, and denote $\Lambda(T) = \{\lambda_j\}_{j=1}^{\infty}$.*

For example, Laplace-Beltrami operators on compact Riemannian manifolds satisfy Definition 1 by Weyl's law (Strauss, 2007, Chapter 11).

As discussed in Subsection A, the Laplacian $\boldsymbol{\Delta}$ need not be a normal matrix. If $\boldsymbol{\Delta}$ is not a normal matrix, we consider an inner product structure on each $L^2(V_n)$ for which $\boldsymbol{\Delta}$ is a normal operator.

[4] A measure is a generalization of the notion of volume. A Borel measure in a topological space is a notion of volume that respects in some sense the topological structure. For example, open sets must have well defined volumes.

[5] Using the notion of volume of a measure space $\mathcal{M}$ it is possible to define integration, and thus define the Lebesgue space of square integrable functions $L^2(\mathcal{M})$.

The Laplacians $\mathcal{L}$ and $\boldsymbol{\Delta}$ define the notion of oscillation on $L^2(\mathcal{M})$ and $L^2(G)$. Namely, the eigenvectors of the Laplacians are seen as the pure harmonics, or Fourier modes. The eigenvalues are seen as an ordering of the pure harmonics, where the larger the eigenvalue corresponding to an eigenvector, the more oscillatory the eigenvector is. Filters are defined as measurable functions $f : \mathbb{C} \rightarrow \mathbb{C}$. Each filter can be manifested in both spaces via functional calculus, where the filter in $L^2(\mathcal{M})$ is defined as $f(\mathcal{L})$, and the filter in $L^2(G)$ is defined as $f(\boldsymbol{\Delta})$.

We suppose that the space G is finite, and thus $L^2(G)$ is finite-dimensional. When $\mathcal{M}$ is infinite, the signal space $L^2(\mathcal{M})$ is infinite-dimensional in general. We consider the finite-dimensional subspace of signal of $L^2(\mathcal{M})$ spanning all of the eigenvectors of $\mathcal{L}$ up to some eigenvalue, as defined next.

Definition 2 *Let $\mathcal{L}$ be a normal operator in $L^2(\mathcal{M})$ with discrete spectrum. Denote the eigenvalues, eigenspaces, and projections upon the eigenspaces of $\mathcal{L}$ by $\{\lambda_j, W_j, P_j\}_{j\in\mathbb{N}}$. For each $\lambda > 0$, we define the λ'th* ***Paley-Wiener*** *space of $\mathcal{M}$ as*

$$PW(\lambda) = \oplus_{j\in\mathbb{N}}\{W_j \mid |\lambda_j| \leq \lambda\}.$$

We denote by $P(\lambda)$ the ***spectral projection*** *upon $PW(\lambda)$, given by*

$$P(\lambda) = \sum_{\lambda_j\in\Lambda(\boldsymbol{\Delta}),\ |\lambda_j|\leq\lambda} P_j.$$

A Paley-Wiener space is interpreted as the space of band-limited signals in the band λ. When $L^2(\mathcal{M})$ is infinite-dimensional we restrict the analysis to a generic Paley-Wiener space $PW(\lambda) \subset L^2(\mathcal{M})$. Namely, transferability is analyzed on signals which are not too oscillatory.

To accommodate a transferability analysis, we consider two mappings that transfer signals from $L^2(\mathcal{M})$ to signals in $L^2(G)$ and back. For each fixed band λ, consider the linear operators

$$S^\lambda : PW(\lambda) \rightarrow L^2(G), \quad R^\lambda : L^2(G) \rightarrow PW(\lambda).$$

We typically think of S^λ as down-sampling or discretization, and R^λ as up-sampling. We thus call S^λ **sampling** and R^λ **interpolation**.

Definition 3 *The* ***transferability error of the filter*** *f (at the band λ), on the signal $s \in PW(\lambda)$, is defined by*

$$\left\| f(\mathcal{L})s - R^\lambda f(\boldsymbol{\Delta}) S^\lambda s \right\|,$$

the ***transferability error of the Laplacian*** *(at the band λ) is defined by*

$$\left\| \mathcal{L}s - R^\lambda \boldsymbol{\Delta} S^\lambda s \right\|,$$

and the ***consistency error*** *(at the band λ) is defined by*

$$\left\| s - R^\lambda S^\lambda s \right\|.$$

What we prove in this section is the following inequality

$$\left\| f(\mathcal{L})s - R^\lambda f(\boldsymbol{\Delta}) S^\lambda s \right\| \leq C_1 \left\| \mathcal{L}s - R^\lambda \boldsymbol{\Delta} S^\lambda s \right\| + C_2 \left\| s - R^\lambda S^\lambda s \right\|$$

up to some constants C_1 and C_2.

3.2 Examples of Transferability Settings

Before we formulate the transferability inequality theorem, let us give three concrete settings of the above transferability analysis. In the first example, which was introduced in Subsections 1.4.1 and 1.4.2, $\mathcal{M}$ is a topological space with a Borel measure. The space G is a graph, where the nodes of G are seen as sample points in $\mathcal{M}$. Sampling general signals in the Lebesgue space $L^2(\mathcal{M})$ is not well defined (unless $\mathcal{M}$ is discrete), since signals in $L^2(\mathcal{M})$ are defined up to a subset of $\mathcal{M}$ of measure zero. To be able to define sampling properly we consider the Paley-Wiener spaces, an approach that generalizes the standard Nyquist—Shannon theory in signal processing in $L^2(\mathbb{R})$. For that we further assume that the Paley-Wiener spaces associated with $\mathcal{L}$ consist of continuous functions (see Definition 18 for more details). Sampling is the operator S^λ that evaluates signals $s \in PW(\lambda) \subset L^2(\mathcal{M})$ at the sample points to obtain a signal on the graph. Similarly to classical digital signal processing, we define the interpolation R^λ as the adjoint of the sampling operator, namely $R^\lambda = S^{\lambda *}$ (see Subsection for more information). Transferability between $\mathcal{M}$ and G is thus seen as the error entailed by operating in the digital domain G instead of the analog domain $\mathcal{M}$.

As a second example, we consider transferability under graph coarsening. Here, $\mathcal{M}$ is a graph, and G is a coarse version of M. In the Graclus algorithm for coarsening (Dhillon et al., 2004), pairs of neighboring nodes in $\mathcal{M}$ with strong weights are collapsed to single nodes in G. Since both $\mathcal{M}$ and G are finite, we consider the whole space $L^2(\mathcal{M})$ as the Paley-Wiener space, and omit the superscript λ in R and S. Given a signal s, coarsening, S, is the operator that assigns the value

$$[Ss](q_{1,2}) = (s(q_1) + s(q_2))/\sqrt{2} \tag{7}$$

to the node $q_{1,2}$ of G with parent nodes from $\mathcal{M}$, q_1 and q_2, that have the signal values $s(q_1)$ and $s(q_2)$ respectively. Piecewise constant interpolation is defined to be $R = S^*$.

The last example is graph perturbation. Here, $\mathcal{M}$ is a graph, and G is a perturbation of $\mathcal{M}$, that is obtain by adding or deleting random edges from $\mathcal{M}$ or perturbing the edge weights. Here we take $S = R = I$. Transferability in this case is called stability.

3.3 Theorem of Transferability Inequality

For the transferability inequality we need the following notations. For a continuous $g : \mathbb{C} \to \mathbb{C}$ and $M \in \mathbb{N}$ denote

$$\|g\|_{\mathcal{L},M} := \max_{0 \leq m \leq M} \{|g(\lambda_m)|\}. \tag{8}$$

For each $\lambda_m \in \Lambda(\mathcal{L})$ denote

$$\mathrm{V}_g(\lambda_m) := \max_{\kappa \in \Lambda(\boldsymbol{\Delta})} \left| \frac{g(\kappa) - g(\lambda_m)}{\kappa - \lambda_m} \right|. \tag{9}$$

Note that for a Lipschitz continuous g with Lipschitz constant D, it follows from $\left|\frac{g(x)-g(y)}{x-y}\right| \leq D$ that $\mathrm{V}_g(\lambda_m) \leq D$. Denote by $\#\{\lambda_j \leq \lambda\}_j$ the number of eigenvalues of $\mathcal{L}$ less or equal to λ, and note that

$$\#\{\lambda_j \leq \lambda\}_j \leq \mathrm{dim} PW(\lambda),$$

where $\mathrm{dim}PW(\lambda)$ is the dimension of $PW(\lambda)$.

Example 1 *For the Laplacian on the d-dimensional torus, we have $\#\{\lambda_j \leq \lambda\}_j = O(\lambda^{1/2})$. For compact Riemannian manifolds and the Laplace-Beltrami operator, by Weyl's law, $\#\{\lambda_j \leq \lambda\}_j \leq \mathrm{dim}PW(\lambda) = O((2\pi)^{-d}\lambda^{d/2})$ where d is the dimension of the manifold (Strauss, 2007, Chapter 11).*

We are now ready to formulate five versions of the transferability inequality.

Theorem 4 *Consider the above setting, and let $\lambda_M > 0$ be a band with $\left\|R^{\lambda_M}\right\| < C$. Let $g : \mathbb{R} \to \mathbb{C}$ be a Lipschitz continuous function with Lipschitz constant D. Let $q = \sum_{m=0}^{M} c_m\phi_m \in PW(\lambda_M) \subset L^2(\mathcal{M})$ have normalized eigenspace components $\phi_m \in W_m$, $m = 0, \ldots, M$. Then the following bounds are satisfied.*

1. *Transferability of $\mathcal{L}$-Fourier modes evaluated in G:*

$$\left\|S^{\lambda_M} g(\mathcal{L})\phi_m - g(\mathbf{\Delta})S^{\lambda_M}\phi_m\right\| \leq \mathrm{V}_g(\lambda_m) \left\|\mathbf{\Delta}S^{\lambda_M}\phi_m - S^{\lambda_M}\lambda_m\phi_m\right\|.$$

2. *Pointwise transferability evaluated in G:*

$$\left\|S^{\lambda_M} g(\mathcal{L})q - g(\mathbf{\Delta})S^{\lambda_M}q\right\| \leq \sum_{m=0}^{M} \mathrm{V}_g(\lambda_m)\left|c_m\right| \left\|S^{\lambda_M}\mathcal{L}\phi_m - \mathbf{\Delta}S^{\lambda_M}\phi_m\right\|.$$

3. *Worst-case transferability evaluated in G:*

$$\begin{aligned}&\left\|S^{\lambda_M} g(\mathcal{L})P(\lambda_M) - g(\mathbf{\Delta})S^{\lambda_M}P(\lambda_M)\right\| \\ &\leq D\sqrt{\#\{\lambda_j \leq \lambda_M\}_j}\left\|S^{\lambda_M}\mathcal{L}P(\lambda_M) - \mathbf{\Delta}S^{\lambda_M}P(\lambda_M)\right\|.\end{aligned}$$

4. *Pointwise transferability evaluated in $\mathcal{M}$:*

$$\begin{aligned}\left\|g(\mathcal{L})q - R^{\lambda_M}g(\mathbf{\Delta})S^{\lambda_M}q\right\| \leq & C\sum_{m=0}^{M} \mathrm{V}_f(\lambda_m)\left|c_m\right| \left\|S^{\lambda_M}\mathcal{L}\phi_m - \mathbf{\Delta}S^{\lambda_M}\phi_m\right\| \\ & + \left\|g\right\|_{\mathcal{L},M}\left\|q - R^{\lambda_M}S^{\lambda_M}q\right\|,\end{aligned}$$

5. *Worst-case transferability evaluated in $\mathcal{M}$:*

$$\begin{aligned}&\left\|g(\mathcal{L})P(\lambda_M) - R^{\lambda_M}g(\mathbf{\Delta})S^{\lambda_M}P(\lambda_M)\right\| \\ &\leq DC\sqrt{\#\{\lambda_j \leq \lambda_M\}_j}\left\|S^{\lambda_M}\mathcal{L}P(\lambda_M) - \mathbf{\Delta}S^{\lambda_M}P(\lambda_M)\right\| \\ &\quad + \left\|g\right\|_{\mathcal{L},M}\left\|P(\lambda_M) - R^{\lambda_M}S^{\lambda_M}P(\lambda_M)\right\|,\end{aligned}$$

Theorem 4 can be seen as a family of bounds, for different choices of Paley-Wiener spaces. Typically, if we choose a small cut-off frequency λ_M, the Laplacian has a lower transferability error (see, e.g., Lemma 36), so we can prove a low approximation error. However, the bounds are true only for the low frequency content of the signal, namely, for the "smooth content." If we choose high λ_M, we can also model "non-smooth" signals, but, on account of a typically higher transferability error of the Laplacian. This principle in choosing the Paley-Wiener space is true for all results presented in this paper which depend on a choice of the cut-off frequency.

We note that at the time of writing this paper, it is still not clear whether the dependency on $\sqrt{\#\{\lambda_j \leq \lambda\}_j}$ in the operator norm bounds 3 and 5 is tight, or just an artifact of the proof.

Let us now study the transferability between two graphs. Consider two graphs G_1 and G_2, with corresponding graph Laplacians $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$, that represent the same phenomenon. Adopting our basic assumption, we thus suppose that both graphs approximate the space $\mathcal{M}$ in the sense that the transferability errors of the Laplacians and the consistency errors are small.

Corollary 5 *Consider a fixed Paley-Wiener space $PW(\lambda_M)$, and for each $n = 1, 2$, suppose $\left\|\mathcal{L}P(\lambda_M) - R_n^{\lambda_M}\boldsymbol{\Delta}_n S_n^{\lambda_M}P(\lambda_M)\right\| \leq \delta$ and $\left\|P(\lambda_M) - R_n^{\lambda_M}S_n^{\lambda_M}P(\lambda_M)\right\| \leq \delta$ for some small $\delta > 0$. Then*

$$\left\|R_1^{\lambda_M} f(\boldsymbol{\Delta}_1)S_1^{\lambda_M}P(\lambda_M) - R_2^{\lambda_M} f(\boldsymbol{\Delta}_2)S_2^{\lambda_M}P(\lambda_M)\right\| = O(\delta). \tag{10}$$

Proof By the triangle inequality we have

$$\begin{aligned}&\left\|R_1^{\lambda_M} f(\boldsymbol{\Delta}_1)S_1^{\lambda_M}P(\lambda_M) - R_2^{\lambda_M} f(\boldsymbol{\Delta}_2)S_2^{\lambda_M}P(\lambda_M)\right\| \\ &\leq \left\|f(\mathcal{L})P(\lambda_M) - R_1^{\lambda_M} f(\boldsymbol{\Delta}_1)S_1^{\lambda_M}P(\lambda_M)\right\| + \left\|f(\mathcal{L})P(\lambda_M) - R_2^{\lambda_M} f(\boldsymbol{\Delta}_2)S_2^{\lambda_M}P(\lambda_M)\right\|.\end{aligned} \tag{11}$$

Thus, (10) follows fromThm.4(5) . ■

A similar result can be obtained in the pointwise analysis.

4. Transferability of Spectral Graph Filters and ConvNets

In this section we study the transferability of spectral graph filters and ConvNets. We formulate general conditions guaranteeing transferability of filters, and then extend the analysis to full convolutional networks. We also give some numerical experiments that showcases transferability.

4.1 Sufficient Conditions for Transferability

In this subsection we consider sufficient conditions for the right-hand-side of the transferability inequality to be small (Theorem 4). The idea is to formulate general conditions, and to later on prove that they are satisfied in the specific setting of graphs discretizing topological spaces. We denote by $\mathbb{R}_+$ the set of non-negative real numbers.

Consider a measure space $\mathcal{M}$ for which $L^2(\mathcal{M})$ is a separable Hilbert space, and let the Laplacian $\mathcal{L}$ be a normal operator in $L^2(\mathcal{M})$ with discrete spectrum. Denote the eigenvalues of $\mathcal{L}$ by λ_j, the eigenprojections by P_j, and the Paley-Wiener spaces $PW(\lambda)$. To accommodate the approximation analysis, we consider a sequence of graphs G_n with d_n vertices and graph Laplacians $\mathbf{\Delta}_n$, such that "$\mathbf{\Delta}_n \xrightarrow[n\rightarrow\infty]{} \mathcal{L}$" in a sense that will be clarified in Definition 9.

By abuse of notation, we denote the set of vertices of G_n also by G_n. We consider an inner product structure on each $L^2(G_n)$ for which $\mathbf{\Delta}_n$ is a normal operator. Denote the eigendecomposition of $\mathbf{\Delta}_n$ by $\{\kappa^n_j, Q^n_j\}_j$, and denote $\Lambda(\mathbf{\Delta}_n) := \{\kappa^n_j\}_j$. For any $\kappa > 0$, denote by $Q_n(\kappa)$ the spectral projection of $\mathbf{\Delta}_n$ defined by

$$Q_n(\kappa) := \sum_{\kappa^n_j\in\Lambda(\mathbf{\Delta}_n),\ \left|\kappa^n_j\right|\leq\kappa} Q^n_j.$$

Generic sampling and interpolation operators are only required to satisfy mild conditions, as defined next.

Definition 6 *Under the above construction, the two mappings*

$$\{S^\lambda_n\}_{n,\lambda} : (n,\lambda)\mapsto S^\lambda_n, \quad \{R^\lambda_n\}_{n,\lambda} : (n,\lambda)\mapsto R^\lambda_n$$

from $\mathbb{N}\times\mathbb{R}_+$ *to the space of linear operators* $L^2(\mathcal{M})\rightarrow L^2(G_n)$ *and* $L^2(G_n)\rightarrow L^2(\mathcal{M})$ *respectively, satisfying for each* $\lambda\geq 0$

$$S^\lambda_n : PW(\lambda)\rightarrow L^2(G_n), \quad R^\lambda_n : L^2(G_n)\rightarrow PW(\lambda),$$

*are called **sampling sequence** and **interpolation sequence** respectively, if the following condition is held.: for every* $\lambda' > \lambda \geq 0$

$$S^{\lambda'}_n P(\lambda) = S^\lambda_n, \quad P(\lambda)R^{\lambda'}_n = R^\lambda_n. \tag{12}$$

Operators S^λ_n *from a sampling sequence are called **sampling operators**, and similarly,* R^λ_n *are called **interpolation operators**.*

For $\lambda' > \lambda$, (12) means that sampling a signal from $PW(\lambda)$ using $S^{\lambda'}_n$ is exactly the same as sampling it using S^λ_n, and in this sense the different sampling operators of a sampling sequence are related to each other. Interpolation operators of an interpolation sequence have a similar interpretation. Given sampling and interpolation operator sequences, by the fact that $PW(\lambda)$ is finite dimensional, S^λ_n and R^λ_n must be bounded for each $n\in\mathbb{N}$ and $\lambda\geq 0$.

In the following, we fix a sampling and interpolation sequence. Next, we define general conditions on $\mathbf{\Delta}_n$, S^λ_n and R^λ_n, and show that these conditions guarantee transferability of spectral graph filters. In Section 5 we give an explicit construction of the sampling and interpolation operators in the DSP setting, where $S^\lambda_n f$ evaluates the signal $f\in PW(\lambda)$ at a set of sample points, viewed as the vertices of G_n. Under that construction, we show in Subsection 5 that the conditions underlying Definitions 7—9 are satisfied.

Definition 7 *The sequence* $\{\{R^\lambda_n, S^\lambda_n\}_n \mid \lambda\in\mathbb{R}\}$ *is called **asymptotically reconstructive** if for any fixed band* λ*,*

$$\lim_{n\rightarrow\infty} R^\lambda_n S^\lambda_n P(\lambda) = P(\lambda). \tag{13}$$

Note that since $PW(\lambda)$ is a finite-dimensional space, the operator norm topology and the strong topology are equivalent, namely

$$\lim_{n\to\infty} \max_{f\in PW(\lambda)} \frac{\left\|f - R_n^\lambda S_n^\lambda f\right\|}{\|f\|} = 0 \quad \Longleftrightarrow \quad \forall f \in PW(\lambda),\ \lim_{n\to\infty} \left\|f - R_n^\lambda S_n^\lambda f\right\| = 0, \tag{14}$$

and the limit in (13) can be defined in either way.

Definition 8 *The sequence $\{\{R_n^\lambda, S_n^\lambda\}_n \mid \lambda \in \mathbb{R}\}$ is called* ***bounded*** *if there exists a global constant $C \geq 1$ such that for any fixed band λ,*

$$\limsup_{n\in\mathbb{N}} \left\|S_n^\lambda\right\| \leq C, \quad \limsup_{n\in\mathbb{N}} \left\|R_n^\lambda\right\| \leq C. \tag{15}$$

where the induced operator norms are with respect to the vector norms in $PW(\lambda)$ and in $L^2(V_n)$.

Boundedness (Definition 8) is a necessary condition for sampling and interpolation to approximate isometries as the resolution of sampling d_n becomes finer, and we typically consider $C = 1$.

Definition 9 *The set of sequences $\{\{\boldsymbol{\Delta}_n, S_n^\lambda\}_n \mid \lambda \in \mathbb{R}\}$ are called* ***convergent*** *to $\mathcal{L}$ if for every fixed band λ,*

$$\lim_{n\to\infty} \left\|S_n^\lambda \mathcal{L} P(\lambda) - \boldsymbol{\Delta}_n S_n^\lambda P(\lambda)\right\| = 0. \tag{16}$$

where the norm in (16) is with respect to $L^2(V_n)$.

In the DSP setting of transferability, for S_n^λ that evaluates the signal at sample points and corresponding R^{λ_n} and $\boldsymbol{\Delta}_n$, boundedness and asymptotic reconstruction (Definitions 7,8) are proved in Proposition 23. Convergence (Definition 9) is proved in Proposition 28 in the DSP setting. We can also treat sampling and interpolation abstractly, allowing other constructions for transforming signals in $L^2(\mathcal{M})$ to graph signals in $L^2(V_n)$. In the abstract setting, sampling and interpolation are assumed to be bounded, asymptotically reconstructive, and graph Laplacians are assumed to be convergent to $\mathcal{L}$. Assuming boundedness, asymptotic reconstruction, and convergence of Laplacians, is permissive in a sense, since we only demand asymptotic properties on the finite-dimensional Paley-Wiener spaces. However, under these assumptions, we are able to prove convergence of spectral filters on band-unlimited signals.

The following proposition proves asymptotic perfect transferability, and is a direct result of the transferability inequality.

Proposition 10 *consider the above setting, and a fixed band $\lambda > 0$. Let S_n^λ, R_n^λ and $\boldsymbol{\Delta}_n$, $n = 1, \ldots, \infty$, be bounded, asymptotically reconstructive, and convergent (Definitions 7-9). Let $g : \mathbb{R} \to \mathbb{C}$ be a Lipschitz continuous function. Then*

$$\begin{aligned}\left\|g(\mathcal{L})P(\lambda) - R_n^\lambda g(\boldsymbol{\Delta}_n) S_n^\lambda P(\lambda)\right\| =& O\Big(\left\|S_n^\lambda \mathcal{L} P(\lambda) - \boldsymbol{\Delta}_n S_n^\lambda P(\lambda)\right\| + \left\|P(\lambda) - R_n^\lambda S_n^\lambda P(\lambda)\right\| \Big) \\ &\xrightarrow[n\to\infty]{} 0.\end{aligned}$$

Proof Denote by M_λ the largest index $M \in \mathbb{N}$ such that $\lambda_M \leq \lambda$. Then by Thr4.(5), by (12) and by the fact that $P(\lambda_m) = P(\lambda)$

$$\begin{aligned}\left\| g(\mathcal{L})P(\lambda) - R_n^\lambda g(\mathbf{\Delta}_n)S_n^\lambda P(\lambda)\right\| = & \left\| P(\lambda_M)g(\mathcal{L})P(\lambda_M) - P(\lambda_M)R_n^\lambda g(\mathbf{\Delta}_n)S_n^\lambda P(\lambda_M)\right\| \\ & + \|g\|_{\mathcal{L},M}\left\| P(\lambda_M) - R_n^{\lambda_M}S_n^{\lambda_M}P(\lambda_M)\right\| \\ \leq & DC\sqrt{\#\{\lambda_j \leq \lambda\}_j}\left\| S_n^{\lambda_M}\mathcal{L}P(\lambda_M) - \mathbf{\Delta}S_n^{\lambda_M}P(\lambda_M)\right\| \\ & + \|g\|_{\mathcal{L},M}\left\| P(\lambda_M) - R_n^{\lambda_M}S_n^{\lambda_M}P(\lambda_M)\right\|\end{aligned}$$

■

Next, we show how to treat band-unlimited signals. Under the conditions of Theorem 10, for each band $\lambda \in \mathbb{N}$, there exists $N_\lambda \in \mathbb{N}$ such that for any $n > N_\lambda$ we have

$$\left\| g(\mathcal{L})P(\lambda) - R_n^\lambda g(\mathbf{\Delta}_n)S_n^\lambda P(\lambda)\right\| < \frac{1}{\lambda}$$

We may choose the sequence $\{N_\lambda\}_{\lambda\in\mathbb{N}}$ increasing. We construct a sequence of bands $\{\psi_n\}$, starting from some index $n_0 > 0$, as follows. For each $\lambda \in \mathbb{N}$, consider N_λ and $N_{\lambda+1}$. For each $N_\lambda < n \leq N_{\lambda+1}$ we define $\psi_n = \lambda$. This gives the following corollary.

Corollary 11 *Under the conditions of Proposition 10, there exists a sequence of bands* $0 < \psi_n \xrightarrow[n\to\infty]{} \infty$ *such that*

$$\lim_{n\to\infty}\left\| g(\mathcal{L}) - R_n^{\psi_n} g(\mathbf{\Delta}_n)S_n^{\psi_n}P(\psi_n)\right\| = 0. \tag{17}$$

A direct result of Corollary 11 is that

$$\lim_{n>m\to\infty}\left\| R_n^{\psi_m} g(\mathbf{\Delta}_n)S_n^{\psi_m}P(\psi_m) - R_m^{\psi_m} g(\mathbf{\Delta}_m)S_m^{\psi_m}P(\psi_m)\right\| = 0. \tag{18}$$

Loosely speaking, the better both $\mathbf{\Delta}_m$ and $\mathbf{\Delta}_j$ approximate $\mathcal{L}$, the larger the band where $g(\mathbf{\Delta}_m)$ and $g(\mathbf{\Delta}_j)$ have approximately the same repercussion.

Last, for the transferability analysis of convolution networks, we also need to assume that sampling approximately commutes with the activation function ρ, in the following sense.

Definition 12 *Consider a measure space $\mathcal{M}$ with a Laplacian $\mathcal{L}$ having a discrete spectrum. Let $P(\lambda)$ be the Paley-Wiener projections corresponding to $\mathcal{L}$. Consider a sequence of graphs G_n, sampling operators S_n^λ from a sampling sequence, and an activation function ρ. Sampling* ***asymptotically commutes with*** *ρ if*

$$\lim_{\lambda\to\infty}\lim_{\lambda'\to\infty}\lim_{n\to\infty}\sup_{f\neq 0}\frac{\left\| \rho(S_n^\lambda P(\lambda)f) - S_n^{\lambda'}P(\lambda')\rho(P(\lambda)f)\right\|}{\|f\|} = 0. \tag{19}$$

In Proposition 26 we prove that in the DSP setting, under natural conditions, sampling asymptotically commutes with ρ for a class of activation functions that include ReLU and the absolute value.

4.2 Transferability of Graph ConvNets

In this subsection we extend the transferability results of the previous subsection from filters to complete ConvNets. Consider two graphs G^j, $j = 1, 2$ and two graph Laplacians $\mathbf{\Delta}_1, \mathbf{\Delta}_2$ approximating the same Laplacian $\mathcal{L}$ in a measure space. Consider a ConvNet with L layers, with or without pooling. In each layer where pooling is performed, the signal is mapped to a signal over a coarsened graph. If pooling is not performed, we define the coarsened graph $G^{j,l}$ at Layer l as the graph of the previous layer. Suppose that each coarsened version of each of the two graphs $G^{j,l}$, where l is the layer, approximates the continuous space in the sense

$$\left\|P(\psi_l) - R^{\psi_l}_{j,l} S^{\psi_l}_{j,l} P(\psi_l)\right\| < \delta \tag{20}$$

$$\left\|S^{\psi_l}_{j,l} \mathcal{L} P(\psi_l) - \mathbf{\Delta}_{j,l} S^{\psi_l}_{j,l} P(\psi_l)\right\| < \delta \tag{21}$$

for some $\delta < 1$. Here, $\mathbf{\Delta}_{j,l}$ is the Laplacian of graph j at Layer l, $S^{\psi_l}_{j,l}, R^{\psi_l}_{j,l}$ are the sampling and interpolation operators of Layer l, and we consider the band ψ^l at each Layer l. Equations (20) and (21) are non-asymptotic versions of Definition 7 and 9.

In each Layer l consider K_l channels. Let

$$\{g^l_{k'k} \mid {\scriptstyle k=1\ldots K_{l-1},\ k'=1\ldots K_l}\}$$

denote the filters of Layer l, and consider the matrix $A^l = \{a^l_{k'k}\}_{k'k} \in \mathbb{R}^{K_l \times K_{l-1}}$. We denote the bias at channel k' and Layer l by $b^l_{k'}$. Here, $b^l_{k'}$ is a scalar signal, namely the signal that assigns the constant value $b^l_{k'}$ to each node. Note that, by abuse of notation, $b^l_{k'}$ denotes both a scalar and a signal. In most common graph ConvNet methods there are no biases, so we typically assume $b^l_{k'} = 0$.

Denote the signals/feature-map at Layer l of the graph ConvNet of graph G^j, by $\{\tilde{f}^{j,l}_{k'}\}_{k'=1}^{K_l}$. The ConvNet maps Layer $l-1$ to Layer l by

$$\{\tilde{f}^{j,l}_{k'}\}_{k'=1}^{K_l} = Y^{j,l}\Big(\rho\Big\{b^l_{k'} + \sum_{k=1}^{K_{l-1}} a^l_{k'k}\ g^l_{k'k}(\mathbf{\Delta}_{j,l-1}) \tilde{f}^{j,l-1}_k\Big\}_{k'=1}^{K_l}\Big), \tag{22}$$

where ρ is an activation function, and $Y^{j,l} : L^2(G^{j,l-1}) \to L^2(G^{j,l})$ is pooling. For the graph ConvNets, the inputs of Layer 1 are $S^{\psi_0}_{j,1} P(\psi_0) f$ for $j = 1, 2$, where $f \in L^2(\mathcal{M})$ is a measure space signal. In the continuous case, we define the measure space ConvNet by

$$\{f^l_{k'}\}_{k'=1}^{K_l} = P(\psi^l)\Big(\rho\big\{P(\psi^{l-1}) b^l_{k'} + \sum_{k=1}^{K_{l-1}} a^l_{k'k}\ g^l_{k'k}(\mathcal{L}) f^{l-1}_k\big\}_{k'=1}^{K_l}\Big), \tag{23}$$

where $\{f^l_{k'}\}_{k'=1}^{K_l}$ is the signal at Layer l. Here, the input $P(\psi_0)f$ of Layer 1 is in $PW(\psi_0)$. To understand the role of the projection $P(\psi_l)$ in (23), note that spaces $PW(\psi_l)$ are not invariant under the activation function ρ in general. Thus, as part of the definition of the ConvNet on $L^2(\mathcal{M})$, after each application of ρ we project the result to $PW(\psi_l)$. Moreover, for typical choices of $\mathcal{L}$, like Laplace-Beltrami operator, the constant signal $b^l_{k'}$ is in $P(0)$, and thus $P(\psi_{l-1}) b^l_{k'} = b^l_{k'}$. More generally, we project $b^l_{k'}$ to the Paley-Wiener space by $P(\psi_{l-1}) b^l_{k'}$.

The graph and measure space ConvNets are defined by iterating formulas (22) and (23) respectively along the layers. We denote the mapping from the input of Layer 1 to Channel k of Layer l of the ConvNet by $\mathcal{N}_k^l$ for the measure space ConvNet, and by $\mathcal{N}_k^{j,l}$ for the graphs ConvNets $j = 1, 2$. Namely

$$f_k^l = \mathcal{N}_k^l P(\psi_0) f, \quad \tilde{f}_k^{j,l} = \mathcal{N}_k^{j,l} S_{j,1}^{\lambda} P(\psi_0) f. \tag{24}$$

We restrict ourselves to contractive activation functions, as defined next.

Definition 13 *The activation function ρ is called* ***contractive*** *if for every $y, z \in \mathbb{C}$ $|\rho(y) - \rho(z)| \leq |y - z|$.*

The contraction property also carries to $L^p(\mathcal{M})$ spaces. Namely, if ρ is contractive, then for every two signals p, g, $\|\rho(p) - \rho(g)\|_p \leq \|p - g\|_p$. For example, the ReLU and the absolute value activation functions are contractive.

We consider a generic pooling operator $Y^{j,l} : L^2(G^j, l-1) \to L^2(G^j, l)$. Typically, $Y^{j,l}$ is the max pooling or l^2 average pooling.

Definition 14 *Suppose that coarsening is done by collapsing sequences of nodes of G to single nodes in G'.* ***Max-pooling*** *is the non-linear operator that assigns the value*

$$[Ys](y) = \max\{s(q_1), \ldots, s(q_K)\}/\sqrt{K} \tag{25}$$

to the node y of G' with parent nodes $q_1, \ldots, q_K$ from G, that have the signal values $s(q_1), \ldots, s(q_K)$ respectively, where $s \in L^2(G)$ is the $\mathbb{R}_+$ valued signal. ***Average pooling*** *is defined similarly by*

$$[Ys](y) = \sqrt{\sum_{k=1}^{K} s^2(q_k)/K} \tag{26}$$

Note that in standard ConvNets of 2D images, pooling is defined via (25) without division by $\sqrt{K}$. We divide max pooling by $\sqrt{K}$ since in the transferability setting it makes sense to normalize the L^2 norm in the coarse graph $G^{j,l}$ (see for example the DSP setting of Subsection 5.1), while in standard ConvNets of 2D images the grid is not normalized. Max pooling is norm-reducing, as defined next.

Definition 15 *Pooling, $Y : L^2(G) \to L^2(G')$, is said to* ***reduce norm*** *if $\|Y(h)\| \leq \|h\|$ for every $h \in L^2(G)$.*

In Theorem 16 we assume that **pooling is consistent with sampling** in the sense that for every layer $l = 1, \ldots, L$ and $j = 1, 2$,

$$\forall f \in PW(\psi_l), \ \left\| Y^{j,l} S_{j,l-1}^{\psi_l} f - S_{j,l}^{\psi_l} f \right\| \leq \delta \|f\| \tag{27}$$

Equation (27) means that sampling to the graph $G^{j,l-1}$ and then pooling to the graph $G^{j,l}$ is approximately the same as sampling to graph $G^{j,l}$ directly.

In the following, we consider normalizations of the components of the ConvNet. In particular, assuming that sampling and interpolation are approximately isometries, we may normalize them with asymptotically small error to $\left\|S^{\psi_l}_{j,l}\right\| = 1, \left\|R^{\psi_l}_{j,l}\right\| = 1$. We also assume that pooling reduces norm,

Suppose that sampling asymptotically commutes with ρ (Definition 12), and let $0 < \delta < 1$ be some tolerance. By (19), it is possible to choose a sequence of bands ψ_l, and fine enough discretizations, guaranteeing

$$\begin{aligned}
&\forall f \in L^2(\mathcal{M}),\ j = 1, 2,\ l = 1, \ldots, L, \\
&\left\|\rho(S^{\psi_{l-1}}_{j,l-1} P(\psi_{l-1}) f) - S^{\psi_l}_{j,l-1} P(\psi_l) \rho(P(\psi_{l-1}) f)\right\| < \delta \left\|f\right\|.
\end{aligned}$$

Note that the band ψ^l increases in l, since the activation function ρ gradually increases the complexity of the signal. This leads us to the non-asymptotic setting of Theorem 16. Note as well that the ConvNet is invariant to multiplying all filters $g^l_{k',k}$ by a constant $\alpha \in \mathbb{R}$ and multiplying A^l by $1/\alpha$. Thus, in Theorem 16 we suppose that all filters are normalized to $\left\|g^l_{k',k}\right\|_\infty = 1$, and the norm of the convolution layers is controlled by A^l.

Theorem 16 *Consider a ConvNet with Lipschitz filters $\{g^l_{k'k} \mid {\scriptstyle k=1\ldots K_{l-1},\ k'=1\ldots K_l}\}$ with Lipschitz constant D at each layer l, normalized to $\left\|g^l_{k',k}\right\|_\infty = 1$, and with A^l satisfying $\left\|A^l\right\|_\infty \leq A$, for some constant $A > 0$. Suppose that the biases satisfy $\left\|b^l_{k'}\right\|_2 \leq B$ for some constant $B > 0$, for $\left\|b^l_{k'}\right\|_2$ denoting the norm of the constant signal $b^l_{k'}$ both in $L^2(G^{j,l})$ and in $L^2(\mathcal{M})$. Consider a contractive activation function ρ (Definition 13). Suppose that $S^{\psi_l}_{j,l}$ and $R^{\psi_l}_{j,l}$ are normalized to $\left\|S^{\psi_l}_{j,l}\right\| = 1, \left\|R^{\psi_l}_{j,l}\right\| = 1$. Let $0 < \delta < 1$ and suppose that for every $j = 1, 2$*

$$\begin{aligned}
\forall l = 0, \ldots, L-1, \quad & \left\|S^{\psi_l}_{j,l} \mathcal{L} P(\psi_l) - \boldsymbol{\Delta}_{j,l} S^{\psi_l}_{j,l} P(\psi_l)\right\| \leq \delta \\
& \left\|P(\psi_L) - R^{\psi_L}_{j,L} S^{\psi_L}_{j,L} P(\psi_L)\right\| \leq \delta \\
\forall f \in PW(\psi_{l-1}),\quad \forall l = 1, \ldots, L, \quad & \left\|\rho(S^{\psi_{l-1}}_{j,l-1} P(\psi_{l-1}) f) - S^{\psi_l}_{j,l-1} P(\psi_l) \rho(P(\psi_{l-1}) f)\right\| < \delta \left\|f\right\| \\
\forall f \in PW(\psi_l),\quad \forall l = 1, \ldots, L, \quad & \left\|Y^{j,l} S^{\psi_l}_{j,l-1} f - S^{\psi_l}_{j,l} f\right\| \leq \delta \left\|f\right\|
\end{aligned} \tag{28}$$

Suppose that pooling reduces norm (Definition 15).

Then, if $A > 1$,

$$\begin{aligned}
&\left\|R^{\psi_L}_{1,L} \mathcal{N}^{1,L}_k S^{\psi_0}_{1,0} P(\psi_0) f - R^{\psi_L}_{2,L} \mathcal{N}^{2,L}_k S^{\psi_0}_{2,L} P(\psi_0) f\right\| \\
&\quad \leq \Big(LD\sqrt{\#\{\lambda_m \leq \psi_L\}_m} + 2L + 2\Big)\Big(A^L \left\|f\right\| + B\frac{A^L - 1}{A - 1}\Big)\delta
\end{aligned} \tag{29}$$

and, if $A = 1$,

$$\begin{aligned}
&\left\|R^{\psi_L}_{1,L} \mathcal{N}^{1,L}_k S^{\psi_0}_{1,0} P(\psi_0) f - R^{\psi_L}_{2,L} \mathcal{N}^{2,L}_k S^{\psi_0}_{2,0} P(\psi_0) f\right\| \\
&\quad \leq \Big(LD\sqrt{\#\{\lambda_m \leq \psi_L\}_m} + 2L + 2\Big)(\left\|f\right\| + LB)\delta.
\end{aligned} \tag{30}$$

The proof of this theorem is in the appendix. Theorem 16 may hint to the importance of regularizing the convolution operators in the infinity norm. The next corollary shows that adding bias increases instability with respect to the depth L, from linear to quadratic.

Corollary 17 *Consider the setting of Theorem 16, with A^l normalized to $\left\|A^l\right\|_\infty = 1$, without biases, namely, $b^l_{k'} = 0$. Then*

$$\left\|R^{\psi_L}_{1,L}\mathcal{N}^{1,L}_k S^{\psi_0}_{1,0}P(\psi_0) - R^{\psi_L}_{2,L}\mathcal{N}^{2,L}_k S^{\psi_0}_{2,0}P(\psi_0)\right\| \leq \Big(LD\sqrt{\#\{\lambda_m \leq \psi_L\}_m} + 2L + 2\Big)\delta. \tag{31}$$

The assumptions of Corollary 17 imply that the ConvNet is contractive. For non-contractive ConvNets, we can simply consider a contractive ConvNet and multiply it by a constant $C > 1$. For such a ConvNet, the bound in (29) is simply multiplied by C.

4.3 Transferability Experiments

In this subsection we showcase transferability of spectral graph methods in practice. We first mention two papers that showcase the transferability of spectral graph ConvNets. In (Bianchi et al., 2021) graph spectral ConvNets are based on rational function filters. The task is graph classification on data sets consisting of many graphs and graph signals. Each graph represents a molecule and its signal represents some node features. The results reported in that paper show that the proposed spectral method obtains state of the art results on these multi-graph settings.

In (Knyazev et al., 2019) different types of graph ConvNets are tested on a machine learning tasks in imaging. Inputs are represented by superpixel images, namely some graphs representing the images. Here, the setting is multi-graph, where different images are represented by different graphs. The reported results suggest that spectral methods are more transferable, dealing better with the multi-graph setting than spatial methods.

Next we present simple experiments that demonstrate transferability. In figure 2 we showcase transferability under coarsening on the Bunny mesh. Here, the graph $\mathcal{M}$ consist of all mesh edges with weight 1, and we consider the normalized Laplacian $\mathcal{L}$. The coarsened version G of $\mathcal{M}$ is computed by the Graclus algorithm (Dhillon et al., 2004). We consider the coarsening operator $S : L^2(\mathcal{M}) \to L^2(G)$ defined as follows. Given a signal $f \in L^2(\mathcal{M})$, for each pair of nodes $x, y \in \mathcal{M}$ which collapse to the node $z \in G$, we define

$$[Sf](z) = \big(f(x) + f(z)\big)/\sqrt{2}.$$

We define the piecewise constant interpolation operator $R : L^2(G) \to L^2(\mathcal{M})$ by $R = S^*$. The **coarsened Laplacian $\boldsymbol{\Delta}$** in $L^2(G)$ is defined by $\boldsymbol{\Delta} = S\mathcal{L}R$. This definition of $\boldsymbol{\Delta}$ is natural when the goal is to promote transferability. We consider a signal $f \in L^2(\mathcal{M})$ and a filter g. The figure compares f, $\mathcal{L}f$, and $g(\mathcal{L})f$ with RSf, $R\boldsymbol{\Delta}Sf$, and $Rg(\boldsymbol{\Delta})Sf$ respectively.

In Figure 3 we showcase the transferability formula Thm.4(1) on the Bunny graph of Figure 2. We consider a filter g with Lipschitz bound D. On the left we plot the Laplacian transferability $\left\|\boldsymbol{\Delta}S^{\lambda_M}\phi_m - S^{\lambda_M}\mathcal{L}\phi_m\right\|$ as a function of the eigenvalue of the eigenvector ϕ_m of $\mathcal{L}$. In the middle we plot the filter transferability $\left\|S^{\lambda_M}g(\mathcal{L})\phi_m - g(\boldsymbol{\Delta})S^{\lambda_M}\phi_m\right\|$ as a function of the Laplacian transferability, with the theoretical bound $y = Dx$ in red. On the

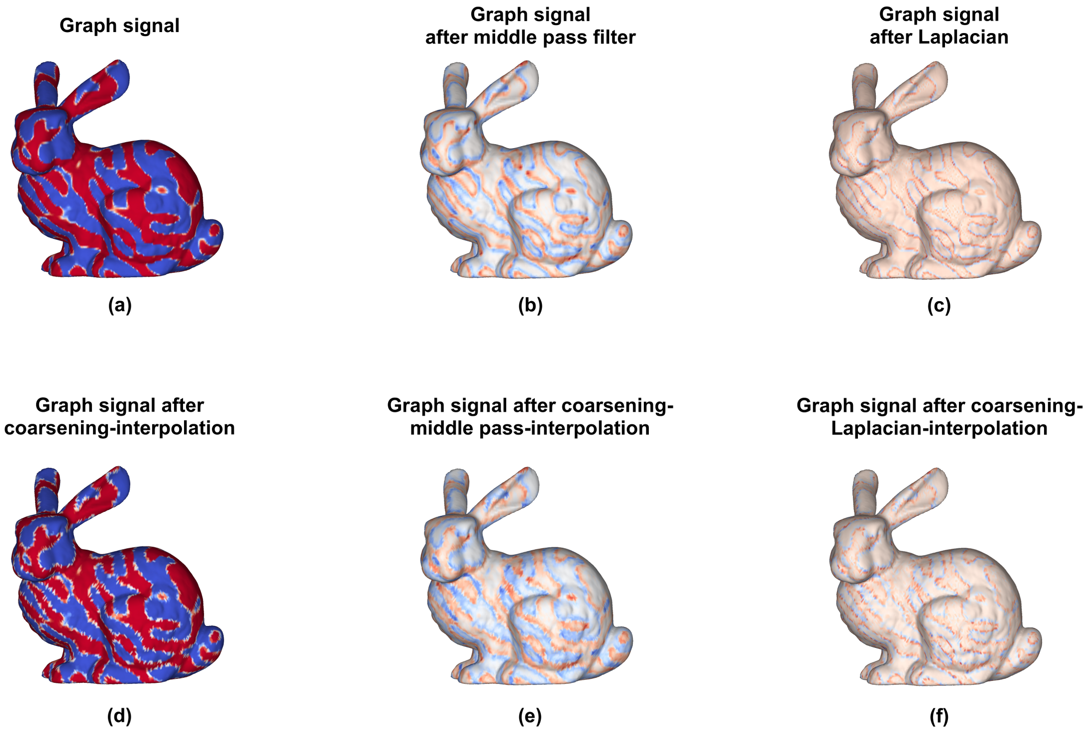

Figure 2: Transferability under coarsening on the Bunny mesh

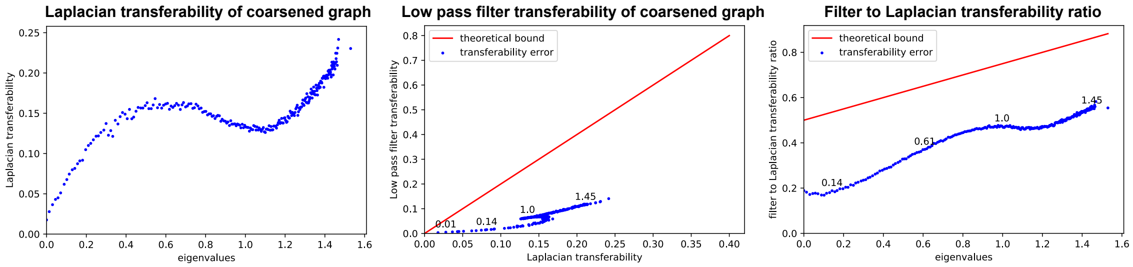

Figure 3: Transferability of $\mathcal{L}$-Fourier modes evaluated in G

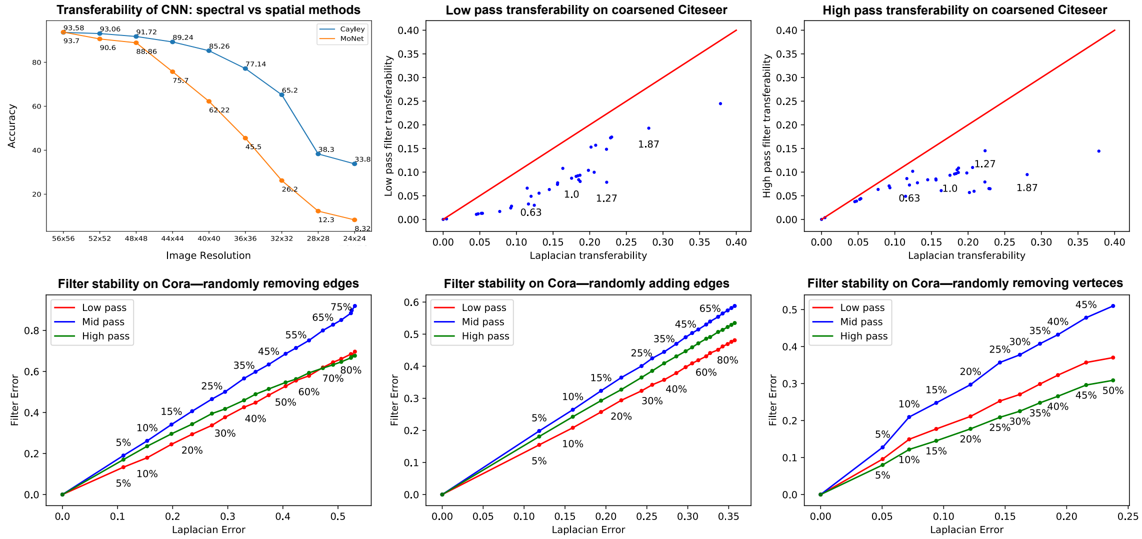

Figure 4: Trasferability experiments

right we plot the filter transferability divided by the Laplacian transferability of eigenvectors ϕ_m as a function of the eigenvalues λ_m. The theoretical bound $V_g(\lambda_m)$ is given in red.

In top-left of Figure 4 we isolate principle transferability form concept-based transferability in MNIST, and compare a spectral graph ConvNet method with a spatial graph ConvNet method. We consider a simple ConvNet architecture based on three convolution layers with max pooling, where the max pooling in the third layer collapses each graph to one node, and two fully connected layers. In CayleyNet, the Cayley polynomial order of all three convolutional layers is 9, and they produce 32, 32, and 64 output features, respectively. In MoNet, all three convolutional layers contain 18 Gaussian kernels, and produce 32, 32 and 64 output features, respectively. Both two models contain 10K parameters. We train the network on MNIST images of one fixed fine resolution ($56X56$) and test on images of various coarse resolutions. The graph Laplacian is given by the central difference approximating second derivative. In this setting, the spectral method, CayleyNet, has higher principle transferability than the spatial method, MoNet. Indeed, its performance degrades slower as we coarsen the grid.

In top-middle and right of Figure 4 we test transferability between the Citeseer graph $\mathcal{M}$ and its coarsened version G. We take the coarsening and interpolation operators S and $R = S^*$ as before. We consider the normalized Laplacian $\mathcal{L}$ on $\mathcal{M}$, and the coarse Laplacian $\boldsymbol{\Delta} = S\mathcal{L}R$ on G. We use low-pass (top-middle) and high-pass (top-right) filters with Lipschitz constant 1. To show transferability, we plot $\|Sf(\mathcal{L})\phi_m - f(\boldsymbol{\Delta})S\phi_m\|$ as a function of $\|S\mathcal{L}\phi_m - \boldsymbol{\Delta}S\phi_m\|$ for various eigenvectors ϕ_m of $\mathcal{L}$ (some corresponding eigenvalues are displayed). All values lie below $y = Dx$, where D is the Lipschitz constant of the corresponding filter. This accords with the transferability inequality Thm.4(1).

In Figure 4 bottom, we test the stability of spectral graph filters in the Cora graph with the normalized Laplacian, for different models of graph perturbations and sub-sampling. We consider three filters: low, mid and high pass. In bottom-left we randomly remove edges, in bottom-middle we randomly add edges, and in bottom-right randomly delete vertices, and compare the filters on the sub-sampled graph. The markers indicate the percentage of edges/vertices that were removed/added. The x axis is the relative error in the Laplacian, and the y axis is the relative error in the filter. The experimental results support the theoretical results on linear stability. All errors are given in Frobenius norm. The Frobenius norm can be seen as the average pointwise error, where the Laplacians and filters are applied on the signals of the standard basis.

5. Transferability of Graph Discretizing Topological Spaces

In this section we develop the DSP setting of transferability, in which graphs are sampled from continuous spaces, as described in Subsections 1.4.1 and 1.4.2. In the classical Nyquist—Shannon approach to digital signal processing, band-limited signals in $L^2(\mathbb{R})$ are discretized to $L^2(\mathbb{Z})$ by sampling them on a grid of appropriate spacing. The original continuous signal can be reconstructed from the discrete signal via interpolation, which is explicitly given as the convolution of the delta train corresponding to the discrete signal with a sinc function. Our goal is to formulate an analogous framework for graphs, where graphs are seen as discretizations of continuous entities, namely topological spaces with Borel measure.

Previous work studied sampling and interpolation in the context of graph signal processing, where the space that is sampled is a discrete graph itself. In (Anis et al., 2014; Chen et al., 2015; Tsitsvero et al., 2016; Puy et al., 2018) sampling is defined by evaluating the graph signal on a subset of vertices, and in (Segarra et al., 2015; Marques et al., 2016) sampling is defined by evaluating the signal on a single vertex, and using repeated applications of the shift operator to aggregate the signal information on this node. In the context of discretizing continuous spaces to graphs, considering graph Laplacians of meshes as discretizations of Laplace-Beltrami operators on Riemannian manifolds is standard. However, manifolds are too restrictive to model the continuous counterparts of general graphs. A more flexible model are more general topological spaces with Borel measure. Treating graph Laplacians as discretizations of metric space Laplacians was considered from a pure mathematics point of view in (Burago et al., 2019). In that work, the convergence of the spectrum of the graph Laplacian to that of the metric space Laplacian was shown under some conditions. However, for our needs, the explicit notion of convergence of Definition 9 is required, and the convergence of the spectrum alone is not sufficient. In (Lovász and Szegedy, 2006), a continuous limit object of graphs was proposed. More accurately, graph vertices are sampled from the continuous space $[0, 1]$, and graph weights are sampled from a measurable weight function $W : [0, 1]^2 \rightarrow [0, 1]$. In contract to this, in our analysis there is a special emphasis on Laplacians, which implicitly model the "geometry" of graphs and topological spaces. We thus bypass the analysis via edge weights, and study directly the discretization of topological Laplacians to graph Laplacians, from an operator theory point of view.

In this section we introduce a discrete signal processing setting, where analog domains are topological spaces, and digital domains are graphs. We present natural conditions, from a signal processing point of view, sufficient for the convergence of the graph Laplacian to the topological space Laplacian in the sense of Definition 9. We also prove asymptotic reconstruction, boundedness, convergence, and asymptotic commutativity of sampling with activation function (Definitions 7, 8, 9 and 12), under these conditions. All proofs are based on quadrature assumptions, stating that certain sums approximate certain integrals. In Subsection 5.4 we prove that the quadrature assumptions are satisfied in high probability, in case graphs are sampled randomly from topological spaces.

5.1 Sampling and Interpolation

We now proceed to give an explicit construction of the sampling and interpolation operators, under which they are asymptotically reconstructive and bounded (Definitions 7 and 8). The approach is similar to the classical Nyquist—Shannon approach to sampling and interpolation.

We start with basic notations and definitions. Let $\mathcal{M}$ be a topological space with a Borel measure μ, such that the volume $\mu(\mathcal{M})$ is finite. We call such a space a **topological-measure space**. Let the Laplacian $\mathcal{L}$ be a normal operator in $L^2(\mathcal{M})$, having discrete spectrum, with eigenvalues $\{\lambda_n\}_{n=1}^{\infty}$ and corresponding eigenvectors $\{\lambda_n, \phi_n\}_{n=1}^{\infty}$. Here, the eigenvectors λ_n are in increasing order of $|\lambda_n|$ and have repetitions if the corresponding eigenspace is more than one dimensional. Denote the Paley-Wiener spaces by $PW(\lambda)$, with projections $P(\lambda)$. Denote by M_λ the index such that λ_{M_λ} is the largest eigenvalue in its absolute value satisfying $\lambda_{M_\lambda} \leq |\lambda|$.

Let

$$G_n = \{x_k^n\}_{k=1}^{N_n} \subset \mathcal{M} \quad , \quad n \in \mathbb{N}$$

be a sequence of **sample sets**, where $N_n \in \mathbb{N}$ for every $n \in \mathbb{N}$. For the following analysis let us fix $n \in \mathbb{N}$. We see G_n as the nodes of a graph. Instead of analyzing the graph Laplacian through the graph adjacency matrix, we directly analyze the graph Laplacian. Consider a diagonalizable operator $\boldsymbol{\Delta}_n$ in each $L^2(G_n)$, that we call graph Laplacian. The graph Laplacian represents the diffusion or shift kernel in $L^2(G_n)$, and hence encapsulates some notion of geometry in $L^2(G_n)$. A non symmetric Laplacian indicates that the space $L^2(G_n)$ samples $L^2(\mathcal{M})$ non-uniformly, as described in Subsection 5.3. Denote the eigendecomposition of $\boldsymbol{\Delta}_n$ with eigenvalues κ_j^n and eigenvector $\boldsymbol{\gamma}_j^n$. Let $\boldsymbol{\Gamma}_n$ be the eigenvector matrix with columns $\boldsymbol{\gamma}_j^n$. Consider the inner product $\langle \mathbf{u}, \mathbf{v} \rangle_{L^2(G_n)} = \mathbf{v}^{\mathrm{H}} \mathbf{B}_n \mathbf{u}$ as defined in (5), with $\mathbf{B}_n = \boldsymbol{\Gamma}_n^{-\mathrm{H}} \boldsymbol{\Gamma}_n^{-1}$. When writing $L^2(G_n)$ we mean the space with the inner product $\langle \mathbf{u}, \mathbf{v} \rangle_{L^2(G_n)}$. Here, for normal $\boldsymbol{\Delta}_n$, $\mathbf{B}_n = \mathbf{I}$, and $\langle \mathbf{u}, \mathbf{v} \rangle_{L^2(G_n)}$ is the standard dot product.

The following construction is defined for a fixed Paley-Wiener space $PW(\lambda)$. We start by defining the **evaluation operator**, that evaluates signals in $PW(\lambda)$ at the sample set G_n. Since general signals in $L^2(\mathcal{M})$ are only defined up to a set of finite measure, to define the evaluation of signals at points, we restrict ourselves to continuous signals. Let $C(\mathcal{M})$ be the Banach space of continuous functions with the infinity norm. The space $C(\mathcal{M})$ is dense in $L^2(\mathcal{M})$. Note that delta functionals that evaluate at a point are well defined on $C(\mathcal{M})$, as elements of the continuous dual $C(\mathcal{M})^*$. Thus, the sampling operator S_n that evaluates at the sample points $\{x_k^n\}_k$ is a well defined bounded operator from $C(\mathcal{M})$ to $L^2(G_n)$. Since in our analysis we work in Paley-Wiener spaces, we consider the natural assumption that the Laplacian respects continuity.

Definition 18 *The Laplacian $\mathcal{L}$ is said to* ***respect continuity*** *if $PW(\lambda)$ is a subspace of $C(\mathcal{M})$ for every $\lambda > 0$.*

Note that Laplace-Beltrami operators on compact manifolds respect continuity, since their domain (L^2 functions with distributional Laplacian in L^2) is a subspace of $C(\mathcal{M})$.

We define the evaluation operator.

Definition 19 *Under the above construction, the* ***evaluation operator*** $\Phi_n^\lambda : PW(\lambda) \rightarrow L^2(G_n)$ *is defined by*

$$\Phi_n^\lambda f = \big(\frac{1}{\sqrt{h_n}} f(x_k^n)\big)_{k=1}^{N_n}, \tag{32}$$

where

$$h_n = \frac{N_n}{\mu(\mathcal{M})} \tag{33}$$

is the density of G_n *in* $\mathcal{M}$.

Consider the Fourier basis $\{\phi_m\}_{m=1}^{M_\lambda}$ of $PW(\lambda)$. Note that (32) can be written in this basis in the matrix form $\mathbf{\Phi}_n^\lambda$, with entries

$$\phi_{k,m} = \frac{1}{\sqrt{h_n}} \phi_m(x_k^n). \tag{34}$$

For a column vector $\mathbf{c} = (c_m)_{m=1}^{M_\lambda}$ and $f = \sum_{m=1}^{M_\lambda} c_m \phi_m$, observe that

$$\Phi_n^\lambda f = \mathbf{\Phi}_n^\lambda \mathbf{c}.$$

When defining sampling and interpolation, one should address the non-uniform density of the sample set entailed by the inner product (5). We thus consider the following definitions of sampling and interpolation.

Definition 20 *Under the above construction,* ***sampling*** $S_n^\lambda : PW(\lambda) \rightarrow L^2(G_n)$ *is defined to be the evaluation operator, with the matrix representation, where the input is in the Fourier basis* $\{\phi_m\}_{m=1}^{M_\lambda}$ *and the output in the standard basis of* $L^2(G_n)$*,*

$$\mathbf{S}_n^\lambda = \mathbf{\Phi}_n, \tag{35}$$

where $\mathbf{\Phi}_n$ *is a matrix with entries (34).* ***Interpolation*** $R_n^\lambda : L^2(G_n) \rightarrow PW(\lambda)$ *is defined as the operator with matrix representation, where the input is in the standard basis of* $L^2(G_n)$ *and the output is in the Fourier basis of* $PW(\lambda)$*,*

$$\mathbf{R}_n^\lambda = \mathbf{\Phi}_n^{\mathrm{H}} \mathbf{B}_n. \tag{36}$$

Claim 21 *The interpolation operator satisfies*

$$R_n^\lambda = (S_n^\lambda)^*. \tag{37}$$

Proof Let us derive a general formula for the adjoint of a linear mapping $PW(\lambda) \rightarrow L^2(G_n)$, represented as a matrix operator $\mathbf{A}$, where $PW(\lambda)$ is represented in the Fourier basis, and $L^2(G_n)$ in the standard basis. Note that the inner product in $PW(\lambda)$, represented in the Fourier basis, is the standard dot product. Thus, for any $\mathbf{c} \in \mathbb{C}^{M_\lambda}$ and $\mathbf{q} \in L^2(G_n) \cong \mathbb{C}^{N_n}$,

$$\langle \mathbf{A}\mathbf{c}, \mathbf{q} \rangle_{L^2(G_n)} = \mathbf{q}^{\mathrm{H}} \mathbf{B}_n \mathbf{A} \mathbf{c} = (\mathbf{A}^H \mathbf{B}_n \mathbf{q})^{\mathrm{H}} \mathbf{c} = \left\langle \mathbf{c}, \mathbf{A}^H \mathbf{B}_n \mathbf{q} \right\rangle_{\mathbb{C}^{M_\lambda}}.$$

Therefore

$$\mathbf{A}^{*} = \mathbf{A}^{H}\mathbf{B}_n.$$

From this, (37) follows as a special case. ■

Next, we would like to find a condition, for $f = \sum_{m=1}^{M_\lambda} c_m\phi_m$, that guarantees

$$\mathbf{R}_n^\lambda \mathbf{S}_n^\lambda \mathbf{c} \xrightarrow[n\rightarrow\infty]{} \mathbf{c}. \tag{38}$$

By requiring (38) for all elements of the Fourier basis, writing (38) using entry-wise limits, and arranging all limits as the entries of a matrix, we obtain the condition

$$\Big(\frac{\mu(\mathcal{M})}{N_n}\Big\langle\Big(\phi_m(x_k^n)\Big)_k, \Big(\phi_{m'}(x_k^n)\Big)_k\Big\rangle_{L^2(G_n)}\Big)_{m,m'} \xrightarrow[n\rightarrow\infty]{} \mathbf{I}. \tag{39}$$

The left hand side of (39) is interpreted as a quadrature approximation of the inner product $\langle\phi_m, \phi_{m'}\rangle_{L^2(\mathcal{M})}$, based on the sample points $\{x_k^n\}_{k=1}^{N_n}$ and their density. We summarize this in a definition.

Definition 22 *Consider the above construction and notations. Denote by $\langle\mathbf{\Phi}_n, \mathbf{\Phi}_n\rangle \in \mathbb{C}^{M_\lambda\times M_\lambda}$ the matrix with entries $\big\langle\Phi_n^\lambda\phi_m, \Phi_n^\lambda\phi_{m'}\big\rangle_{L^2(G_n)}$. The pair $\{G_n, \mathbf{\Delta}_n\}_{n=1}^\infty$ is called a **quadrature sequence with respect to reconstruction**, if*

$$\langle\mathbf{\Phi}_n, \mathbf{\Phi}_n\rangle \xrightarrow[n\rightarrow\infty]{} \mathbf{I}. \tag{40}$$

Next, we prove that quadrature sequences are asymptotically reconstructive and bounded.

Proposition 23 *Consider the above construction and notations, with $\{G_n, \mathbf{\Delta}_n\}_{n=1}^\infty$ a quadrature sequence and $\mathcal{L}$ that respects continuity. Then sampling and interpolation are asymptotically reconstructive and bounded (Definitions 7 and 8), with bound $C = 1$ in Definition 8.*

Proof The proof of Definition 7 is given by the above analysis. For Definition 8, Definition 22 asserts that $\mathbf{S}_n^\lambda$ approximates an isometric embedding. More accurately, for two vectors $\mathbf{c}_1, \mathbf{c}_2$ of Fourier coefficients, by (40)

$$\Big\langle\mathbf{S}_n^\lambda\mathbf{c}_1, \mathbf{S}_n^\lambda\mathbf{c}_2\Big\rangle = \mathbf{c}_2^{\mathrm{H}}(\mathbf{S}_n^\lambda)^{\mathrm{H}}\mathbf{B}_n\mathbf{S}_n^\lambda\mathbf{c}_1$$

$$= \mathbf{c}_2^{\mathrm{H}}\,\langle\mathbf{\Phi}_n, \mathbf{\Phi}_n\rangle\,\mathbf{c}_1.$$

For $\mathbf{c}_1 = \mathbf{c}_2 = \mathbf{c}$ we have

$$\Big\|\mathbf{S}_n^\lambda\Big\| = \Big\|\langle\mathbf{\Phi}_n, \mathbf{\Phi}_n\rangle^{1/2}\Big\|.$$

Thus, since $PW(\lambda)$ has a fixed finite dimension with-respect-to n, and since convergence in matrix norm is equivalent to entry-wise convergence, by Definition 22 we have

$$\Big\|\mathbf{S}_n^\lambda\Big\| = \Big\|\langle\mathbf{\Phi}_n, \mathbf{\Phi}_n\rangle^{1/2}\Big\| \xrightarrow[n\rightarrow\infty]{} 1.$$

Finally, by Claim 21, $\mathbf{R}_n^\lambda = (\mathbf{S}_n^\lambda)^*$, and thus $\|\mathbf{R}_n^\lambda\| = \|\mathbf{S}_n^\lambda\|$. ■

5.2 Asymptotic Commutativity of Sampling and Activation Functions

In this section we prove that Sampling asymptotically commutes with the activation function (Definition 12) under some quadrature conditions. Definition 12 involves a term of the form

$$\left\|\rho(S_n^{\lambda}P(\lambda)f) - S_n^{\lambda'}P(\lambda')\rho(P(\lambda)f)\right\|. \tag{41}$$

Let us first show how to swap the order between sampling and ρ in $\rho(S_n^{\lambda}P(\lambda)f)$. For every continuous $\rho : \mathbb{C} \to \mathbb{C}$ and $f \in C(\mathcal{M})$, we also have $\rho(f) \in C(\mathcal{M})$. Moreover, $S_n\rho(f) = \rho(S_n f)$ for every continuous f. Thus, assuming that $\mathcal{L}$ respects continuity, sampling $S_n^{\lambda} = S_n$ does not depend on λ, and $\rho(S_n^{\lambda}P(\lambda)f) = \rho(S_nP(\lambda)f) = S_n\rho(P(\lambda)f)$ for any continuous activation function ρ. As a result, for continuous ρ, (41) takes the form

$$\left\|\rho(S_n^{\lambda}P(\lambda)f) - S_n^{\lambda'}P(\lambda')\rho(P(\lambda)f)\right\| = \left\|S_n\rho(P(\lambda)f) - S_nP(\lambda')\rho(P(\lambda)f)\right\|. \tag{42}$$

The right hand side of (42) can be seen as a quadrature approximation of $\|\rho(P(\lambda)f) - P(\lambda')\rho(P(\lambda)f)\|$, which leads us to the following assumption.

Definition 24 *The sampling operators $\{S_n^{\lambda}\}_{\lambda>0}$ are said to be* ***quadrature with respect to the continuous activation function*** *ρ, if $\mathcal{L}$ respects continuity, and for every $f \in L^2(\mathcal{M})$ and $\lambda' > \lambda > 0$,*

$$\lim_{n\to\infty} \left\|S_n\rho(P(\lambda)f) - S_nP(\lambda')\rho(P(\lambda)f)\right\| = \left\|\rho(P(\lambda)f) - P(\lambda')\rho(P(\lambda)f)\right\|.$$

Next, we focus on a common class of activation functions, that include ReLU, absolute value, and absolute value or ReLU of the real or imaginary part of a complex number.

Definition 25 *Consider the field $\mathbb{R}$ or $\mathbb{C}$, and denote it by $\mathbb{F}$. The continuous activation function $\rho : \mathbb{F} \to \mathbb{F}$ is called* ***positively homogeneous*** *of degree 1, if for every $z \in \mathbb{F}$ and every real $c \geq 0$,*

$$\rho(cz) = c\rho(z).$$

Proposition 26 *Consider a DSP framework, quadrature with respect to reconstruction. Consider a contractive positively homogeneous activation function ρ of degree 1. Suppose that $\mathcal{L}$ respects continuity and that the sampling operators are quadrature with respect to the continuous activation function ρ. Then sampling asymptotically commutes with ρ (Definition 12).*

The proof is in Appendix B.3.

5.3 Convergence of Sampled Laplacians to Topological Space Laplacians

In this subsection we discuss different definitions of topological Laplacians and their discretizations to graph Laplacians via sampling. We show convergence of the graph Laplacians to the topological-measure Laplacians, in the sense of Definition 9, under a quadrature assumption.

Assume that $\mathcal{M}$ is a compact metric space with finite Borel measure $\mu(\mathcal{M}) < \infty$. Since such a measure space is a probability space up to normalization, we assume that

$\mu(\mathcal{M})=1$. Let $S_r(x_0), B_r(x_0)$ denote the sphere and ball or radius r about x_0 respectively. One definition of the Laplacian in the Euclidean space of dimension d is

$$\mathcal{L}f(x_0) := \lim_{r\rightarrow 0}\frac{2d}{r^2}\Big(A\big(S_r(x_0)\big)^{-1}\int_{S_r(x_0)} f(x)dx - f(x_0)\Big).$$

By integrating on the radius r', from 0 to r, with weights $V\big(S_{r'}(x_0)\big)^{-1}A\big(S_{r'}(x_0)\big)$, and using the mean value theorem for integrals, we obtain the equivalent definition

$$\mathcal{L}f(x_0) = \lim_{r\rightarrow 0} V\big(B_r(x_0)\big)^{-1}\int_{B_r(x_0)}\frac{2d}{\left|x-x_0\right|^2}\big(f(x)-f(x_0)\big)dx.$$

Another equivalent definition for the Laplace-Beltrami operator on manifolds of dimension d is

$$\mathcal{L}f(x_0) = \lim_{r\rightarrow 0}(2d+2)V\big(B_r(x_0)\big)^{-1}r^{-2}\int_{B_r(x_0)}\big(f(x)-f(x_0)\big)dx.$$

This motivates two classes of Laplacians in general metric-measure spaces. First, an infinitesimal definition

$$\mathcal{L}f(x_0) = \lim_{r\rightarrow 0} V\big(B_r(x_0)\big)^{-1}r^{-2}\int_{B_r(x_0)} H(x_0,x)\big(f(x)-f(x_0)\big)dx, \tag{43}$$

where a prototypical example is $H(x_0,x)=1$, for which (43) are termed **Korevaar-Schoen type energies** (Korevaar and Schoen, 1993). Second, a non-infinitesimal definition

$$\mathcal{L}f(x_0) = \int_{\mathcal{M}} H(x_0,x)\big(f(x)-f(x_0)\big)dx, \tag{44}$$

where a prototypical example is $H(x_0,x)=V\big(B_r(x_0)\big)^{-1}r^{-2}\chi_{B_r(x_0)}$ for some fixed radius r. Here, $\chi_{B_r(x_0)}$ is the characteristic function of the ball $B_r(x_0)$. Formulas (43) and (44) define symmetric operators in case $H(x,x_0)=H(x_0,x)$. Indeed, (44) is a sum of an integral and a multiplicative operator, both symmetric. Moreover, the symmetric property is preserved under limits in (43), since the limit commutes with the inner product.

In (Burago et al., 2019) it was shown, under some mild conditions, that (44) with $H(x,x_0)=r^{-2}\chi_{B_r(x_0)}$ is a self-adjoint operator with spectrum supported in $[0,2r]$. Moreover, the part of the spectrum in $[0,\ r)$ is discrete, and the eigenvalues of the sampled Laplacian in $[0,\ r)$ converge to the eigenvalues of the continuous Laplacian, assuming that sampling becomes denser in n in some sense.

The advantage of Laplacians of the form (44) is that they are readily discretizable on sample sets, by approximating the integral in (44) by a sum over the sample set. Suppose that H is symmetric ($H(x,x_0)=H(x_0,x)$), and consider a continuous weight function $w:\mathcal{M}\rightarrow\mathbb{R}_+$. For a detailed explanation of the role of w we refer to Subsection 5.4. Given a sample set $G_n=\{x^n_k\}_{k-1}^{N_n}$, define the discrete Laplacian $\mathbf{\Delta}_n$ acting on a vector $\mathbf{q}$ by

$$[\mathbf{\Delta}_n\mathbf{q}]_k = \frac{1}{\sqrt{N_n}}\sum_{k'=1}^{N_m}\frac{1}{w(x^n_{k'})}H(x^n_k,x^n_{k'})q_{k'}. \tag{45}$$

For $q_{k'} = f(x^n_{k'})$, (45) is interpreted as a quadrature approximation of (44). It is easy to show that the inner product (5) under which $\mathbf{\Delta}_n$ is self-adjoint is based on

$$\mathbf{B}_n = \mathrm{diag}\{\frac{1}{N_n w(x^n_k)}\}_{k=1}^{N_n}, \tag{46}$$

where $\mathbf{A} = \mathrm{diag}\{v_j\}_{j=1}^N$ is the diagonal matrix with diagonal entries $a_{j,j} = v_j$.

For our analysis, we relax the assumption that $\mathcal{M}$ is a compact metric space to a compact topological space. We further assume that the Laplacian $\mathcal{L}$ has discrete spectrum in the sense of Definition 1. However, for continuous H on a compact topological space $\mathcal{M}$, any Laplacian (44) is bounded, and thus has a discrete spectrum in the sense of Definition 1 only if the range of $\mathcal{L}$ is finite-dimensional. We thus approximate Laplacians $\mathcal{L}$ having discrete spectrum in two steps. First, by a finite-dimensional Laplacian of the form (44), and then, by the discretization (45).

The approximation of $\mathcal{L}$ by a finite-dimensional Laplacian works as follows.
Let $\{\lambda_m, \phi_m\}_{m=1}^{\infty}$ be the eigendecomposition of $\mathcal{L}$, and $\overline{\lambda}$ be some large band. Denote $\overline{M} = M_{\overline{\lambda}}$. We define the integral operator

$$\mathcal{L}^{\overline{\lambda}} f(x_0) = \int_x H(x_0, x) f(x) dx \tag{47}$$

based on the kernel

$$H_{\overline{\lambda}}(x_0, x) = \sum_{m=1}^{\overline{M}} \phi_m(x_0) \lambda_m \overline{\phi_m(x)}. \tag{48}$$

It is easy to see that

$$\mathcal{L}^{\overline{\lambda}} = \mathcal{L} P(\overline{\lambda}). \tag{49}$$

Therefore, for every $f \in L^2(\mathcal{M})$, we have $\lim_{\overline{\lambda} \rightarrow \infty} \mathcal{L}^{\overline{\lambda}} f = \mathcal{L} f$. Moreover, by (49) for every $f \in PW(\lambda)$ with $\lambda < \overline{\lambda}$, we have $\mathcal{L}^{\overline{\lambda}} f = \mathcal{L} f$.

We then treat the total approximation of $\mathcal{L}$ by a graph Laplacian using some sort of a diagonal extraction method. This is explained in Theorem 32 of Section 5.4. For now, let us focus on the non-asymptotic Laplacian $\mathcal{L}^{\overline{\lambda}}$ of (47) with discrete spectrum, denoted by abuse of notation by $\mathcal{L}$ where λ is fixed. To guarantee that the sequence of graph Laplacians as sampling operators are convergent (Definition 9) we consider the following quadrature assumption.

Definition 27 *Under the above construction, $G_n = \{x^n_k\}_{k-1}^{N_n}$ is a **quadrature sequence with respect to** $\mathcal{L}$, if for every $P(\lambda) f \in PW(\lambda)$*

$$\lim_{n \rightarrow \infty} \left\| S^{\lambda}_n \mathcal{L} P(\lambda) f - \mathbf{\Delta}_n S^{\lambda}_n P(\lambda) f \right\|_{L^2(G_n)} = 0.$$

Proposition 28 *Consider the above construction, with radon space $\mathcal{M}$, Laplacian $\mathcal{L}$ with discrete spectrum, and Paley-Wiener projections $P(\lambda)$. Consider a sampling sequence $\{S^{\lambda}_n\}_{n,\lambda}$ based on the sample points G_n, $n = 1, \ldots, \infty$, where G_n is quadrature sequence with respect to $\mathcal{L}$. Then $\mathbf{\Delta}_n$ converges to $\mathcal{L}$ in the sense of Definition 9.*

Proof

The operator $A_n = S_n^\lambda \mathcal{L} - \boldsymbol{\Delta}_n S_n^\lambda$ maps the M_λ dimensional space $PW(\lambda)$ to an M_λ dimensional space $W_n \subset L^2(G_n)$ containing the space $A_n PW(\lambda)$. Consider an isometric isomorphism $Q_n : W_n \to PW(\lambda)$. The operators $Q_n A_n : PW(\lambda) \to PW(\lambda)$ converge to zero as $n \to \infty$ in the strong topology, and since $PW(\lambda)$ is finite-dimensional, $Q_n A_n$ converge to zero also in the operator norm topology. Thus, since Q_n preserves norm, A_n converges to zero in the operator norm topology, which proves convergence as defined in Definition 9. ■

5.4 Transferability of Random Graph Laplacians

In this section we show that under some setting of random sampling of Laplacians $\mathcal{L}$ that respect continuity, graph Laplacian, sampling operators, and interpolation operators are asymptotically reconstructive, bounded, convergent, and sampling asymptotically commutes with the activation function (Definitions 7,8,9 and 12). To model the arbitrariness in which graphs can be sampled from topological-measure spaces, we suppose that the sample points $\{x_k^n\}_{k-1}^{N_n}$ are chosen at random. This allows us to treat the graph Laplacians as Monte-Carlo approximations of the topological-measure Laplacian.

Let $f = P(\lambda)f \in PW(\lambda)$. Consider a weighted μ measure, μ_w, defined for measurable sets $X \subset \mathcal{M}$ by

$$\mu_w(X) := \int_X w(x) d\mu(x). \tag{50}$$

Here, the weight function $w : \mathcal{M} \to \mathbb{R}$ is positive, continuous, and satisfies

$$\int_{\mathcal{M}} w(x) d\mu(x) = 1.$$

We take $\{x_k^n\}_{k-1}^{N_n}$ as random points in the probability space $\{\mathcal{M}, \mu_w\}$.

Definition 29 *Let $\{\mathcal{M}, \mu\}$ be a compact topological-measure space with $\mu(\mathcal{M}) = 1$. Let the weighted measure μ_w satisfy (50). Let $\mathcal{L}$ be a symmetric Laplacian of the form (44), such that $H \in L^2(\mathcal{M}^2)$. Suppose that $\mathcal{L}$ respects continuity and has discrete spectrum. Let $\{x_k^n\}_{k-1}^{N_n}$ be N_n random points from the probability space $\{\mathcal{M}, \mu_w\}$. The **random sampled Laplacian** $\boldsymbol{\Delta}_n$ is a random variable $\{\mathcal{M}^{N_n}; \mu_w^{N_n}\} \to \mathbb{C}^{N_n \times N_n}$, defined by (45) for the random samples $\{x_k^n\}_{k-1}^{N_n}$. The **random sampling and interpolation operators** S_n^λ, R_n^λ are defined as in Definition 20 on the random points $\{x_k^n\}_{k-1}^{N_n}$, with the inner product structure (46) of $L^2(G_n)$.*

For Theorem 32 below, we need one more assumption on ρ and $\mathcal{L}$. Let us consider for motivation the standard Laplacian $\mathcal{L}$ on the unit circle, and the ReLU activation function. Consider the classical Fourier basis $\{\phi_n\}_{n=-\infty}^{\infty}$. Any $f \in PW(\lambda)$ is smooth, and $\rho(f)$ is piecewise smooth and continuous. Thus $\rho(f)$ can be differentiated term-by-term, and

$$\|\partial_x \rho(f)\|_2^2 = 4\pi^2 \sum_{n=-\infty}^{\infty} n^2 \left|\langle \rho(f), \phi_n \rangle\right|^2 .$$

On the other hand, observe that for ReLU

$$\|\rho(f)\|_2 \leq \|f\|_2 \ , \quad \|\partial_x \rho(f)\|_2 \leq \|\partial_x f\|_2 \, . \tag{51}$$

Thus

$$\sum_{n=-\infty}^{\infty} n^2 \left|\langle \rho(f), \phi_n\rangle\right|^2 \leq \sum_{n=-M_\lambda}^{M_\lambda} n^2 \left|\langle f, \phi_n\rangle\right|^2 \leq M_\lambda^2 \|f\|_2^2 \, . \tag{52}$$

We can now show the following claim

Claim 30 *The ReLU function ρ is a continuous mapping of signals from $PW(\lambda)$ to signals in the norm*

$$\|h\|_{1+\kappa,2} = \sqrt{\left|\langle h, \phi_0\rangle\right|^2 + \sum_{n=-\infty}^{\infty} |n|^{1+\kappa} \left|\langle h, \phi_n\rangle\right|^2} \tag{53}$$

for any $0 < \kappa < 1$.

The proof of this claim in in Appendix B.6.

This analysis motivates the following definition in the general case.

Definition 31 *The activation function ρ is said to* ***preserve spectral decay*** *if there exists $\kappa > 0$ such that for every λ, the activation function ρ applied on signals from $PW(\lambda)$ is continuous in the norm*

$$\|h\|_{\kappa,2} = \sqrt{\sum_{n=1}^{\infty} |n|^{1+\kappa} \|\phi_n\|_\infty^2 \left|\langle h, \phi_n\rangle\right|^2}. \tag{54}$$

Note that in the finite-dimensional domain $PW(\lambda)$, all norms are equivalent. Thus, for ρ that preserves spectral decay,

$$\lim_{\|f-g\|_2 \to 0} \sqrt{\sum_{n=1}^{\infty} |n|^{1+\kappa} \|\phi_n\|_\infty^2 \left|\langle \rho(f) - \rho(g), \phi_n\rangle\right|^2} = 0, \tag{55}$$

where the limit is over $f, g \in PW(\lambda)$.

Preservation of spectral decay is interpreted as follows. Applying ρ on a band-limited signal $f \in PW(\lambda)$ results in a continuous signal which is not band-limited and in general has frequency coefficients in all frequencies. Namely, after applying ρ on f, which decays rapidly in the frequency domain, $\rho(f)$ is not guaranteed to decay rapidly. However, under Definition 31, $\rho(f)$ is guaranteed to have some decay rate in the frequency domain, since the weighted sum (54), with weights increasing to ∞ in frequency, is finite.

The following notation is used in the asymptotic analysis in Theorem 32. For any $M \in \mathbb{N}$ denote

$$\left\|\boldsymbol{\lambda}^M\right\|_1 = \sum_{m=1}^{M} |\lambda_m| \, . \tag{56}$$

Theorem 32 *Let $\{\mathcal{M}, \mu\}$ be a probability topological-measure space, and μ_w another measure satisfying (50) with positive and continuous w. Let $\mathcal{L}$ be a topological-measure Laplacian with discrete spectrum that respects continuity. Let ρ be a contractive positively homogeneous of degree 1 activation function that preserves spectral decay. Consider a sequence of random μ_w sample sets $\{x_k^n\}_{n=1}^{N_n}$, $n \in \mathbb{N}$, with $N_n \xrightarrow[n\to\infty]{} \infty$. Then, for every series of bands $\overline{\lambda}_n \xrightarrow[n\to\infty]{} \infty$, such that $\left\|\boldsymbol{\lambda}^{M_{\overline{\lambda}_n}}\right\|_1 = o(N_n^{1/2})$, and random sampled Laplacians $\boldsymbol{\Delta}_n = \boldsymbol{\Delta}_n^{\overline{\lambda}_n}$ with $\mathcal{L}^{\overline{\lambda}_n}$ defined by (47) and (48), and for every $\delta > 0$, in probability 1 there exists a subsequence $n_m \subset \mathbb{N}$ such that the following holds:*

i *for every $n \in \mathbb{N}$ we have $n \in \{n_m\}_{m\in\mathbb{N}}$ in probability more than $(1-\delta)$, and*

ii *the sampled Laplacians $\{\boldsymbol{\Delta}_{n_m}\}_m$ satisfy Definitions 7,8,9 and 12.*

Remark 33 *The sequence of random sample sets is treated formally in the following fashion. The basis of the topology of a sequence of topological spaces is defined as follows. A generic set in the basis of the topology is constructed by choosing finitely many indexes and picking an open set for each of the corresponding spaces. For each of the rest of the indexes we pick the whole corresponding probability space. The measure of such sets is the product of the measures of the sets of the finite subsequence.*

By Theorems 4 and 17, Theorem 32 is interpreted as follows. If $\boldsymbol{\Delta}_n$ are sampled from $\mathcal{L}$ by drawing N_n random sample points and sampling band-limited approximations of $\mathcal{L}$, where the bands do not increase too fast with respect to N_n, then graph filters and ConvNets approximate topological-measure filters and ConvNets. Therefore, graph filters and ConvNets are transferable. The explicit bounds on different transferability terms in (28) of Theorem 16 are given in Appendix B.4.

Last, let us use the results in Theorem 32 and Appendix B.4 to derive non-asymptotic bounds for the transferability error of filters.

Proposition 34 *Consider the setting of Theorem 32, where we choose λ_n such that*

$$\left\|\boldsymbol{\lambda}^{M_{\overline{\lambda}_n}}\right\|_1 \leq B N_n^{1/2-\alpha},$$

where $B > 0$ is some constant, and $\alpha \in (0, 1/2]$. Let g be a Lipschitz continuous filter, with Lipschitz constant D, and let $\|g\|_{\mathcal{L},M}$ be as defined in (8). Denote $w_{\min} = \min_{x\in\mathcal{M}} w(x)$. Then, for each n, with probability more than $1-2\delta$,

$$\begin{aligned}&\left\|g(\mathcal{L})P(\lambda) - R_n^{\lambda} g(\boldsymbol{\Delta}_n) S_n^{\lambda} P(\lambda)\right\| \\ &\leq M_{\lambda}\Big(2DBw_{\min}^{-1} \max_{m\leq M_{\lambda}} \|\phi_m\|_{\infty} N_n^{-\alpha} + \|g\|_{\mathcal{L},M}\, w_{\min}^{-1/2} \max_{m\leq M_{\lambda}} \|\phi_m\|_{\infty}^2 N_n^{-1/2}\Big)\delta^{-1/2}.\end{aligned} \tag{57}$$

In Proposition 34, different choices of $\alpha \in (0, 1/2]$ correspond to different choices of the Laplacian discretization. The choice $\alpha = 1/2$ means that we discretize a fixed Paley-Wiener projection of $\mathcal{L}$, and the closer α is to 0, the faster the band of $\mathcal{L}$ that we approximate goes to infinity in n.

6. Conclusion

In this paper, we proved that spectral graph filters and ConvNets are transferable. We took the philosophical point of view in which a ConvNet is called transferable, if, whenever two graphs represent the same phenomenon, the ConvNet has approximately the same repercussion on both graphs. We modeled mathematically "graph representing a phenomenon" as a graph which is sampled from an underlying "continuous" Borel space. Here, sampling is treated very broadly, and two examples are sampling by evaluating at sample points, and graph coarsening. We modeled mathematically "ConvNet having approximately the same repercussion" via the sampling-interpolation approach. Using this model, we were able to prove that spectral ConvNets are transferable. It is interesting to note that, after the publication of the current paper, Nilsson and Bresson (2020) tested ChebNet, a spectral ConvNet, on a set of multi-graph benchmark problems, with the goal of testing our results experimentally. The results showed that ChebNet outperforms vanilla spatial methods, especially in settings where the graphs are synthetically generated from an underlying continuous model. This validates that spectral methods indeed have competitive transferability capabilities in practice.

We believe that our paradigm of treating transferability by modeling "graphs representing the same phenomenon" and "ConvNet having the same repercussion" is a good starting point for any future research on graph ConvNet transferability. Such research should focus on modeling these concepts mathematically, justifying the model experimentally or heuristically, and proving corresponding transferability error bounds.

Acknowledgments

R.L. acknowledges support by the DFG SPP 1798 "Compressed Sensing in Information Processing" through Project Massive MIMO-II.

M.B. is partially supported by the ERC Consolidator grant No. 724228 (LEMAN)

G.K. acknowledges partial support by the Berlin Mathematics Research Center MATH+ through Project EF1x1, the DFG SPP 1798 "Compressed Sensing in Information Processing" through Project Massive MIMO-II, and the BMBF through Project MaGriDo.

Appendix A. Laplacians of Directed Graphs as Normal Operators

Next we explain how functional calculus applies as-is to non-normal matrices, even though the theory is defined only for normal operators. As a result, spectral filters can be defined on directed graphs represented by non-symmetric adjacency matrices.

Every finite-dimensional normal operator has an eigendecomposition with complex eigenvalues and orthonormal eigenvectors. Functional calculus applies to finite-dimensional normal operators by (2), and is canonical in the sense that it is equivalent to compute a rational function of a normal operator by (2), or by compositions, linear combinations, and inversions by (3). On the other hand, any diagonalizable matrix can be seen as a normal operator, considering an appropriate inner product. Moreover, almost any matrix is diagonalizable. Eigendecomposition and functional calculus are theories of self-adjoint/unitary/normal operators, which need not be represented by symmetric/orthonormal/normal matrices. Thus, spectral graph theory applies also to directed graphs. Note that no eigendecomposition is ever calculated in practice, and all computations in applying filters (compositions, linear combinations, and inversions) are algebraic and do not depend on the inner product structure. Thus, the theory applies as-is on directed graphs, with no extra considerations. We thus focus on finite-dimensional normal Laplacian operators, which can represent non-symmetric Laplacian matrices on directed graphs.

Given an $N \times N$ diagonalizable matrix $\mathbf{A}$ with eigenvectors $\{\boldsymbol{\gamma}_k\}_{k=1}^N$, consider the matrix $\mathbf{\Gamma}$ comprising the eigenvectors as columns. Define the inner product

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{v}^{\mathrm{H}} \mathbf{B} \mathbf{u}, \tag{58}$$

where $\mathbf{B} = \mathbf{\Gamma}^{-\mathrm{H}} \mathbf{\Gamma}^{-1}$ is symmetric, $\mathbf{u}$ and $\mathbf{v}$ are given as column vectors, and for a matrix $\mathbf{C} = (c_{m,k})_{n,m} \in \mathbb{C}^{N \times N}$, the Hermitian transpose $\mathbf{C}^{\mathrm{H}}$ is the matrix consisting of entries $c^{\mathrm{H}}_{m,k} = \overline{c_{k,m}}$. It is easy to see that (58) defines an inner product for which $\mathbf{A}$ is normal. Consider an operator A represented by the matrix $\mathbf{A}$. The adjoint A^* of an operator A is defined to be the unique operator such that

$$\forall \mathbf{u}, \mathbf{v} \in \mathbb{C}^d, \quad \langle A\mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{u}, A^* \mathbf{v} \rangle .$$

By the equality

$$\mathbf{v}^{\mathrm{H}} \mathbf{B} \mathbf{A} \mathbf{u} = \mathbf{v}^{\mathrm{H}} \mathbf{B} \mathbf{A} \mathbf{B}^{-1} \mathbf{B} \mathbf{u} = \left(\mathbf{B}^{-1} \mathbf{A}^{\mathrm{H}} \mathbf{B} \mathbf{v} \right)^{\mathrm{H}} \mathbf{B} \mathbf{u},$$

the matrix representation of the adjoint A^* is given by

$$\mathbf{A}^* = \mathbf{B}^{-1} \mathbf{A}^{\mathrm{H}} \mathbf{B}. \tag{59}$$

Thus, an operator is self-adjoint if $\mathbf{B}^{-1}\mathbf{A}^{\mathrm{H}}\mathbf{B} = \mathbf{A}$, unitary if $\mathbf{B}^{-1}\mathbf{A}^{\mathrm{H}}\mathbf{B} = \mathbf{A}^{-1}$, and normal if

$$\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^{\mathrm{H}}\mathbf{B} = \mathbf{B}^{-1}\mathbf{A}^{\mathrm{H}}\mathbf{B}\mathbf{A}.$$

Note the difference between transpose and adjoint, and between symmetric/orthonormal matrices and self-adjoint/unitary operators: a non-symmetric matrix may represent a self-adjoint operator. To emphasize this difference, we opt in this paper for a Hilbert space formulation of inner products and basis expansions, over the more commonly used formulation in the graph signal processing community of matrix products and dot products.

The eigenvalues and eigenspaces of a diagonalizable matrix, and the eigenvalues and eigenspaces of the corresponding normal operator, are identical. Indeed, eigenvalues and eigenspaces are defined algebraically, independently of the inner product structure. If the eigenvalues of the matrix are real or in $e^{i\mathbb{R}}$, then the corresponding operator is self-adjoint or unitary respectively.

Appendix B. Proofs

B.1 Proof of Theorem 4

By linearity and finite dimension of $PW(\lambda)$, we start with a signal $\phi_m \in PW(\lambda_M)$ which is an eigenvector of $\mathcal{L}$ corresponding to the eigenvalue λ_j, and then generalize to linear combinations. Let Q_k be the projection upon the eigenspace of $\boldsymbol{\Delta}$ corresponding to the eigenvalue κ_k. Then, by $\mathcal{L}\phi_m = \lambda_m\phi_m$,

$$\boldsymbol{\Delta} S^{\lambda_M}\phi_m - S^{\lambda_M}\mathcal{L}\phi_m = \sum_k \kappa_k Q_k S^{\lambda_M}\phi_m - \lambda_m S^{\lambda_M}\sum_k Q_k\phi_m = \sum_k (\kappa_k - \lambda_m) Q_k S^{\lambda_M}\phi_m.$$

By orthogonality of the projections $\{Q_k\}_k$,

$$\left\|\sum_k \kappa_k Q_k S^{\lambda_M}\phi_m - \lambda_m S^{\lambda_M}\phi_m\right\|^2 = \sum_k \left|\kappa_k - \lambda_m\right|^2 \left\|Q_k S^{\lambda_M}\phi_m\right\|^2 \tag{60}$$

Now, similarly to the derivation of (60), by functional calculus and by (9),

$$\begin{aligned}\left\|f(\boldsymbol{\Delta})S^{\lambda_M}\phi_m - S^{\lambda_M}f(\mathcal{L})\phi_m\right\|^2 =& \sum_k \left|f(\kappa_k) - f(\lambda_m)\right|^2 \left\|Q_k S^{\lambda_M}\phi_m\right\|^2 \\ =& \sum_k \left|\frac{f(\kappa_k) - f(\lambda_m)}{\kappa_k - \lambda_m}\right|^2 \left|\kappa_k - \lambda_m\right|^2 \left\|Q_k S^{\lambda_M}\phi_m\right\|^2 \\ \leq& \mathrm{V}_f(\lambda_m)^2 \sum_k \left|\kappa_k - \lambda_m\right|^2 \left\|Q_k S^{\lambda_M}\phi_m\right\|^2 \qquad (61) \\ =& \mathrm{V}_f(\lambda_m)^2 \left\|\sum_k \kappa_k Q_k S^{\lambda_M}\phi_m - \lambda_m S^{\lambda_M}\phi_m\right\|^2 \\ =& \mathrm{V}_f(\lambda_m)^2 \left\|\boldsymbol{\Delta} S^{\lambda_M}\phi_m - S^{\lambda_M}\mathcal{L}\phi_m\right\|^2,\end{aligned}$$

which proves Thm.4(1).

Now, for $q = \sum_m c_m\phi_m$, we have

$$\left\|f(\boldsymbol{\Delta})S^{\lambda_M}P(\lambda_M)q - S^{\lambda_M}f(\mathcal{L})P(\lambda_M)q\right\| = \left\|\sum_{m=1}^M c_m\Big(f(\boldsymbol{\Delta})S^{\lambda_M} - S^{\lambda_M}f(\mathcal{L})\Big)\phi_m\right\|.$$

By the triangle inequality and Thm.4(1),

$$\begin{aligned}\left\|f(\boldsymbol{\Delta})S^{\lambda_M}P(\lambda_M)q - S^{\lambda_M}f(\mathcal{L})P(\lambda_M)q\right\| \leq& \sum_{m=1}^M |c_m| \left\|\Big(f(\boldsymbol{\Delta})S^{\lambda_M}\phi_m - S^{\lambda_M}f(\mathcal{L})\phi_m\Big)\right\| \\ \leq& \sum_{m=1}^M |c_m|\, \mathrm{V}_f(\lambda_m) \left\|\boldsymbol{\Delta} S^{\lambda_M}\phi_m - S^{\lambda_M}\mathcal{L}\phi_m\right\|\end{aligned} \tag{62}$$

which proves Thm.4(2). Moreover, by (62), by $\mathrm{V}_f(\lambda_m) \leq D$ by $\|\phi_m\| = 1$ and by Hölder's inequality,

$$\left\|f(\boldsymbol{\Delta})S^{\lambda_M}P(\lambda_M)q - S^{\lambda_M}f(\mathcal{L})P(\lambda_M)q\right\| \leq \|q\|_1 D \left\|\boldsymbol{\Delta} S^{\lambda_M}P(\lambda_M) - S^{\lambda_M}\mathcal{L}P(\lambda_M)\right\|. \tag{63}$$

Here, $\|q\|_1 := \sum_{m=1}^{M} |c_m|$. By Cauchy–Schwarz inequality we have

$$\|q\|_1 \leq \|q\|_2 \sqrt{M},$$

which proves Thm.4(3).

By the triangle inequality

$$\begin{aligned}
&\left\| f(\mathcal{L})P(\lambda_M) - R^{\lambda_M} f(\mathbf{\Delta}) S^{\lambda_M} P(\lambda_M) \right\| \\
&\leq \left\| f(\mathcal{L})P(\lambda_M) - R^{\lambda_M} S^{\lambda_M} f(\mathcal{L})P(\lambda_M) \right\| \\
&\quad + \left\| R^{\lambda_M} S^{\lambda_M} f(\mathcal{L})P(\lambda_M) - R^{\lambda_M} f(\mathbf{\Delta}) S^{\lambda_M} P(\lambda_M) \right\| \\
&\leq \left\| P(\lambda_M) - R^{\lambda_M} S^{\lambda_M} P(\lambda_M) \right\| \left\| f(\mathcal{L})P(\lambda_M) \right\| \\
&\quad + \left\| R^{\lambda_M} \right\| \left\| S^{\lambda_M} f(\mathcal{L})P(\lambda_M) - f(\mathbf{\Delta}_n) S^{\lambda_M} P(\lambda_M) \right\|.
\end{aligned}$$

Note that by assumption

$$\left\| R^{\lambda_M} \right\| \leq C$$

and, by the diagonal form of $f(\mathcal{L})P(\lambda_M)$,

$$\left\| f(\mathcal{L})P(\lambda_M) \right\| \leq \|f\|_{\mathcal{L},M},$$

which gives Thm.4(5). A similar use of the triangle inequality gives Thm.4(4).

B.2 Proof of Theorem 16

First we show that, if $A \neq 1$, then

$$\begin{aligned}
\left\| \mathcal{N}_k^l P(\psi^0) f \right\| &\leq A^l \left\| P(\psi^0) f \right\| + \frac{A^l - 1}{A - 1} B \quad \text{for all } f \in L^2(\mathcal{M}) \\
\left\| \mathcal{N}_k^{j,l} \tilde{f} \right\| &\leq A^l \left\| \tilde{f} \right\| + \frac{A^l - 1}{A - 1} B \quad \text{for all } \tilde{f} \in L^2(G^{j,0})
\end{aligned}$$

and if $A = 1$

$$\begin{aligned}
\left\| \mathcal{N}_k^l P(\psi_0) f \right\| &\leq \left\| P(\psi_0) f \right\| + (l-1)B \quad \text{for all } f \in L^2(\mathcal{M}) \\
\left\| \mathcal{N}_k^{j,l} \tilde{f} \right\| &\leq \left\| \tilde{f} \right\| + (l-1)B \quad \text{for all } \tilde{f} \in L^2(G^{j,0})
\end{aligned}$$

for every l, k and $j = 1, 2$. We next focus on $\mathcal{N}_k^l$, and remark that we can use similar arguments for $\mathcal{N}_k^{j,l}$. Note that $\left\| g_{k',k}^l(\mathcal{L}) \right\| \leq \left\| g_{k',k}^l \right\|_\infty \leq 1$ for every l, k, k'. Moreover,

$$\begin{aligned}
\left\| \sum_{k=1}^{K_{l-1}} a_{k'k}^l \; g_{k'k}^l(\mathcal{L}) f_k^{l-1} + b_{k'}^l \right\| &\leq \sum_{k=1}^{K_{l-1}} \left| a_{k'k}^l \right| \left\| g_{k'k}^l(\mathcal{L}) f_k^{l-1} \right\| + \left\| b_{k'}^l \right\| \\
&\leq \sum_{k=1}^{K_{l-1}} \left| a_{k'k}^l \right| \left\| f_k^{l-1} \right\| + B \\
&\leq A \max_k \left\| f_k^{l-1} \right\| + B.
\end{aligned}$$

Moreover,

$$\left\| \rho\Big(\sum_{k=1}^{K_{l-1}} a^l_{k'k}\ g^l_{k'k}(\mathcal{L}) f^{l-1}_k + b^l_{k'} \Big) \right\| \leq \left\| \sum_{k=1}^{K_{l-1}} a^l_{k'k}\ g^l_{k'k}(\mathcal{L}) f^{l-1}_k + b^l_{k'} \right\| \leq A \max_k \left\| f^{l-1}_k \right\| + B.$$

Fianlly, using the fact that pooling and projection decreases norm by assumption implies This shows that

$$\max_k \left\| f^l_k \right\| \leq A \max_k \left\| f^{l-1}_k \right\| + B.$$

Thus, by solving this recursive sequence for $A \neq 1$ we get

$$\max_k \left\| f^l_k \right\| \leq A^l \left\| P(\psi_0) f \right\| + \frac{A^l - 1}{A - 1} B. \tag{64}$$

or for $A = 1$

$$\max_k \left\| f^l_k \right\| \leq \left\| P(\psi_0) f \right\| + (l-1)B.$$

Let us now prove (29) and (30), starting with $f \in L^2(\mathcal{M})$ at the input of Layer 0. The error in one convolution $g^l_{k'k}$, between the continuous and the discrete signals $j = 1, 2$, satisfies

$$\begin{aligned}
&\left\| S^{\psi_{l-1}}_{j,l-1} g^l_{k'k}(\mathcal{L}) P(\psi_{l-1}) f^{l-1}_k - g^l_{k'k}(\mathbf{\Delta}_{j,l-1}) \tilde{f}^{j,l-1}_k \right\| \\
&\leq \left\| S^{\psi_{l-1}}_{j,l-1} g^l_{k'k}(\mathcal{L}) P(\psi_{l-1}) f^{l-1}_k - g^l_{k'k}(\mathbf{\Delta}_{j,l-1}) S^{\psi_{l-1}}_{j,l-1} P(\psi_{l-1}) f^{l-1}_k \right\| \\
&\quad + \left\| g^l_{k'k}(\mathbf{\Delta}_{j,l-1}) S^{\psi_{l-1}}_{j,l-1} P(\psi_{l-1}) f^{l-1}_k - g^l_{k'k}(\mathbf{\Delta}_{j,l-1}) \tilde{f}^{j,l-1}_k \right\| \\
&\leq \left\| S^{\psi_{l-1}}_{j,l-1} g^l_{k'k}(\mathcal{L}) P(\psi_{l-1}) - g^l_{k'k}(\mathbf{\Delta}_{j,l-1}) S^{\psi_{l-1}}_{j,l-1} P(\psi_{l-1}) \right\| \left\| f^{l-1}_k \right\| \\
&\quad + \left\| g^l_{k'k}(\mathbf{\Delta}_{j,l-1}) \right\| \left\| S^{\psi_{l-1}}_{j,l-1} P(\psi_{l-1}) f^{l-1}_k - \tilde{f}^{j,l-1}_k \right\|.
\end{aligned}$$

Thus, by Thm.3.(4), and by $\left\| g^l_{k'k}(\mathbf{\Delta}_{j,l-1}) \right\| \leq \left\| g^l_{k'k} \right\|_\infty = 1$,

$$\begin{aligned}
&\left\| S^{\psi_{l-1}}_{j,l-1} g^l_{k'k}(\mathcal{L}) P(\psi_{l-1}) f^{l-1}_k - g^l_{k'k}(\mathbf{\Delta}_{j,l-1}) \tilde{f}^{j,l-1}_k \right\| \\
&\leq D(\psi_L) \delta \left\| f^{l-1}_k \right\| + \left\| S^{\psi_{l-1}}_{j,l-1} P(\psi_{l-1}) f^{l-1}_k - \tilde{f}^{j,l-1}_k \right\|,
\end{aligned} \tag{65}$$

where $D(\psi_L) = D\sqrt{\#\{\lambda_m \leq \psi_L\}_m}$.

Now, the error in the output of the network, before pooling, is

$$
\begin{aligned}
&\left\| S^{\psi_l}_{j,l-1} P(\psi_l)\rho\Big(\sum_{k=1}^{K_{l-1}} a^l_{k'k}\ g^l_{k'k}(\mathcal{L})P(\psi_{l-1})f_k^{l-1}\Big) - \rho\Big(\sum_{k=1}^{K_{l-1}} a^l_{k'k} g^l_{k'k}(\boldsymbol{\Delta}_{j,l-1})\tilde{f}_k^{j,l-1}\Big)\right\| \\
&\leq \left\| \rho\Big(S^{\psi_{l-1}}_{j,l-1}\sum_{k=1}^{K_{l-1}} a^l_{k'k}\ g^l_{k'k}(\mathcal{L})P(\psi_{l-1})f_k^{l-1}\Big) - S^{\psi_l}_{j,l-1}P(\psi_l)\rho\Big(\sum_{k=1}^{K_{l-1}} a^l_{k'k}\ g^l_{k'k}(\mathcal{L})P(\psi_{l-1})f_k^{l-1}\Big)\right\| \\
&\quad + \left\| \rho\Big(S^{\psi_{l-1}}_{j,l-1}\sum_{k=1}^{K_{l-1}} a^l_{k'k}\ g^l_{k'k}(\mathcal{L})P(\psi_{l-1})f_k^{l-1}\Big) - \rho\Big(\sum_{k=1}^{K_{l-1}} a^l_{k'k} g^l_{k'k}(\boldsymbol{\Delta}_{j,l-1})\tilde{f}_k^{j,l-1}\Big)\right\| \\
&\leq \delta \left\| \sum_{k=1}^{K_{l-1}} a^l_{k'k}\ g^l_{k'k}(\mathcal{L})P(\psi_{l-1})f_k^{l-1}\right\| \\
&\quad + \left\| S^{\psi_{l-1}}_{j,l-1}\sum_{k=1}^{K_{l-1}} a^l_{k'k}\ g^l_{k'k}(\mathcal{L})P(\psi_{l-1})f_k^{l-1} - \sum_{k=1}^{K_{l-1}} a^l_{k'k}\ g^l_{k'k}(\boldsymbol{\Delta}_{j,l-1})\tilde{f}_k^{j,l-1}\right\| \\
&\leq \delta A^l \left\| P(\psi_0)f\right\| + \delta\frac{A^l-1}{A-1}B + \sum_{k=1}^{K_{l-1}}\left|a^l_{k'k}\right| \left\| S^{\psi_{l-1}}_{j,l-1} g^l_{k'k}(\mathcal{L})P(\psi_{l-1})f_k^{l-1} - g^l_{k'k}(\boldsymbol{\Delta}_{j,l-1})\tilde{f}_k^{j,l-1}\right\|
\end{aligned}
$$

or for $A=1$

$$
\begin{aligned}
&\left\| S^{\psi_l}_{j,l-1} P(\psi_l)\rho\Big(\sum_{k=1}^{K_{l-1}} a^l_{k'k}\ g^l_{k'k}(\mathcal{L})P(\psi_{l-1})f_k^{l-1}\Big) - \rho\Big(\sum_{k=1}^{K_{l-1}} a^l_{k'k} g^l_{k'k}(\boldsymbol{\Delta}_{j,l-1})\tilde{f}_k^{j,l-1}\Big)\right\| \\
&\quad \leq \delta \left\| P(\psi_0)f\right\| + \delta(l-1)B + \sum_{k=1}^{K_{l-1}}\left|a^l_{k'k}\right| \left\| S^{\psi_{l-1}}_{j,l-1} g^l_{k'k}(\mathcal{L})P(\psi_{l-1})f_k^{l-1} - g^l_{k'k}(\boldsymbol{\Delta}_{j,l-1})\tilde{f}_k^{j,l-1}\right\|
\end{aligned}
$$

Therefore, by (65)

$$
\begin{aligned}
&\left\| S^{\psi_l}_{j,l-1} P(\psi_l)\rho\Big(\sum_{k=1}^{K_{l-1}} a^l_{k'k}\ g^l_{k'k}(\mathcal{L})P(\psi_{l-1})f_k^{l-1}\Big) - \rho\Big(\sum_{k=1}^{K_{l-1}} a^l_{k'k} g^l_{k'k}(\boldsymbol{\Delta}_{j,l-1})\tilde{f}_k^{j,l-1}\Big)\right\| \\
&\leq \delta A^l \left\| P(\psi_0)f\right\| + \delta\frac{A^l-1}{A-1}B + A\max_k\Big\{ D(\psi_L)\delta\left\| f_k^{l-1}\right\| + \left\| S^{\psi_{l-1}}_{j,l-1}P(\psi_{l-1})f_k^{l-1} - \tilde{f}_k^{j,l-1}\right\|\Big\}.
\end{aligned}
$$

The error after pooling takes the form

$$\begin{aligned}
&\left\| S^{\psi_l}_{j,l} P(\psi_l) \rho\Big(\sum_{k=1}^{K_{l-1}} a^l_{k'k} \; g^l_{k'k}(\mathcal{L}) P(\psi_{l-1}) f^{l-1}_k \Big) - Y^{j,l} \rho\Big(\sum_{k=1}^{K_{l-1}} a^l_{k'k} g^l_{k'k}(\mathbf{\Delta}_{j,l-1}) \tilde{f}^{j,l-1}_k \Big) \right\| \\
&\leq \Bigg\| S^{\psi_l}_{j,l} P(\psi_l) \rho\Big(\sum_{k=1}^{K_{l-1}} a^l_{k'k} \; g^l_{k'k}(\mathcal{L}) P(\psi_{l-1}) f^{l-1}_k \Big) \\
&\qquad\qquad\qquad\qquad - Y^{j,l} S^{\psi_l}_{j,l-1} P(\psi_l) \rho\Big(\sum_{k=1}^{K_{l-1}} a^l_{k'k} \; g^l_{k'k}(\mathcal{L}) P(\psi_{l-1}) f^{l-1}_k \Big) \Bigg\| \\
&+ \left\| Y^{j,l} S^{\psi_l}_{j,l-1} P(\psi_l) \rho\Big(\sum_{k=1}^{K_{l-1}} a^l_{k'k} \; g^l_{k'k}(\mathcal{L}) P(\psi_{l-1}) f^{l-1}_k \Big) - Y^{j,l} \rho\Big(\sum_{k=1}^{K_{l-1}} a^l_{k'k} g^l_{k'k}(\mathbf{\Delta}_{j,l-1}) \tilde{f}^{j,l-1}_k \Big) \right\| \\
&\leq \delta \left\| \rho\Big(\sum_{k=1}^{K_{l-1}} a^l_{k'k} \; g^l_{k'k}(\mathcal{L}) P(\psi_{l-1}) f^{l-1}_k \Big) \right\| \\
&+ \left\| S^{\psi_l}_{j,l-1} P(\psi_l) \rho\Big(\sum_{k=1}^{K_{l-1}} a^l_{k'k} \; g^l_{k'k}(\mathcal{L}) P(\psi_{l-1}) f^{l-1}_k \Big) - \rho\Big(\sum_{k=1}^{K_{l-1}} a^l_{k'k} g^l_{k'k}(\mathbf{\Delta}_{j,l-1}) \tilde{f}^{j,l-1}_k \Big) \right\| \\
&\leq 2\delta A^l \left\| P(\psi_0) f \right\| + 2\delta \frac{A^l - 1}{A - 1} B + A \max_k \Big\{ D(\psi_L) \delta \left\| f^{l-1}_k \right\| + \left\| S^{\psi_{l-1}}_{j,l-1} P(\psi_{l-1}) f^{l-1}_k - \tilde{f}^{j,l-1}_k \right\| \Big\}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\left\| S^{\psi_l}_{j,l} P(\psi_l) f^l_{k'} - \tilde{f}^{j,l}_{k'} \right\| \\
&\leq (D(\psi_L) + 2)\delta \Big(A^l \left\| P(\psi_0) f \right\| + \frac{A^l - 1}{A - 1} B \Big) + A \max_k \left\| S^{\psi_{l-1}}_{j,l-1} P(\psi_{l-1}) f^{l-1}_k - \tilde{f}^{j,l-1}_k \right\|.
\end{aligned}$$

By solving this recurrent sequence, we obtain for $A > 1$

$$\left\| S^{\psi_L}_{j,L} \mathcal{N}^L_k P(\psi_0) f - \mathcal{N}^{j,L}_k S^{\psi_0}_{j,L} P(\psi_0) f \right\| \leq L(D(\psi_L) + 2)\delta \Big(A^L \left\| f \right\| + B \frac{A^L - 1}{A - 1} \Big).$$

For $A = 1$ we get

$$\left\| S^{\psi_L}_{j,L} \mathcal{N}^L_k P(\psi_0) f - \mathcal{N}^{j,L}_k S^{\psi_0}_{j,0} P(\psi_0) f \right\| \leq L(D(\psi_L) + 2)\delta \Big(\left\| f \right\| + LB \Big)$$

Finally,

$$\begin{aligned}
&\left\| \mathcal{N}^L_k P(\psi_0) f - R^{\psi_L}_{j,L} \mathcal{N}^{j,L}_k S^{\psi_0}_{j,1} P(\psi_0) f \right\| \\
&\leq \left\| \mathcal{N}^L_k P(\psi_0) f - R^{\psi_L}_{j,L} S^{\psi_L}_{j,L} \mathcal{N}^L_k P(\psi_0) f \right\| + \left\| R^{\psi_L}_{j,L} S^{\psi_L}_{j,L} \mathcal{N}^L_k P(\psi_0) f - R^{\psi_L}_{j,L} \mathcal{N}^{j,L}_k S^{\psi_0}_{j,L} P(\psi_0) f \right\| \\
&\leq \left\| P(\psi_L) - R^{\psi_L}_{j,L} S^{\psi_L}_{j,L} P(\psi_L) \right\| \left\| \mathcal{N}^L_k P(\psi_0) f \right\| + \left\| R^{\psi_L}_{j,L} \right\| \left\| S^{\psi_L}_{j,l} \mathcal{N}^L_k P(\psi_0) f - \mathcal{N}^{j,L}_k S^{\psi_1}_{j,L} P(\psi_0) f \right\| \\
&\leq L\Big(D\sqrt{\#\{\lambda_m \leq \psi_L\}_m} + 2 \Big)\delta \Big(A^L \left\| f \right\| + B \frac{A^L - 1}{A - 1} \Big) + \Big(A^L \left\| f \right\| + \frac{A^L - 1}{A - 1} B \Big)\delta.
\end{aligned}$$

This shows that

$$\begin{aligned}&\left\| R^{\psi_L}_{1,L}\mathcal{N}^{1,L}_k S^{\psi_0}_{1,0}P(\psi_0)f - R^{\psi_L}_{2,L}\mathcal{N}^{2,L}_k S^{\psi_0}_{2,L}P(\psi_0)f \right\| \\ &\leq \Big(LD\sqrt{\#\{\lambda_m \leq \psi_L\}_m} + 2L + 2\Big)\Big(A^L \left\|f\right\| + B\frac{A^L-1}{A-1}\Big)\delta,\end{aligned}$$

and similarly for $A = 1$.

B.3 Proof of Proposition 26

Lemma 35 *Consider the setting of Proposition 26. Then*

$$\lim_{n\to\infty}\sup_{f\neq 0}\frac{\|S_n\rho(P(\lambda)f) - S_nP(\lambda')\rho(P(\lambda)f)\| - \|\rho(P(\lambda)f) - P(\lambda')\rho(P(\lambda)f)\|}{\|P(\lambda)f\|} = 0 \qquad (66)$$

$$\lim_{\lambda'\to\infty}\sup_{f\neq 0}\frac{\|\rho(P(\lambda)f) - P(\lambda')\rho(P(\lambda)f)\|}{\|P(\lambda)f\|} = 0 \qquad (67)$$

Proof We first prove (66). Observe that any nonzero vector in $PW(\lambda)$ can be written as cf, where $c > 0$ is a real scalar, and $f \in PW(\lambda)$ has norm 1. Now, by the positive homogeneity of ρ,

$$\begin{aligned}&\frac{\|S_n\rho(cP(\lambda)f) - S_nP(\lambda')\rho(cP(\lambda)f)\| - \|\rho(cP(\lambda)f) - P(\lambda')\rho(cP(\lambda)f)\|}{\|cP(\lambda)f\|} \\ &= \left\|S_n\rho(P(\lambda)f) - S_nP(\lambda')\rho(P(\lambda)f)\right\| - \left\|\rho(P(\lambda)f) - P(\lambda')\rho(P(\lambda)f)\right\|.\end{aligned}$$

Thus, (66) is equivalent to

$$\lim_{n\to\infty}\sup_{P(\lambda)f\in\mathcal{S}(\lambda)}\left\|S_n\rho(P(\lambda)f) - S_nP(\lambda')\rho(P(\lambda)f)\right\| - \left\|\rho(P(\lambda)f) - P(\lambda')\rho(P(\lambda)f)\right\| = 0$$

where $\mathcal{S}(\lambda)$ is the unit sphere in $PW(\lambda)$. Note that the mapping $F_n : \mathcal{S}(\lambda) \to \mathbb{R}$ defined by

$$\begin{aligned}F_n\big(P(\lambda)f\big) &= \left\|S_n\rho(P(\lambda)f) - S_nP(\lambda')\rho(P(\lambda)f)\right\| - \left\|\rho(P(\lambda)f) - P(\lambda')\rho(P(\lambda)f)\right\| \\ &= \left\|S_n\big(I - P(\lambda')\big)\rho(P(\lambda)f)\right\| - \left\|\big(I - P(\lambda')\big)\rho(P(\lambda)f)\right\|\end{aligned}$$

is Lipschitz continuous in $P(\lambda)f$ for large enough n. Indeed, by $\|I - P(\lambda')\| = 1$ and contraction of ρ,

$$\begin{aligned}\left|F_n\big(P(\lambda)f_1\big) - F_n\big(P(\lambda)f_2\big)\right| &\leq \left\|S_n\big(I - P(\lambda')\big)\rho(P(\lambda)f_1) - S_n\big(I - P(\lambda')\big)\rho(P(\lambda)f_2)\right\| \\ &\quad + \left\|\big(I - P(\lambda')\big)\rho(P(\lambda)f_1) - \big(I - P(\lambda')\big)\rho(P(\lambda)f_2)\right\| \\ &\leq (C+1)\left\|P(\lambda)f_1 - P(\lambda)f_2\right\|,\end{aligned}$$

where C is the bound of $\left\|S^{\lambda}_n\right\|$, guaranteed by Proposition 23, and can be chosen $C = 2$ for large enough n. Note that the Lipschitz constants of F_n are uniformly bounded by $D = 3$.

Observe that by Definition 24, F_n converges to 0 pointwise as $n \to \infty$. Our goal is to show uniform convergence. Since the domain $\mathcal{S}(\lambda)$ of F_n is compact, F_n obtains a maximum for each n. Denote

$$P(\lambda)f_n = \underset{P(\lambda)f\in\mathcal{S}(\lambda)}{\operatorname{argmax}}\ F_n(P(\lambda)f).$$

Suppose that $\lim_{n\to\infty} F_n(P(\lambda)f_n)$ does not exist, or converges to a nonzero limit. Since $\mathcal{S}(\lambda)$ is compact, and F_n uniformly bounded by $2D$, there is a subsequence $P(\lambda)f_{n_m}$ converging to some $P(\lambda)f_\infty \in \mathcal{S}(\lambda)$, such that

$$\lim_{m\to\infty} F_{n_m}(P(\lambda)f_{n_m}) = A > 0.$$

Now, for every $\epsilon > 0$ there is a large enough M, such that, for every $m > M$,

$$\begin{aligned} |F_{n_m}(P(\lambda)f_\infty) - A| &\leq |F_{n_m}(P(\lambda)f_\infty) - F_{n_m}(P(\lambda)f_{n_m})| + \epsilon/2 \\ &\leq D \left\| P(\lambda)f_\infty - P(\lambda)f_{n_m} \right\| + \epsilon/2 < \epsilon. \end{aligned}$$

By picking $\epsilon = A/3$, this contradicts the fact that $\lim_{n\to\infty} F_n(P(\lambda)f_\infty) = 0$, guaranteed by Definition 24.

Similarly, for (67),

$$\sup_{f\neq 0} \frac{\|\rho(P(\lambda)f) - P(\lambda')\rho(P(\lambda)f)\|}{\|P(\lambda)f\|} = \sup_{P(\lambda)f\in\mathcal{S}(\lambda)} \left\| \big(I - P(\lambda')\big)\rho(P(\lambda)f) \right\|.$$

For a fixed f, the fact that $\big(I - P(\lambda')\big)\rho(P(\lambda)f)$ is the tail in the expansion of $\rho(P(\lambda)f)$ in the eigenbasis of $\mathcal{L}$, we have

$$\lim_{\lambda'\to\infty} \left\| \big(I - P(\lambda')\big)\rho(P(\lambda)f) \right\| = 0 \quad \text{for all } P(\lambda)f \in \mathcal{S}(\lambda). \tag{68}$$

The uniform convergence of (67) is derived from the pointwise convergence of (68) in the same procedure as above.

■

Proof [Proof of Proposition 26] By Lemma 35

$$\begin{aligned} &\lim_{\lambda'\to\infty}\lim_{n\to\infty}\sup_{f\neq 0} \frac{\|S_n\rho(P(\lambda)f) - S_nP(\lambda')\rho(P(\lambda)f)\|}{\|P(\lambda)f\|} \\ &\leq \lim_{\lambda'\to\infty}\lim_{n\to\infty}\sup_{f\neq 0} \frac{\|S_n\rho(P(\lambda)f) - S_nP(\lambda')\rho(P(\lambda)f)\| - \|\rho(P(\lambda)f) - P(\lambda')\rho(P(\lambda)f)\|}{\|P(\lambda)f\|} \\ &\quad + \lim_{\lambda'\to\infty}\sup_{f\neq 0} \frac{\|\rho(P(\lambda)f) - P(\lambda')\rho(P(\lambda)f)\|}{\|P(\lambda)f\|} = 0. \end{aligned}$$

Now, the limit as $\lambda \to \infty$ follows trivially. ■

B.4 Proof of Theorem 32

We prove Theorem 32 using three lemmas.

Lemma 36 *Let $f \in PW(\lambda)$. Let $\{\mathcal{M}, \mu\}$ be a compact topological-measure space with $\mu(\mathcal{M}) = 1$. Consider the weighted measure μ_w satisfying (50). Let $\mathcal{L}$ be a Laplacian of the* *form (44), such that $H \in L^2(\mathcal{M}^2; \mu \times \mu)$. Suppose that $\mathcal{L}$ respects continuity. Let $\mathbf{\Delta}_n$ be a random sampled Laplacian. Let*

$$C = \frac{1}{w_{\min}} \|H\|_{L^2(\mathcal{M}^2;\mu\times\mu)} C_\lambda \tag{69}$$

for $w_{\min} = \min_{x\in\mathcal{M}} w(x)$, and C_λ is the constant such that

$$\forall g \in PW(\lambda). \quad \|g\|_\infty \leq C_\lambda \|g\|_2 \,, \tag{70}$$

guaranteed by the fact that $PW(\lambda)$ is finite-dimensional.

Then for every $\delta > 0$, in probability more than $(1 - \delta)$,

$$\left\| S_n^\lambda \mathcal{L} P(\lambda) - \mathbf{\Delta}_n S_n^\lambda P(\lambda) \right\|_{L^2(G_n)} \leq C \delta^{-1/2} N_n^{-1/2}. \tag{71}$$

where the induced norm is for operators $L^2(\mathcal{M}; \mu) \to L^2(G_n)$.

Proof Let $f \in PW(\lambda)$, and note that f is continuous since $\mathcal{L}$ respects continuity. For a fixed $x_0 \in \mathcal{M}$, consider the random variable $F_{x_0} : \{\mathcal{M}; \mu_w\} \to \mathbb{C}$ defined by

$$F_{x_0}(x) = \frac{1}{w(x)} H(x_0, x) f(x). \tag{72}$$

By (44) and (50), the expected value of F_{x_0} is

$$\mathrm{E}(F_{x_0}) = \mathcal{L} f(x_0). \tag{73}$$

Consider N_n i.i.d random variables (72), denoted by

$$F_{x_0;k'} = \frac{1}{w(x^n_{k'})} H(x_0, x^n_{k'}) f(x^n_{k'}), \quad k' = 1, \ldots, N_n.$$

Let

$$F_{x_0}^{N_n} = \frac{1}{N_n} \sum_{k'=1}^{N_n} F_{x_0;k'}. \tag{74}$$

By (73) we have

$$\mathrm{E}\Big(F_{x_0}^{N_n}\Big) = \mathcal{L} f(x_0)$$

On the other hand, the realization of the sum in (74) can be written for $x_0 = x^n_k$ as

$$F_{x^n_k}^{N_n} = \sum_{k'=1}^{N_m} \frac{1}{w(x^n_{k'})} H(x^n_k, x^n_{k'}) f(x^n_{k'}) dx = [\mathbf{\Delta}_n S_n^\lambda f]_k. \tag{75}$$

This shows that the graph Laplacians coincide on average with the topological-measure Laplacian.

Next let us analyze the average mean square error over $x_0 \in \mathcal{M}$. In the following, Fubini's theorem follows the fact that $\mathcal{M}$ is compact and all integrands are continuous. Hence,

$$\begin{aligned}
&\mathrm{E}\left\| F^{N_n}_{(\cdot)} - \mathcal{L}f \right\|^2_{L^2(\mathcal{M})} \\
&= \iint_{x_1,\ldots,x_n} \int_{x_0} \left| F^{N_n}_{x_0}(x^n_1,\ldots,x^n_{N_n}) - [\mathcal{L}f](x_0) \right|^2 dx_0 \; w(x^n_1)dx^n_1 \cdot w(x^n_{N_n})dx^n_{N_n} \\
&= \int_{x_0} \iint_{x_1,\ldots,x_n} \left| F^{N_n}_{x_0}(x^n_1,\ldots,x^n_{N_n}) - [\mathcal{L}f](x_0) \right|^2 w(x^n_1)dx^n_1 \cdot w(x^n_{N_n})dx^n_{N_n} \; dx_0 \\
&= \int_{x_0} \mathrm{Var}F^{N_n}_{x_0} dx_0 = \int_{x_0} \frac{\mathrm{Var}F_{x_0}}{N_n} dx_0 = \frac{\left\|\mathrm{Var}F_{(\cdot)}\right\|_1}{N_n}
\end{aligned}$$

Next, we prove that prove $\mathrm{Var}F_{(\cdot)} \in L^1(\mathcal{M})$, and bound $\left\|\mathrm{Var}F_{(\cdot)}\right\|_1$. We have

$$\mathrm{Var}F_{x_0} \leq \int_x \left|F_{x_0}(x)\right|^2 w(x)dx.$$

This yields

$$\begin{aligned}
\left\|\mathrm{Var}F_{(\cdot)}\right\|_1 &\leq \int_{x_0} \int_x \left|F_{x_0}(x)\right|^2 w(x)dxdx_0 \\
&= \int_{x_0} \int_x \frac{1}{w(x)} \left|H(x_0,x)\right|^2 \left|f(x)\right|^2 dxdx_0.
\end{aligned}$$

Thus

$$\begin{aligned}
\left\|\mathrm{Var}F_{(\cdot)}\right\|_1 &\leq \left\| \frac{1}{\sqrt{w(\cdot)}} H(\cdot,\cdot\cdot) \right\|^2_{L^2(\mathcal{M}^2)} \left\|f\right\|^2_\infty \\
&\leq \frac{1}{w_{\min}} \left\|H\right\|^2_{L^2(\mathcal{M}^2)} \left\|f\right\|^2_\infty
\end{aligned}$$

This proves that the expected mean square error satisfies

$$\mathrm{E}\left\| F^{N_n}_{(\cdot)} - \mathcal{L}f \right\|^2_{L^2(\mathcal{M})} \leq \frac{1}{w_{\min}} \left\|H\right\|^2_{L^2(\mathcal{M}^2)} \left\|f\right\|^2_\infty \frac{1}{N_n}. \tag{76}$$

To obtain a convergence result in high probability, we can use theorems on concentration of measure, like Markov's, Chebyshev's or Bernstein's inequalities. For Lemma 36, we consider Markov's inequality, that states that for a random variable X with finite non-zero expected value

$$\Pr\Big(X \geq \frac{\mathrm{E}(X)}{\delta}\Big) \leq \delta$$

for any $0 < \delta < 1$. In our case, by (76), Markov's inequality states that in probability more than $(1-\delta)$

$$\left\| F^{N_n}_{(\cdot)} - \mathcal{L}f \right\|_{L^2(\mathcal{M})} \leq \frac{1}{\sqrt{w_{\min}}} \left\|H\right\|_{L^2(\mathcal{M}^2)} \left\|f\right\|_{L^\infty(\mathcal{M})} \frac{1}{\sqrt{N_n}} \frac{1}{\sqrt{\delta}}. \tag{77}$$

This means that for every k,

$$\left|F^{N_n}_{x^n_k} - \mathcal{L}f(x^n_k)\right| \leq C_\lambda \frac{1}{\sqrt{w_{\min}}} \left\|H\right\|_{L^2(\mathcal{M}^2)} \left\|f\right\|_{L^\infty(\mathcal{M})} \frac{1}{\sqrt{N_n}} \frac{1}{\sqrt{\delta}}. \tag{78}$$

We finally conclude that, by the inner product structure (46) of $L^2(G_n)$, and by (75)

$$\left\|\mathbf{\Delta}_n S^\lambda_n f - S^\lambda_n \mathcal{L}f\right\|_{L^2(G_n)} = \sqrt{\frac{1}{N_n}\sum_{k=1}^{N_n} \frac{1}{w(x^n_k)}\left|F^{N_n}_{x^n_k} - \mathcal{L}f(x^n_k)\right|^2} \leq C N_n^{-1/2} \delta^{-1/2} \left\|f\right\|_{L^2(\mathcal{M})}$$

where C is given in (69).

■

Denote by $\|A\|_{F(\mathbb{C}^{M\times M})}$ the Frobenius norm of the matrix $A \in \mathbb{C}^{M\times M}$.

Lemma 37 *Let $\{\mathcal{M},\mu\}$ be a compact topological-measure space with $\mu(\mathcal{M}) = 1$. Let μ_w be a weighted measure satisfying (50). Let $\mathcal{L}$ be a Laplacian of the form (44), such that $H \in L^2(\mathcal{M}^2)$. Suppose that $\mathcal{L}$ respects continuity. Let $\mathbf{S}^\lambda_n$ and $\mathbf{R}^\lambda_n$ be random sampling and interpolation operators. Consider the corresponding random variable $\langle \mathbf{\Phi}_n, \mathbf{\Phi}_n\rangle$ given in Definition 22 on the random sample points. Then for every $\delta > 0$, in probability more than $(1-\delta)$*

$$\left\|\langle \mathbf{\Phi}_n, \mathbf{\Phi}_n\rangle - \mathbf{I}\right\|_{F(\mathbb{C}^{M_\lambda \times M_\lambda})} \leq C\delta^{-1/2} N_n^{-1/2}. \tag{79}$$

Here,

$$C = \frac{M_\lambda}{\sqrt{w_{\min}}} \max_{m \leq M_\lambda} \left\|\phi_m\right\|^2_\infty,$$

and $M_\lambda = \dim(PW(\lambda))$ as before.

Proof For fixed $m, m' \in \mathcal{M}$, consider the random variable $F_{m,m'} : \{\mathcal{M};\mu_w\} \to \mathbb{C}$ defined by

$$F_{m,m'}(x) = \frac{1}{w(x)} \phi_m(x)\overline{\phi_{m'}(x)}. \tag{80}$$

By (80) and (50), the expected value of F_{x_0} is

$$\mathrm{E}(F_{x_0}) = \langle \phi_m, \phi_{m'}\rangle = \delta_{m,m'}, \tag{81}$$

where the Kronecker delta $\delta_{m,m'}$ is 1 if $m = m'$ and 0 otherwise.

Consider N_n i.i.d random variables (80), denoted by

$$F_{m,m';k'} = \frac{1}{w(x^n_{k'})} \phi_m(x^n_{k'})\overline{\phi_{m'}(x^n_{k'})}, \quad k' = 1, \ldots, N_n.$$

Let

$$F^{N_n}_{m,m'} = \frac{1}{N_n}\sum_{k'=1}^{N_n} F_{m,m';k'}. \tag{82}$$

By (81) we have

$$\mathrm{E}\Big(F^{N_n}_{m,m'}\Big) = \langle \phi_m, \phi_{m'}\rangle.$$

On the other hand, the realization of the sum in (82) can be written as

$$F^{N_n}_{m,m'} = [\langle \boldsymbol{\Phi}_{\mathbf{n}}, \boldsymbol{\Phi}_{\mathbf{n}}\rangle]_{m,m'}. \tag{83}$$

This shows that $\langle \boldsymbol{\Phi}_{\mathbf{n}}, \boldsymbol{\Phi}_{\mathbf{n}}\rangle$ coincide on average with $\mathbf{I}$.

Next let us analyze the average mean square error over $m, m' \in \mathcal{M}$. For a matrix $\mathbf{A} = (a_{m,m'})_{m,m'}$, denote

$$\left\|\mathbf{A}\right\|_{\mathrm{F}} = \sqrt{\sum_{m,m'} \left|a_{m,m'}\right|^2}\ , \quad \left\|\mathbf{A}\right\|_{\mathrm{F},1} = \sum_{m,m'} \left|a_{m,m'}\right|.$$

We have

$$\begin{aligned}
&\mathrm{E}\left\|\langle \boldsymbol{\Phi}_{\mathbf{n}}, \boldsymbol{\Phi}_{\mathbf{n}}\rangle - \mathbf{I}\right\|_{\mathrm{F}}^2 \\
&= \iint_{x_1,\ldots,x_n} \sum_{m,m'} \left|F^{N_n}_{m,m'}(x^n_1,\ldots,x^n_{N_n}) - \delta_{m,m'}\right|^2 w(x^n_1)dx^n_1 \cdot w(x^n_{N_n})dx^n_{N_n} \\
&= \sum_{m,m'} \iint_{x_1,\ldots,x_n} \left|F^{N_n}_{m,m'}(x^n_1,\ldots,x^n_{N_n}) - \delta_{m,m'}\right|^2 w(x^n_1)dx^n_1 \cdot w(x^n_{N_n})dx^n_{N_n} \\
&= \sum_{m,m'} \mathrm{Var}F^{N_n}_{m,m'} = \sum_{m,m'} \frac{\mathrm{Var}F_{m,m'}}{N_n} = \frac{\left\|\mathrm{Var}F_{(\cdot)}\right\|_{\mathrm{F},1}}{N_n}
\end{aligned}$$

Next, we bound $\left\|\mathrm{Var}F_{(\cdot)}\right\|_{\mathrm{F},1}$. We have

$$\mathrm{Var}F_{m,m'} \leq \int_x \frac{1}{w(x)} \left|\phi_m(x)\phi_{m'}(x)\right|^2 w(x)dx,$$

so

$$\left\|\mathrm{Var}F_{(\cdot)}\right\|_{\mathrm{F},1} \leq \frac{M_\lambda^2}{w_{\min}} \max_m \left\|\phi_m\right\|_\infty^4$$

This proves that the expected mean square error satisfies

$$\mathrm{E}\left\|\langle \boldsymbol{\Phi}_{\mathbf{n}}, \boldsymbol{\Phi}_{\mathbf{n}}\rangle - \mathbf{I}\right\|_{\mathrm{F}}^2 \leq \frac{M_\lambda^2}{w_{\min}} \max_{m \leq M_\lambda} \left\|\phi_m\right\|_\infty^4 \frac{1}{N_n}. \tag{84}$$

Finally, by Markov's inequality, in probability more than $(1-\delta)$

$$\left\|\langle \boldsymbol{\Phi}_{\mathbf{n}}, \boldsymbol{\Phi}_{\mathbf{n}}\rangle - \mathbf{I}\right\|_{\mathrm{F}} \leq \frac{M_\lambda}{\sqrt{w_{\min}}} \max_{m \leq M_\lambda} \left\|\phi_m\right\|_\infty^2 \frac{1}{\sqrt{N_n}} \frac{1}{\sqrt{\delta}}. \tag{85}$$

■

Before formulating the last Monte-Carlo lemma, we require two more lemmas.

Lemma 38 *Let $\mathcal{S}(\lambda)$ be the unit $L^2(\mathcal{M})$ sphere in $PW(\lambda)$, and let ρ be a contractive positively homogeneous of order 1 activation function that preserves spectral decay. Then*

$$\mathcal{S}(\lambda) \ni f \mapsto \big(I - P(\lambda')\big)\rho(f)$$

is continuous as a mapping $\mathcal{S}(\lambda) \to L^\infty(\mathcal{M})$.

Proof

Let $f, g \in PW(\lambda)$. Consider the following calculation for any $M_2 > M_1 > M_{\lambda'}$.

$$\begin{aligned}
&\left\| \sum_{m=M_1}^{M_2} \langle \rho(f) - \rho(g), \phi_m \rangle \phi_m \right\|_\infty \\
&\leq \sum_{m=M_1}^{M_2} |\langle \rho(f) - \rho(g), \phi_m \rangle| \, \|\phi_m\|_\infty \\
&= \sum_{m=M_1}^{M_2} \|\phi_m\|_\infty \, |\langle \rho(f), \phi_m \rangle - \langle \rho(g), \phi_m \rangle| \qquad (86) \\
&= \sum_{m=M_1}^{M_2} m^{-1/2-\kappa/2} \, \|\phi_m\|_\infty \left| m^{1/2+\kappa/2} \langle \rho(f), \phi_m \rangle - m^{1/2+\kappa/2} \langle \rho(g), \phi_m \rangle \right| \\
&\leq R \sqrt{\sum_{m=M_1}^{\infty} \|\phi_m\|_\infty^2 \, m^{1+\kappa} \, |\langle \rho(f), \phi_m \rangle - \langle \rho(g), \phi_m \rangle|^2},
\end{aligned}$$

where

$$R = \sqrt{\sum_{m=1}^{\infty} m^{-1-2\kappa}}.$$

By (55),

$$\lim_{M_1 \to \infty} \sqrt{\sum_{m=M_1}^{\infty} \|\phi_m\|_\infty^2 \, m^{1+\kappa} \, |\langle \rho(f), \phi_m \rangle - \langle \rho(g), \phi_m \rangle|^2} = 0.$$

Therefore

$$\Big\{ \sum_{m=M_{\lambda'}}^{M} \langle \rho(f) - \rho(g), \phi_m \rangle \phi_m \Big\}_{M=M_{\lambda'}}^{\infty} \tag{87}$$

is a Cauchy sequence in $L^\infty(\mathcal{M})$, and thus converges in $L^\infty(\mathcal{M})$ to a limit we denote by

$$\sum_{m=M_{\lambda'}}^{\infty} \langle \rho(f) - \rho(g), \phi_m \rangle \phi_m. \tag{88}$$

The series (87) also converges in $L^2(\mathcal{M})$, to $(I - P(\lambda'))(\rho(f) - \rho(g))$. Since convergence in $L^2(\mathcal{M})$ implies pointwise convergence of a subsequence almost everywhere, we must have

$$\sum_{m=M_{\lambda'}}^{\infty} \langle \rho(f) - \rho(g), \phi_m \rangle \phi_m = (I - P(\lambda'))(\rho(f) - \rho(g)),$$

with convergence in $L^\infty(\mathcal{M})$. By conservation of bounds under limits, and by (86), we now have

$$\begin{aligned}
&\left\|(I - P(\lambda'))\rho(f) - P(\lambda_M)(I - P(\lambda'))\rho(g)\right\|_\infty \\
&= \left\|(I - P(\lambda'))(\rho(f) - \rho(g))\right\|_\infty \\
&\leq R\sqrt{\sum_{m=M_{\lambda'}}^{\infty} m^{1+\kappa} \left\|\phi_m\right\|_\infty^2 \left|\langle \rho(f) - \rho(g), \phi_m\rangle\right|^2}.
\end{aligned} \tag{89}$$

Last, the continuity of $(I - P(\lambda'))\rho(f)$ as a mapping $\mathcal{S}(\lambda) \to L^\infty(\mathcal{M})$ follows from (89) and (55). ■

By Lemma 38, $\left\|\big(I - P(\lambda')\big)\rho(f)\right\|_\infty$ has a maximal value in the compact domain $\mathcal{S}(\lambda)$ that we denote by $C_{\lambda'}$. For the next proposition we also need the following simple observation.

Lemma 39 *Let $A, B \geq 0$ such that $\left|A^2 - B^2\right| < \kappa$. Then $|A - B| < \sqrt{\kappa}$.*

Proof The equation $\left|A^2 - B^2\right| < \kappa$ is equivalent to

$$B^2 - \kappa < A^2 < B^2 + \kappa$$

or

$$\sqrt{B^2 - \kappa} < A < \sqrt{B^2 + \kappa}. \tag{90}$$

As a result

$$\sqrt{B^2} - \sqrt{\kappa} < A < \sqrt{B^2} + \sqrt{\kappa}$$

or equivalently

$$|A - B| < \sqrt{\kappa}.$$

■

Lemma 40 *Let $\{\mathcal{M}, \mu\}$ be a compact topological-measure space with $\mu(\mathcal{M}) = 1$. Consider the weighted measure μ_w satisfying (50), and a random sample set $\{x_k^n\}_{n=1}^{N_n}$ from $\{\mathcal{M}, \mu_w\}$. Consider a Laplacian $\mathcal{L}$ with eigenbasis $\{\phi_m\}$ as before. Suppose that the activation function ρ is contractive, positively homogeneous of order 1, and preserves spectral decay. Suppose that $\mathcal{L}$ respects continuity. Then for every $\delta > 0$, in probability more than $(1 - \delta)$*

$$\max_{f \in PW(\lambda)} \frac{\left\|S_n\rho(f) - S_n^\lambda P(\lambda')\rho(f)\right\|_{L^2(G_n)} - \left\|\rho(f) - P(\lambda')\rho(f)\right\|_{L^2(G_n)}}{\left\|P(\lambda)f\right\|_{L^2(\mathcal{M};\mu)}} \leq \frac{1}{w_{\min}^{1/4}} C_{\lambda'} \frac{1}{N_n^{1/4}} \frac{1}{\delta^{1/4}}, \tag{91}$$

where

$$C_{\lambda'} = \max_{f \in \mathcal{S}(\lambda)} \left\|\big(I - P(\lambda')\big)\rho(f)\right\|_\infty$$

and $\mathcal{S}(\lambda)$ is the unit sphere in $PW(\lambda)$.

Proof

First, since ρ is positively homogeneous of order 1, the maximum in (91) is equal to

$$\max_{f\in\mathcal{S}(\lambda)} \left\| S_n\rho(f) - S_n^{\lambda}P(\lambda')\rho(f)\right\|_{L^2(V^2)} - \left\|\rho(f) - P(\lambda')\rho(f)\right\|_{L^2(V^2)}. \tag{92}$$

Consider the random variable $F : \{\mathcal{M};\mu_w\} \rightarrow \mathbb{C}$ defined by

$$F(x) = \frac{1}{w(x)} \left|\big(\rho(f(x)) - P(\lambda')\rho(f(x))\big)\right|^2. \tag{93}$$

By (93) and (50), the expected value of F is

$$\mathrm{E}(F) = \left\|\rho(f) - P(\lambda')\rho(f)\right\|_2^2. \tag{94}$$

Consider N_n i.i.d random variables (93), denoted by

$$F_{k'} = \frac{1}{w(x^n_{k'})} \left|\big(\rho(f(x^n_{k'})) - P(\lambda')\rho(f(x^n_{k'}))\big)\right|^2, \quad k' = 1, \ldots, N_n.$$

Let

$$F^{N_n} = \frac{1}{N_n}\sum_{k'=1}^{N_n} F_{k'}. \tag{95}$$

By (94) we have

$$\mathrm{E}\Big(F^{N_n}\Big) = \left\|\rho(f) - P(\lambda')\rho(f)\right\|_2^2.$$

On the other hand, the realization of the sum in (95) can be written as

$$F^{N_n} = \left\|S_n\rho(P(\lambda)f) - S_nP(\lambda')\rho(P(\lambda)f)\right\|_{L^2(V^2)}^2. \tag{96}$$

This shows that on average (92) is zero.

Next let us analyze the expected error of (92).

$$\begin{aligned}
&\mathrm{E}\left|F^{N_n} - \left\|\rho(f) - P(\lambda')\rho(f)\right\|_2^2\right|^2\\
&= \iint_{x_1,\ldots,x_n} \left|F^{N_n}(x_1^n,\ldots,x^n_{N_n}) - \left\|\rho(f) - P(\lambda')\rho(f)\right\|_2^2\right|^2 w(x_1^n)dx_1^n\cdot w(x^n_{N_n})dx^n_{N_n}\\
&= \mathrm{Var}F^{N_n} = \frac{\mathrm{Var}F}{N_n}.
\end{aligned}$$

We have

$$\begin{aligned}
\mathrm{Var}F &\leq \int_x |F(x)|^2\, w(x)dx\\
&= \int_x \frac{1}{w(x)}\left|\big(\rho(f(x)) - P(\lambda')\rho(f(x))\big)\right|^4 dx\\
&\leq \frac{1}{w_{\min}}\left\|\big(I - P(\lambda')\big)\rho(f(x))\right\|_4^4\\
&\leq \frac{1}{w_{\min}}\left\|\big(I - P(\lambda')\big)\rho(f(x))\right\|_\infty^4 \leq \frac{1}{w_{\min}}C_{\lambda'}^4.
\end{aligned} \tag{97}$$

By (97), Markov's inequality states that in probability more than $(1-\delta)$

$$\left|F^{N_n} - \left\|\rho(f) - P(\lambda')\rho(f)\right\|_2^2\right| \leq \frac{1}{\sqrt{w_{\min}}} C_{\lambda'}^2 \frac{1}{\sqrt{N_n}} \frac{1}{\sqrt{\delta}}. \tag{98}$$

This shows, by Lemma 39 and (96), that

$$\max_{f\in PW(\lambda)} \frac{\left\|S_n\rho(f) - S_n^{\lambda}P(\lambda')\rho(f)\right\|_{L^2(V^2)} - \left\|\rho(f) - P(\lambda')\rho(f)\right\|_{L^2(V^2)}}{\|f\|_2} \leq \frac{1}{w_{\min}^{1/4}} C_{\lambda'} \frac{1}{N_n^{1/4}} \frac{1}{\delta^{1/4}}.$$

■

Proof [Proof of Theorem 32] We apply Lemmas 36,37 and 40 with failure probability $\delta/3$. Then, with probability more than $(1-\delta)$ the bounds (71), (79) and (91) are satisfied simultaneously. We thus consider the subsequence n_m that contains any n independently in probability more than $(1-\delta)$, for which the bounds (71), (79) and (91) are deterministic. Note that the sequence n_m is infinite in probability 1.

Denote $\overline{M}_n = M_{\overline{\lambda}_n}$. By assumption $\left\|\boldsymbol{\lambda}^{M_{\overline{\lambda}_n}}\right\|_1 = o(N_n^{1/2})$, where $\left\|\boldsymbol{\lambda}^{\overline{M}_n}\right\|_1$ is defined in (56). Let us analyze the dependency of the bounds (71), (79) and (91) on $\overline{M}_n$ and N_n. Note that the dependency of (71), (79) and (91) on λ does not affect the validity of Definitions 9 and 12, and 22. The asymptotic analysis in $\overline{M}_n$ and N_n in these definitions is for fixed λ.

The bound (71) depends on $\overline{M}_n$ as follows:

$$\begin{aligned}
\|H\|_2^2 &= \int_x\int_{x_0}\left|\sum_{m=1}^{\overline{M}_n}\phi_m(x_0)\lambda_m\phi_m(x)\right|^2 dx_0dx \\
&\leq \Big(\sum_{m=1}^{\overline{M}_n}|\lambda_m|\sqrt{\int_x\int_{x_0}|\phi_m(x_0)\phi_m(x)|^2\,dx_0dx}\Big)^2 \\
&= \Big(\sum_{m=1}^{\overline{M}_n}|\lambda_m|\sqrt{\int_{x_0}|\phi_m(x_0)|^2\,dx_0}\sqrt{\int_x|\phi_m(x)|^2\,dx}\Big)^2 \\
&= \Big(\sum_{m=1}^{\overline{M}_n}|\lambda_m|\Big)^2 = \left\|\boldsymbol{\lambda}^{\overline{M}_n}\right\|_1^2.
\end{aligned}$$

Thus, since the bound (71) also depend multiplicatively on $N_n^{-1/2}$, any choice of M_n such that $\left\|\boldsymbol{\lambda}^{M_{\overline{\lambda}_n}}\right\|_1 = o(N_n^{1/2})$ makes the bound converge to zero as $n\rightarrow\infty$, and guarantees Definition 9.

Note that the bounds (79) and (91) do not depend on M_n. The bound (79) proves that Definition 22 is satisfied for the subsequence n_m, which proves Definitions 7 and 8. For the relation between the bound (91) and Definition 12, we use Lemma 35, where (66) converges to zero in the subsequence n_m, and (67) converges to zero deterministically. This proves Definition 12 for the subsequence n_m.

■

B.5 Proof of Proposition 34

By the proof of Proposition 10,

$$\begin{aligned}\left\|g(\mathcal{L})P(\lambda) - R_n^{\lambda} g(\mathbf{\Delta}_n) S_n^{\lambda} P(\lambda)\right\| \leq & DC\sqrt{\#\{\lambda_j \leq \lambda\}_j}\left\|S_n^{\lambda_M}\mathcal{L}P(\lambda_M) - \mathbf{\Delta} S_n^{\lambda_M} P(\lambda_M)\right\| \\ & + \|g\|_{\mathcal{L},M}\left\|P(\lambda_M) - R_n^{\lambda_M} S_n^{\lambda_M} P(\lambda_M)\right\|.\end{aligned} \tag{99}$$

By the proof of Theorem 32,

$$\|H_{\lambda_n}\|_2 \leq \left\|\boldsymbol{\lambda}^{\overline{M}_n}\right\|_1 \leq BN_n^{1/2-\alpha}.$$

By Lemma 36, in probability more than $(1-\delta)$,

$$\left\|S_n^{\lambda}\mathcal{L}P(\lambda) - \mathbf{\Delta}_n S_n^{\lambda} P(\lambda)\right\|_{L^2(G_n)} \leq \frac{1}{w_{\min}}\|H_{\lambda_n}\|_{L^2(\mathcal{M}^2;\mu\times\mu)} C_{\lambda}\delta^{-1/2}N_n^{-1/2}.$$

where C_λ is defined by (70)

$$\forall g \in PW(\lambda). \quad \|g\|_\infty \leq C_\lambda \|g\|_2.$$

So

$$\begin{aligned}\left\|S_n^{\lambda}\mathcal{L}P(\lambda) - \mathbf{\Delta}_n S_n^{\lambda} P(\lambda)\right\|_{L^2(G_n)} &\leq \frac{1}{w_{\min}} BN_n^{1/2-\alpha} C_\lambda \delta^{-1/2} N_n^{-1/2} \\ &= \frac{B}{w_{\min}} C_\lambda \delta^{-1/2} N_n^{-\alpha}.\end{aligned} \tag{100}$$

We can bound C_λ as follows. Let $g = \sum_{n=1}^{M_\lambda} c_n\phi_n$. For every x in $\mathcal{M}$,

$$\begin{aligned}\|g\|_2 = \sqrt{\sum_{n=1}^{M_\lambda}|c_n|^2} &\geq M_\lambda^{-1/2}\sum_{n=1}^{M_\lambda}|c_n| \\ &\geq \frac{1}{\max_{x,n\in\{1,\ldots,M_\lambda\}}|\phi_n(x)|} M_\lambda^{-1/2}\sum_{n=1}^{M_\lambda}|c_n\phi_n(x)| \\ &\geq \frac{1}{\max_{m\leq M_\lambda}\|\phi_m\|_\infty} M_\lambda^{-1/2}\left|\sum_{n=1}^{M_\lambda} c_n\phi_n(x)\right| \\ &= \frac{1}{\max_{m\leq M_\lambda}\|\phi_m\|_\infty} M_\lambda^{-1/2}|g(x)|.\end{aligned}$$

Hence,

$$\|g\|_2 \geq \frac{1}{\max_{m\leq M_\lambda}\|\phi_m\|_\infty} M_\lambda^{-1/2}\|g\|_\infty.$$

Therefore, the optimal bound C_λ satisfies

$$C_\lambda \leq \max_{m\leq M_\lambda}\|\phi_m\|_\infty M_\lambda^{1/2}. \tag{101}$$

Note that for classical Fourier bases we have $\max_{m\leq M_\lambda} \|\phi_m\|_\infty = 1$.

Next we analyze the second term in the bound of Proposition 10. By (38), we can write in the basis $\{\phi_m\}_{m=1}^{M_\lambda}$

$$\mathbf{R}_n^\lambda \mathbf{S}_n^\lambda \mathbf{c} = \langle \mathbf{\Phi}_n, \mathbf{\Phi}_n \rangle \mathbf{c}. \tag{102}$$

By Lemma 37, in probability more than $(1-\delta)$,

$$\|\langle \mathbf{\Phi}_n, \mathbf{\Phi}_n\rangle - \mathbf{I}\|_{F(\mathbb{C}^{M_\lambda\times M_\lambda})} \leq C'\delta^{-1/2}N_n^{-1/2}. \tag{103}$$

Here,

$$C' = \frac{M_\lambda}{\sqrt{w_{\min}}} \max_{m\leq M_\lambda} \|\phi_m\|_\infty^2 ,$$

and $M_\lambda = \dim(PW(\lambda))$ as before.

By the fact that the induced l_2 norm is bounded by the Frobenius norm,

$$\left\| P(\lambda) - R_n^\lambda S_n^\lambda P(\lambda) \right\|_2 \leq \|\langle \mathbf{\Phi}_n, \mathbf{\Phi}_n\rangle - \mathbf{I}\|_{F(\mathbb{C}^{M_\lambda\times M_\lambda})}. \tag{104}$$

We now plug (100), (101), (103) and (104) in (99). The bound C on the norm of interpolation $\|R_n^\lambda\|$ is close to 1 by (102), (103), and the fact that $\|R_n^\lambda\| = \|(S_n^\lambda)^*\| = \|S_n^\lambda\|$. Therefore, for large enough n, $C < 2$ (in the same event of (103)). This leads to (57) in probability more than $1-2\delta$, since, in the worst case, not satisfying the bound (100) and not satisfying the bound (103) are disjoint events.

B.6 Proof of Claim 30

Let $\epsilon > 0$ and $f \in PW(\lambda)$. Let $g \in PW(\lambda)$ such that $\|f-g\|_2 < 1$. For any $N\in\mathbb{N}$

$$\begin{aligned}\sum_{|n|>N} n^{1+\kappa} \left|\langle \rho(g), \phi_n\rangle\right|^2 &= \sum_{|n|>N} |n|^{-1+\kappa} n^2 \left|\langle \rho(g), \phi_n\rangle\right|^2 \\ &\leq N^{-1+\kappa} \sum_{n=-\infty}^{\infty} n^2 \left|\langle \rho(g), \phi_n\rangle\right|^2 \\ &\leq N^{-1+\kappa} M_\lambda^2 \|g\|_2^2 \leq N^{-1+\kappa} M_\lambda^2 (\|f\|_2^2 + 1).\end{aligned}$$

Similarly,

$$\sum_{|n|>N} |n|^{1+\kappa} \left|\langle \rho(g), \phi_n\rangle\right|^2 \leq N^{-1+\kappa} M_\lambda^2 (\|f\|_2^2 + 1).$$

Now, choose $N = N(\epsilon)$ such that $N^{-1+\kappa}M_\lambda^2(\|f\|_2^2+1) < \epsilon/8$. Moreover, choose $\delta < \frac{\epsilon}{2N(\epsilon)^{1+\kappa}}$. Now, if $\|f-g\| < \min\{\delta, 1\}$ we have

$$\sum_{n=-N}^{N} n^{1+\kappa} \left|\langle \rho(f)-\rho(g), \phi_n\rangle\right|^2 \leq N^{1+\kappa} \sum_{n=-\infty}^{\infty} \left|\langle \rho(f)-\rho(g), \phi_n\rangle\right|^2 = N^{1+\kappa} \|\rho(f)-\rho(g)\|_2^2$$

and by the fact that ρ is contractive,

$$\sum_{n=-N}^{N} n^{1+\kappa} \left|\langle \rho(f)-\rho(g), \phi_n\rangle\right|^2 \leq N^{1+\kappa} \|f-g\|_2^2 < \epsilon/2.$$

Altogether,

$$\|\rho(f) - \rho(g)\|_{1+\kappa,2}^2 \leq \sum_{n=-N}^{N} |n|^{1+\kappa} |\langle \rho(f) - \rho(g), \phi_n \rangle|^2$$
$$+ 4 \max \left\{ \sum_{|n|>N} |n|^{1+\kappa} |\langle \rho(f), \phi_n \rangle|^2 , \sum_{|n|>N} |n|^{1+\kappa} |\langle \rho(g), \phi_n \rangle|^2 \right\} < \epsilon,$$

which proves continuity.

References

A. Anis, A. Gadde, and A. Ortega. Towards a sampling theorem for signals on arbitrary graphs. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3864–3868, May 2014. doi: 10.1109/ICASSP.2014.6854325.

Y. Bi, A. Chadha, A. Abbas, E. Bourtsoulatze, and Y. Andreopoulos. Graph-based object classification for neuromorphic vision sensing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 491–501, 2019.

F. M. Bianchi, D. Grattarola, L. Livi, and C. Alippi. Graph neural networks with convolutional arma filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

M. Boguñá, I. Bonamassa, M. De Domenico, S. Havlin, D. Krioukov, and M. Á. Serrano. Network geometry. *Nature Reviews Physics*, 3:114–135, 2021.

C. Borgs, J.T. Chayes, L. Lovász, V.T. Sós, and K. Vesztergombi. Convergent sequences of dense graphs i: Subgraph frequencies, metric properties and testing. *Advances in Mathematics*, 219(6):1801–1851, 2008. ISSN 0001-8708. doi: https://doi.org/10.1016/j.aim.2008.07.008.

M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, July 2017. ISSN 1053-5888. doi: 10.1109/MSP.2017.2693418.

J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, Aug 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.230.

J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun. Spectral networks and locally connected networks on graphs. *Proc. ICLR*, 2013.

D. Burago, S. Ivanov, and Y. Kurylev. Spectral stability of metric-measure laplacians. *Israel Journal of Mathematics*, 232(1):125–158, 2019.

Y. Cai, L. Ge, J. Liu, J. Cai, T. Cham, J. Yuan, and N. M. Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2272–2281, 2019.

F. Chen, S. Pan, J. Jiang, H. Huo, and G. Long. Dagcn: Dual attention graph convolutional networks, 2019.

S. Chen, R. Varma, A. Sandryhaila, and J. Kovačević. Discrete signal processing on graphs: Sampling theory. *IEEE Transactions on Signal Processing*, 63(24):6510–6523, Dec 2015. ISSN 1053-587X. doi: 10.1109/TSP.2015.2469645.

M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proc. NIPS*, 2016.

I. S. Dhillon, Y. Guan, and B. Kulis. Kernel k-means: Spectral clustering and normalized cuts. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 551–556, New York, NY, USA, 2004. ACM. ISBN 1-58113-888-1.

M. Fey, J. E. Lenssen, F. Weichert, and H. Müller. Splinecnn: Fast geometric deep learning with continuous b-spline kernels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 869–877, 2018.

F. Gama, A. G. Marques, G. Leus, and A. Ribeiro. Convolutional neural network architectures for signals supported on graphs. *IEEE Transactions on Signal Processing*, 67(4): 1034–1049, 2018.

F. Gama, J. Bruna, and A. Ribeiro. Stability of graph scattering transforms. *Advances in Neural Information Processing Systems*, 32:8038–8048, 2019a.

F. Gama, A. Ribeiro, and J. Bruna. Diffusion scattering transforms on graphs. In *in Proceedings of Int. Conf. Learning Representations*, 2019b.

F. Gama, J. Bruna, and A. Ribeiro. Stability properties of graph neural networks. *IEEE Transactions on Signal Processing*, 68:5680–5695, 2020.

M. Gori, G. Monfardini, and F. Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734 vol. 2, July 2005. doi: 10.1109/IJCNN.2005.1555942.

E. Isufi, A. Loukas, A. Simonetto, and G. Leus. Autoregressive moving average graph filtering. *IEEE Transactions on Signal Processing*, 65(2):274–288, Jan 2017a. ISSN 1053-587X. doi: 10.1109/TSP.2016.2614793.

E. Isufi, A. Loukas, A. Simonetto, and G. Leus. Filtering random graph processes over random time-varying graphs. *IEEE Transactions on Signal Processing*, 65(EPFL-ARTICLE-230521):4406–4421, 2017b.

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *In Proceedings of International Conference on Learning Representations (ICLR)*, 2017.

B. Knyazev, X. Lin, M. R. Amer, and G. W. Taylor. Image classification with hierarchical multigraph networks. *arXiv preprint arXiv:1907.09000 [cs.CV]*, 2019.

N. J. Korevaar and R. M. Schoen. Sobolev spaces and harmonic maps for metric space targets. *Comm. Anal. Geom.*, 1(4):39–75, 1993.

I. Kostrikov, Z. Jiang, D. Panozzo, D. Zorin, and J. Bruna. Surface networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Boguñá. Hyperbolic geometry of complex networks. *Phys. Rev. E*, 82:036106, Sep 2010. doi: 10.1103/PhysRevE.82.036106.

R. Levie, E. Isufi, and G. Kutyniok. On the transferability of spectral graph filters. In *2019 13th International conference on Sampling Theory and Applications (SampTA)*, pages 1–5. IEEE, 2019a.

R. Levie, F. Monti, X. Bresson, and M. M. Bronstein. Cayleynets: Graph convolutional neural networks with complex rational spectral filters. *IEEE Transactions on Signal Processing*, 67(1):97–109, Jan 2019b. ISSN 1053-587X. doi: 10.1109/TSP.2018.2879624.

L. Lovász and B. Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933 – 957, 2006. ISSN 0095-8956. doi: https://doi.org/10.1016/j.jctb.2006.05.002.

H. Maron, H. Ben-Hamu, N. Shamir, and Y. Lipman. Invariant and equivariant graph networks. *arXiv preprint arXiv:1812.09902*, 2018.

A. G. Marques, S. Segarra, G. Leus, and A. Ribeiro. Sampling of graph signals with successive local aggregations. *IEEE Transactions on Signal Processing*, 64(7):1832–1843, April 2016. ISSN 1053-587X. doi: 10.1109/TSP.2015.2507546.

F. Monti, D. Boscaini, J. Masci, E. Rodolà, J. Svoboda, and M. M. Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5425–5434, 2017.

A. Nilsson and X. Bresson. An experimental study of the transferability of spectral graph networks, 2020.

A. Ortega, P. Frossard, J. Kovačević, J. M. F. Moura, and P. Vandergheynst. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5):808–828, 2018.

G. Puy, N. Tremblay, R. Gribonval, and P. Vandergheynst. Random sampling of bandlimited signals on graphs. *Applied and Computational Harmonic Analysis*, 44(2):446–475, 2018.

F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, Jan 2009. ISSN 1045-9227. doi: 10.1109/TNN.2008.2005605.

S. Segarra, A. G. Marques, G. Leus, and A. Ribeiro. Interpolation of graph signals using shift-invariant graph filters. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 210–214, Aug 2015. doi: 10.1109/EUSIPCO.2015.7362375.

S. Segarra, A. G. Marques, and A. Ribeiro. Optimal graph-filter design and applications to distributed linear network operators. *IEEE Transactions on Signal Processing*, 65(15): 4117–4131, 2017.

M. A. Serrano, D Krioukov, and M. Boguñá. Self-similarity of complex networks and hidden metric spaces. *Phys. Rev. Lett.*, 100:078701, Feb 2008. doi: 10.1103/PhysRevLett.100.078701.

C. Song, S. Havlin, and H. A. Makse. Self-similarity of complex networks. *Nature*, 433(27): 392–395, 2005.

W. A. Strauss. *Partial Differential Equations: An Introduction, 2nd Edition.* Wiley, 2007.

G. Te, W. Hu, A. Zheng, and Z. Guo. Rgcnn: Regularized graph cnn for point cloud segmentation. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 746–754, 2018.

M. Tsitsvero, S. Barbarossa, and P. Di Lorenzo. Signals on graphs: Uncertainty principle and sampling. *IEEE Transactions on Signal Processing*, 64(18):4845–4860, Sep. 2016. ISSN 1053-587X. doi: 10.1109/TSP.2016.2573748.

Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 2020.

ZD. Zou and G. Lerman. Graph convolutional neural networks via scattering. *Applied and Computational Harmonic Analysis*, 49(3):1046–1074, 2020.