Politecnico di Torino

Master's Program in Mathematical Engineering
Master's Thesis

Spatial Networks in Geometric Deep Learning

Raffaele Paolino

Supervisors

Prof. Enrico Magli — Prof. Gitta Kutyniok
Dr. Giulia Fracastoro — Dr. Ron Levie

December 2021

Acknowledgements

This work has been developed during my stay at the Ludwig-Maximilians-Universität in Munich in 2021. I would like to thank Prof. Enrico Magli and Dr. Giulia Fracastoro for the opportunity of writing my master thesis abroad. This period has been challenging yet extremely educational.

A heartfelt thanks and my deepest gratitude to Prof. Gitta Kutyniok, who has welcomed me in her group and supported me all along this journey, and to Dr. Ron Levie, who has been an inexhaustible source of suggestions and ideas. I am looking forward to continuing working with you!

I would like to thank all the people that worked behind the scenes: Tamara, whose help in solving bureaucratic problems has been fundamental, and all the members of the "Mathematical Foundations of Artificial Intelligence" group for their warm welcome and the fruitful discussions.

To my flatmates Maria, Sophie, Franzi for making the shared life light and funny
To my friends Vale, Vicky, Fla, Ray, Ale that lighten up these hard times
To Fra, for constantly listening to me rant
To Anna, for being home, for being roots
To Melania, whose love goes beyond the kilometres between us
To Valentina, for her sympathetic ear
To my parents, for their support
To my brother Carmine, who motivates me to leave my comfort zone
To Simon, for his unconditional affection and encouragement
I dedicate this work.

Table of Contents

Chapter 1

Introduction

In recent years, convolutional neural networks (CNNs) have attracted considerable attention in the scientific community because of their ability to solve various tasks, especially in the realm of image analysis (e.g., image classification, image segmentation, medical diagnostics). The breakthrough happened in 2012, when a CNN called AlexNet was used for the first time in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC2012). The aim of the competition is to evaluate performance of algorithms for the task of image classification and object detection: a set of 1.2 million images is given, hand labeled with the presence or absence of one thousand object categories; the algorithm is then asked to correctly identify which object are present and where they are located in the picture. AlexNet obtained a top-5 error of 17% [1], ca. ten percentage points less than the previous established result. Since then, CNNs have been extensively used for other tasks where the domain is inherently Euclidean, mainly one dimensional sequences (text, time series) and two dimensional grids (images, videos). From a mathematical perspective, a CNN implements an operation called convolution that aggregates the information coming from local neighbours. Stacking multiple convolutional layers and combining them with non-linearities allows to extract multi-scale localized features to better represent the input for the task at hand without any human supervision.

The term "Geometric Deep Learning", first introduced in [2] and generalized in [3], refers to the geometric unification of several machine learning algorithms on different domains (grids, graphs, groups, manifolds, gauges). An example is represented by graph convolutional neural networks (ConvGNNs), a generalization of convolutional neural networks for graph-structured data [2, 4, 5]. Graphs arise naturally when relationships (edges) between entities (nodes) have to be modeled. For instance, in social networks, friendships (edges) are the way opinions are spread among the users (nodes); in e-commerce systems, the relationship between users and products can be exploited in order to give better recommendations; in

chemistry, protein-protein interactions need to be studied for drug design. Those three examples show that the use of graphs as a mathematical model is wide and it arises in real world applications. The tasks to be performed on relational data can, in general, be categorized as:

- Node-level task: the graph is fixed and the neural network is asked to learn representations for individual nodes. Such representation is learned in a semi-supervised or unsupervised fashion. In the former case, labels of a subset of nodes are known; in the latter case the labels are not known, and the training procedure tries to minimize an auxiliary loss function. An example of node-level task is given by the prediction of voting intention in social networks.

- Edge-level task: there is a fixed graph or multi-graph, and the neural network is asked to predict a value for pair of nodes. Usually, the neural network extracts node-level features and uses them to reconstruct the edges. For instance, recommender system are asked to predict the affinity between users and products.

- Graph-level task: the neural network is asked to predict a value for each graph in the dataset, and the graphs can vary in size and shape. The neural network extracts node-level features and summarizes them via a pooling layer to learn a representation of each graph. One possible example is to predict the properties (e.g., toxicity) of molecular compounds on human body.

The importance of powerful ConvGNNs can be especially seen in the last example: drug design is a long and expensive process that can take decades. Moreover, the set of synthetizable molecules is large, and the ability to test experimental drugs in vitro is limited. Such limitation can be overcome by means of models, such as ConvGNNs, that could accurately predict the properties of molecular compounds in order to focus the attention on a small set of candidates. For instance, in [6] the antibiotic properties of a molecular compound, called halicin, have been predicted by a graph neural network.

Another important application is fake news detection. Receiving misleading information can affect not just a single person, but the whole society. This is seen nowadays: in a world plagued by a pandemic, deceptive articles about the harmfulness of COVID-19 as well as the safety and side effects of vaccines is extremely dangerous [7]. Incorrect medical information can adversely affect health and delay proper treatment, possibly leading to an overload of the health-care system and economic losses. Misleading information could be also used intentionally to inflame social conflicts: some examples are discussed in recent news articles [8, 9, 10]. Therefore, models that could promptly and accurately distinguish between real and fake news could be used in order to block the latter from being spread.

There are several challenges when concepts from Euclidean data are transferred to graph-structured data; one of these is the lacking of a well-posed definition of direction. While in an image it is easy to identify the top left and bottom right corners, this is not possible in a generic graph because there is no inherent order among its nodes. Convolution for images can be seen as a fixed-size sliding window applied to the pixel domain; therefore, the generalization of convolution on graphs is not straightforward because the number of neighbours is variable and unbounded. Historically, two approaches have been followed, leading to spatial ConvGNNs and spectral ConvGNNs. In spatial ConvGNNs, graph convolution is defined in the spatial domain as the aggregation of feature information coming from the neighbours of each node. Examples of spatial ConvGNNs are presented in [11, 12], where information of nodes is propagated iteratively until a stable fixed point is reached.

In spectral ConvGNNs, graph convolution is defined in the frequency domain by means of the eigendecomposition of a self-adjoint operator, called graph Laplacian. Even though spectral ConvGNNs have a theoretical foundation, relying on spectral graph theory and signal processing, the need to compute an eigendecomposition make them computationally expensive and slow: a fast Fourier transform algorithm for generic graphs does not exist, and eigendecomposition is usually unstable under graph perturbations. The gap between spatial and spectral ConvGNNs has been bridged by [13, 14, 15], where the spectral convolution is performed in the spatial domain. For instance, [14] parametrizes the spectral filters as Chebyshev polynomials of the graph Laplacian, while [15] parametrizes the spectral filters as rational functions of the graph Laplacian.

Another important difference between Euclidean and graph-structured domains is that graphs can vary in size and shape. This becomes a main concern when one deals with stability of ConvGNNs. Loosely speaking, stability is a desirable property for which a small change in the input causes a small change in the output. In order to study stability of ConvGNNs, a notion of proximity for graphs is required: if two graphs are "near", a stable ConvGNN will produce a similar output. It is not possible to use the algebraic characterization of graphs (e.g., adjacency matrix, Laplacian matrix) to estimate their proximity, since the graphs could have different number of nodes. Therefore, new ways of characterizing graphs should be developed. One possible solution is shown in [16]: every convergent sequence of graphs converges to a limit object called "graphon" and every "graphon" is the limit object of a convergent sequence of graphs. Hence, as done in [17], one could consider the graphs similar if they belong to the same "graphon" family. A different approach is presented in [18]: graphs are thought as discretizations of topological spaces; hence, graphs of different sizes are similar if they discretize the same space.

1.1 Structure of the Work

The present work leverages on this second approach via the notion of "spatial networks", also called "random geometric graphs" [19]: a set of points is randomly sampled from a region of space, and any two points are linked if their distance is less than a specified value. Such space can be a generic topological space equipped with a metric, useful to identify balls, and a measure, useful to compute their volume. The concepts of metric and measure are revisited in Sections 2.1 to 2.2 respectively. The definition of a measure gives a way to construct a theory of integration and to define integral operators. The theory of integral operators is revisited in Section 2.3. The graph approximation of the topological metric measure space is obtained via a sampling procedure: the definitions of random variables and distributions are revisited in Section 2.4.

The construction of a graph from a metric measure space is explained in Chapter 3. In particular, in Section 3.1 the definition of Differential Laplacian is given in terms of an integral operator which is generalized in Section 3.2 to generic metric measure spaces. Such operator can be approximated by means of naive Monte-Carlo method. The sampling procedure, studied in Section 3.3, identifies a sample set that constitutes the nodes of the graph, while edges are built accordingly to the integral kernel, usually a normalized indicator of balls. In Section 3.4 the developed theory is used to obtain the common definitions of graph Laplacians, such as combinatorial, random walk and symmetric normalized Laplacians.

While Chapter 3 focuses on how to obtain a graph from a latent topological space, Chapter 4 focuses on the inverse problem, i.e. which properties of the latent space can be inferred from the topology of the graph. In real scenarios the latent topological space is not known, nor the sampling procedure that generated the graph. In Section 4.1 it is shown that in some cases the perturbation introduced by non-uniform sampling is small, hence, structural properties of the graph (e.g., degree of a node) approximates intrinsic properties of the latent space (e.g., measure of the ball centered at that node). In Section 4.2 a simple model, namely the unit circle model, is introduced, and applied in Sections 4.2.1 to 4.2.2 as the latent space for common citation networks (Cora, Pubmed, Citeseer) to improve performances of semi-supervised node classification.

In Chapter 5 it is shown that polynomial spectral graph filters are linearly stable to edge perturbations. While in [20, 18] stability bounds are given in terms of functional norms, in this work point-wise upper bounds are obtained. Point-wise upper bounds are useful as they allow to identify the nodes that could cause stability problems.

1.2 Contributions

The contribution of this work is to provide a common mathematical framework to different graph Laplacians that can be encountered in the literature. Indeed, the general definition of metric measure Laplacian in Chapter 3 gives a way to retrieve the usual definitions of graph Laplacian as well as to construct more of them by choosing the appropriate integral kernel. This is shown, for instance, in Section 3.4.5 where a novel graph Laplacian, called "symmetric normalized K-Laplacian", is built and its properties analyzed. The graph Laplacian identifies how convolution on the graph is performed; therefore, a correct choice is crucial to guarantee good performance of ConvGNNs, as shown in Section 4.2.1.

Several papers deal with the problem of "learning the Laplacian", i.e. finding the best Laplacian for the task at hand: [21] uses a virtual adjacency matrix obtained by learning a distance function over features of nodes; [22] introduces a parametrized family of graph Laplacians that unifies the commonly used ones. None of them, however, takes into account the properties of the latent space, that is the underlying geometry of the graph. In this work, the "learning the Laplacian" paradigma is seen in terms of learning the sampling density. In order to do so, a latent space must be introduced. Euclidean spaces are thought to be not suited for graph embedding: [23, 24] suggests that spherical or hyperbolic spaces are the natural latent spaces for graphs. In this work, the simple spherical space represented by the unit circle is studied in Section 4.2.

Another contribution is the debunking of the common belief that spectral graph filters are not transferable. Following the same line of reasoning in [18], in this work it is shown that the error of spectral filters is intimately related to the error of the Laplacian. Differently from the above-mentioned paper, the bounds on the error are not provided in terms of functional norms but point-wise. Point-wise upper bounds can be used for the identification of critical nodes for stability: this information is lost when one considers functional norms.

Chapter 2

Preliminaries

In this chapter, some useful concepts are revised. In particular, in Section 2.1 the definition of metric space is given, and some particular cases are analyzed (e.g., normed space, inner product space, ultrametric space). In Section 2.2 the concept of measure space is analyzed, leading to the construction of abstract integration theory. In Section 2.3 two kinds of operators, namely Hilbert-Schmidt and multiplication operators, are studied. In Section 2.4, the basics of probability theory are introduced such as random variable and distribution, as well as the mathematical foundations of naive Monte-Carlo methods. Finally, in Section 2.5 the theory of graph convolution is developed, and used to build a ConvGNN.

2.1 Metric Spaces

A metric (also called distance) is a function that quantifies how far two points of a set are.

Definition 2.1 (Metric)**.** Given a set $\mathcal{V}$, a *metric* d is a real valued function defined on $\mathcal{V}$ that satisfies the following properties

$$
\begin{aligned}
&d(x,y) \geq 0\,, && \text{(non-negativity)}\\
&d(x,y) = 0 \iff x = y\,, && \text{(identity of indiscernibles)}\\
&d(x,y) = d(y,x)\,, && \text{(symmetry)}\\
&d(x,y) \leq d(x,z) + d(z,y)\,, && \text{(weak triangle ineq.)}
\end{aligned}
$$

for all $x, y, z \in \mathcal{V}$. The pair $(\mathcal{V}, d)$ is called *metric space*.

In the following, some particular cases of metric space will be defined.

Definition 2.2 (Norm)**.** Given a vector space $\mathcal{V}$ over a field $\mathbb{F}$, a *norm* on $\mathcal{V}$ is a function $\|\cdot\|_{\mathcal{V}} : \mathcal{V} \to \mathbb{R}$, that satisfies the following properties

$$\begin{aligned} &\|x\|_{\mathcal{V}} \geq 0\,, && \text{(non-negativity)}\\ &\|x\|_{\mathcal{V}} = 0 \iff x = 0\,, && \text{(identity of indiscernibles)}\\ &\|\alpha x\|_{\mathcal{V}} = |\alpha|\|x\|_{\mathcal{V}}\,, && \text{(homogeneity)}\\ &\|x+y\|_{\mathcal{V}} \leq \|x\|_{\mathcal{V}} + \|y\|_{\mathcal{V}}\,, && \text{(triangle ineq.)} \end{aligned}$$

for all $x, y \in \mathcal{V}$, $\alpha \in \mathbb{F}$. The pair $(\mathcal{V}, \|\cdot\|_{\mathcal{V}})$ is called *normed space*.

A normed space is a metric space because the norm induces a metric on $\mathcal{V}$ defined as $d(x, y) = \|x - y\|_{\mathcal{V}}$.

Definition 2.3 (Inner Product)**.** Given a vector space $\mathcal{V}$ over a field $\mathbb{F}$, an *inner product* on $\mathcal{V}$ is a function $\langle\cdot,\cdot\rangle_{\mathcal{V}} : \mathcal{V} \times \mathcal{V} \to \mathbb{F}$, that satisfies the following properties

$$\begin{aligned} &\langle \alpha x + \beta y, z\rangle_{\mathcal{V}} = \alpha\langle x, z\rangle_{\mathcal{V}} + \beta\langle y, z\rangle_{\mathcal{V}}\,, && \text{(linearity)}\\ &\langle x, z\rangle_{\mathcal{V}} = \overline{\langle z, x\rangle_{\mathcal{V}}}\,, && \text{(conjugate-symmetry)}\\ &\langle x, x\rangle_{\mathcal{V}} \geq 0 \text{ if } x \neq 0\,, && \text{(positive-definiteness)} \end{aligned}$$

for all $x, y, z \in \mathcal{V}$, $\alpha, \beta \in \mathbb{F}$. The pair $(\mathcal{V}, \langle\cdot,\cdot\rangle_{\mathcal{V}})$ is called *inner product space*.

An inner product space is a metric space, because the inner product induces a norm on $\mathcal{V}$ defined as $\|x\|_{\mathcal{V}} = \sqrt{\langle x, x\rangle_{\mathcal{V}}}$, hence a metric. An example of inner product space is shown next.

Example 2.1 (Euclidean Space)**.** The Euclidean space of dimension n, denoted by $\mathbb{R}^{\mathrm{n}}$, is the space of all tuples of length n

$$\mathbb{R}^{\mathrm{n}} = \{\mathbf{x} = (x_1, \ldots, x_n) : x_i \in \mathbb{R}\,,\ \forall 1 \leq i \leq n\}\,,$$

equipped with the inner product

$$\langle \mathbf{x}, \mathbf{y}\rangle = \sum_{i=1}^{n} x_i y_i\,.$$

Another important class of metric spaces can be built strengthening the weak triangle ineq. [25].

Definition 2.4 (Ultrametric)**.** Given a metric d on a set $\mathcal{V}$, if d satisfies

$$d(x, z) \leq \max\{d(x, y), d(y, z)\}\,, \qquad \text{(strong triangle ineq.)}$$

for all $x, y, z \in \mathcal{V}$, then d is said to be an *ultrametric* (or *non-Archimedean metric*) on $\mathcal{V}$, and the pair $(\mathcal{V}, d)$ is called *ultrametric space*.

The definition of ultrametric space is well posed because the strong triangle ineq. implies the weak triangle ineq.

$$\begin{aligned} d(x,z) &\leq \max\{d(x,y), d(y,z)\} \leq \max\{d(x,y), d(y,z)\} + \min\{d(x,y), d(y,z)\} \\ &= d(x,y) + d(y,z)\,. \end{aligned}$$

Ultrametric spaces arise naturally in applications, as the following example shows.

Example 2.2 (Space of Infinite Sequences)**.** Consider the set of all infinitely long words from a discrete alphabet A

$$\mathcal{A} = \{\mathbf{a} \coloneqq \{a_i\}_{i\in\mathbb{N}} : a_i \in A\}\ ,$$

equipped with the metric

$$d(\mathbf{a},\mathbf{b}) = \exp\left(-\inf\{i \in \mathbb{N} : a_i \neq b_i\}\right)\ .$$

Intuitively, d is lower for words that share a common long prefix.

It can be proved that $(\mathcal{A}, d)$ is an ultrametric space: the strong triangle ineq. can be verified as follows.

Given two positive integers $N > n$ and three sequences $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathcal{A}$ such that $d(\mathbf{a},\mathbf{b}) = \exp(-N)$ and $d(\mathbf{a},\mathbf{c}) = \exp(-n)$, it can be noted that $a_n \neq c_n$

$$\begin{array}{cccccccccccc} \mathbf{b} & = & a_1 & a_2 & \cdots & a_{n-1} & a_n & a_{n+1} & \cdots & a_{N-1} & b_N & b_{N+1} & \cdots \\ \mathbf{c} & = & a_1 & a_2 & \cdots & a_{n-1} & c_n & c_{n+1} & \cdots & c_{N-1} & c_N & c_{N+1} & \cdots \end{array}$$

therefore, $d(\mathbf{b},\mathbf{c}) = \exp(-n) = \max\{d(\mathbf{a},\mathbf{b}), d(\mathbf{a},\mathbf{c})\}$. In the case $N = n$

$$\begin{array}{ccccccccc} \mathbf{b} & = & a_1 & a_2 & \cdots & a_{n-1} & b_n & b_{n+1} & \cdots \\ \mathbf{c} & = & a_1 & a_2 & \cdots & a_{n-1} & c_n & c_{n+1} & \cdots \end{array}$$

the two sequences could or could not have the same value in the n-th position; hence, the minimum index at which the two sequences differ is greater or equal than n and $d(\mathbf{b},\mathbf{c}) \leq \exp(-n) = \max\{d(\mathbf{a},\mathbf{b}), d(\mathbf{a},\mathbf{c})\}$.

Even though an ultrametric is a metric, the consequences on the topology of the space are counter-intuitive, as the following statement shows.

Theorem 2.1 (Balls in Ultrametric Spaces)**.** *Suppose $(\mathcal{V}, d)$ is an ultrametric space; define the ball of radius $r \geq 0$ and center $z \in \mathcal{V}$ as $B_r(z) = \{v \in \mathcal{V} : d(z,v) < r\}$, then*

1. *every point inside a ball is its center, i.e. $y \in B_r(x) \implies B_r(y) = B_r(x)$;*

2. *intersecting balls are contained in each other, i.e. if* $0 \leq m \leq M$*, either* $B_m(x) \cap B_M(y) = \emptyset$ *or* $B_m(x) \subset B_M(y)$.

Proof of (1). Fix $y \in \mathrm{B}_r(x)$ and consider $z \in \mathrm{B}_r(x)$, $w \in \mathrm{B}_r(y)$. It holds

$$d(y,z) \leq \max\{d(x,y), d(x,z)\} < r \implies z \in \mathrm{B}_r(y)\,,$$
$$d(x,w) \leq \max\{d(x,y), d(y,w)\} < r \implies w \in \mathrm{B}_r(x)\,,$$

hence, $\mathrm{B}_r(x) \subset \mathrm{B}_r(y)$ because z is arbitrary, and $\mathrm{B}_r(y) \subset \mathrm{B}_r(x)$ because w is arbitrary, from which the thesis follows. □

Proof of (2). Fix two points $x, y \in \mathcal{V}$ and fix two radii $0 \leq m \leq M$ such that the balls $\mathrm{B}_m(x)$, $\mathrm{B}_M(y)$ have no empty intersection $\mathrm{B}_m(x) \cap \mathrm{B}_M y \neq \emptyset$. Consider a point lying in the intersection $z \in \mathrm{B}_m(x) \cap \mathrm{B}_M(y)$. Take $w \in \mathrm{B}_m(x)$, it holds

$$d(w,y) \leq \max\{d(w,z), d(z,y)\} = M\,,$$

therefore, $w \in \mathrm{B}_M(y)$ implies $\mathrm{B}_m(x) \subset \mathrm{B}_M(y)$. □

2.2 Measure Theory

While a metric is a function that quantifies how close two points of a set $\mathcal{V}$ are, a measure is a way to quantify how big subsets of $\mathcal{V}$ are. However, it is not possible to assign a measure to all the subsets of $\mathcal{V}$, particularly if $\mathcal{V}$ is a continuous space; therefore, it is necessary to introduce the concept of σ-algebra.

Definition 2.5 (Measurable Space)**.** Given a set $\mathcal{V}$, a σ-algebra Σ on $\mathcal{V}$ is a collection of subsets of $\mathcal{V}$ such that

$$\mathcal{V} \in \Sigma\,,$$
$$E \in \Sigma \implies \mathcal{V} \setminus E \in \Sigma\,, \qquad \text{(closure under complement)}$$
$$\forall \{E_n\}_{n\in\mathbb{N}} : E_n \in \Sigma \implies \bigcup_{n\in\mathbb{N}} E_n \in \Sigma\,. \qquad \text{(closure under countable union)}$$

The elements of Σ are called *measurable sets*, and the pair $(\mathcal{V}, \Sigma)$ is called *measurable space* (or *Borel space*).

The elements of Σ are the only subsets of $\mathcal{V}$ that it is possible to measure. Intuitively, the requirements on Σ guarantees that it is possible to compute the measure of a set if the measure of its complement is known, or if it can be decomposed in finitely or countably many pieces whose measure is known. The properties a measure should satisfy are presented next.

Definition 2.6 (Measure)**.** Given a measurable space $(\mathcal{V}, \Sigma)$, a set function $\mu : \Sigma \to [0, +\infty]$ is called *measure* on $(\mathcal{V}, \Sigma)$ if

$$\mu(\emptyset) = 0\,,$$

$$\forall \{E_n\}_{n\in\mathbb{N}} \,:\, E_i \cap E_j = \emptyset, i \neq j \implies \mu\left(\bigcup_{n\in\mathbb{N}} E_n\right) = \sum_{n\in\mathbb{N}} \mu(E_n)\,. \quad (\sigma\text{-additivity})$$

The triple $(\mathcal{V}, \Sigma, \mu)$ is called *measure space*.

A non trivial example of measure is the Lebesgue measure on $\mathbb{R}^{\mathrm{n}}$. The most general way to define it is due to Carathéodory [26]; however, a more intuitive construction can be found in [27].

Example 2.3 (Lebesgue Measure on Euclidean Spaces)**.** An elementary set $E \subset \mathbb{R}^{\mathrm{n}}$ is a cartesian product of closed intervals

$$E = \prod_{i=1}^{n} [a_i, b_i]\,,\ a_i < b_i\,,$$

whose measure is defined as the product of the length of the intervals

$$\mu(E) \coloneqq \prod_{i=1}^{n} b_i - a_i\,.$$

An arbitrary subset $S \subset \mathbb{R}^{\mathrm{n}}$, can be approximated from without by the union of elementary sets, hence, its outer measure can be defined as

$$\mu^*(S) \coloneqq \inf\left\{\sum_{k=1}^{\infty} \mu(E_k), E_k \subset \mathbb{R}^{\mathrm{n}} \text{ elementary set } \forall 1 \leq k \leq \infty, A \subset \bigcup_{k=1}^{\infty} E_k\right\}\,,$$

and from within by compact sets, hence, its inner measure can be defined as

$$\mu_*(S) \coloneqq \sup\left\{\mu^*(K), K \subset \mathbb{R}^{\mathrm{n}} \text{ compact set, } A \supset K\right\}\,,$$

It is easy to see that $\mu_*(S) \leq \mu^*(S)$ since K is a subset of S, thus, a covering for S is also a covering for K. The set S is said to be Lebesgue-measurable if

$$\mu_*(S) = \mu(S) = \mu^*(S) < \infty\,,$$

with Lebesgue measure $\mu(S)$. If the outer measure of S is infinite, then S is said to be Lebesgue measurable if

$$\mu_*(S \cap M) = \mu(S \cap M) = \mu^*(S \cap M) < \infty\,,\ \forall M \,:\, \mu_*(M) = \mu^*(M) < \infty\,,$$

with Lebesgue measure

$$\mu(S) = \sup\left\{\mu(S \cap M) \,:\, \mu_*(M) = \mu^*(M) < \infty\right\}\,.$$

The Euclidean space $\mathbb{R}^n$ is an inner product space; therefore, it can be equipped with a metric d induced by the inner product. As the previous example shows, $\mathbb{R}^n$ can be equipped with a measure μ. The space $(\mathbb{R}^n, d, \mu)$ is a member of an interesting family of spaces whose definition follows [28].

Definition 2.7 (Uniformly Distributed Measure Space)**.** A measure space $(\mathcal{V}, \Sigma, \mu)$ equipped with a metric d is said to be a *uniformly distributed measure space* if the measure of an open ball depends only on its radius and not on its centre

$$0 < \mu(\mathrm{B}_r(x)) = \mu(\mathrm{B}_r(y)) < \infty\,,\ \forall x, y \in \mathcal{V}\,,\ 0 < r < \infty\,.$$

In other terms, in a uniformly distributed measure space all the balls with same radius have same measure.

The definition of a measure μ on a set $\mathcal{V}$ allows to construct a theory of integration. As done with measurable sets, one could wonder which are the functions that are worth integrating.

Definition 2.8 (Measurable Function)**.** Given two measurable space $(\mathcal{V}_1, \Sigma_1)$, $(\mathcal{V}_2, \Sigma_2)$, a function $X : (\mathcal{V}_1, \Sigma_1) \to (\mathcal{V}_2, \Sigma_2)$ is called *measurable* (relative to Σ_1 and Σ_2) if the counter-image of measurable sets is a measurable set, i.e.

$$X^{-1}(E) \in \Sigma_1\,,\ \forall E \in \Sigma_2\,.$$

The following example shows how to define the integral of functions defined on $\mathbb{R}^n$ using Lebesgue measure.

Example 2.4 (Lebesgue Integral on $\mathbb{R}^n$)**.** The integral, of a generic measurable function $f : \mathbb{R}^n \to \mathbb{R}$ can be defined in steps. The first step is to define the integral for simple measurable functions, i.e. measurable functions that take on only a finite number of positive values

$$s = \sum_{i=1}^{N} a_i \mathbb{1}_{A_i}\,,$$

where $a_i \in [0, +\infty)$ and A_i are measurable and disjoint. For this class of functions, the integral is defined as

$$\int s\,\mathrm{d}\mu = \sum_{i=1}^{N} a_i\,\mu(A_i)\,.$$

The second step is to define the integral for positive measurable functions. Suppose $f : \mathbb{R}^n \to [0, +\infty]$ is a measurable function, then the integral of f is

$$\int f\,\mathrm{d}\mu = \sup\left\{\int s\,\mathrm{d}\mu : s \text{ simple measurable function},\ s \le f\right\}\,.$$

An arbitrary measurable function $f : \mathbb{R}^n \to [-\infty, \infty]$ can be decomposed as $f = f^+ - f^-$ where $f^+ = \max(0, f)$ is the positive part and $f^- = -\min(0, f)$ is the

negative part. Both positive and negative parts are positive, hence, their integrals are well defined. In case both integrals are finite, f is said to be "integrable" and its integral is

$$\int f \,\mathrm{d}\mu = \int f^+ \,\mathrm{d}\mu - \int f^- \,\mathrm{d}\mu \,.$$

The collection of all integrable functions is denoted as $\mathrm{L}^1(\mathbb{R}^n)$.

Such construction could be generalized to real-valued functions defined on a generic measure space $(\mathcal{V}, \Sigma, \mu)$. However, the space of all integrable functions is a particular case of a more general class of functions that is defined below.

Definition 2.9 (Lebesgue Spaces)**.** Let $(\mathcal{V}, \Sigma, \mu)$ be a measure space; the space $\mathrm{L}^p(\mathcal{V})$ is the set of all functions $f : \mathcal{V} \to \mathbb{R}$ such that

$$\|f\|_{\mathrm{L}^p(\mathcal{V})} \coloneqq \begin{cases} \left(\int\limits_{\mathcal{V}} |f|^p \,\mathrm{d}\mu \right)^{\frac{1}{p}}, & 1 \leq p < \infty \\ \sup\limits_{\mathcal{V}} \operatorname{ess} |f| \,, & p = \infty \end{cases}$$

is finite.

The following result gives a relationship between different Lebesgue spaces.

Theorem 2.2 (Hölder Inequality)**.** *Let $(\mathcal{V}, \Sigma, \mu)$ be a measure space; let $f \in \mathrm{L}^p(\mathcal{V})$ and $g \in \mathrm{L}^q(\mathcal{V})$ such that $p^{-1} + q^{-1} = 1$, it holds*

$$\|f\, g\|_{\mathrm{L}^1(\mathcal{V})} \leq \|f\|_{\mathrm{L}^p(\mathcal{V})} \|g\|_{\mathrm{L}^q(\mathcal{V})}$$

In particular, if $\mu(\mathcal{V})$ is finite, the previous theorem states that the Lebesgue spaces are encapsulated

$$\mathrm{L}^p(\mathcal{V}) \subset \mathrm{L}^q(\mathcal{V}) \,,\ 1 \leq q < p \leq \infty \,.$$

The importance of Lebesgue integral is that it behaves well with limit processes, as stated in the following theorem.

Theorem 2.3 (Lebesgue's Dominated Convergence Theorem)**.** *Let $(\mathcal{V}, \Sigma, \mu)$ be a measure space; Let $\{f_k\}_{k \in \mathbb{N}}$ be a collection of real-valued measurable functions; let $g \in \mathrm{L}^1(\mathcal{V})$ be a positive function such that $f_k(x) \leq g(x)$ for almost every x. Suppose $\lim_{k\to\infty} f_k(x)$ exists for almost every x, then*

$$\int \lim_{k\to\infty} f_k \,\mathrm{d}\mu = \lim_{k\to\infty} \int f_k \,\mathrm{d}\mu \,.$$

2.3 Operator Theory

Operator theory is a branch of functional analysis that studies linear operators defined on function spaces, such as Lebesgue Spaces. Two types of operator will be analyzed in this section: Hilbert-Schmidt operators and multiplication operators.

Definition 2.10 (Hilbert-Schmidt Operator)**.** Let $(\mathcal{V}, \Sigma, \mu)$ be a measure space; let $K \in \mathrm{L}^2(\mathcal{V} \times \mathcal{V})$ be a square-integrable kernel; the *Hilbert-Schmidt operator* $\mathcal{K}$ is defined as

$$\begin{aligned} \mathcal{K} : \mathrm{L}^2(\mathcal{V}) &\to \mathrm{L}^2(\mathcal{V}) \,, \\ u &\mapsto \mathcal{K}u \,, \ \mathcal{K}u(x) = \int_{\mathcal{V}} K(x,y)\, u(y)\, \mathrm{d}\mu(y) \,. \end{aligned}$$

The following theorem analyzes the properties of a generic Hilbert-Schmidt operator.

Theorem 2.4 (Properties of Hilbert-Schmidt Operators)**.** *Let $(\mathcal{V}, \Sigma, \mu)$ be a measure space, the Hilbert-Schmidt Operator $\mathcal{K}$ is*

1. *linear and bounded;*
2. *self-adjoint, provided that K is symmetric, i.e. $K(x,y) = K(y,x)$;*
3. *compact.*

Proof of (1). Linearity is a direct consequence of linearity of the integral; continuity is proven using Hölder Inequality

$$\begin{aligned} |\mathcal{K}u(x)|^2 = \left| \int_{\mathcal{V}} K(x,y)\, u(y)\ \mathrm{d}\mu(y) \right|^2 &\leq \left(\int_{\mathcal{V}} |K(x,y)|\, |u(y)|\ \mathrm{d}\mu(y) \right)^2 \\ &\leq \|u\|^2_{\mathrm{L}^2(\mathcal{V})} \int_{\mathcal{V}} |K(x,y)|^2\ \mathrm{d}\mu(y) \,. \end{aligned}$$

Integrating both sides on $\mathcal{V}$ and taking the square root

$$\|\mathcal{K}u\|_{\mathrm{L}^2(\mathcal{V})} \leq \|K\|_{\mathrm{L}^2(\mathcal{V}\times\mathcal{V})}\, \|u\|_{\mathrm{L}^2(\mathcal{V})} \,,$$

the thesis follows. □

Proof of (2). Consider the scalar product in $\mathrm{L}^2(\mathcal{V})$, applying Fubini's theorem and the symmetry of the kernel it holds

$$\langle \mathcal{K}u, v \rangle_{\mathrm{L}^2(\mathcal{V})} = \int_{\mathcal{V}} v(x) \int_{\mathcal{V}} K(x,y)\, u(y)\, \mathrm{d}\mu(y)\, \mathrm{d}\mu(x)$$

$$
\begin{aligned}
&= \int_{\mathcal{V}} \int_{\mathcal{V}} K(x,y)\, u(y) v(x)\, \mathrm{d}\mu(y)\, \mathrm{d}\mu(x) \\
&= \int_{\mathcal{V}} u(y) \int_{\mathcal{V}} K(y,x) v(x)\, \mathrm{d}\mu(x)\, \mathrm{d}\mu(y) \\
&= \langle u, \mathcal{K} v \rangle_{\mathrm{L}^2(\mathcal{V})} \ .
\end{aligned}
$$

□

Proof of (3). The Hilbert-Schmidt operator $\mathcal{K}$ is compact because it is the limit of finite rank operators. Specifically, let $\{\Phi_i(x)\}_{i=1}^{\infty}$ be a complete orthonormal basis of $\mathrm{L}^2(\mathcal{V})$, then the product$\{\Phi_i(x)\,\Phi_j(y)\}_{i,j=1}^{\infty}$ is a complete orthonormal basis of $\mathrm{L}^2(\mathcal{V}\times\mathcal{V})$. Consider the approximated kernel

$$
K_N(x,y) \coloneqq \sum_{i=1}^{N}\sum_{j=1}^{N} k_{ij}\, \Phi_i(x)\, \Phi_j(y)\,,
$$

by Parseval's identity it holds

$$
\infty > \|K\|^2_{\mathrm{L}^2(\mathcal{V}\times\mathcal{V})} = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} |k_{ij}|^2\,,
$$

and, as a consequence

$$
\|K - K_N\|^2_{\mathrm{L}^2(\mathcal{V}\times\mathcal{V})} = \sum_{i=N+1}^{\infty}\sum_{j=N+1}^{\infty} |k_{ij}|^2 \xrightarrow[N\to\infty]{} 0\,,
$$

The approximated Hilbert-Schmidt operator

$$
\mathcal{K}_N u(x) \coloneqq \int_{\mathcal{V}} K_N(x,y) u(y)\, \mathrm{d}\mu(y) = \sum_{i=1}^{N} \underbrace{\sum_{j=1}^{N} k_{ij} \int_{\mathcal{V}} \Phi_j(y)\, u(y)\, \mathrm{d}\mu(y)}_{\in\mathbb{R}}\ \Phi_i(x)\,,
$$

is a finite rank operator, because it is an element of the span $\left(\{\Phi_i\}_{i=1}^{N}\right)$. Due to Hölder Inequality

$$
\begin{aligned}
\|(\mathcal{K} - \mathcal{K}_N)\, u\|^2_{\mathrm{L}^2(\mathcal{V})} &= \int_{\mathcal{V}} \left| \int_{\mathcal{V}} (K(x,y) - K_N(x,y)) u(y)\, \mathrm{d}\mu(y) \right|^2 \mathrm{d}\mu(x) \\
&\leq \|K - K_N\|^2_{\mathrm{L}^2(\mathcal{V}\times\mathcal{V})}\ \|u\|^2_{\mathrm{L}^2(\mathcal{V})}\,,
\end{aligned}
$$

hence

$$
\|\mathcal{K} - \mathcal{K}_N\|_{\mathrm{L}^2(\mathcal{V})\to\mathrm{L}^2(\mathcal{V})} \leq \|K - K_N\|_{\mathrm{L}^2(\mathcal{V}\times\mathcal{V})} \xrightarrow[N\to\infty]{} 0\,.
$$

□

Another important class of operators is represented by multiplication operators whose definition is given below.

Definition 2.11 (Multiplication Operator)**.** Let $(\mathcal{V}, \Sigma, \mu)$ be a measure space; let $M \in \mathrm{L}^\infty(\mathcal{V})$ be essentially bounded; the multiplication operator $\mathcal{M}$ is defined as

$$\begin{aligned} \mathcal{M} : \mathrm{L}^2(\mathcal{V}) &\to \mathrm{L}^2(\mathcal{V}) \,, \\ u &\mapsto \mathcal{M}u \,,\ \mathcal{M}u(x) = M(x)\, u(x) \end{aligned}$$

In the following theorem some of the properties of a generic multiplication operator are presented.

Theorem 2.5 (Properties of Multiplication Operators)**.** *Let $(\mathcal{V}, \Sigma, \mu)$ be a measure space, the Multiplication Operator $\mathcal{M}$ is*

1. *linear and bounded;*
2. *self-adjoint;*
3. *compact if and only if M is null almost everywhere.*

Proof of (1). Linearity is trivial; continuity is proven using Hölder Inequality

$$\begin{aligned} \|\mathcal{M}u\|_{\mathrm{L}^2(\mathcal{V})} &= \left(\int_{\mathcal{V}} |\mathcal{M}u(x)|^2 \,\mathrm{d}\mu(x) \right)^{\frac{1}{2}} = \left(\int_{\mathcal{V}} \left| M(x)|^2 |u(x) \right|^2 \mathrm{d}\mu(x) \right)^{\frac{1}{2}} \\ &\leq \|M\|_{\mathrm{L}^\infty(\mathcal{V})} \, \|u\|_{\mathrm{L}^2(\mathcal{V})} \,. \end{aligned}$$

□

Proof of (2). Consider the scalar product in $\mathrm{L}^2(\mathcal{V})$, it holds

$$\langle \mathcal{M}u, v \rangle_{\mathrm{L}^2(\mathcal{V})} = \int_{\mathcal{V}} v(x)\, M(x)\, u(x) \,\mathrm{d}\mu(x) = \langle u, \mathcal{M}v \rangle_{\mathrm{L}^2(\mathcal{V})} \,.$$

□

Proof of (3). The "if" is trivial; therefore, only the "only if" will be proven. Suppose $M(x)$ is not null and consider the set of points $\mathcal{S}_0 = \{x : |M(x)| > \epsilon\}$, then $\mu(\mathcal{S}_0) > 0$. Consider a sequence of subsets $\{\mathcal{S}_n\}_{n=0}^\infty$ of $\mathcal{V}_\epsilon$ such that

$$\mathcal{S}_n \subset \mathcal{S}_{n-1} \,,\ \mu(\mathcal{S}_n) = 2^{-n}\, \mu(\mathcal{S}_0) \,,$$

and the sequences of functions $\{f_n\}_{n=0}^\infty$ defined as

$$f_n(x) = 2^{n/2}\, \mathbb{1}_{\mathcal{S}_n}(x) \,.$$

Let $m \geq 1$, it holds

$$\begin{aligned}
\|\mathcal{M}\, f_n - \mathcal{M}\, f_{n+m}\|^2_{\mathrm{L}^2(\mathcal{V})} &= \int_{\mathcal{V}} |M(x)(f_n(x) - f_{n+m}(x))|^2 \,\mathrm{d}\mu(x) \\
&\geq \epsilon^2 \int_{\mathcal{S}_n \setminus \mathcal{S}_{n+m}} \left|2^{n/2}\right|^2 \mathrm{d}\mu(x) \\
&+ \epsilon^2 \int_{\mathcal{S}_{n+m}} \left|2^{n/2} - 2^{(n+m)/2}\right|^2 \mathrm{d}\mu(x) \\
&= \epsilon^2\, 2^n\, \mu(\mathcal{S}_n \setminus \mathcal{S}_{n+m}) \\
&+ \epsilon^2 \left(2^n + 2^{n+m} - 2^{n+1+m/2}\right) \mu(\mathcal{S}_{n+m}) \\
&= 2 \left(1 - 2^{-m/2}\right) \epsilon^2 \mu(\mathcal{S}_0) \\
&\geq \left(2 - \sqrt{2}\right) \epsilon^2 \mu(\mathcal{S}_0)\,.
\end{aligned}$$

The sequence $\{f_n\}_{n=0}^{\infty}$ is bounded because $\|f_n\|_2 = \mu(\mathcal{S}_0)$, but the sequence $\{\mathcal{M} f_n\}_{n=0}^{\infty}$ is not a Cauchy-sequence; therefore, it has no convergent subsequence and the thesis follows. □

2.4 Probability Theory

An important class of measure spaces are the so called probability spaces, that are spaces with unit measure.

Definition 2.12 (Probability Space)**.** Given a measure space $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathbb{P}$ is said to be a *probability measure* on Ω if $\mathbb{P}[\Omega] = 1$. The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is then called a *probability space*, the elements of $\mathcal{F}$ are called *events* and Ω is called *sample space*.

While probability theory can be studied in a measure theoretic setting, the concepts, as well as the notation, and the interpretation are different. For instance, measurable functions are called random variables.

Definition 2.13 (Random Variable)**.** A measurable function defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is called *random variable*.

Usually, when dealing with a random variable $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{V}, \Sigma)$ one wants to compute the probability that X takes its value in a given subset $E \in \Sigma$, i.e.

$$\mathbb{P}[X^{-1}(E)] = (\mathbb{P} \circ X^{-1})[E] = \mathbb{P}[X \in E] = \mathbb{P}[\{\omega \in \Omega : X(\omega) \in E\}]\,.$$

Definition 2.14 (Distribution of a Random Variable)**.** A random variable X defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $(\mathcal{V}, \Sigma)$ induces a measure on $(\mathcal{V}, \Sigma)$

$$\mathbb{P}^X[E] = \mathbb{P}[\{\omega \in \Omega : X(\omega) \in E\}]\,,\ \forall E \in \Sigma$$

called *distribution* of the random variable X.

Something more can be said in case the random variable takes on real values, as shown in the following example.

Example 2.5 (Distribution of a Real Valued Random Variable)**.** Let X be a real valued r.v.; let $\mathcal{B}$ be the Borel σ-algebra of $\mathbb{R}$, i.e. the smallest sigma algebra containing the open sets of $\mathbb{R}$; the distribution of X is

$$F_X(x) = \mathbb{P}^X[(-\infty, x]] = \mathbb{P}[X^{-1}(-\infty, x]] = \mathbb{P}[\{\omega \in \Omega \,:\, X(\omega) \leq x\}] = \mathbb{P}[X \leq x]\,.$$

When exists f_x such that

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\,\mathrm{d}\mu(y)\,,\ \forall x \in \mathbb{R}\,,$$

f_X is called *probability density function* of the random variable X.

The integral of a random variable is called expectation, and it is defined as follows.

Definition 2.15 (Expected Value of a Random Variable)**.** A random variable X defined on $(\Omega, \Sigma, \mathbb{P})$ with values in $(\mathcal{V}, \Sigma)$ has a finite expectation (or is integrable) if both $\mathbb{E}[X^+], \mathbb{E}[X^-]$ are finite. In this case, the *expected value* is

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]\,,$$

also written as $\int_\Omega X\,\mathrm{d}\mathbb{P}$.

The following theorem is useful to understand the behaviour of a random variable X. In particular, it gives an upper bound on the probability that X will take values at a certain distance form its mean.

Theorem 2.6 (Bienaymé-Chebyshev Inequality)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let X be a random variable with finite expected value and finite centered p-moment $\mathbb{E}\left[|X - \mathbb{E}[X]|^p\right]$. Then for any real number $k > 0$, it holds*

$$\mathbb{P}\left[|X - \mathbb{E}[X]| > k\right] \leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^p]}{k^p}$$

Proof. Let g be a measurable function, non-decreasing and non-negative such that $g(k)$ is not null

$$\mathbb{P}\left[|X - \mathbb{E}[X]| > k\right] = \int_\Omega \mathbb{1}_{|X - \mathbb{E}[X]| > k}\,\mathrm{d}\mathbb{P}$$

$$
\begin{aligned}
(g(k) \neq 0) &= \frac{1}{g(k)} \int_{\Omega} g(k) \mathbb{1}_{|X-\mathbb{E}[X]|>k} \, \mathrm{d}\mathbb{P} \\
(g \text{ non-decreasing}) &\leq \frac{1}{g(k)} \int_{\Omega} g\left(|X-\mathbb{E}[X]|\right) \mathbb{1}_{|X-\mathbb{E}[X]|>k} \, \mathrm{d}\mathbb{P} \\
(g \geq 0) &\leq \frac{1}{g(k)} \int_{\Omega} g\left(|X-\mathbb{E}[X]|\right) \mathrm{d}\mathbb{P} \\
&= \frac{\mathbb{E}\left[g\left(|X-\mathbb{E}[X]|\right)\right]}{g(k)} \,,
\end{aligned}
$$

the thesis follows with $g(k) = k^p$. □

Two important concepts in probability theory are presented next.

Definition 2.16 (Independent Random Variables)**.** Let X_1, X_2 be two random variables that take values on $(\mathcal{V}_1, \Sigma_1)$, $(\mathcal{V}_2, \Sigma_2)$ respectively; they are said to be *independent* if

$$
\mathbb{P}[(X \in A) \cap (Y \in B)] = \mathbb{P}[X \in A] \, \mathbb{P}[Y \in B] \,, \, \forall A \in \Sigma_1, \, \forall B \in \Sigma_2 \,.
$$

Loosely speaking, two random variables are independent if the knowledge of Y does not affect the probabilities that X will take on certain values.

Definition 2.17 (Identically Distributed Random Variables)**.** Let X_1, X_2 be two random variables that take values on the same measurable space $(\mathcal{V}, \Sigma)$; they are said to be identically distributed if

$$
\mathbb{P}[X_1 \in A] = \mathbb{P}[X_2 \in A] \,, \, \forall A \in \Sigma \,.
$$

The previous definitions are necessary to state the following theorem. It represents the theoretical foundation of the naive Monte-Carlo method, a method to approximate integrals via finite sums.

Theorem 2.7 (Strong Law of Large Numbers)**.** *Let $\{X_i\}_{i=1}^N$ be a sequence of independent random variables identically distributed to an integrable random variable X. Define the sample mean as*

$$
\overline{X}_N = \sum_{i=1}^{N} \frac{X_i}{N} \,,
$$

it holds

$$
\lim_{N \to \infty} \overline{X}_N = \mathbb{E}[X] \ a. \ s. \,.
$$

When approximating an integral via a finite sum using the Strong Law of Large Numbers, it is necessary to sample from the random variable X. In some cases, the problem can be reduced to sample from a uniform distribution, as shown in the next result.

Theorem 2.8 (Inverse Transform Sampling)**.** *Given a real-valued random variable X with cumulative distribution function F_X and a uniform random variable U supported in $(0,1)$, then $F_X^{-1}(U)$ has the same distribution of X.*

Proof. We have to show that the cumulative distribution function of $F_X^{-1}(U)$ and X are the same

$$\mathbb{P}[F_X^{-1}(U) \leq x] = \mathbb{P}[U \leq F_X(x)] = F_X(x) = \mathbb{P}[X \leq x]\,,$$

where the last equality is guaranteed by the fact that $F_X(x) \in (0,1)$. □

However, in some cases the inverse cumulative distribution function is not known. The following theorem gives an alternative way to get a sample from X.

Theorem 2.9 (Acceptance-Rejection Sampling)**.** *Given a random variable X with bounded probability density function f_X and bounded support (a,b). Suppose $M = \max_{x\in(a,b)} f_X(x)$, then*

- *draw a sample $y \sim \mathrm{Unif}(a,b)$;*
- *draw a sample $u \sim \mathrm{Unif}(0,1)$, independent of y;*
- *accept y if $u < f_X(y)/M$, otherwise reject.*

gives a sample from X.

2.5 Spectral Graph Theory

A graph is a mathematical object that represents interactions between entities.

Definition 2.18 (Graph)**.** A *graph* $\mathcal{G}$ is a pair $(\mathcal{V},\mathcal{E})$ where $\mathcal{V}$ represents the set of vertices and $\mathcal{E} \subset \mathcal{V}\times\mathcal{V}$ represents the set of links. A link is an ordered pair (u,v) such that $u,v \in \mathcal{V}$. A graph is said to be *undirected* if $(u,v) \in \mathcal{E} \iff (v,u) \in \mathcal{E}$.

The most common way to represent a graph is using its adjacency matrix. The adjacency matrix indicates whether two points are connected or not.

Definition 2.19 (Adjacency Matrix)**.** Given a graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$ with N nodes, the *adjacency matrix* $\mathbf{A} \in \{0,1\}^{N\times N}$ is the matrix defined as

$$[\mathbf{A}]_{ij} = 1 \iff (i,j) \in \mathcal{E}\,.$$

A general algebraic representation of a graph is introduces in the following definition.

Definition 2.20 (Graph Shift Operator)**.** Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with N nodes, a *graph shift operator* is a matrix $\mathbf{L} \in \mathbb{R}^{N \times N}$ that respects the connectivity of the graph, i.e.

$$[\mathbf{L}]_{ij} = 0\,,\ (i,j) \notin \mathcal{E}\,,\ i \neq j\,.$$

In order to introduce the most common choices for a graph shift operator, it is useful to define the degree matrix.

Definition 2.21 (Degree Matrix)**.** Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with N nodes and adjacency matrix $\mathbf{A}$, the *degree matrix* $\mathbf{D} \in \mathbb{R}^{N \times N}$ is the diagonal matrix defined as

$$\mathbf{D} = \mathrm{diag}(\mathbf{A}\mathbf{1})\,,$$

where $\mathbf{1}$ is the constant N-dimensional unit vector.

The degree of a node is the number of its neighbours. A node whose degree is 0 is said to be isolated. The presence of such nodes could make the degree matrix not invertible. A possible solution is to set $[\mathbf{D}^{-1}]_{ii} = 0$ if $[\mathbf{D}]_{ii} = 0$. With this convention, the graph shift operators

$$\begin{aligned}
\mathbf{L}^{\mathrm{c}} &= \mathbf{D} - \mathbf{A}\,, && \text{(combinatorial Laplacian)}\\
\mathbf{L}^{\mathrm{rw}} &= \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}\,, && \text{(random walk Laplacian)}\\
\mathbf{L}^{\mathrm{sn}} &= \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\,, && \text{(symmetric normalized Laplacian)}
\end{aligned}$$

are well defined.

2.5.1 Graph Convolution

In digital signal processing theory, a filter is a convolution operator. Filtering is a fundamental operation in the analysis of signals, because it amplifies or attenuates frequencies, and can be performed in the frequency domain by a point-wise multiplication.

Given two signals $f \in \mathrm{L}^p(\mathbb{R}^{\mathrm{N}})$ and $g \in \mathrm{L}^q(\mathbb{R}^{\mathrm{N}})$, the convolution $f \star g \in \mathrm{L}^r(\mathbb{R}^{\mathrm{N}})$ is the function

$$(f \star g)(\mathbf{x}) = \int_{\mathbb{R}^{\mathrm{N}}} f(\mathbf{y}) g(\mathbf{y} - \mathbf{x})\,\mathrm{d}\mathbf{y}\,,$$

where r is such that $p^{-1} + q^{-1} - r^{-1} = 1$. Due to the convolution theorem, the convolution can be computed by a point-wise product of the Fourier transforms of the two signals

$$(f \star g)(\mathbf{x}) = \mathcal{F}^{-1}\left[\mathcal{F}[f](\mathbf{s})\,\mathcal{F}[g](\mathbf{s})\right](\mathbf{x})\,,$$

where $\mathcal{F}$ represents the Fourier transform

$$\mathcal{F}[f](\mathbf{s}) = \int_{\mathbb{R}^N} f(\mathbf{x}) \exp\left(-\mathrm{i}\mathbf{x}^{\mathrm{T}}\mathbf{s}\right) \mathrm{d}\mathbf{x}$$

and $\mathcal{F}^{-1}$ the inverse Fourier transform

$$\mathcal{F}^{-1}[f](\mathbf{x}) = \frac{1}{(2\pi)^N} \int_{\mathbb{R}^N} f(\mathbf{x}) \exp\left(\mathrm{i}\mathbf{x}^{\mathrm{T}}\mathbf{s}\right) \mathrm{d}\mathbf{s}$$

The Fourier transform can be seen as a scalar product of the signal f against the elements of the Fourier basis; each element of the Fourier basis satisfies an eigenvalue-type equation

$$\Delta u(\mathbf{x}) = -k^2 u(\mathbf{x})$$

where Δ is the Laplacian operator.

It is possible to define the convolution for signals on graphs in an similar way.

Definition 2.22 (Fourier Transform of Graph Signals)**.** Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with N nodes. Let $\mathbf{L} \in \mathbb{R}^{N \times N}$ be an Hermitian Graph Shift Operator with eigen-decomposition (Φ, Λ), i.e. $\mathbf{L} = \Phi\Lambda\Phi^{\mathrm{T}}$. The *graph Fourier transform* of a signal $\mathbf{x} : \mathcal{V} \to \mathbb{R}$ is defined as

$$\mathcal{F}[\mathbf{x}] \coloneqq \Phi^{\mathrm{T}}\mathbf{x}$$

and the *inverse graph Fourier transform* as

$$\mathcal{F}^{-1}[\mathbf{x}] \coloneqq \Phi\mathbf{x}$$

Mimicking the convolution theorem, the graph convolution can be defined as follows.

Definition 2.23 (Graph Convolution)**.** Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with N nodes. Let $\mathbf{L} \in \mathbb{R}^{N \times N}$ be an Hermitian Graph Shift Operator with eigen-decomposition (Φ, Λ). Let $\mathbf{h}, \mathbf{x} : \mathcal{V} \to \mathbb{R}$ be two graph signals, the *graph convolution* $\mathbf{h} \star \mathbf{x}$ is defined as

$$\mathbf{h} \star \mathbf{x} \coloneqq \mathcal{F}^{-1}[\mathcal{F}[\mathbf{h}] \odot \mathcal{F}[\mathbf{x}]] = \Phi \,\mathrm{diag}(\Phi^{\mathrm{T}}\mathbf{h})\, \Phi^{\mathrm{T}}\mathbf{x}\,,$$

where $\odot$ represents the Hadamard product.

If $\mathbf{h}$ is interpreted as a graph filter, then $\mathbf{H} = \mathrm{diag}\left(\Phi^{\mathrm{T}}\mathbf{x}\right)$ represents its frequency response. In order to ensure that that the filter respects the structure of the graph, $\mathbf{H}$ can be parametrized as a function of the eigenvalues of $\mathbf{L}$, i.e. $\mathbf{H} = h(\Lambda)$. If h is a polynomial, or more generally a rational function, then

$$\mathbf{h} \star \mathbf{x} = \Phi\, h(\Lambda)\, \Phi^{\mathrm{T}}\mathbf{x} = h(\mathbf{L})\mathbf{x}\,.$$

Definition 2.24 (Polynomial Spectral Filters)**.** Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with N nodes. Let $\mathbf{L} \in \mathbb{R}^{N \times N}$ be a Graph Shift Operator. Let h be a polynomial function, then $h(\mathbf{L})$ is called *polynomial spectral filter*.

2.5.2 Graph Convolutional Neural Network

The definition of a Graph Convolution, gives a way to define graph convolutional layer; the simplest choices are provided in [13, 14], here presented as definitions.

Definition 2.25 (GCNConv)**.** Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with N nodes, adjacency matrix $\mathbf{A}$ and degree matrix $\mathbf{D}$. Let $\mathbf{X} \in \mathbb{R}^{N \times d}$ be the input feature matrix, and $\mathbf{W} \in \mathbb{R}^{d \times d'}$ the learnable weight matrix, then the *GCNConv* layer is defined as

$$\mathbf{X}' = (\mathbf{D} + \mathbf{I})^{-\frac{1}{2}}(\mathbf{A} + \mathbf{I})(\mathbf{D} + \mathbf{I})^{-\frac{1}{2}}\mathbf{X}\mathbf{W}\,.$$

The GCNConv layer aggregates information from local neighbours. However, it is possible to aggregate information up to the K-hop neighbours, where K can be chosen arbitrarly. This is done, for example, by the ChebConv layer.

Definition 2.26 (ChebConv)**.** Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with N nodes and Graph Shift Operator $\mathbf{L}$. Let $\mathbf{X} \in \mathbb{R}^{N \times d}$ be the input feature matrix. Let $K \in \mathbb{N}$ and $W^{(k)} \in \mathbb{R}^{d \times d'}$ be a learnable matrix for all $k \in \{1, \dots, K\}$, then the *ChebConv* layer is defined as

$$\mathbf{X}' = \sum_{k=0}^{K} \mathbf{Z}^{(k)}\mathbf{X}\mathbf{W}^{(k)}\,,$$

where $\mathbf{Z}^{(k)}$ is the k-th Chebischev polynomial of the normalized graph shift operator $\mathbf{L}$

$$\begin{cases} \mathbf{Z}^{(0)} = \mathbf{I}\,, \\ \mathbf{Z}^{(1)} = \dfrac{2}{\lambda_{\max}}\mathbf{L} - \mathbf{I}\,, \\ \mathbf{Z}^{(k)} = 2\left(\dfrac{2}{\lambda_{\max}}\mathbf{L} - \mathbf{I}\right)\mathbf{Z}^{(k-1)} - \mathbf{Z}^{(k-2)}\,,\ k \geq 1 \end{cases}.$$

The normalization $2\,\mathbf{L}/\lambda_{\max} - \mathbf{I}$ guarantees numerical stability of matrix powers.

Stacking multiple graph convolutional layers, and combining them with non-linear function gives a way to construct a Graph Convolutional Neural Network, such as the following.

Definition 2.27 (GCN)**.** Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with N nodes, adjacency matrix $\mathbf{A}$ and degree matrix $\mathbf{D}$. Let $\mathbf{X} \in \mathbb{R}^{N \times d}$ be the input feature matrix. Let L be the total number of layers, $\sigma^{(l)} : \mathbb{R} \to \mathbb{R}$ the entry-wise non-linear function at layer l, $\mathbf{W}^{(l)} \in \mathbb{R}^{d_{l-1} \times d_l}$ the learnable weight matrix at layer l, then the *GCN* is defined as

$$\begin{cases} \mathbf{X}^{(0)} = \mathbf{X}\,,\ d_0 = d \\ \mathbf{X}^{(l)} = \sigma^{(l)}\left((\mathbf{D} + \mathbf{I})^{-\frac{1}{2}}(\mathbf{A} + \mathbf{I})(\mathbf{D} + \mathbf{I})^{-\frac{1}{2}}\mathbf{X}^{(l-1)}\mathbf{W}^{(l)}\right)\,,\ l \in \{1, \dots, L\} \end{cases} \tag{2.1}$$

Chapter 3

Graph Approximation of Metric Measure Spaces

The aim of this chapter is to explain how a graph can be built from a metric measure space. In particular, in Section 3.1 an integral representation of the Differential Laplacian is given. Such representation is generalized in Section 3.2 to a topological space $\mathcal{V}$ equipped with a metric d and a measure μ. In Section 3.3 the Laplacian is random sampled, leading to a graph approximation of $\mathcal{V}$.

3.1 Differential Laplacian

Definition 3.1 (Differential Laplacian)**.** The Laplacian $\Delta : \mathrm{C}^2(\Omega) \to \mathrm{C}^0(\Omega)$ maps twice continuously differentiable functions defined on an open set $\Omega \subset \mathbb{R}^{\mathrm{n}}$ to the sum of its unmixed second derivatives, i.e.

$$u(\mathbf{x}) \mapsto \sum_{i=1}^{n} \frac{\partial^2 u}{\partial x_i^2}(\mathbf{x}), \ \forall \mathbf{x} \in \Omega .$$

Roughly speaking, the Laplacian of a function u at a point $\mathbf{x}$ measures by how much the average value of u over small balls centered at $\mathbf{x}$ deviates from $u(\mathbf{x})$, as stated in the following theorem.

Theorem 3.1 (Integral Form of Differential Laplacian)**.** *Let $u \in \mathrm{C}^2(\Omega)$ be a twice continuously differentiable function defined on an open set $\Omega \subseteq \mathbb{R}^{\mathrm{n}}$, then*

$$\Delta u(\mathbf{x}) = \lim_{r \to 0} \frac{2(n+2)}{vol(B_1(\mathbf{0}))\, r^{n+2}} \int\limits_{B_r(\mathbf{x})} (u(\mathbf{y}) - u(\mathbf{x}))\, \mathrm{d}\mathbf{y} .$$

Proof. Due to the hypothesis $u \in \mathrm{C}^2(\Omega)$, u can be Taylor-expanded up to the second order in a neighborhood of $\mathbf{x}$

$$u(\mathbf{y}) - u(\mathbf{x}) = \nabla u(\mathbf{x})^{\mathrm{T}}(\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^{\mathrm{T}}\mathrm{H}_u(\mathbf{x})(\mathbf{y}-\mathbf{x}) + o(\|\mathbf{y}-\mathbf{x}\|^2),$$

where

$$\left(\mathrm{H}_u(\mathbf{x})\right)_{ij} = \left(\frac{\partial^2 u}{\partial x_i \partial x_j}(\mathbf{x})\right)_{ij},$$

is the Hessian matrix of u computed in $\mathbf{x}$. Integrating both sides over a ball of radius r centered in $\mathbf{x}$, it holds

$$\begin{aligned}
\int_{\mathrm{B}_r(\mathbf{x})} (u(\mathbf{y}) - u(\mathbf{x}))\,\mathrm{d}\mathbf{y} = &\int_{\mathrm{B}_r(\mathbf{x})} \nabla u(\mathbf{x})^{\mathrm{T}}(\mathbf{y}-\mathbf{x})\,\mathrm{d}\mathbf{y} \\
&+ \frac{1}{2}\int_{\mathrm{B}_r(\mathbf{x})} (\mathbf{y}-\mathbf{x})^{\mathrm{T}}\mathrm{H}_u(\mathbf{x})(\mathbf{y}-\mathbf{x})\,\mathrm{d}\mathbf{y} \\
&+ \int_{\mathrm{B}_r(\mathbf{x})} o(\|\mathbf{y}-\mathbf{x}\|^2)\,\mathrm{d}\mathbf{y} \\
= &\int_{\mathrm{B}_r(\mathbf{0})} \nabla u(\mathbf{x})^{\mathrm{T}}\mathbf{z}\,\mathrm{d}\mathbf{z} \\
&+ \frac{1}{2}\int_{\mathrm{B}_r(\mathbf{0})} \mathbf{z}^{\mathrm{T}}\mathrm{H}_u(\mathbf{x})\mathbf{z}\,\mathrm{d}\mathbf{z} \\
&+ \int_{\mathrm{B}_r(\mathbf{0})} o(\|\mathbf{z}\|^2)\,\mathrm{d}\mathbf{z},
\end{aligned}$$

where the last equality comes from the change of variable $\mathbf{y}-\mathbf{x} = \mathbf{z}$. For the symmetry of $\mathrm{B}_r(\mathbf{0})$, all the odd functions give no contribution to the final result

$$\int_{\mathrm{B}_r(\mathbf{x})} (u(\mathbf{y}) - u(\mathbf{x}))\,\mathrm{d}\mathbf{y} = \frac{1}{2}\sum_{i=1}^{n} \frac{\partial^2 u}{\partial x_i^2}(\mathbf{x}) \int_{\mathrm{B}_r(\mathbf{0})} z_i^2\,\mathrm{d}\mathbf{z} + \int_{\mathrm{B}_r(\mathbf{0})} o(\|\mathbf{z}\|^2)\,\mathrm{d}\mathbf{z},$$

while the even terms give all the same contribution equal to

$$\int_{\mathrm{B}_r(\mathbf{0})} z_i^2\,\mathrm{d}\mathbf{z} = \frac{1}{n}\sum_{i=1}^{n} \int_{\mathrm{B}_r(\mathbf{0})} z_i^2\,\mathrm{d}\mathbf{z}.$$

The last equality introduces a radially symmetric function $\sum_i z_i^2 = \rho^2$ that is constant on each spherical shell. Denote $\mathrm{S}_r(\mathbf{0})$ the sphere of radius r in $\mathbb{R}^{\mathrm{n}}$ centered

at $\mathbf{0}$

$$\begin{aligned}\int_{\mathrm{B}_r(\mathbf{x})} (u(\mathbf{y}) - u(\mathbf{x}))\,\mathrm{d}\mathbf{y} =& \sum_{i=1}^{n} \frac{\partial^2 u}{\partial x_i^2}(\mathbf{x}) \frac{1}{2\,n} \int_0^r \rho^2 \,\mathrm{area}(\mathrm{S}_\rho(\mathbf{0}))\,\mathrm{d}\rho \\ &+ \int_0^r o(\rho^2)\,\mathrm{area}(\mathrm{S}_\rho(\mathbf{0}))\,\mathrm{d}\rho \\ =& \sum_{i=1}^{n} \frac{\partial^2 u}{\partial x_i^2}(\mathbf{x}) \frac{\mathrm{area}(\mathrm{S}_1(\mathbf{0}))}{2\,n} \int_0^r \rho^{n+1}\,\mathrm{d}\rho \\ &+ \mathrm{area}(\mathrm{S}_1(\mathbf{0})) \int_0^r o(\rho^2)\rho^{n-1}\,\mathrm{d}\rho \\ =& \sum_{i=1}^{n} \frac{\partial^2 u}{\partial x_i^2}(\mathbf{x}) \frac{\mathrm{area}(\mathrm{S}_1(\mathbf{0}))}{2\,n} \frac{r^{n+2}}{n+2} \\ &+ \mathrm{area}(\mathrm{S}_1(\mathbf{0}))\, o(\rho^{n+2})\,\mathrm{d}\rho\,,\end{aligned}$$

where

$$\mathrm{area}(\mathrm{S}_\rho(\mathbf{0})) = \mathrm{area}(\mathrm{S}_1(\mathbf{0}))\,\rho^{n-1}\,.$$

Recalling that the volume of the n-dimensional unit ball is linked to the area of its shell by the following formula

$$\mathrm{vol}(\mathrm{B}_1(\mathbf{0})) = \frac{\mathrm{area}(\mathrm{S}_1(\mathbf{0}))}{n}\,,$$

it holds

$$\int_{\mathrm{B}_r(\mathbf{x})} (u(\mathbf{y}) - u(\mathbf{x}))\,\mathrm{d}\mathbf{y} = \Delta u(\mathbf{x}) \frac{\mathrm{vol}(\mathrm{B}_1(\mathbf{0}))}{2} \frac{r^{n+2}}{n+2} + o(r^{n+2})\,.$$

Finally, isolating the Laplacian and taking the limit

$$\Delta u(\mathbf{x}) = \lim_{r \to 0^+} \frac{2(n+2)}{\mathrm{vol}(\mathrm{B}_1(\mathbf{0}))\, r^{n+2}} \int_{\mathrm{B}_r(\mathbf{x})} (u(\mathbf{y}) - u(\mathbf{x}))\,\mathrm{d}\mathbf{y}\,,$$

the thesis follows. □

3.2 Metric Measure Laplacian

The Integral Form of Differential Laplacian necessitates a way to identify balls and to compute their volumes. A possible generalization is to define the Laplacian

operator in a topological space $\mathcal{V}$ equipped with a metric d, useful to identify balls, and a measure μ, useful to compute the volume of balls.

Such approach has already been followed in [29]. On a generic metric measure space, the authors define a Laplacian operator, called ρ-Laplacian, as

$$\mathcal{L}_\rho u(x) = \frac{1}{\mu\left(\mathrm{B}_\rho(x)\right)\,\rho^2} \int_{\mathcal{V}} \mathbb{1}_{\mathrm{B}_\rho(x)}(y)(u(y)-u(x))\,\mathrm{d}\mu(y)\,,$$

for a fixed small $\rho > 0$. While on a Riemannian manifold the ρ-Laplacian tends to the Laplace-Beltrami operator as ρ goes to zero, in a generic metric measure space, for instance a discrete space, the concept of limit could be not well defined; this is why the limit does not appear in the definition of ρ-Laplacian. The same applies to the normalization factor $2(n+2)$, that could be hard to generalize.

Following a similar line of reasoning, a possible generalization of the Differential Laplacian is given below.

Definition 3.2 (K-Laplacian)**.** Let $(\mathcal{V}, d, \mu)$ be a compact metric measure space of finite measure $\mu(\mathcal{V}) < \infty$; let $K \in \mathrm{L}^2(\mathcal{V}\times\mathcal{V})$ be a square-integrable kernel such that

$$\operatorname*{ess\,sup}_{x\in\mathcal{V}} \int_{\mathcal{V}} K(x,y)\,\mathrm{d}\mu(y) < \infty\,.$$

The *K-Laplacian operator* $\mathcal{L}_K : \mathrm{L}^2(\mathcal{V}) \to \mathrm{L}^2(\mathcal{V})$ is defined as

$$\mathcal{L}_K u(x) = \int_{\mathcal{V}} K(x,y)\,(u(y)-u(x))\,\mathrm{d}\mu(y)\,.$$

The K-Laplacian can be rewritten as

$$\mathcal{L}_K u(x) = \int_{\mathcal{V}} K(x,y)\,u(y)\;\mathrm{d}\mu(y) - \int_{\mathcal{V}} K(x,y)\;\mathrm{d}\mu(y) u(x)\,,$$

where the first addend is an Hilbert-Schmidt Operator that quantifies the contribution of the neighborhood of x to the value of the Laplacian, and the second addend is a Multiplication Operator that quantifies the contribution of x to the value of the Laplacian. A more general definition decouples the two contributions.

Definition 3.3 (MK-Laplacian)**.** Let $(\mathcal{V}, d, \mu)$ be a compact metric measure space of finite measure, i.e. $\mu(\mathcal{V}) < \infty$; let $K \in \mathrm{L}^2(\mathcal{V}\times\mathcal{V})$ be a square-integrable kernel; let $M \in \mathrm{L}^\infty(\mathcal{V})$. The *MK-Laplacian operator* $\mathcal{L}_{K,M} : \mathrm{L}^2(\mathcal{V}) \to \mathrm{L}^2(\mathcal{V})$ is defined as

$$\mathcal{L}_{K,M} u(x) = \int_{\mathcal{V}} K(x,y)\,u(y)\,\mathrm{d}\mu(y) - M(x)\,u(x)\,. \tag{3.1}$$

To see that MK-Laplacian is more general than K-Laplacian, it can be noted that the K-Laplacian can be retrieved setting

$$M(x) = \int_{\mathcal{V}} K(x,y)\,\mathrm{d}\mu(y)\,.$$

While the kernel in the Integral Form of Differential Laplacian is a normalized indicator of balls, Definitions 3.2 to 3.3 give the freedom to choose it. Some possibilities are the following:

$$K^{\mathrm{c}}(x,y) \coloneqq \mathbb{1}_{\mathrm{B}_r(x)}(y)\,, \qquad \text{(combinatorial kernel)}$$

$$K^{\mathrm{rw}}(x,y) \coloneqq \frac{\mathbb{1}_{\mathrm{B}_r(x)}(y)}{\mu(\mathrm{B}_r(x))}\,, \qquad \text{(random walk kernel)}$$

$$K^{\mathrm{sn}}(x,y) \coloneqq \frac{\mathbb{1}_{\mathrm{B}_r(x)}(y)}{\sqrt{\mu(\mathrm{B}_r(x))}\sqrt{\mu(\mathrm{B}_r(y))}}\,. \qquad \text{(symmetric normalized kernel)}$$

3.3 Random Sampled Laplacian

As pointed out in [18], the advantage of K- and MK-Laplacians is that they are readily discretizable: the integral can be approximated by a finite sum over sample sets. If the sample points are chosen at random, the finite sum approximation is a Monte-Carlo approximation.

Definition 3.4 (Random Sampled Laplacian)**.** Let $(\mathcal{V}, d, \mu)$ be a continuous metric-measure space of finite measure $\mu(\mathcal{V}) < \infty$. Let $\rho : \mathcal{V} \to (0, +\infty)$ be a continuous, positive function, bounded away from 0 and ∞, and satisfying

$$\int_{\mathcal{V}} \rho(y)\,\mathrm{d}\mu(y) = 1\,.$$

Let $\mathbf{x} = \{x_i\}_{i=1}^N$ be a i.i.d. random sample from ρ. The *random sampled MK-Laplacian* $\mathcal{L}_{K,M,\rho,\mathbf{x}}$ is defined as

$$\mathcal{L}_{K,M,\rho,\mathbf{x}} u(x_i) \coloneqq \frac{1}{N}\sum_{j=1}^{N} \frac{K(x_i,x_j)}{\rho(x_j)} u(x_j) - M(x_i)\,u(x_i)\,.$$

In a similar way the *random sampled K-Laplacian* $\mathcal{L}_{K,\rho,\mathbf{x}}$ is defined as

$$\mathcal{L}_{K,\rho,\mathbf{x}} u(x_i) \coloneqq \frac{1}{N}\sum_{j=1}^{N} \frac{K(x_i,x_j)}{\rho(x_j)} (u(x_j) - u(x_i))\,.$$

Theorem 3.2 (Convergence of Random Sampled MK-Laplacian)**.** *The random sampled MK-Laplacian converges in probability, in* L^2 *and almost surely to the MK-Laplacian, provided that*

$$\operatorname*{ess\,sup}_{y\in\mathcal{V}} \frac{K(x,y)^2}{\rho(y)} < \infty\,,\ \forall x \in \mathcal{V}\,.$$

Moreover, the convergence rate is $\mathcal{O}\left(N^{-\frac{1}{2}}\right)$.

Proof. Let $\mathbf{x} = \{x_i\}_{i=1}^N$ be an i.i.d. random sample from ρ, where ρ is the sampling density. The finite sum approximation $\mathcal{L}_{K,M,\rho,\mathbf{x}}u(x_i)$ of $\mathcal{L}_{K,M}u(x_i)$ is itself a random variable that depends on the random sample $\mathbf{x}$. Conditioned on $x_i = x$, the expected value is

$$\mathbb{E}\left[\mathcal{L}_{K,M,\rho,\mathbf{x}}u(x)\right] = \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left[\frac{1}{\rho(x_j)}K(x,x_j)\,u(x_j)\right] - M(x)u(x) = \mathcal{L}_{K,M}u(x)\,.$$

Since the random variables $\{x_j\}_{j=1}^N$ are i.i.d. to y, then also the random variables

$$\left\{\frac{K(x,x_j)}{\rho(x_j)}\,u(x_j)\right\}_{j=1}^{N}$$

are i.i.d., hence,

$$\begin{aligned}
\operatorname{var}\left[\mathcal{L}_{K,M,\rho,\mathbf{x}}u(x)\right] &= \operatorname{var}\left[\frac{1}{N}\sum_{j=1}^{N}\frac{K(x,x_j)}{\rho(x_j)}u(x_j) - M(x)\,u(x)\right] \\
&= \frac{1}{N}\operatorname{var}\left[\frac{K(x,y)}{\rho(y)}u(y)\right] \\
&\le \frac{1}{N}\mathbb{E}\left[\left|\frac{K(x,y)}{\rho(y)}u(y)\right|^2\right] \\
&= \frac{1}{N}\int_{\mathcal{V}}\left|\frac{K(x,y)}{\rho(y)}u(y)\right|^2\rho(y)\,\mathrm{d}\mu(y) \\
&\le \frac{1}{N}\left\|\frac{K(x,\cdot)^2}{\rho(\cdot)}\right\|_{\mathrm{L}^\infty(\mathcal{V})}\|u\|^2_{\mathrm{L}^2(\mathcal{V})}\,,
\end{aligned}$$

therefore the series converges in $\mathrm{L}^2(\mathcal{V})$ as $N\to\infty$, thus in probability . Finally, the series converges almost surely by the "Two-Series Theorem" [30].

Using the Bienaymé-Chebyshev Inequality the rate of convergence can be quantified. In probability $(1-\epsilon)$ it holds

$$\left|\mathcal{L}_{K,M,\rho,\mathbf{x}}u(x) - \mathcal{L}_{K,M}u(x)\right| \le \frac{1}{\sqrt{N\,\epsilon}}\sqrt{\left\|\frac{K(x,\cdot)^2}{\rho(\cdot)}\right\|_{\mathrm{L}^\infty(\mathcal{V})}}\|u\|_{\mathrm{L}^2(\mathcal{V})}\,,$$

i.e. the deviation goes to 0 as $\mathcal{O}\left(N^{-\frac{1}{2}}\right)$. □

The same result holds for a K-Laplacian with a different constant. Using Young's inequality for products, in probability $(1-\epsilon)$

$$\left|\mathcal{L}_{K,\rho,\mathbf{x}}u(x) - \mathcal{L}_K u(x)\right| \leq \sqrt{\frac{2}{N\,\epsilon}\left\|\frac{K(x,\cdot)^2}{\rho(\cdot)}\right\|_{\mathrm{L}^\infty(\mathcal{V})}\left(\|u\|^2_{\mathrm{L}^2(\mathcal{V})} + \mu(\mathcal{V})|u(x)|^2\right)}\,.$$

The different bound can be explained by the fact that the multiplication operator in the K-Laplacian must be approximated too via a finite sum.

3.4 Graph Laplacian

The random sample $\mathbf{x} = \{x_i\}_{i=1}^N$ can be interpreted as the discrete approximation of the continuous metric-measure space $\mathcal{V}$; the density value $\rho(x_i)$ can be interpreted as the "relative likelihood" of picking x_i from $\mathcal{V}$; the kernel $K(x_i, x_j)$ quantifies the relationship between the sampled points x_i and x_j. Such relationship between sampled points can be described by a graph object $\mathcal{G}_N = (\mathcal{V}_N, \mathcal{E}_N)$: the set of nodes $\mathcal{V}_N$ is the set of sampled points $\{x_i\}_{i=1}^N$; the set of edges $\mathcal{E}_N$ is the set of all pairs $\{(x_i, x_j)\}_{i,j=1}^N$ for which the integral kernel is not null, i.e. $K(x_i, x_j) \neq 0$. Graphs of this type are known in the Literature as "spatial networks", or "random geometric graphs" [19].

The Monte-Carlo approximation of the continuous metric measure Laplacian gives a graph approximation of the continuous space $\mathcal{V}$; the Random Sampled Laplacian can be interpreted as the associated Graph Shift Operator. One could wonder if this approach can retrieve the usual definition of graph Laplacians. The affirmative answer will be proven in the next sections.

3.4.1 Adjacency Matrix

Let K be the combinatorial kernel; let M be the null function. Consider the associated MK-Laplacian

$$\mathcal{L}_{K^{\mathrm{c}},0}u(x) = \int_{\mathcal{V}} \mathbb{1}_{\mathrm{B}_r(x)}(y)\, u(y)\, \mathrm{d}\mu(y)\,,$$

the corresponding random sampled MK-Laplacian is

$$\mathcal{L}_{K^{\mathrm{c}},0,\rho,\mathbf{x}}u(x_i) = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbb{1}_{\mathrm{B}_r(x_i)}(x_j)}{\rho(x_j)} u(x_j)\, \mathrm{d}\mu(y)\,.$$

The indicator function gives a natural way to define the adjacency matrix

$$[\mathbf{A}]_{ij} \coloneqq \begin{cases} 1 & d(x_i, x_j) < r\,,\ i \neq j \\ 0 & \text{otherwise} \end{cases}\,. \tag{3.2}$$

Such definition does not include self-loops, even though a distance function satisfies the identity of indiscernibles. Usually, in the Literature, self-loops are kept separated from the proper links; hence, the same will be done in this work. Define the weight matrix and the vector signal as

$$\mathbf{P} = \operatorname{diag}\left(\{\rho(x_i)\}_{i=1}^N\right)\,, \tag{3.3}$$

$$\mathbf{u} = \{u(x_i)\}_{i=1}^N\,, \tag{3.4}$$

the random sampled MK-Laplacian can be re-written in matrix notation as

$$\mathcal{L}_{K^{\mathrm{c}},0,\rho,\mathbf{x}}\mathbf{u} = \frac{1}{N}(\mathbf{A}+\mathbf{I})\,\mathbf{P}^{-1}\mathbf{u}\,.$$

3.4.2 Combinatorial Laplacian

Let K be the combinatorial kernel; let $M(x)$ be the measure of the ball centered at x and radius r. Consider the associated MK-Laplacian

$$\mathcal{L}_{K^{\mathrm{c}},M}u(x) = \int_{\mathcal{V}} \mathbb{1}_{\mathrm{B}_r(x)}(y)\,u(y)\,\mathrm{d}\mu(y) - \mu(\mathrm{B}_r(x))\,u(x)\,.$$

It should be noted that this particular choice for K and M give rise to a K-Laplacian because

$$M(x) = \int_{\mathcal{V}} K^{\mathrm{c}}(x,y)\,\mathrm{d}\mu(y) = \int_{\mathcal{V}} \mathbb{1}_{\mathrm{B}_r(x)}(y)\,\mathrm{d}\mu(y) = \mu(\mathrm{B}_r(x))\,,$$

hence, the corresponding random sampled Laplacian can be written as

$$\mathcal{L}_{K^{\mathrm{c}},\rho,\mathbf{x}}u(x_i) = \frac{1}{N}\sum_{j=1}^{N}\frac{\mathbb{1}_{\mathrm{B}_r(x_i)}(x_j)}{\rho(x_j)}(u(x_j)-u(x_i))\,.$$

Define the adjacency matrix, the weight matrix and the vector signal as in Equations (3.2) to (3.4), then, the random sampled Laplacian can be re-written in matrix notation as

$$\begin{aligned}\mathcal{L}_{K^{\mathrm{c}},\rho,\mathbf{x}}\mathbf{u} &= \frac{1}{N}\left((\mathbf{A}+\mathbf{I})\mathbf{P}^{-1} - \operatorname{diag}\left((\mathbf{A}+\mathbf{I})\mathbf{P}^{-1}\mathbf{1}\right)\right)\mathbf{u} \\ &= \frac{1}{N}\left(\mathbf{A}\mathbf{P}^{-1} - \operatorname{diag}\left(\mathbf{A}\mathbf{P}^{-1}\mathbf{1}\right)\right)\mathbf{u}\,,\end{aligned}$$

where the last equality is justified by the fact that $\mathbf{P}$ is a diagonal matrix, hence, $\mathbf{P}^{-1} = \operatorname{diag}(\mathbf{P}^{-1}\mathbf{1})$. Define the ρ-degree matrix to be

$$\mathbf{D}_\rho = \operatorname{diag}(\mathbf{A}\mathbf{P}^{-1}\mathbf{1})\,. \tag{3.5}$$

When the sampling is uniform, the ρ-degree matrix is the actual degree matrix as in Definition 2.21. Therefore, in case of uniform sampling the ρ-degree matrix will be called degree matrix and denoted by $\mathbf{D}$. One could wonder if it is possible to factor $\mathbf{P}^{-1}$ out, i.e. if it holds $\mathbf{D}_\rho = \mathbf{D}\mathbf{P}^{-1}$; in general this does not hold, as shown in the following theorem.

Theorem 3.3 (Commutative Property of diag(·) Operator)**.** *Let* $\mathbf{A}$ *be a real, symmetric matrix with non-negative entries and let* $\mathbf{P}$ *be a real, diagonal matrix with non-negative entries, it holds*

$$\operatorname{diag}(\mathbf{P}\mathbf{A}\mathbf{1}) \underset{(1)}{=} \mathbf{P}\,\operatorname{diag}(\mathbf{A}\mathbf{1}) \underset{(2)}{=} \operatorname{diag}(\mathbf{A}\mathbf{1})\,\mathbf{P}\,.$$

Moreover,

$$\operatorname{diag}(\mathbf{A}\mathbf{P}\mathbf{1}) \underset{(3)}{=} \operatorname{diag}(\mathbf{P}\mathbf{A}\mathbf{1})\,,$$

holds if and only if $\mathbf{A}$ *has the form*

$$\mathbf{A} = \sum_{k=1}^{n} \sum_{\substack{i \geq k+1 \\ p_i = p_k}} [\mathbf{A}]_{ki}(\mathbf{e}_k\mathbf{e}_i^{\mathrm{T}} + \mathbf{e}_i\mathbf{e}_k^{\mathrm{T}}) + \sum_{k=1}^{n} [\mathbf{A}]_{kk}\mathbf{e}_k\mathbf{e}_k^{\mathrm{T}}\,.$$

Proof. Equality (2) is trivial, since diagonal matrices commutes; therefore, we will prove only (1). Consider

$$\begin{aligned}
[\operatorname{diag}(\mathbf{P}\mathbf{A}\mathbf{1})]_{ii} = [\mathbf{P}\mathbf{A}\mathbf{1}]_i &= \sum_{j=1}^{n} [\mathbf{P}]_{ij}[\mathbf{A}\mathbf{1}]_j = \sum_{j=1}^{n}\sum_{k=1}^{n} [\mathbf{P}]_{ij}[\mathbf{A}]_{jk} \\
&= \sum_{k=1}^{n} [\mathbf{P}]_{ii}[\mathbf{A}]_{ik} = [\mathbf{P}]_{ii}\sum_{k=1}^{n}[\mathbf{A}]_{ik} = [\mathbf{P}\,\operatorname{diag}(\mathbf{A}\mathbf{1})]_{ii}\,.
\end{aligned}$$

In order to prove (3), we note that $\mathbf{P}$ can be decomposed as $\mathbf{P} = \sum_{i=1}^{n} [\mathbf{P}]_{ii}\mathbf{e}_i\mathbf{e}_i^{\mathrm{T}}$. Therefore

$$\begin{aligned}
0 = [\operatorname{diag}(\mathbf{A}\mathbf{P}\mathbf{1}) - \operatorname{diag}(\mathbf{P}\mathbf{A}\mathbf{1})]_{kk} &= \operatorname{diag}\left(\sum_{i=1}^{n} [\mathbf{P}]_{ii}(\mathbf{A}\mathbf{e}_i\mathbf{e}_i^{\mathrm{T}}\mathbf{1} - \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}}\mathbf{A}\mathbf{1})\right)_{kk} \\
&= \left[\sum_{i=1}^{n} [\mathbf{P}]_{ii}\left(\mathbf{A}\mathbf{e}_i - \sum_{j=1}^{n}[\mathbf{A}]_{ij}\mathbf{e}_i)\right)\right]_k
\end{aligned}$$

$$= \sum_{i=1}^{n} [\mathbf{P}]_{ii}[\mathbf{A}]_{ki} - [\mathbf{P}]_{kk} \sum_{j=1}^{n} [\mathbf{A}]_{kj}$$
$$= \sum_{i=1}^{n} ([\mathbf{P}]_{ii} - [\mathbf{P}]_{kk})[\mathbf{A}]_{ki}\,,$$

must hold for all values of k. Consider the indices $k_1, k_2, \ldots, k_n$ corresponding to the values $[\mathbf{P}]_{k_1k_1} \leq [\mathbf{P}]_{k_2k_2} \leq \cdots \leq [\mathbf{P}]_{k_nk_n}$, then

$$0 = \sum_{i=1}^{n} \underbrace{([\mathbf{P}]_{ii} - [\mathbf{P}]_{k_1k_1})}_{\geq 0}[\mathbf{A}]_{k_1i}\,,$$

then $[\mathbf{A}]_{k_1i} = 0$ for each i such that $[\mathbf{P}]_{ii} > [\mathbf{P}]_{k_1k_1}$. Take the index k_2 and consider

$$0 = \sum_{i=1}^{n} \underbrace{([\mathbf{P}]_{ii} - [\mathbf{P}]_{k_2k_2})}_{\geq 0}[\mathbf{A}]_{k_2i}$$
$$= \sum_{\substack{i=1 \\ i \neq k_1}}^{n} \underbrace{([\mathbf{P}]_{ii} - [\mathbf{P}]_{k_2k_2})}_{\geq 0}[\mathbf{A}]_{k_2i} + \underbrace{([\mathbf{P}]_{k_1k_1} - [\mathbf{P}]_{k_2k_2})[\mathbf{A}]_{k_2k_1}}_{=0}\,.$$

The second addend is 0 because $[\mathbf{P}]_{k_2k_2}$ can be either equal to $[\mathbf{P}]_{k_1k_1}$, in which case the difference is null, or $[\mathbf{P}]_{k_2k_2} > [\mathbf{P}]_{k_1k_1}$, in which case from the previous step $[\mathbf{A}]_{k_2k_1} = 0$. Therefore $[\mathbf{A}]_{k_1i} = 0$ for each i such that $[\mathbf{P}]_{ii} > [\mathbf{P}]_{k_2k_2}$. By finite induction, the thesis holds when $\mathbf{A}$ has null entries in position (i, j) whenever $[\mathbf{P}]_{ii} \neq [\mathbf{P}]_{jj}$. □

When the sampling is uniform, the random sampled Laplacian is

$$\mathcal{L}_{K^{\mathrm{c}},1,\mathbf{x}}\mathbf{u} = \frac{1}{N}\,(\mathbf{A} - \mathbf{D})\,\mathbf{u}\,,$$

hence, the usual definition of combinatorial Laplacian is retrieved.

3.4.3 Random Walk Laplacian

Let K be the random walk kernel: let M be the unit function. Consider the associated MK-Laplacian

$$\mathcal{L}_{K^{\mathrm{rw}},1}u(x) = \int_{\mathcal{V}} \frac{\mathbb{1}_{\mathrm{B}_r(x)}(y)}{\mu(\mathrm{B}_r(x))}u(y)\,\mathrm{d}\mu(y) - u(x)\,.$$

Also in this case the choice for K and M gives rise to a K-Laplacian because

$$M(x) = \int_{\mathcal{V}} K^{\mathrm{rw}}(x, y)\,\mathrm{d}\mu(y) = \int_{\mathcal{V}} \frac{\mathbb{1}_{\mathrm{B}_r(x)}(y)}{\mu(\mathrm{B}_r(x))}\,\mathrm{d}\mu(y) = 1\,.$$

The corresponding random sampled Laplacian has the functional form

$$\mathcal{L}_{K^{\mathrm{rw}},\rho,\mathbf{x}}u(x_i) = \left(\sum_{k=1}^{N}\frac{\mathbb{1}_{\mathrm{B}_r(x_i)}(x_k)}{\rho(x_k)}\right)^{-1}\left(\sum_{j=1}^{N}\frac{\mathbb{1}_{\mathrm{B}_r(x_i)}(x_j)}{\rho(x_j)}u(x_j)\right) - u(x_i)\,.$$

Define the adjacency matrix, the weight matrix, the signal vector and the ρ-degree matrix as in Equations (3.2) to (3.5), then in matrix notation

$$\begin{aligned}\mathcal{L}_{K^{\mathrm{rw}},\rho,\mathbf{x}}\mathbf{u} &= \left(\mathrm{diag}\left((\mathbf{A}+\mathbf{I})\mathbf{P}^{-1}\mathbf{1}\right)^{-1}(\mathbf{A}+\mathbf{I})\mathbf{P}^{-1} - \mathbf{I}\right)\mathbf{u}\\ &= \left((\mathbf{D}_\rho+\mathbf{P}^{-1})^{-1}(\mathbf{A}+\mathbf{I})\mathbf{P}^{-1} - \mathbf{I}\right)\mathbf{u}\,.\end{aligned}$$

When the sampling is uniform

$$\mathcal{L}_{K^{\mathrm{rw}},1,\mathbf{x}}\mathbf{u} = \left((\mathbf{D}+\mathbf{I})^{-1}(\mathbf{A}+\mathbf{I}) - \mathbf{I}\right)\mathbf{u}\,,$$

the usual definition of random walk Laplacian is retrieved.

Example 3.1 (Equivalence between random-walk and combinatorial Laplacian in uniformly distributed measure spaces)**.** In case of a Uniformly Distributed Measure Space, the combinatorial Laplacian and the random-walk Laplacian differs only for a multiplicative constant

$$\begin{aligned}\mathcal{L}_{K^{\mathrm{rw}}}u(x) &= \int_{\mathcal{V}}\frac{\mathbb{1}_{\mathrm{B}_r(x)}}{\mu(\mathrm{B}_r(x))}(u(y)-u(x))\,\mathrm{d}\mu(y)\\ &= \frac{1}{\mu_r}\int_{\mathcal{V}}\mathbb{1}_{\mathrm{B}_r(x)}(u(y)-u(x))\,\mathrm{d}\mu(y)\\ &= \frac{1}{\mu_r}\mathcal{L}_{K^{\mathrm{c}}}u(x)\,.\end{aligned}$$

3.4.4 Symmetric Normalized Laplacian

Let K be the symmetric normalized kernel; let M be the unit function. Consider the associated MK-Laplacian

$$\mathcal{L}_{K^{\mathrm{sn}},1}u(x) = \int_{\mathcal{V}}\frac{\mathbb{1}_{\mathrm{B}_r(x)}(y)}{\sqrt{\mu(\mathrm{B}_r(x))}\sqrt{\mu(\mathrm{B}_r(y))}}u(y)\,\mathrm{d}\mu(y) - u(x)\,,$$

the corresponding random sampled Laplacian has the functional form

$$\mathcal{L}_{K^{\mathrm{sn}},1,\rho,\mathbf{x}}u(x_i) = \sum_{j=1}^{N}\left(\sum_{k=1}^{N}\frac{\mathbb{1}_{\mathrm{B}_r(x_i)}(x_k)}{\rho(x_k)}\right)^{-\frac{1}{2}}\frac{\mathbb{1}_{\mathrm{B}_r(x_i)}(x_j)}{\rho(x_j)}\left(\sum_{k=1}^{N}\frac{\mathbb{1}_{\mathrm{B}_\rho(x_j)}(x_k)}{\rho(x_k)}\right)^{-\frac{1}{2}}u(x_j)$$

$$- u(x_i)\,.$$

Define the adjacency matrix, the weight matrix, the signal vector and the ρ-degree matrix as in Equations (3.2) to (3.5), then in matrix notation

$$\mathcal{L}_{K^{\mathrm{sn}},1,\rho,\mathbf{x}}\mathbf{u} = \left((\mathbf{D}_\rho + \mathbf{P}^{-1})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})\mathbf{P}^{-1}(\mathbf{D}_\rho + \mathbf{P}^{-1})^{-\frac{1}{2}} - \mathbf{I}\right)\mathbf{u}\,.$$

When the sampling is uniform

$$\mathcal{L}_{K^{\mathrm{sn}},1,1,\mathbf{x}}\mathbf{u} = \left((\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})(\mathbf{D}+\mathbf{I})^{-\frac{1}{2}} - \mathbf{I}\right)\mathbf{u}\,.$$

the usual definition of symmetric normalized Laplacian is retrieved.

Example 3.2 (Equivalence between random-walk and symmetric-normalized Laplacian for ultrametric spaces)**.** If $(\mathcal{V}, d)$ is an Ultrametric, then, as stated in Theorem 2.1, $y \in \mathrm{B}_r(x)$ implies $\mathrm{B}_r(x) = \mathrm{B}_r(y)$, hence, $\sqrt{\mu(\mathrm{B}_r(x))} = \sqrt{\mu(\mathrm{B}_r(y))}$ and

$$\begin{aligned}
\mathcal{L}_{K^{\mathrm{sn}},1}u(x) &= \int_{\mathcal{V}} \frac{\mathbb{1}_{\mathrm{B}_r(x)}(y)}{\sqrt{\mu(\mathrm{B}_r(x))}\sqrt{\mu(\mathrm{B}_r(y))}}u(y)\,\mathrm{d}\mu(y) - u(x) \\
&= \int_{\mathcal{V}} \frac{\mathbb{1}_{\mathrm{B}_r(x)}(y)}{\mu(\mathrm{B}_r(x))}u(y)\,\mathrm{d}\mu(y) - u(x) \\
&= \mathcal{L}_{K^{\mathrm{rw}}}u(x)\,.
\end{aligned}$$

The graph approximation of an ultrametric space is composed of single nodes or complete connected components because the characteristic functions of $\mathrm{B}_r(y)$ and $\mathrm{B}_r(x)$ are equal, hence, x and y are connected to the same nodes in the graph.

Example 3.3 (Equivalence between combinatorial and symmetric-normalized Laplacian for uniformly distributed measure spaces)**.** If the continuous metric measure space is a Uniformly Distributed Measure Space, the symmetric-normalized Laplacian is equivalent to both the random-walk and the combinatorial Laplacians

$$\begin{aligned}
\mathcal{L}_{K^{\mathrm{sn}},1}u(x) &= \int_{\mathcal{V}} \frac{\mathbb{1}_{\mathrm{B}_r(x)}(y)}{\sqrt{\mu(\mathrm{B}_r(x))}\sqrt{\mu(\mathrm{B}_r(y))}}u(y)\,\mathrm{d}\mu(y) - u(x) \\
&= \int_{\mathcal{V}} \frac{\mathbb{1}_{\mathrm{B}_r(x)}(y)}{\mu(\mathrm{B}_r(x)}u(y)\,\mathrm{d}\mu(y) - u(x) \\
&= \mathcal{L}_{K^{\mathrm{rw}}} \\
&= \frac{1}{\mu_r}\mathcal{L}_{K^{\mathrm{c}}}\,.
\end{aligned}$$

3.4.5 Symmetric Normalized K-Laplacian

Let K be the symmetric normalized kernel. Consider the associated K-Laplacian

$$\mathcal{L}_{K^{\text{sn}}} u(x) = \int_{\mathcal{V}} \frac{\mathbb{1}_{\mathrm{B}_r(x)}(y)}{\sqrt{\mu(\mathrm{B}_r(x))}\sqrt{\mu(\mathrm{B}_r(y))}}(u(y) - u(x))\,\mathrm{d}\mu(y)\,.$$

Define the adjacency matrix, the density matrix, the signal vector and the ρ-degree matrix as in Equations (3.2) to (3.5), then the corresponding random sampled Laplacian can be written as

$$\begin{aligned}\mathcal{L}_{K^{\text{sn}},\rho,\mathbf{x}}\mathbf{u} =& (\mathbf{D}_\rho + \mathbf{P}^{-1})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})\mathbf{P}^{-1}(\mathbf{D}_\rho + \mathbf{P}^{-1})^{-\frac{1}{2}}\mathbf{u} \\ &- \operatorname{diag}\left((\mathbf{D}_\rho + \mathbf{P}^{-1})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})\mathbf{P}^{-1}(\mathbf{D}_\rho + \mathbf{P}^{-1})^{-\frac{1}{2}}\mathbf{1}\right)\mathbf{u}\,.\end{aligned}$$

Theorem 3.3 alone would be sufficient to prove that the symmetric normalized K-Laplacian and the symmetric normalized Laplacian do not coincide because $(\mathbf{D}_\rho + \mathbf{P}^{-1})^{-\frac{1}{2}}$ cannot be factored out the $\operatorname{diag}(\cdot)$ operator. For instance, as a consequence of Theorem 3.3 when the sampling is uniform, $(\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}$ can be factored out just when a node is linked to nodes with the same degree, hence, when the graph is k-regular.

In [31] is proved that the spectral radius of the symmetric normalized Laplacian is 2; this is in general not true for the symmetric normalized K-Laplacian as shown in the following theorem for the uniform sampling case.

Theorem 3.4 (Bound on the Eigenvalues of symmetric normalized K-Laplacian)**.** *Let λ be en eigenvalue of $\mathcal{L}_{K^{sn},1,\mathbf{x}}$, then $|\lambda| \leq \sqrt{2\,N}$.*

Proof. In order to keep the notation light, in the following $\mathbf{D} \leftarrow \mathbf{D} + \mathbf{I}$ and $\mathbf{A} \leftarrow \mathbf{A} + \mathbf{I}$. The eigenvalues can be characterized via the Rayleigh quotient

$$\frac{\left\langle \mathbf{u}, \left(\operatorname{diag}\left(\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\mathbf{1}\right) - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\right)\mathbf{u}\right\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle}\,.$$

Using Theorem 3.3, and considering $\mathbf{u} = \mathbf{D}^{\frac{1}{2}}\mathbf{v}$ the previous formula can be rewritten as

$$\frac{\left\langle \mathbf{D}^{\frac{1}{2}}\mathbf{v}, \left(\operatorname{diag}\left(\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\mathbf{1}\right) - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\right)\mathbf{v}\right\rangle}{\left\langle \mathbf{D}^{\frac{1}{2}}\mathbf{v}, \mathbf{D}^{\frac{1}{2}}\mathbf{v}\right\rangle} = \frac{\mathbf{v}^{\mathrm{T}}\left(\operatorname{diag}\left(\mathbf{D}^{\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\mathbf{1}\right) - \mathbf{A}\right)\mathbf{v}}{\mathbf{v}^{\mathrm{T}}\mathbf{D}\mathbf{v}}\,.$$

Using the symmetry of $\mathbf{A}$, the numerator can be rewritten as

$$\sum_{i,j}[\mathbf{v}]_i^2\sqrt{\frac{[\mathbf{D}]_{ii}}{[\mathbf{D}]_{jj}}}[\mathbf{A}]_{ij} - \sum_{i,j}[\mathbf{v}]_i[\mathbf{A}]_{ij}[\mathbf{v}]_j$$

$$
\begin{aligned}
&= \frac{1}{2}\sum_{i,j}[\mathbf{v}]_i^2\sqrt{\frac{[\mathbf{D}]_{ii}}{[\mathbf{D}]_{jj}}}[\mathbf{A}]_{ij} + \frac{1}{2}\sum_{i,j}[\mathbf{v}]_j^2\sqrt{\frac{[\mathbf{D}]_{jj}}{[\mathbf{D}]_{ii}}}[\mathbf{A}]_{ij} - \sum_{i,j}[\mathbf{v}]_i[\mathbf{A}]_{ij}[\mathbf{v}]_j \\
&= \frac{1}{2}\left(\sum_{i,j}[\mathbf{v}]_i[\mathbf{A}]_{ij}\left(\sqrt{\frac{[\mathbf{D}]_{ii}}{[\mathbf{D}]_{jj}}}[\mathbf{v}]_i - [\mathbf{v}]_j\right) - \sum_{i,j}[\mathbf{v}]_j[\mathbf{A}]_{ij}\left([\mathbf{v}]_i - \sqrt{\frac{[\mathbf{D}]_{jj}}{[\mathbf{D}]_{ii}}}[\mathbf{v}]_j\right)\right) \\
&= \frac{1}{2}\sum_{i,j}\left(\frac{[\mathbf{v}]_i}{\sqrt{[\mathbf{D}]_{jj}}} - \frac{[\mathbf{v}]_j}{\sqrt{[\mathbf{D}]_{ii}}}\right)[\mathbf{A}]_{ij}\left(\sqrt{[\mathbf{D}]_{ii}}[\mathbf{v}]_i - \sqrt{[\mathbf{D}]_{jj}}[\mathbf{v}]_j\right) \\
&= \frac{1}{2}\sum_{i,j}\frac{[\mathbf{A}]_{ij}}{\sqrt{[\mathbf{D}]_{ii}[\mathbf{D}]_{jj}}}\left(\sqrt{[\mathbf{D}]_{ii}}[\mathbf{v}]_i - \sqrt{[\mathbf{D}]_{jj}}[\mathbf{v}]_j\right)^2 \\
&\leq \sum_{i,j}\frac{[\mathbf{A}]_{ij}}{\sqrt{[\mathbf{D}]_{ii}[\mathbf{D}]_{jj}}}\left([\mathbf{D}]_{ii}[\mathbf{v}]_i^2 + [\mathbf{D}]_{jj}[\mathbf{v}]_j^2\right) \\
&= 2\sum_{i,j}[\mathbf{A}]_{ij}\sqrt{\frac{[\mathbf{D}]_{ii}}{[\mathbf{D}]_{jj}}}[\mathbf{v}]_i^2 \\
&\leq \sqrt{2\,N}\sum_i[\mathbf{D}]_{ii}[\mathbf{v}]_i^2 \\
&= \sqrt{2\,N}\,\mathbf{v}^{\mathrm{T}}\mathbf{D}\mathbf{v}\,,
\end{aligned}
$$

where the last inequality is justified by the fact that a node is either isolated with degree 1 (self-loop) or not isolated with degree ≥ 2. □

To show the differences between the two definitions, two examples are given.

Example 3.4 (Line Graph)**.** Consider a line graph with $N = 3$ nodes. The adjacency matrix with self-loops $\mathbf{A} + \mathbf{I}$ is a tridiagonal matrix

$$
\mathbf{A} + \mathbf{I} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \; \mathbf{D} + \mathbf{I} = \operatorname{diag}\begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix},
$$

hence

$$
\begin{aligned}
(\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})(\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}\mathbf{1} &= \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{3}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{pmatrix}\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{3}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{2} & \frac{1}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{3} & \frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{6}} & \frac{1}{2} \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}
\end{aligned}
$$

$$= \begin{pmatrix} \frac{1}{2} + \frac{1}{\sqrt{6}} \\ \frac{1}{3} + \frac{1}{\sqrt{6}} \\ \frac{1}{2} + \frac{1}{\sqrt{6}} \end{pmatrix},$$

and the difference between the novel and the usual symmetric normalized Laplacians is

$$\mathbf{I} - \operatorname{diag}\left((\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})(\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}\mathbf{1}\right) = \operatorname{diag}\begin{pmatrix} \frac{1}{2} - \frac{1}{\sqrt{6}} \\ \frac{2}{3} - \frac{1}{\sqrt{6}} \\ \frac{1}{2} - \frac{1}{\sqrt{6}} \end{pmatrix}.$$

For $N \geq 4$ nodes, it can be shown that

$$\mathbf{I} - \operatorname{diag}\left((\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})(\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}\mathbf{1}\right) = \operatorname{diag}\begin{pmatrix} \frac{1}{2} - \frac{1}{\sqrt{6}} \\ \frac{1}{3} - \frac{1}{\sqrt{6}} \\ \mathbf{0}_{N-4} \\ \frac{1}{3} - \frac{1}{\sqrt{6}} \\ \frac{1}{2} - \frac{1}{\sqrt{6}} \end{pmatrix},$$

where $\mathbf{0}_{N-4}$ is the $(N-4)$-dimensional null vector, and $N-4$ is exactly the number of nodes whose neighbours have the same degree.

The previous example analyses a graph where the degree of a node is either 1 or 2, hence independent of the total number of nodes N. This is the reason why the spectral radius of the difference between the two differently defined Laplacians does not depend on N. In the following example a star graph is considered. In such a graph, the degree of a node is either 1 or $N-1$; as a consequence, the difference between the two Laplacians is unbounded. The example also shows that the bound in Theorem 3.4 is asymptotically tight.

Example 3.5 (Star graph)**.** Consider $N \geq 4$ (for $N \in \{2,3\}$ the star graph is a line graph), the adjacency matrix can be partitioned as

$$\mathbf{A}+\mathbf{I} = \begin{pmatrix} 1 & \mathbf{1}^{\mathrm{T}} \\ \mathbf{1} & \mathbf{I} \end{pmatrix}, \ \mathbf{D}+\mathbf{I} = \begin{pmatrix} N & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & 2\,\mathbf{I} \end{pmatrix},$$

where $\mathbf{0}$, $\mathbf{1}$ are the null and unit vector, and $\mathbf{I}$ is the identity matrix. It holds

$$(\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})(\mathbf{D}+\mathbf{I})^{-\frac{1}{2}} = \begin{pmatrix} \frac{1}{\sqrt{N}} & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & \frac{1}{\sqrt{2}}\mathbf{I} \end{pmatrix} \begin{pmatrix} 1 & \mathbf{1}^{\mathrm{T}} \\ \mathbf{1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{N}} & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & \frac{1}{\sqrt{2}}\mathbf{I} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{N} & \frac{1}{\sqrt{2\,N}}\mathbf{1}^{\mathrm{T}} \\ \frac{1}{\sqrt{2\,N}}\mathbf{1} & \frac{1}{2}\mathbf{I} \end{pmatrix}.$$

The difference between the novel and the usual symmetric normalized Laplacian is

$$\mathbf{I} - \operatorname{diag}\left((\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})(\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}\mathbf{1}\right) = \operatorname{diag}\begin{pmatrix} 1 - \frac{1}{N} - \frac{N-1}{\sqrt{2\,N}} \\ (1 - \frac{1}{2} - \frac{1}{\sqrt{2\,N}})\mathbf{1} \end{pmatrix}.$$

Moreover

$$(\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})(\mathbf{D}+\mathbf{I})^{-\frac{1}{2}} - \operatorname{diag}\left((\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})(\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}\mathbf{1}\right)$$
$$= \begin{pmatrix} -\frac{N-1}{\sqrt{2\,N}} & \frac{1}{\sqrt{2\,N}}\mathbf{1}^{\mathrm{T}} \\ \frac{1}{\sqrt{2\,N}}\mathbf{1} & -\frac{1}{\sqrt{2\,N}}\mathbf{I} \end{pmatrix},$$

and it is easy to see that

$$\begin{pmatrix} -\frac{N-1}{\sqrt{2\,N}} \\ \frac{1}{\sqrt{2\,N}}\mathbf{1} \end{pmatrix},$$

is an eigenvector with corresponding eigenvalue $-\sqrt{N/2}$.

3.5 Some Facts on the MK- and K-Laplacian

The theory developed so far allows to consider under a unified approach several different graph Laplacians that can be encountered in the Literature. The benefit of doing so is the ability to understand what happens in the continuous metric measure space when a graph Laplacian is chosen. Indeed, some operations are better acknowledged when they are seen in the underlying space. For instance, consider a linear combination of MK-Laplacians, then

$$\alpha\mathcal{L}_{K_1,M_1} + \beta\mathcal{L}_{K_2,M_2} = \mathcal{L}_{\alpha K_1+\beta K_2,\alpha M_1+\beta M_2}\,, \tag{3.6}$$

for $\alpha, \beta \in \mathbb{R}$ and for all K_1, K_2, M_1, M_2 satisfying the hypothesis of Definition 3.3. The previous equation states that the MK-Laplacian's space can be equipped with a vector space structure. It is natural, therefore, to give the following definitions.

Definition 3.5 (Space of MK-Laplacians)**.** The space of all MK-Laplacians

$$\mathcal{L}_{\cdot,\cdot} = \{\mathcal{L}_{K,M} \,:\, K, M \text{ satisfying Definition 3.3}\}\,,$$

is a vector space.

As a consequence, the same structure is preserved by Random Sampled Laplacians if the sample set is fixed

$$\alpha\mathcal{L}_{K_1,M_1,\rho,\mathbf{x}} + \beta\mathcal{L}_{K_2,M_2,\rho,\mathbf{x}} = \mathcal{L}_{\alpha K_1+\beta K_2,\alpha M_1+\beta M_2,\rho,\mathbf{x}}\,,$$

hence, the following definition is well-posed.

Definition 3.6 (Space of Random Sampled Laplacians)**.** The space of all Random Sampled Laplacian

$$\mathcal{L}_{\cdot,\cdot,\rho,\mathbf{x}} = \{\mathcal{L}_{K,M,\rho,\mathbf{x}} \,:\, \mathcal{L}_{K,M} \in \mathcal{L}_{\cdot,\cdot}\}\,,$$

is a vector space.

One could wonder if there are proper vector subspace of $\mathcal{L}_{\cdot,\cdot}$. It is easy to see that if M is either the null function or equal to $\int_{\mathcal{V}} K(x,y)\,\mathrm{d}\mu(y)$, this is indeed the case. The same vector structure is held by the random sampled versions if seen as subsets of $\mathcal{L}_{\cdot,\cdot,\rho,\mathbf{x}}$.

Example 3.6 (Learning the Laplacian)**.** Let K, M be

$$\begin{aligned} K(x,y) &= m_2\,(\mu(\mathrm{B}_r(x)) + \alpha)^{e_2}\,\mathbb{1}_{\mathrm{B}_r(x)}(y)\,(\mu(\mathrm{B}_r(y)) + \alpha)^{e_3}\,,\\ M(x) &= m_1\,(\mu(\mathrm{B}_r(x)) + \alpha)^{e_1} + m_2\,\alpha\,(\mu(\mathrm{B}_r(x)) + \alpha)^{e_2+e_3} + m_3\,, \end{aligned}$$

then, if the sampling is uniform, in matrix notation

$$\begin{aligned} \mathcal{L}_{K,M,1,\mathbf{x}} =& m_2\,\left(\frac{1}{N}(\mathbf{D}+\mathbf{I}) + \alpha\mathbf{I}\right)^{e_2}\,\frac{1}{N}(\mathbf{A}+\mathbf{I})\,\left(\frac{1}{N}(\mathbf{D}+\mathbf{I}) + \alpha\mathbf{I}\right)^{e_3}\\ &+ m_1\,\left(\frac{1}{N}(\mathbf{D}+\mathbf{I}) + \alpha\mathbf{I}\right)^{e_1} + m_2\,\alpha\,\left(\frac{1}{N}(\mathbf{D}+\mathbf{I}) + \alpha\mathbf{I}\right)^{e_2+e_3} + m_3\mathbf{I}\,. \end{aligned}$$

Denote by

$$\tilde{m}_1 = \frac{m_1}{N^{e_1}}\,,\ \tilde{m}_2 = \frac{m_2}{N^{e_2+e_3+1}}\,,\ \tilde{\alpha} = \alpha\,N + 1\,,$$

then

$$\mathcal{L}_{K,M,1,\mathbf{x}} = \tilde{m}_1\,(\mathbf{D} + \tilde{\alpha}\mathbf{I})^{e_1} + \tilde{m}_2\,(\mathbf{D} + \tilde{\alpha}\mathbf{I})^{e_2}\,(\mathbf{A} + \tilde{\alpha}\mathbf{I})\,(\mathbf{D} + \tilde{\alpha}\mathbf{I})^{e_3} + m_3\mathbf{I}$$

The previous formula is the parametrized family of Laplacians introduced in [22]; therefore, the theory developed in this chapter generalizes well to already known facts.

The next result characterize $\mathcal{L}_{\cdot,\rho,\mathbf{x}}$ as a subset of 0-sum matrices, i.e. the space of matrices whose rows sum up to 0.

Theorem 3.5 (Kernel of Random Sampled K-Laplacian)**.** *Each vector* $\mathbf{u}$ *that is constant in each connected component of* $\mathcal{G}_N$ *is an eigenvector of* $\mathcal{L}_{K,\rho,\mathbf{x}}$ *with corresponding eigenvalue* 0*.*

Proof. If the graph $\mathcal{G}_N$ has m connected components, after suitable relabeling of the nodes, the kernel matrix $\mathbf{K}$ can be partitioned in a diagonal block matrix

$$\mathbf{K} = \operatorname{diag}\begin{pmatrix} \mathbf{K}_1 \\ \mathbf{K}_2 \\ \vdots \\ \mathbf{K}_m \end{pmatrix},$$

where each $\mathbf{K}_i$ is a square matrix of dimension d_i. The inverse weight matrix $\mathbf{P}^{-1}$ can be partitioned accordingly, as well as the Random Sampled Laplacian

$$\mathcal{L}_{K,\rho,\mathbf{x}} = \frac{1}{N} \operatorname{diag}\begin{pmatrix} \mathbf{K}_1\mathbf{P}_1^{-1} - \operatorname{diag}\left(\mathbf{K}_1\mathbf{P}_1^{-1}\mathbf{1}_{\mathbf{d_1}}\right) \\ \mathbf{K}_2\mathbf{P}_2^{-1} - \operatorname{diag}\left(\mathbf{K}_2\mathbf{P}_2^{-1}\mathbf{1}_{\mathbf{d_2}}\right) \\ \vdots \\ \mathbf{K}_m\mathbf{P}_m^{-1} - \operatorname{diag}\left(\mathbf{K}_m\mathbf{P}_m^{-1}\mathbf{1}_{\mathbf{d_m}}\right) \end{pmatrix}.$$

The vectors

$$\begin{pmatrix} \mathbf{1}_{d_1} \\ \mathbf{0}_{d_2} \\ \vdots \\ \mathbf{0}_{d_m} \end{pmatrix} \quad \begin{pmatrix} \mathbf{0}_{d_1} \\ \mathbf{1}_{d_2} \\ \vdots \\ \mathbf{0}_{d_m} \end{pmatrix} \quad \dots \quad \begin{pmatrix} \mathbf{0}_{d_1} \\ \mathbf{0}_{d_2} \\ \vdots \\ \mathbf{1}_{d_m} \end{pmatrix}$$

are all eigenvectors with corresponding eigenvalue 0; hence, their linear combination is still an eigenvector with corresponding eigenvalue 0. This is a consequence of the fact that constant functions are in the kernel of the K-Laplacian, regardless of the choice of the integral kernel K.

□

The dimension of $\mathcal{L}_{\cdot,1,\mathbf{x}}$ is easy to compute if K is symmetric.

Theorem 3.6 (Dimension of $\mathcal{L}_{\cdot,1,\mathbf{x}}$)**.** *Let* K *be a symmetric kernel that satisfies Definition 3.2, then*

$$\dim(\mathcal{L}_{\cdot,1,\mathbf{x}}) = \frac{N(N-1)}{2}. \tag{3.7}$$

Proof. To show that the dimension is $N(N-1)/2$, we build $N(N-1)/2$ linearly independent 0-mean matrices. Using the fact that an element of $\mathbf{L} \in \mathcal{L}_{\cdot,1,\mathbf{x}}$ is a 0-sum matrix, consider

$$\mathbf{E}_{ij} = (\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^{\mathrm{T}}, \quad i \in \{1, \dots, N\}, j \in \{i+1, \dots, N\},$$

then $\mathbf{L}$ can be decomposed as

$$\mathbf{L} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} [\mathbf{L}]_{ij} \mathbf{E}_{ij} \,,$$

hence, $\{\mathbf{E}_{ij}\}$ is a (non orthonormal) basis for the space $\mathcal{L}_{\cdot,1,\mathbf{x}}$. □

Table 3.1: Random Sampled Laplacian depending on the choice of K, M. The matrix $\mathbf{D}_\rho = \mathrm{diag}(\mathbf{A}\mathbf{P}^{-1}\mathbf{1})$ is the ρ-degree matrix; when the sampling density is uniform, it is exactly the degree matrix $\mathbf{D} = \mathrm{diag}(\mathbf{A}\mathbf{1})$.

$K(x,y)$	$M(x)$		$\mathcal{L}_{K,M,\rho,\mathbf{x}}$
$K^{\mathrm{c}}(x,y)$	0	$\frac{1}{N}(\mathbf{A}+\mathbf{I})\mathbf{P}^{-1}$	
	1		$-\mathbf{I}$
	$\int K(x,y)\,\mathrm{d}\mu(y)$		$-\frac{1}{N}(\mathbf{D}_\rho + \mathbf{P}^{-1})$
$K^{\mathrm{rw}}(x,y)$	0	$(\mathbf{D}_\rho + \mathbf{P}^{-1})^{-1}(\mathbf{A}+\mathbf{I})\mathbf{P}^{-1}$	
	1		$-\mathbf{I}$
	$\int K(x,y)\,\mathrm{d}\mu(y)$		$-\mathbf{I}$
$K^{\mathrm{sn}}(x,y)$	0	$(\mathbf{D}_\rho + \mathbf{P}^{-1})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})\mathbf{P}^{-1}(\mathbf{D}_\rho + \mathbf{P}^{-1})^{-\frac{1}{2}}$	
	1		$-\mathbf{I}$
	$\int K(x,y)\,\mathrm{d}\mu(y)$		$-\mathrm{diag}\left((\mathbf{D}_\rho + \mathbf{P}^{-1})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})\mathbf{P}^{-1}(\mathbf{D}_\rho + \mathbf{P}^{-1})^{-\frac{1}{2}}\mathbf{1}\right)$

Chapter 4

Networks in Latent Geometry

In real scenarios, the metric measure space $(\mathcal{V}, d, \mu)$ as well as the sampling density ρ are not known; the only available piece of information is the topology of the graph. If no a-priori knowledge is accessible, one could ignore the weights and suppose the sampling is uniform: this is covered in Section 4.1. On the contrary, one could make assumptions on the latent space: in Section 4.2 a simple model for the latent space is studied, and used in Section 4.2.1 to approximate the sampling density and in Section 4.2.2 to learn it.

4.1 Ignoring the Density

Given a graph $\mathcal{G}_N = (\mathcal{V}_N, \mathcal{E}_N)$ with adjacency matrix $\mathbf{A}$, define the observed degree as the number of neighbours of each node

$$\mathbf{D} = \mathrm{diag}(\mathbf{A}\mathbf{1})\,, \tag{4.1}$$

and the real degree to be

$$\mathbf{D}_\rho = \mathrm{diag}(\mathbf{A}\mathbf{P}^{-1}\mathbf{1})\,. \tag{4.2}$$

While the real degree $\mathbf{D}_\rho$ comes up naturally when one samples a metric measure space, in reality the sampling density ρ is not known; therefore, one could simply replace $\mathbf{P}$ by $\mathbf{I}$, and $\mathbf{D}_\rho$ by $\mathbf{D}$. What happens in doing so is explained in the following theorem.

Theorem 4.1 (Convergence of Real and Observed Degree)**.** *Given a graph $\mathcal{G}_N = (\mathcal{V}_N, \mathcal{E}_N)$ with N nodes, adjacency matrix $\mathbf{A}$, real degree $\mathbf{D}_\rho$ and observed degree $\mathbf{D}$,*

it holds

$$\frac{[\mathbf{D}_\rho + \mathbf{P}^{-1}]_{ii}}{N} \xrightarrow[N\to\infty]{} \mu(B_r(x_i)) \,, \frac{[\mathbf{D} + \mathbf{I}]_{ii}}{N} \xrightarrow[N\to\infty]{} \mathbb{P}_\rho[B_r(x_i)]$$

where ρ is the sampling density and $\{x_i\}_{i=1}^N \subset \mathcal{V}$ are the sampled points.

Proof. The proof relies on the Strong Law of Large Numbers; in particular, due to $\{x_i\}_{i=1}^N$ be an i.i.d. sample from ρ, it holds

$$\frac{1}{N}[\mathbf{D}_\rho + \mathbf{P}^{-1}]_{ii} = \frac{1}{N}\sum_{j=1}^{N} \frac{\mathbb{1}_{\mathrm{B}_r(x_i)}(x_j)}{\rho(x_j)} \xrightarrow[N\to\infty]{} \mathbb{E}\left[\frac{\mathbb{1}_{\mathrm{B}_r(x_i)}(y)}{\rho(y)}\right] = \mu\left(\mathrm{B}_r(x_i)\right) \,,$$

and

$$\frac{1}{N}[\mathbf{D} + \mathbf{I}]_{ii} = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}_{\mathrm{B}_r(x_i)}(x_j) \xrightarrow[N\to\infty]{} \mathbb{E}\left[\mathbb{1}_{\mathrm{B}_r(x_i)}(y)\right] = \mathbb{P}_\rho\left[\mathrm{B}_r(x_i)\right] \,.$$

□

The previous result states that the knowledge of the weights $\mathbf{P}$ is important to soften the distortion introduced by sampling. However, in certain cases the distortion is little; in order to prove this, the following result is needed.

Theorem 4.2 (Integral Mean Value Theorem)**.** *Let ρ be a continuous density function over $\mathcal{V}$, then there exist a function $c : \mathcal{V} \to \mathcal{V}$ such that $c(x) \in B_r(x)$ and*

$$(\rho \circ c)(x) = \frac{1}{\mu(B_r(x))} \int_{\mathcal{V}} \mathbb{1}_{B_r(x)}(y)\, \rho(y)\, \mathrm{d}\mu(y) = \frac{\mathbb{P}_\rho[B_r(x)]}{\mu(B_r(x))} \,.$$

Loosely speaking, c maps each point x to the point $c(x)$ that satisfies the mean value property of ρ. What the previous theorem suggests is that if ρ is "almost" uniform, the distortion introduced by sampling is little; hence, the observed degree is a good approximation of the real degree. However, the importance of this theorem will be seen in the following sections, in which the "observed" graph shift operators will be studied .

4.1.1 Observed Combinatorial Laplacian

While the combinatorial Laplacian tends to

$$\frac{1}{N}\left[(\mathbf{A}\mathbf{P}^{-1} - \mathbf{D}_\rho)\mathbf{u}\right]_i \xrightarrow[N\to\infty]{} \int_{\mathcal{V}} \mathbb{1}_{\mathrm{B}_r(x_i)}(y)\, (u(y) - u(x_i))\, \mathrm{d}\mu(y) \,,$$

the observed one tends to

$$\frac{1}{N}\left[(\mathbf{A} - \mathbf{D})\mathbf{u}\right]_i \xrightarrow[N\to\infty]{} \int_{\mathcal{V}} \mathbb{1}_{\mathrm{B}_r(x_i)}(y)\, \rho(y)\, (u(y) - u(x_i))\, \mathrm{d}\mu(y) \,,$$

hence, it is not possible to remove the distortion introduced by the sampling density ρ.

4.1.2 Observed Random Walk Laplacian

While the random walk Laplacian tends to

$$\left[((\mathbf{D}_\rho + \mathbf{P}^{-1})^{-1}(\mathbf{A}+\mathbf{I})\mathbf{P}^{-1} - \mathbf{I})\mathbf{u}\right]_i \xrightarrow[N\to\infty]{} \int_{\mathcal{V}} \frac{\mathbb{1}_{\mathrm{B}_r(x_i)}(y)}{\mu(\mathrm{B}_r(x_i))} \left(u(y) - u(x_i)\right) \mathrm{d}\mu(y)\,,$$

the observed one tends to

$$\begin{aligned}\left[\left((\mathbf{D}+\mathbf{I})^{-1}(\mathbf{A}+\mathbf{I}) - \mathbf{I}\right)\mathbf{u}\right]_i &\xrightarrow[N\to\infty]{} \int_{\mathcal{V}} \frac{\mathbb{1}_{\mathrm{B}_r(x_i)}(y)}{\mathbb{P}[\mathrm{B}_r(x_i)]}\,\rho(y)\,u(y)\,\mathrm{d}\mu(y) - u(x_i) \\ &= \int_{\mathcal{V}} \frac{\mathbb{1}_{\mathrm{B}_r(x_i)}(y)}{\mu(\mathrm{B}_r(x_i))}\,\frac{\rho(y)}{(\rho\circ c)(x_i)}\,u(y)\,\mathrm{d}\mu(y) - u(x_i)\,.\end{aligned}$$

where the equality is due to the Integral Mean Value Theorem.

If ρ is continuously differentiable, a first order Taylor expansion in a neighborhood of x_i leads to

$$\frac{\rho(y)}{(\rho\circ c)(x_i)} \leq 1 + 2\,r\,\frac{\|\rho'(x_i)\|}{\rho(x_i)} + \mathrm{o}(r) \xrightarrow[r\to 0^+]{} 1\,,$$

while, if just the denominator is expanded in a neighborhood of y

$$\begin{aligned}\frac{\rho(y)}{(\rho\circ c)(x_i)} &\geq \frac{1}{1 + 2\,r\,\dfrac{\|\rho'(y)\|}{\rho(y)} + \mathrm{o}(r)} \\ &\geq \frac{1}{1 + 2\,r\,\max\limits_{y\in\mathrm{B}_r(x_i)}\dfrac{\|\rho'(y)\|}{\rho(y)} + \mathrm{o}(r)} \xrightarrow[r\to 0^+]{} 1\,,\end{aligned}$$

where the maximum exists because $\mathcal{V}$ is compact and ρ', ρ are continuous. The previous bounds state that, in the limit of r going to zero, the distortion introduced by a non uniform sampling is corrected.

4.1.3 Observed Symmetric Normalized K-Laplacian

The same reasoning of the previous section can be applied to the observed symmetric-normalized K-Laplacian. The i^{th} component of

$$\left((\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})(\mathbf{D}+\mathbf{I})^{-\frac{1}{2}} - \mathrm{diag}\left((\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})(\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}\,\mathbf{1}\right)\right)\mathbf{u}\,,$$

converges for $N\to\infty$ to

$$\int_{\mathcal{V}} \frac{\mathbb{1}_{\mathrm{B}_r(x_i)}(y)}{\sqrt{\mathbb{P}_\rho\left[\mathrm{B}_r(x_i)\right]}\sqrt{\mathbb{P}_\rho\left[\mathrm{B}_r(y)\right]}}\rho(y)(u(y)-u(x_i))\,\mathrm{d}\mu(y)\,. \tag{4.3}$$

Using the Integral Mean Value Theorem, the previous equation can be written as

$$\int_{\mathcal{V}} \frac{\mathbb{1}_{\mathrm{B}_r(x)}(y)}{\sqrt{\mu(\mathrm{B}_r(x_i))}\sqrt{\mu(\mathrm{B}_r(y))}} \frac{\rho(y)}{\sqrt{(\rho \circ c)\,(x_i)}\sqrt{(\rho \circ c)\,(y)}}(u(y) - u(x_i))\,\mathrm{d}\mu(y)\,.$$

As done previously, a first-order approximation leads to

$$\frac{\rho(y)}{\sqrt{(\rho \circ c)(x_i)}\sqrt{(\rho \circ c)(y)}} \leq 1 + \frac{3}{2}\, r \frac{\|\rho'(x_i)\|}{\rho(x_i)} + \mathrm{o}(r) \xrightarrow[r\to 0^+]{} 1$$

$$\frac{\rho(y)}{\sqrt{(\rho \circ c)(x_i)}\sqrt{(\rho \circ c)(y)}} \geq \frac{1}{1 + \frac{3}{2}\, r \max\limits_{y \in \mathrm{B}_r(x_i)} \frac{\|\rho'(y)\|}{\rho(y)} + \mathrm{o}(r)} \xrightarrow[r\to 0^+]{} 1\,,$$

hence, once again, the distortion of the non-uniform sampling is softened in the limit of r going to zero. The same applies to the observed symmetric normalized Laplacian.

4.2 Learning the Density: Unit Circle Model

The simplest choice of metric measure space is the unit circle

$$\mathbb{S}^1 = \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\|_2 = 1\}\,,$$

equipped with geodesic distance

$$\mathring{d}(\mathbf{x}, \mathbf{y}) = \arccos(\mathbf{x}^{\mathrm{T}}\mathbf{y})\,,\ \forall \mathbf{x}, \mathbf{y} \in \mathbb{S}^1\,,$$

and measure

$$\mathring{\mu}(B_r(\mathbf{x})) = 2\,\min\{r, \pi\}\,,$$

The unit circle $\mathbb{S}^1$ is intimately related to the segment $[-\pi, \pi)$ equipped with distance

$$\bar{d}(x, y) = |x - y|\,,$$

and measure

$$\bar{\mu}(B_r(x)) = \min\{\pi, x + r\} - \max\{-\pi, x - r\}\,.$$

Indeed, the maps

$$\begin{aligned} \varphi : [-\pi, \pi) &\to \mathbb{S}^1\,, \\ x &\mapsto \mathbf{x} = \Big(\cos(x),\ \ \sin(x)\Big)^{\mathrm{T}}\,, \\ \varphi^{-1} : \mathbb{S}^1 &\to [-\pi, \pi)\,, \\ \mathbf{x} &\mapsto x = \arctan\left(\frac{[\mathbf{x}]_2}{[\mathbf{x}]_1}\right)\,, \end{aligned}$$

bring a point from one space to the other. The geodesic distance between points on $\mathbb{S}^1$ can be computed without using explicitly $\mathring{d}$, because

$$\begin{aligned}\mathring{d}(\mathbf{x},\mathbf{y}) &= \arccos(\mathbf{x}^{\mathrm{T}}\mathbf{y}) = \arccos(\cos(x)\cos(y)+\sin(x)\sin(y)) \\ &= \arccos(\cos(x-y)) = |x-y| + 2\,\min\{0,\pi-|x-y|\} \\ &= \bar{d}(x,y) + 2\,\min\{0,\pi-\bar{d}(x,y)\}\,.\end{aligned}$$

Hence, one could work only on $[-\pi,\pi)$, reducing the overall computational cost (for example, when pairwise distances need to be computed). As a remark, the difference between $(\mathbb{S}^1,\mathring{d},\mathring{\mu})$ and $([-\pi,\pi),\bar{d},\bar{\mu})$ is that the first space is a Uniformly Distributed Measure Space because the measure of the balls is a function of the radius only

$$\begin{aligned}\mathring{\mu}(B_r(x)) &= 2\,\min\{r,\pi\}\,, \\ \bar{\mu}(B_r(x)) &= \begin{cases} 2\,\min\{r,\pi\} & |x| \le |\pi - r| \\ -|x| + r + \pi & |x| \ge |\pi - r| \end{cases}.\end{aligned}$$

In order to build graph approximations of $\mathbb{S}^1$, probability densities on $[-\pi,\pi)$ should be constructed. The von Mises distribution is a continuous probability distribution whose probability density function is defined as

$$\rho(x;\mu,\kappa) = \frac{\exp\{\kappa\cos(x-\mu)\}}{2\,\pi I_0(\kappa)}\,\mathbb{1}_{[-\pi,\pi)}(x)\,,$$

where x represents the angle, I_0 is the modified Bessel function of order 0, μ is the maximum point of the density and κ gives information about the variance: if $\kappa = 0$ the distribution is uniform, while if κ increases the distribution becomes more concentrated on the value of μ. Define the cumulative distribution function as

$$F(x;\mu,\kappa) = \int_{-\infty}^{x} \rho(y;\mu,\kappa)\,\mathrm{d}\mu(y)\,,$$

and the probability of balls $\mathbb{P}_\rho[\mathrm{B}_r(x)]$ as

$$\bar{F}(x;\mu,\kappa,r) = F(x+r,\mu,\kappa) - F(x-r,\mu,\kappa)\,, \tag{4.4}$$

for $([-\pi,\pi),\bar{d},\bar{\mu})$, and

$$\begin{aligned}\mathring{F}(x;\mu,\kappa,r) =& \bar{F}(x;\mu,\kappa,r) + F(x+r-2\pi,\mu,\kappa) \\ &+ (1 - F(x-r+2\pi,\mu,\kappa))\,,\end{aligned} \tag{4.5}$$

for $(\mathbb{S}^1,\mathring{d},\mathring{\mu})$, where the second and third addends take into account the periodicity of the circle.

A generic probability density function on $[-\pi, \pi)$ can be constructed as follows. Given a set of coefficients $\mathbf{c} = \{c_k\}_{k=1}^K$, the function

$$\rho_{\mathbf{c}}(x) = c_0 + \sum_{k=1}^{K} c_k \, \sin(k\,(x-\mu))\,,$$

is 2π-periodic. In order to be a consistent probability density function, it must be positive and have a unit integral. It can be observed that

$$-|c_k| \leq c_k \, \sin(k\,(x-\mu)) \leq |c_k|$$

hence, non negativity is guaranteed if

$$\rho_{\mathbf{c}}(x) = \sum_{k=1}^{K} c_k \, \sin(k\,(x-\mu)) + \sum_{k=0}^{K} |c_k|\,.$$

Regarding unit integral, it can be noted that

$$\int_{-\pi}^{\pi} \rho_{\mathbf{c}}(s)\,\mathrm{d}s = 2\,\pi \, \sum_{k=0}^{K} |c_k|\,,$$

therefore, a proper probability density function on the circle is

$$\rho_{\mathbf{c}}(x) = \sum_{k=1}^{K} \tilde{c}_k \, \sin(k\,(x-\mu)) + \frac{1}{2\pi}\,,\ \tilde{c}_k = \frac{c_k}{2\,\pi \, \sum\limits_{j=0}^{K} |c_j|}\,.$$

The corresponding cumulative distribution function is

$$F_{\mathbf{c}}(x) = -\sum_{k=1}^{K} \tilde{c}_k \, \frac{\cos(k\,(\pi+\mu)) - \cos(k\,(x-\mu))}{k} + \frac{x+\pi}{2\pi}\,,$$

and the probability of the balls can be computed as in Equations (4.4) to (4.5). If $\mathbf{c}$ is allowed to be a random vector of random length, $\rho_{\mathbf{c}}$ will be a random variable whose outcome is a probability density function that identifies a random variable.

A similar construction can be repeated for the sum of generic 2π-periodic functions: in this case the offset, that guarantees non negativity, and the normalization constant, that guarantees unit integral, will change. For example, one could obtain a generalization of the Von Mises distribution as

$$\rho_{\mathbf{c}}(x) = \frac{\dfrac{c_0}{2\,\pi} + \sum\limits_{k=1}^{K} c_k \, \dfrac{\exp(\kappa \cos(k\,(x-\mu)))}{2\,\pi\,\mathrm{I}_0(\kappa)} - \exp(\kappa) \sum\limits_{k=1}^{K} \min\left\{0, \dfrac{c_k}{2\,\pi\,\mathrm{I}_0(\kappa)}\right\}}{c_0 + \sum\limits_{j=1}^{K} \left(c_j - \exp(\kappa)\, \min\left\{0, \dfrac{c_j}{\mathrm{I}_0(\kappa)}\right\}\right)}\,.$$

In this case, a closed formula for the cumulative distribution function is not known, therefore, Acceptance-Rejection Sampling should be performed.

4.2.1 Approximating the Density

Consider a semi-supervised node classification task. If the sampling density is known, one could use the real quantities

$$\begin{cases} \mathbf{X}^{(0)} = \mathbf{X}\,, \\ \mathbf{X}^{(l)} = \sigma^{(l)}\left((\mathbf{D}_\rho + \mathbf{P}^{-1})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})\mathbf{P}^{-1}(\mathbf{D}_\rho + \mathbf{P}^{-1})^{-\frac{1}{2}}\mathbf{X}^{(l-1)}\mathbf{W}^{(l)}\right)\,,\ l = 1,\dots,L\,, \end{cases}$$

where $\mathbf{X} \in \mathbb{R}^{N\times d_0}$ is the input node features, $\mathbf{W}^{(l)} \in \mathbb{R}^{d_{l-1}\times d_l}$ is the learnable weight matrix at layer l, $\sigma^{(l)} : \mathbb{R} \to \mathbb{R}$ is the entry-wise non-linear functions at layer l, and $\mathbf{X}^{(L)}$ is the output of the neural network. If one is not willing to make assumptions on the underlying latent space, the observed quantities could be used as explained in Section 4.1

$$\begin{cases} \mathbf{X}^{(0)} = \mathbf{X}\,, \\ \mathbf{X}^{(l)} = \sigma^{(l)}\left((\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})(\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}\mathbf{X}^{(l-1)}\mathbf{W}^{(l)}\right)\,,\ l = 1,\dots,L\,, \end{cases} \tag{CNet-1}$$

Suppose the underlying latent space is $(\mathbb{S}^1, \mathring{d}, \mathring{\mu})$, then for Theorem 4.1, if N is sufficiently large

$$(\mathbf{D}_\rho + \mathbf{P}^{-1})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})\mathbf{P}^{-1}(\mathbf{D}_\rho + \mathbf{P}^{-1})^{-\frac{1}{2}}\mathbf{X}^{(i-1)}\mathbf{W}^{(i)} \approx \frac{1}{2\,N\,r}(\mathbf{A}+\mathbf{I})\mathbf{P}^{-1}\mathbf{X}^{(i-1)}\mathbf{W}^{(i)}\,.$$

From Theorem 4.2, if the sampling probability is not oscillating too fast, the probability of the balls $\mathbb{P}_\rho[\mathrm{B}_r(x)]$ and the sampling density $\rho(x)$ are related by a multiplicative constant, i.e. $\rho(x) \approx \mathbb{P}_\rho[\mathrm{B}_r(x)]/(2\,r)$ and $\mathbf{P} \approx (\mathbf{D}+\mathbf{I})/(2\,N\,r)$, allowing to remove the dependence on r

$$(\mathbf{D}_\rho + \mathbf{P}^{-1})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})\mathbf{P}^{-1}(\mathbf{D}_\rho + \mathbf{P}^{-1})^{-\frac{1}{2}}\mathbf{X}^{(i-1)}\mathbf{W}^{(i)} \approx (\mathbf{A}+\mathbf{I})(\mathbf{D}+\mathbf{I})^{-1}\mathbf{X}^{(i-1)}\mathbf{W}^{(i)}\,,$$

and obtaining the neural network

$$\begin{cases} \mathbf{X}^{(0)} = \mathbf{X}\,, \\ \mathbf{X}^{(l)} = \sigma^{(l)}\left((\mathbf{A}+\mathbf{I})(\mathbf{D}+\mathbf{I})^{-1}\mathbf{X}^{(l-1)}\mathbf{W}^{(l)}\right)\,,\ l = 1,\dots,L\,, \end{cases} \tag{CNet-2}$$

In Figure 4.2, CNet-2 and CNet-1 are applied on real citation networks. A citation network is a graph whose nodes are scientific publications and links are citations between the documents. To each node is attributed a binary input feature vector representing the presence or absence of certain words. More specifically:

- the "Cora" dataset consists of 2708 nodes, 5429 links and 1433 words; each node is classified into one of 7 classes;

- the "Citeseer" dataset consists of 3327 nodes, 4732 links and 3703 words; each node is classified into one of 6 classes;
- the "Pubmed" dataset consists of 19717 nodes, 44338 links and 500 words; each node is classified into one of 3 classes.

The task is to predict the class of the nodes using, in training phase, just a subset of them. One could wonder why such hypotheses on the latent space are plausible; the reason relies on the distribution of the observed degree, shown in Figure 4.1. The observed degree is concentrated towards small values, meaning that each node is connected to few nodes, hence, r is small.

As a remark, the hypotheses on the latent space lead to a new Graph Shift Operator, namely $(\mathbf{A}+\mathbf{I})(\mathbf{D}+\mathbf{I})^{-1}$. Such Graph Shift Operator can be generalized to any other latent space because it can be seen as the observed random sampled of a 0K-Laplacian defined as

$$\mathcal{L}_{K,0}u(x) = \int_{\mathcal{V}} \frac{\mathbb{1}_{\mathrm{B}_r(x)}(y)}{\mu(\mathrm{B}_r(y))} u(y)\,\mathrm{d}\mu(y)\,.$$

4.2.2 Learning the Density

The theoretical findings of Section 4.1 are important: if r is small and ρ does not oscillate too much, the observed quantities approximate well the real ones. However, in real scenarios, this piece of information is not accessible.

Instead of ignoring the density as done in Section 4.1, it could be learnt. For instance, one could be tempted to use the results from Section 4.2.1 to build a ConvGNN of type

$$\begin{cases} \mathbf{X}^{(0)} = \mathbf{X}\,, \\ \mathbf{X}^{(l)} = \sigma^{(i)}\left((\mathbf{A}+\mathbf{I})\tilde{\mathbf{P}}^{-1}\mathbf{X}^{(l-1)}\mathbf{W}^{(l)}\right)\,,\ l=1,\dots,L\,, \end{cases} \tag{CNet-3}$$

where $\tilde{\mathbf{P}}^{-1} \in \mathbb{R}^{N\times N}$ is a learnable diagonal matrix of order N. This approach has two main drawbacks: 1) it can only be applied if the graph is fixed, because $\tilde{\mathbf{P}}^{-1}$ depends on the number of nodes, and 2) it can suffer dramatically of overfitting.

In order to avoid the abovementioned disadvantages, a graph neural network, could be used to predict the density ρ from the observed degree, because GNNs are known to be transferable between graphs of different sizes and equivariant to node re-indexing. The architecture used is a GCN

$$\begin{cases} \mathbf{P}^{(0)} = \mathbf{D}+\mathbf{I}\,, \\ \mathbf{P}^{(l)} = \sigma_P^{(l)}\left((\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})(\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}\mathbf{P}^{(l-1)}\mathbf{W_P}^{(l)}\right)\,,\ l=1,\dots,L_P\,, \end{cases} \tag{PNet}$$

and it will be trained on a geometric dataset $\mathcal{D}$ made of graphs approximations of the unit circle. The geometric dataset $\mathcal{D}$ is built in a way that each element is coming from a different density, as explained in the following.

1. For $i = 1, \dots, |\mathcal{D}|$ fix a radius r_i.
2. Define a probability density function ρ_i following the construction in Section 4.2.
3. From ρ_i draw a i.i.d. random sample θ_{ij}, $j = 1, \dots, N$.
4. Connect all the pairs $(\theta_{ij}, \theta_{ik})$ such that $\mathring{d}(\theta_{ij}, \theta_{ik}) \leq r_i$, for all $j, k = 1, \dots, N$.

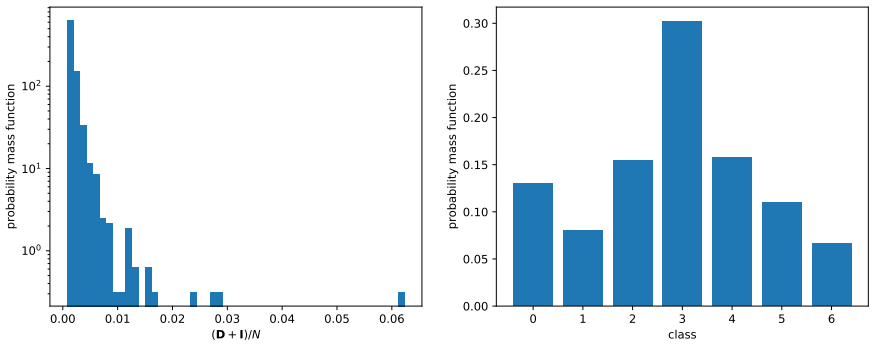
The node classification task is performed by a GCN similar to CNet-3

$$\begin{cases} \mathbf{X}^{(0)} = \mathbf{X}\,, \\ \mathbf{X}^{(l)} = \sigma^{(l)}\left((\mathbf{A}+\mathbf{I})\left(\mathbf{P}^{(L_P)}\right)^{-1}\mathbf{X}^{(l-1)}\mathbf{W}^{(l)}\right)\,,\ l = 1, \dots, L\,, \end{cases} \tag{CNet-4}$$

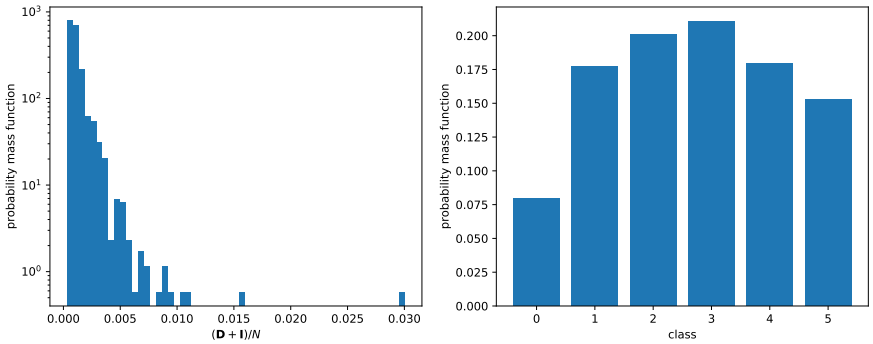
where $\mathbf{P}^{(L_P)}$ is the output of PNet. PNet is trained exclusively on the geometric dataset $\mathcal{D}$, hence, Cora, Citeseer, and Pubmed are out-of-distribution. In Figure 4.2 an interesting phenomena happens: after a certain epoch, the test loss of CNet-4 increases, while the average accuracy does not decrease. The explanation relies on the fact that accuracy and loss are not perfectly inversely correlated, and the Cross Entropy loss is unbounded: few misclassifications could cause an increment in the loss while the network keeps learning from the dataset. From Table 4.1, as well as Figure 4.2, it can be noted that CNet-4 reaches the best accuracy on all the citation networks, with the smallest variance and requiring fewer number of iterations.

Figure 4.1: Degree and class distribution of some citation networks.

(a) Dataset: Cora.



(b) Dataset: Citeseer.



(c) Dataset: Pubmed.

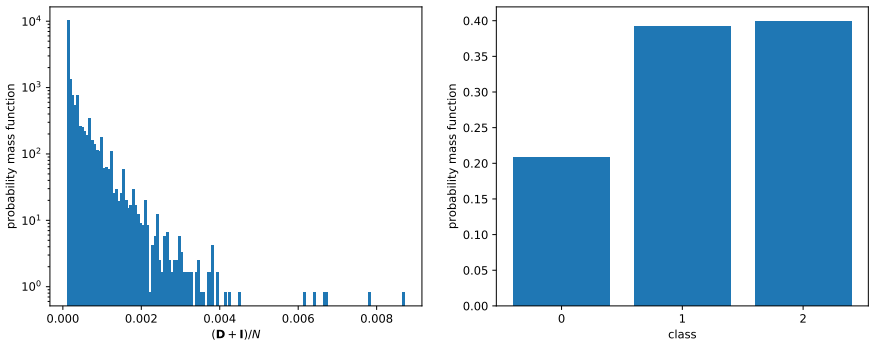
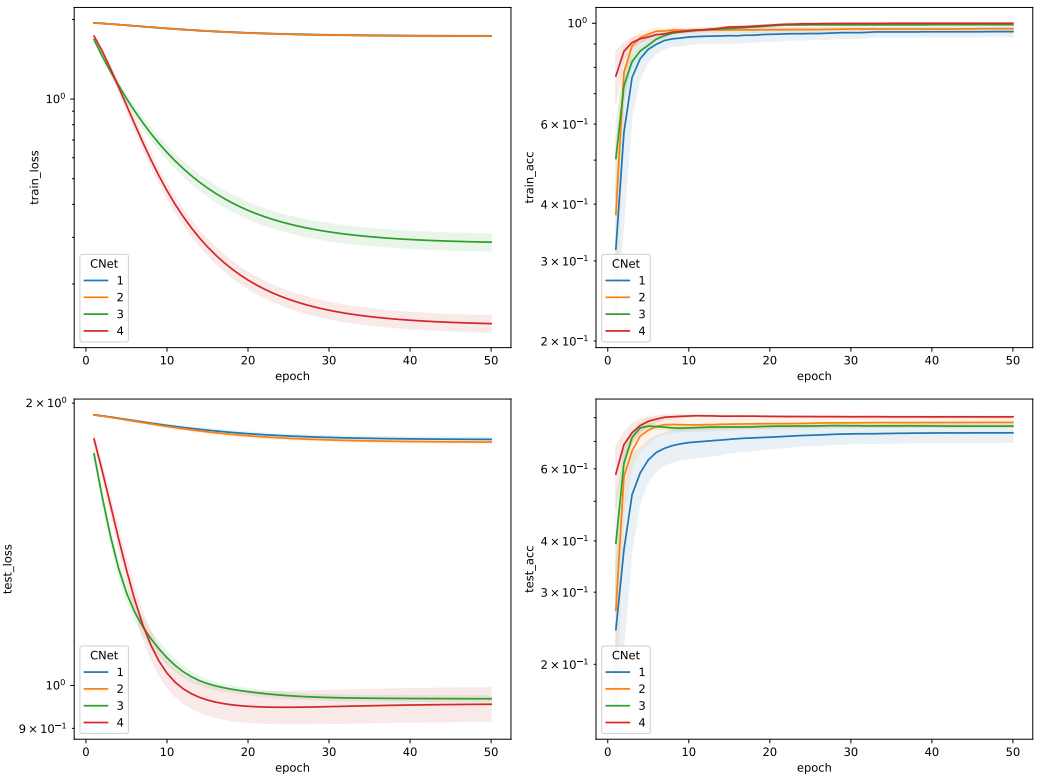
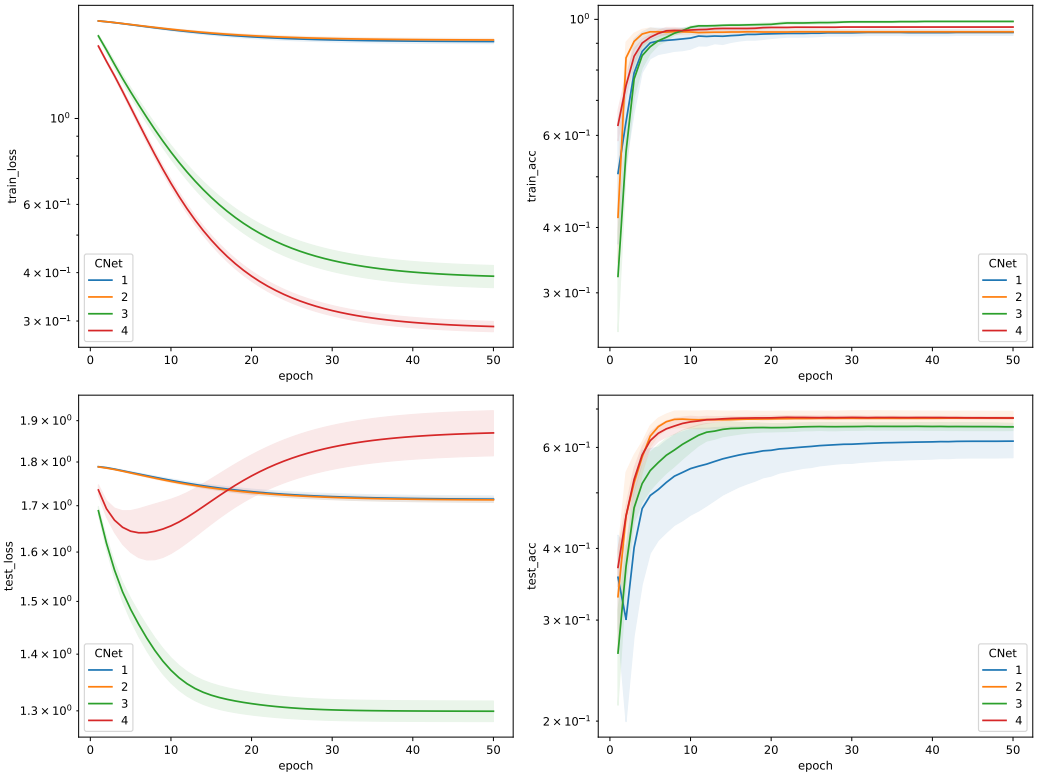


Figure 4.2: Node classification task using CNet-1, CNet-2, CNet-3, CNet-4 with $L = 2$, $d_1 = 64$, $\sigma^{(1)} = \text{ReLU}$. The depicted curves are average curves of 10 repetitions, while the coloured area represent the standard deviation. The Cross Entropy Loss has been used, as well as the Adam optimizer with an exponential scheduler (decay rate equal to 0.9, initial learning rate equal to 0.01). When using CNet-3, the number of hidden layers is not 64 but less, in order to keep the comparison fair and the number of parameters comparable to the other networks: the number of hidden channels is computed as $\lfloor d_1 - N/(d_0+d_2+2) \rfloor$. A general trend is that CNet-2 has smaller variance of CNet-1, and greater average test accuracy in the early stage of training. However, CNet-4 guarantees better performances.

(a) Dataset: Cora, $N = 2708$, $d_0 = 1433$, $d_2 = 7$. The number of training nodes is 140, the number of isolated nodes is 0. CNet-2 has a higher average accuracy both on training nodes and on testing nodes, as well as smaller variance than CNet-1; CNet-4 outperform the other GNNs.



(b) Dataset: Citeseer, $N = 3327$, $d_0 = 3703$, $d_2 = 6$. The number of training nodes is 120, the number of isolated nodes is 48. CNet-2 is better than the other GNNs because it guarantees better generalizability: higher average accuracy and smaller variance.



(c) Dataset: Pubmed, $N = 19717$, $d_0 = 500$, $d_2 = 3$. The number of training nodes is 60, the number of isolated nodes is 0. CNet-4 outperforms the other GNNs. CNet-2 has better performances than CNet-1 in the early stages of training.

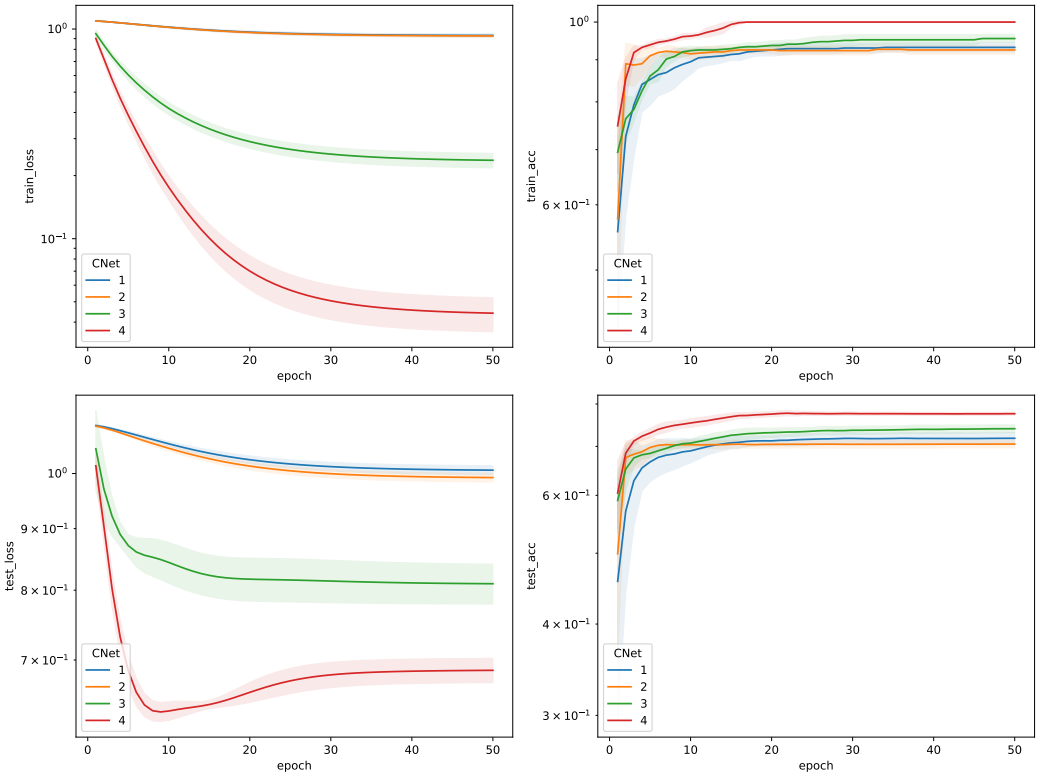


Table 4.1: Maximal test accuracy: comparisons between GCNs. CNet-4 reaches the best average accuracy on all the datasets, smaller variance, and requiring a smaller number of epochs.

		Dataset		
		Cora	Citeseer	Pubmed
CNet-1	Epoch	50	50	50
	Mean	0.735	0.615	0.718
	Std. dev.	0.038	0.040	0.013
CNet-2	Epoch	50	45	50
	Mean	0.779	0.675	0.705
	Std. dev.	0.019	0.019	0.008
CNet-3	Epoch	28	32	50
	Mean	0.765	0.653	0.740
	Std. dev.	0.005	0.010	0.009
CNet-4	Epoch	11	22	22
	Mean	0.809	0.677	0.777
	Std. dev.	0.004	0.005	0.006

4.2.3 Barycenter Task

In real scenarios, the sampling density ρ is not known, hence, training a neural network against it is not possible. It is interesting to study the capability of the network to learn ρ while it is trained on other task. A "toy-example" can be constructed as follows.

Given a closed curve $\mathcal{C} \subset \mathbb{R}^2$, the barycenter (or center of mass) can be computed as the line integral

$$\mathbf{C} = \left(\int_{\mathcal{C}} \mathrm{d}s\right)^{-1} \int_{\mathcal{C}} (x_1, x_2)^{\mathrm{T}} \, \mathrm{d}s \, .$$

However, if a parametrization of $\mathcal{C}$ is not known but only a sample $\{(x_{1j}, x_{2j})\}_{j=1}^{N}$ is accessible, one could approximate the barycenter as

$$\mathbf{C}_1 = \frac{1}{N} \sum_{j=1}^{N} (x_{1j}, x_{2j})^{\mathrm{T}} \, . \tag{4.6}$$

If the sample is drawn according to a probability density function ρ, a better way to approximate the barycenter is

$$\mathbf{C}_\rho = \left(\sum_{k=1}^{N} \frac{1}{\rho(x_{1k}, x_{2k})} \right)^{-1} \sum_{j=1}^{N} \frac{1}{\rho(x_{1j}, x_{2j})} (x_{1j}, x_{2j})^{\mathrm{T}} \, . \tag{4.7}$$

Equation (4.7) tells that points more likely to be sampled will have a smaller contribution on the final value of the barycenter. The knowledge of the density is important to obtain a good approximation of the barycenter; therefore, one could wonder if the GNN is able to learn it while being trained to learn $\mathbf{C}$. The dataset $\mathcal{D}$ on which the GNN is trained is composed of graph approximations of circles; each graph approximation is generated as follows:

1. Define a probability density function ρ following the construction in Section 4.2.
2. From ρ draw a random sample θ_j, $j = 1, \ldots, N$.
3. Draw a center $\mathbf{C}$ and two radii r, R.
4. Compute the coordinates $\mathbf{x}_j = \mathbf{C} + R\,(\cos(\theta_j), \sin(\theta_j))^{\mathrm{T}}$.
5. Connect all the pairs $(\mathbf{x}_j, \mathbf{x}_k)$ such that $\mathring{d}(\mathbf{x}_j, \mathbf{x}_k) \leq r$, for all $j, k = 1, \ldots, N$.

The architecture of the network combines 3 GCNConv layers and 1 ChebConv layer of order 0; hence, the part of the network that learns the density is much

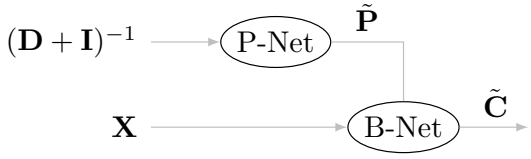
more expressive than the one that computes the final output

$$\begin{cases} \mathbf{P}^{(0)} = (\mathbf{D}+\mathbf{I})^{-1}\,, \\ \mathbf{P}^{(l)} = \sigma_P^{(l)}\left((\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})(\mathbf{D}+\mathbf{I})^{-\frac{1}{2}}\mathbf{P}^{(l-1)}\mathbf{W}_P^{(l)}\right)\,,\ l=1,2,3, \\ \tilde{\mathbf{P}} = \left(\mathbf{P}^{(3)}\right)^{-1}\,, \\ \tilde{\mathbf{C}} = \left(\mathbf{P}^{(3)}\right)^{\mathrm{T}}\mathbf{X}\,\mathbf{W}_C\,, \end{cases} \tag{BPNet}$$

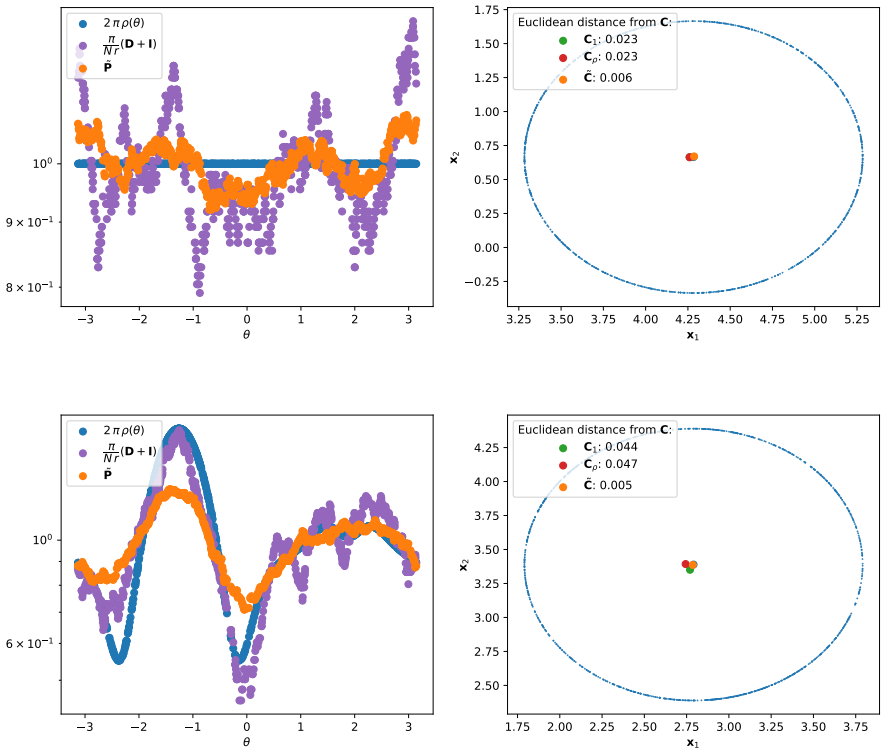
where $\mathbf{D} \in \mathbb{R}^{N\times N}$ is the degree matrix of the network, $\mathbf{X} \in \mathbb{R}^{N\times d_0}$ is the matrix containing the coordinates of the points. In Figure 4.3 is shown that the network learns to reproduce the behaviour of the density ρ; the difference between the true density and the learnt one can be explained as the ability of the network to learn the best representation of ρ for the task at hand.

Figure 4.3: Barycenter task: approximate the center of circles with a graph neural network.

(a) Scheme of BPNet: PNet is composed of 3 GCNConv layers followed by ReLU activation functions; BNet is composed of 1 ChebConv layer, $K = 1$ and a global mean pooling layer. Hence, PNet is more expressive than BNet. The loss is L_1.



(b) Some results from the validation set. The learnt $\tilde{\mathbf{P}}$ is able to reproduce the behaviour of the density ρ. What PNet learns is tailored to in task at hand: therefore, it is not surprising that the learnt density is somehow different from the real one.



Chapter 5

Stability of Polynomial Spectral Graph Filters

A desirable property of graph convolutional neural networks is stability. Loosely speaking, a ConvGNN is stable if a small change in the input cause a small change in the output. In this case inputs are graphs; hence, a small change in the input is a change in the topological structure of the graph. This includes deletion or addition of edges and nodes that prevents one from using the algebraic representation of a Graph Shift Operator.

In the setting in which graphs are considered as finite approximations of a metric measure space $(\mathcal{V}, d, \mu)$, two graphs are near if they are sampled from the same underlying space. The problem is then cast into the latent space; for instance, edge perturbations can be thought as if they were generated by a kernel perturbation, while nodes perturbation are caused by different sampling procedure.

There is a pletora of scientific publications that deal with stability of graph neural networks. In [32], the stability of polynomial spectral graph filters w.r.t. the change in the normalized Laplacian matrix is analyzed. In [33] the stability of ConvGNNs under rewiring between high degree nodes is analyzed. In [34] interpretable stability bounds are given in terms of structural properties of the graph and properties of the edge perturbation. In [35, 36], the authors analyze stability and approximation power of GNNs on random graphs. In [37, 38], the authors prove stability of GNNs with integral Lipschitz filters. In [39] the theory of graphon signal processing is introduced, and used in [17] to analyze transferability of GNNs across graphs. In [20] the authors analyze the linear stability of spectral graph filters in the Cayley smoothness space, followed up by [18] where transferability of spectral ConvGNNs between graphs of different size and topology is studied.

Even though the analysis is inspired to [18], the stability bounds are given point-wise and not in functional norms. Therefore, it is necessary to study filters

that preserve continuity.

Definition 5.1 (Preserves Continuity)**.** The metric measure Laplacian $\mathcal{L}$ is said to "preserve continuity" if $u \in \mathrm{C}^0(\mathcal{V}, \|\cdot|_\infty)$ implies $\mathcal{L}u \in \mathrm{C}^0(\mathcal{V}, \|\cdot|_\infty)$, hence, if continuous functions are mapped in continuous functions.

The following proposition will show that, under reasonable assumption, the MK-Laplacian preserves continuity.

Theorem 5.1 (MK-Laplacian preserves continuity)**.** *Let $(\mathcal{V}, d, \mu)$ be a compact metric measure space of finite measure $\mu(\mathcal{V}) < \infty$. The MK-Laplacian where K is either the combinatorial, random walk or symmetric normalized kernel preserves continuity, provided that M is continuous.*

Proof. The proof relies on Lebesgue's Dominated Convergence Theorem; in particular, let $\{x_n\}_n \subset \mathcal{V}$ be a sequence converging to $x \in \mathcal{V}$, then $\mathbb{1}_{\mathrm{B}_r(x_n)}(y)u(y) \to \mathbb{1}_{\mathrm{B}_r(x)}(y)u(y)$ and $\mathbb{1}_{\mathrm{B}_r(x_n)}(y)|u(y)| \leq |u(y)|$ with $\int_{\mathcal{V}}|u(y)|\,\mathrm{d}\mu(y) < \infty$; the thesis follows. □

The theorem is valid also for the K-Laplacian, due to the continuity of the function $x \to \mu(\mathrm{B}_r(x))$.

Another important preliminary result is the following.

Theorem 5.2 (Decomposition of Difference of Powers)**.** *Let $\mathcal{A}, \mathcal{B} : \mathcal{X} \to X$ be two linear bounded operators, then $\mathcal{A}^s - \mathcal{B}^s$ is a linear bounded operator and*

$$\mathcal{A}^s - \mathcal{B}^s = \sum_{j=0}^{s-1} \mathcal{A}^{s-j-1} (\mathcal{A} - \mathcal{B}) \mathcal{B}^j \tag{5.1}$$

Proof. The linearity is trivial to prove. The continuity of $\mathcal{A}^s$, $\mathcal{B}^s$ comes from submultiplicativity of norm operator; the continuity of $\mathcal{A}^s - \mathcal{B}^s$ comes from the structure of vector space of bounded linear operators. Simple algebraic manipulations leads to

$$\begin{aligned}
\sum_{j=0}^{s-1} \mathcal{A}^{s-j-1} (\mathcal{A} - \mathcal{B}) \mathcal{B}^j =& \sum_{j=0}^{s-1} \mathcal{A}^{s-j} \mathcal{B}^j - \sum_{n=0}^{j-1} \mathcal{A}^{s-j-1} \mathcal{B}^{j+1} \\
=& \mathcal{A}^s + \sum_{j=1}^{s-1} \mathcal{A}^{s-j} \mathcal{B}^j - \sum_{j=0}^{s-2} \mathcal{A}^{s-j-1} \mathcal{B}^{j+1} - \mathcal{B}^s \\
=& \mathcal{A}^s \pm \sum_{j=1}^{s-1} \mathcal{A}^{s-j} \mathcal{B}^j - \mathcal{B}^s \\
=& \mathcal{A}^s - \mathcal{B}^s ,
\end{aligned}$$

□

5.1 Kernel Perturbation

Let $\mathcal{L}_{K,M}$ be a MK-Laplacian that preserves continuity, it holds

$$\begin{aligned}|\mathcal{L}_{K,M}u(x)| &\leq \int_{\mathcal{V}} |K(x,y)\,u(y)|\,\mathrm{d}\mu(y) + |M(x)\,u(x)| \\ &\leq \|K(x,\cdot)\|_{\mathrm{L}^2(\mathcal{V})}\,\|u\|_{\mathrm{L}^2(\mathcal{V})} + |M(x)|\,|u(x)|\,,\end{aligned}$$

and

$$\|\mathcal{L}_{K,M}u\|_{\mathrm{L}^2(\mathcal{V})} \leq 2\left(\|K\|_{\mathrm{L}^2(\mathcal{V}\times\mathcal{V})} + \|M\|_{\mathrm{L}^\infty(\mathcal{V})}\right)\|u\|_{\mathrm{L}^2(\mathcal{V})}\,.$$

Denote $c_K(x) \coloneqq \|K(x,\cdot)\|_{\mathrm{L}^2(\mathcal{V})}$, $c_M(x) \coloneqq |M(x)|$, $c \coloneqq 2\|K\|_{\mathrm{L}^2(\mathcal{V}\times\mathcal{V})} + 2\|M\|_{\mathrm{L}^\infty(\mathcal{V})}$, then

$$\begin{aligned}&|\mathcal{L}^s_{K,M}u(x)| \leq c_K(x)\sum_{j=1}^{s} c^{s-j}\,c_M(x)^{j-1}\|u\|_{\mathrm{L}^2(\mathcal{V})} + c_M(x)^s|u(x)|\,,\\ &\|\mathcal{L}^s_{K,M}u\|_{\mathrm{L}^2(\mathcal{V})} \leq c^s\,\|u\|_{\mathrm{L}^2(\mathcal{V})}\,.\end{aligned}$$

Consider now two MK-Laplacian $\mathcal{L}_{K_1,M_1}$ and $\mathcal{L}_{K_2,M_2}$, then by Theorem 5.2 and Equation (3.6) it holds

$$\begin{aligned}\mathcal{L}^S_{K_1,M_1} - \mathcal{L}^S_{K_2,M_2} &= \sum_{s=0}^{S-1}\mathcal{L}^{S-s-1}_{K_1,M_1}\left(\mathcal{L}_{K_1,M_1} - \mathcal{L}_{K_2,M_2}\right)\mathcal{L}^s_{K_2,M_2}\\ &= \sum_{s=0}^{S-1}\mathcal{L}^{S-s-1}_{K_1,M_1}\,\mathcal{L}_{K_1-K_2,M_1-M_2}\,\mathcal{L}^s_{K_2,M_2}\,,\end{aligned}$$

Denote by

$$\begin{cases} c_{K_i}(x) \coloneqq \|K_i(x,\cdot)\|_{\mathrm{L}^2(\mathcal{V})} \\ c_{M_i}(x) \coloneqq |M_i(x)| \\ c_i \coloneqq 2\|K_i\|_{\mathrm{L}^2(\mathcal{V}\times\mathcal{V})} + 2\|M_i\|_{\mathrm{L}^\infty(\mathcal{V})} \end{cases}, \; i \in \{1,2\}\,,$$

and by

$$\begin{cases} c_{K_1-K_2}(x) \coloneqq \|K_1(x,\cdot) - K_2(x,\cdot)\|_{\mathrm{L}^2(\mathcal{V})} \\ c_{M_1-M_2}(x) \coloneqq |M_1(x) - M_2(x)| \\ c \coloneqq 2\|K_1 - K_2\|_{\mathrm{L}^2(\mathcal{V}\times\mathcal{V})} + 2\|M_1 - M_2\|_{\mathrm{L}^\infty(\mathcal{V})} \end{cases},$$

then

$$\left|\mathcal{L}^{S-s-1}_{K_1,M_1}\,\mathcal{L}_{K_1-K_2,M_1-M_2}\,\mathcal{L}^s_{K_2,M_2}u(x)\right| \leq c\,c_2^s\,c_{K_1}(x)\sum_{j=1}^{S-s-1} c_1^{S-s-1-j}\,c_{M_1}(x)^{j-1}\,\|u\|_{\mathrm{L}^2(\mathcal{V})}$$

$$
\begin{aligned}
&+ c_{K_1-K_2}(x)\, c_{M_1}(x)^{S-s-1}\, c_2^s \|u\|_{\mathrm{L}^2(\mathcal{V})} \\
&+ c_{M_1-M_2}(x)\, c_{K_2}(x) \sum_{j=1}^{s} c_2^{s-j}\, c_{M_2}(x)^{j-1} \|u\|_{\mathrm{L}^2(\mathcal{V})} \\
&+ c_{M_1-M_2}(x)\, c_{M_2}(x)^s |u(x)|\,,
\end{aligned}
$$

hence,

$$
\begin{aligned}
\left|\left(\mathcal{L}^S_{K_1,M_1} - \mathcal{L}^S_{K_2,M_2}\right) u(x)\right| &\leq \sum_{s=0}^{S-1} \left|\mathcal{L}^{S-s-1}_{K_1,M_1}\, \mathcal{L}_{K_1-K_2,M_1-M_2}\, \mathcal{L}^s_{K_2,M_2} u(x)\right| \\
&\leq c \sum_{s=0}^{S-1} c_2^s\, c_{K_1}(x) \sum_{j=1}^{S-s-1} c_1^{S-s-1-j}\, c_{M_1}(x)^{j-1}\, \|u\|_{\mathrm{L}^2(\mathcal{V})} \\
&+ c_{K_1-K_2}(x) \sum_{s=0}^{S-1} c_{M_1}(x)^{S-s-1}\, c_2^s \|u\|_{\mathrm{L}^2(\mathcal{V})} \\
&+ c_{M_1-M_2}(x)\, c_{K_2}(x) \sum_{s=0}^{S-1} \sum_{j=1}^{s} c_2^{s-j}\, c_{M_2}(x)^{j-1} \|u\|_{\mathrm{L}^2(\mathcal{V})} \\
&+ c_{M_1-M_2}(x) \sum_{s=0}^{S-1} c_{M_2}(x)^s |u(x)|\,,
\end{aligned}
$$

from which the polynomial filters are linearly stable in the metric measure space

$$
\begin{aligned}
&\left|\sum_{S=1}^{P} h_S \left(\mathcal{L}^S_{K_1,M_1} - \mathcal{L}^S_{K_2,M_2}\right) u(x)\right| \\
&\leq c \sum_{S=1}^{P} |h_S| \sum_{s=0}^{S-1} c_2^s\, c_{K_1}(x) \sum_{j=1}^{S-s-1} c_1^{S-s-1-j}\, c_{M_1}(x)^{j-1}\, \|u\|_{\mathrm{L}^2(\mathcal{V})} \\
&+ c_{K_1-K_2}(x) \sum_{S=1}^{P} |h_S| \sum_{s=0}^{S-1} c_{M_1}(x)^{S-s-1}\, c_2^s \|u\|_{\mathrm{L}^2(\mathcal{V})} \\
&+ c_{M_1-M_2}(x)\, c_{K_2}(x) \sum_{S=1}^{P} |h_S| \sum_{s=0}^{S-1} \sum_{j=1}^{s} c_2^{s-j}\, c_{M_2}(x)^{j-1} \|u\|_{\mathrm{L}^2(\mathcal{V})} \\
&+ c_{M_1-M_2}(x) \sum_{S=1}^{P} |h_S| \sum_{s=0}^{S-1} c_{M_2}(x)^s |u(x)|\,,
\end{aligned}
$$

5.2 Edge Perturbation

The edges of a spatial network depends on the support of the kernel; therefore, edge addition or deletion could be considered as inheritance of kernel perturbation.

For Theorem 3.2, in probability $(1-\epsilon)$ it holds

$$\left|\mathcal{L}_{K,M}u(x) - \mathcal{L}_{K,M,\rho,\mathbf{x}}u(x)\right| \leq N^{-\frac{1}{2}}\epsilon^{-\frac{1}{2}}\left\|\frac{K(x,\cdot)^2}{\rho(\cdot)}\right\|_{\mathrm{L}^\infty(\mathcal{V})}\|u\|_{\mathrm{L}^2(\mathcal{V})}\,,$$

hence

$$\begin{aligned}\left|\left(\mathcal{L}_{K_1,M_1,\rho,\mathbf{x}} - \mathcal{L}_{K_2,M_2,\rho,\mathbf{x}}\right)u(x)\right| &= \left|\mathcal{L}_{K_1,M_1,\rho,\mathbf{x}}u(x) - \mathcal{L}_{K_1,M_1}u(x)\right| \\ &+ \left|\mathcal{L}_{K_2,M_2,}u(x) - \mathcal{L}_{K_2,M_2,\rho,\mathbf{x}}u(x)\right| \\ &+ \left|\mathcal{L}_{K_2-K_1,M_2-M_1,}u(x)\right| \\ &\leq N^{-\frac{1}{2}}\epsilon^{-\frac{1}{2}}\left\|\frac{K_1(x,\cdot)^2}{\rho(\cdot)}\right\|_{\mathrm{L}^\infty(\mathcal{V})}\|u\|_{\mathrm{L}^2(\mathcal{V})} \\ &+ N^{-\frac{1}{2}}\epsilon^{-\frac{1}{2}}\left\|\frac{K_2(x,\cdot)^2}{\rho(\cdot)}\right\|_{\mathrm{L}^\infty(\mathcal{V})}\|u\|_{\mathrm{L}^2(\mathcal{V})} \\ &+ c_{K_1-K_2}(x)\,\|u\|_{\mathrm{L}^2(\mathcal{V})} + c_{M_1-M_2}(x)|u(x)|\,,\end{aligned}$$

Example 5.1 (Interpretable Bounds)**.** Let ρ be the uniform density function, fix the random sample $\mathbf{x}$ and consider two different K-Laplacian $\mathcal{L}_{K_1}$, $\mathcal{L}_{K_2}$; the corresponding Random Sampled Laplacians are

$$\mathcal{L}_{K_i,1,\mathbf{x}}\mathbf{u} = \frac{1}{N}(\mathbf{K}_i - \mathrm{diag}(\mathbf{K}_i\mathbf{1}))\mathbf{u}\,,\ i\in\{1,2\}\,.$$

Due to the fact that only the kernel changes, the two graph approximations differ only edge-wise. It is natural to define $\mathcal{E}_a$ as the set of added edges, $\mathcal{E}_d$ as the set of deleted edges, $\mathcal{E}_m$ as the set of modified edges and $\mathcal{E}_u$ as the set of unmodified edges

$$\begin{aligned}\mathcal{E}_a &= \{(x_i,x_j) : K_1(x_i,x_j) = 0, K_2(x_i,x_j)\neq 0\} \\ \mathcal{E}_d &= \{(x_i,x_j) : K_1(x_i,x_j) \neq 0, K_2(x_i,x_j) = 0\} \\ \mathcal{E}_m &= \{(x_i,x_j) : K_1(x_i,x_j) \neq 0, K_2(x_i,x_j)\neq 0, K_1(x_i,x_j)\neq K_2(x_i,x_j)\} \\ \mathcal{E}_u &= \{(x_i,x_j) : K_1(x_i,x_j) = K_2(x_i,x_j)\}\end{aligned}$$

The set $\{\mathcal{E}_a,\mathcal{E}_d,\mathcal{E}_m,\mathcal{E}_u\}$ is a partition of $\mathcal{V}_N\times\mathcal{V}_N$; using Theorem 3.6 the difference between the two random sampled Laplacians can be decomposed as

$$\begin{aligned}\Delta &= \frac{1}{N}\left((\mathbf{K}_1-\mathbf{K}_2) - \mathrm{diag}\left((\mathbf{K}_1-\mathbf{K}_2)\mathbf{1}\right)\right) \\ &= -\sum_{(x_i,x_j)\in\mathcal{E}_a} K_2(x_i,x_j)\mathbf{E}_{ij} + \sum_{(x_i,x_j)\in\mathcal{E}_d} K_1(x_i,x_j)\mathbf{E}_{ij}\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{(x_i,x_j)\in\mathcal{E}_m} (K_1(x_i,x_j) - K_2(x_i,x_j))\mathbf{E}_{ij} \\
&= \sum_{(x_i,x_j)\in\mathcal{E}_a\cup\mathcal{E}_d\cup\mathcal{E}_m} (K_1(x_i,x_j) - K_2(x_i,x_j))\mathbf{E}_{ij}\,,
\end{aligned}
$$

where $\mathbf{E}_{ij} = (\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^{\mathrm{T}}$ subtracts the weight of an edge from position (i,j) and (j,i) and adds it to the diagonal position (i,i) and (j,j). The sum of the rows of $\mathbf{E}_{ij}$ is 0; moreover, $\mathbf{E}_{ij}$ is a symmetric matrix, hence, the perturbation matrix

$$
\mathbf{E} = \sum_{(x_i,x_j)\in\mathcal{E}_a\cup\mathcal{E}_d\cup\mathcal{E}_m} \mathbf{E}_{ij}\,,
$$

preserves all the properties and by Gershgorin's Circle Theorem its spectral radius is less or equal than

$$
\max_i \left| [\mathbf{E}]_{ii} + \operatorname{sign}\left([\mathbf{E}]_{ii}\right) \sum_{j\neq i} |[\mathbf{E}]_{ij}| \right| .
$$

The difference $K_1(x_i,x_j) - K_2(x_i,x_j)$ is bounded by the perturbation constant $p = \max_{i,j}|K_1(x_i,x_j) - K_2(x_i,x_j)|$, hence

$$
\|\Delta\| \leq p\|\mathbf{E}\| \leq \begin{cases} p \max\limits_i \left| [\mathbf{E}]_{ii} + \operatorname{sign}\left([\mathbf{E}]_{ii}\right) \sum\limits_{j\neq i} |[\mathbf{E}]_{ij}| \right| \\ 2\,p\,|\mathcal{E}_a \cup \mathcal{E}_d \cup \mathcal{E}_m| \end{cases},
$$

where

$$
\operatorname{sign}(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases}.
$$

To show the difference between the two bounds, consider a line graph with $N = 3$ nodes; adding an edge between node 1 and node 3 generates a cycle graph. The perturbation matrix $\mathbf{E} = \mathbf{E}_{13}$ has spectral radius equal to 2. If now the edge between node 2 and node 3 is removed, a new line graph is generated. From the original line graph, the perturbation matrix $\mathbf{E} = \mathbf{E}_{13} - \mathbf{E}_{23}$ has spectral radius equal to $\sqrt{3}$. The perturbation matrix has the form

$$
\mathbf{E} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & -1 & 1 \\ -1 & 1 & 0 \end{pmatrix}
$$

hence the bound 2 is tighter than $2\,p\,|\mathcal{E}_a \cup \mathcal{E}_d| = 4$. Preliminarly, let's introduce the function

$$
\varphi\,:\,\mathbf{E} \mapsto \max_i \left| [\mathbf{E}]_{ii} + \operatorname{sign}\left([\mathbf{E}]_{ii}\right) \sum_{j\neq i} |[\mathbf{E}]_{ij}| \right| ,
$$

then, for polynomial filters it holds

$$\begin{aligned}\left\|h(\mathcal{L}_{K_1,1,\mathbf{x}}) - h(\mathcal{L}_{K_2,1,\mathbf{x}})\right\| &= \left\|\sum_{m=1}^{M} h_m \sum_{n=0}^{m-1} \mathcal{L}_{K_1,1,\mathbf{x}}^{m-n-1} \mathcal{L}_{K_1-K_2,1,\mathbf{x}} \mathcal{L}_{K_2,1,\mathbf{x}}^{n}\right\| \\ &\leq p\,\varphi(\mathbf{E}) \sum_{m=1}^{M} \left(\frac{1}{N}\right)^m |h_m| \sum_{n=0}^{m-1} \varphi(\mathbf{K}_1)^{m-n-1} \varphi(\mathbf{K}_2)^n \\ &\leq p\,\varphi(\mathbf{E}) \sum_{m=1}^{M} \left(\frac{1}{N}\right)^m |h_m|\, m\, \max\left\{\varphi(\mathbf{K}_1), \varphi(\mathbf{K}_2)\right\}^{m-1} .\end{aligned}$$

From the previous bound we can draw some conclusions: the change in the polynomial filter is small if the perturbation of the kernel p is small, or if $\varphi(\mathbf{E})$ is small, meaning that the perturbations are not concentrated on a single node. Moreover, from the last inequality, if nodes with small degree are perturbed, then $\varphi(\mathbf{K}_1)$ and $\varphi(\mathbf{K}_2)$ are similar, hence, the maximum between them does not increase much; viceversa, if nodes with high degree are perturbed, the maximum can increase leading to a looser bound. Similar results are drawn in [34].

Chapter 6

Conclusions and Future Developments

Justified by the inability to use the common algebraic characterizations of graphs to compare them, the theory developed so far allows to cast the problem in a continuous metric measure space. Exploiting the underlying geometry of a graph proved to be successful in the semi-supervised node classification task. This is not an isolated fact: usually great advances are made if the underlying geometry of a problem is recognized. The effort, then, is the correct identification of the latent space. It is well known that networks that show a high hierarchical structure, such as social and biological networks, are better represented by hyperbolic spaces [24]. A simple model for the hyperbolic space is the Poincaré disk, a disk centered at the origin with unit radius, equipped with a distance that increases exponentially towards the shell. In this context, the unit circle model is preliminary and preparatory to the analysis of the Poincaré disk. Indeed, while the sampling procedure of the former requires to sample an angle $\theta \in [-\pi, \pi)$, the sampling procedure of the second requires to sample an angle, as in the previous case, and a radius $r \in (0, 1)$. One challenge that could arise, that is also a core difference between hyperbolic spaces and spherical spaces, is the non-compactness of the former.

Another possible development of the theory would be the study of a broader class of spectral filters. While this work focuses mainly on polynomial spectral filters, in [18] spectral filters are defined by means of functional calculus [40]. Loosely speaking, functional calculus is the theory of applying a continuous function to an operator, a generalization of the theory of functions of matrices [41]. In order to give point-wise bounds, one should study under which conditions functional calculus preserves continuity, as defined in Definition 5.1.

Even though the theory presented in this work was developed in order to be able to compare graphs of different size, this aspect has not been properly analyzed.

The present work focuses mainly on edge perturbation as seen as a consequence of a kernel perturbation. The addition or deletion of nodes can be thought as a consequence of a different sampling procedure. However, the problem can also be reformulated as a kernel perturbation problem. For instance, a deleted node is a node that loses all of its connections and becomes isolated, while an added node can be thought as an isolated node that suddenly builds connections.

Future developments of the current work, as they are presented in this section, are meant to explore which are the best models to represent real networks. Spatial networks are known to be highly modular and assortative, features that are also shared by social networks. Hence, such generative model, combined with the correct choice of latent space and metric measure Laplacian, will hopefully prove to boost performance of ConvGNNs.

Bibliography

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. «ImageNet Classification with Deep Convolutional Neural Networks». In: *Communications of the ACM* 60.6 (May 2017), pp. 84–90. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3065386 (cit. on p. 1).

[2] Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. «Geometric Deep Learning: Going beyond Euclidean Data». In: *IEEE Signal Processing Magazine* 34.4 (July 2017), pp. 18–42. ISSN: 1053-5888, 1558-0792. DOI: 10.1109/MSP.2017.2693418 (cit. on p. 1).

[3] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. «Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges». In: *arXiv:2104.13478 [cs, stat]* (May 2021). arXiv: 2104.13478 [cs, stat] (cit. on p. 1).

[4] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. «A Comprehensive Survey on Graph Neural Networks». In: *IEEE Transactions on Neural Networks and Learning Systems* 32.1 (Jan. 2021), pp. 4–24. ISSN: 2162-237X, 2162-2388. DOI: 10.1109/TNNLS.2020.2978386. arXiv: 1901.00596 (cit. on p. 1).

[5] William L. Hamilton. «Graph Representation Learning». In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 14.3 (Sept. 2020), pp. 1–159. ISSN: 1939-4608, 1939-4616. DOI: 10.2200/S01045ED1V01Y202009AIM046 (cit. on p. 1).

[6] Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackermann, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. «A Deep Learning Approach to Antibiotic Discovery». In: *Cell* 180.4 (Feb. 2020), 688–702.e13. ISSN: 00928674. DOI: 10.1016/j.cell.2020.01.021 (cit. on p. 2).

[7] Rubal Kanozia and Ritu Arya. «"Fake News", Religion, and COVID-19 Vaccine Hesitancy in India, Pakistan, and Bangladesh». In: *Media Asia* 48.4 (Oct. 2021), pp. 313–321. ISSN: 0129-6612. DOI: 10.1080/01296612.2021.1921963 (cit. on p. 2).

[8] Marc Fisher, John Woodrow Cox, and Peter Hermann. «Pizzagate: From Rumor, to Hashtag, to Gunfire in D.C.» In: *Washington Post* (2016-12-06T08:34-500). ISSN: 0190-8286 (cit. on p. 2).

[9] «In Myanmar, Fake News Spread on Facebook Stokes Ethnic Violence». In: *The World from PRX* () (cit. on p. 2).

[10] Sheera Frenkel. «Lies on Social Media Inflame Israeli-Palestinian Conflict». In: *The New York Times* (May 2021). ISSN: 0362-4331 (cit. on p. 2).

[11] Marco Gori, Gabriele Monfardini, and Franco Scarselli. «A New Model for Earning In». In: (), p. 6 (cit. on p. 3).

[12] F. Scarselli, M. Gori, Ah Chung Tsoi, M. Hagenbuchner, and G. Monfardini. «The Graph Neural Network Model». In: *IEEE Transactions on Neural Networks* 20.1 (Jan. 2009), pp. 61–80. ISSN: 1045-9227, 1941-0093. DOI: 10.1109/TNN.2008.2005605 (cit. on p. 3).

[13] Thomas N. Kipf and Max Welling. «Semi-Supervised Classification with Graph Convolutional Networks». In: *arXiv:1609.02907 [cs, stat]* (Feb. 2017). arXiv: 1609.02907 [cs, stat] (cit. on pp. 3, 22).

[14] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. «Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering». In: *arXiv:1606.09375 [cs, stat]* (Feb. 2017). arXiv: 1606.09375 [cs, stat] (cit. on pp. 3, 22).

[15] Ron Levie, Federico Monti, Xavier Bresson, and Michael M. Bronstein. «CayleyNets: Graph Convolutional Neural Networks With Complex Rational Spectral Filters». In: *IEEE Transactions on Signal Processing* 67.1 (Jan. 2019), pp. 97–109. ISSN: 1053-587X, 1941-0476. DOI: 10.1109/TSP.2018.2879624 (cit. on p. 3).

[16] László Lovász. *Large Networks and Graph Limits*. Vol. 60. Colloquium Publications. Providence, Rhode Island: American Mathematical Society, Dec. 2012. ISBN: 978-0-8218-9085-1 978-1-4704-1583-9. DOI: 10.1090/coll/060 (cit. on p. 3).

[17] Luana Ruiz, Luiz F. O. Chamon, and Alejandro Ribeiro. «Graphon Neural Networks and the Transferability of Graph Neural Networks». In: *arXiv:2006.03548 [cs, stat]* (Oct. 2020). arXiv: 2006.03548 [cs, stat] (cit. on pp. 3, 59).

[18] Ron Levie, Wei Huang, Lorenzo Bucci, Michael M. Bronstein, and Gitta Kutyniok. «Transferability of Spectral Graph Convolutional Neural Networks». In: *arXiv:1907.12972 [cs, stat]* (Mar. 2020). arXiv: 1907.12972 [cs, stat] (cit. on pp. 3–5, 27, 59, 66).

[19] Mathew Penrose. *Random Geometric Graphs*. Oxford Studies in Probability 5. Oxford ; New York: Oxford University Press, 2003. ISBN: 978-0-19-850626-3 (cit. on pp. 4, 29).

[20] Ron Levie, Elvin Isufi, and Gitta Kutyniok. «On the Transferability of Spectral Graph Filters». In: *arXiv:1901.10524 [cs, stat]* (Jan. 2019). arXiv: 1901.10524 [cs, stat] (cit. on pp. 4, 59).

[21] Ruoyu Li, Sheng Wang, Feiyun Zhu, and Junzhou Huang. «Adaptive Graph Convolutional Neural Networks». In: *arXiv:1801.03226 [cs, stat]* (Jan. 2018). arXiv: 1801.03226 [cs, stat] (cit. on p. 5).

[22] George Dasoulas, Johannes Lutzeyer, and Michalis Vazirgiannis. «Learning Parametrised Graph Shift Operators». In: *arXiv:2101.10050 [cs, stat]* (Apr. 2021). arXiv: 2101.10050 [cs, stat] (cit. on pp. 5, 39).

[23] Marian Boguna, Ivan Bonamassa, Manlio De Domenico, Shlomo Havlin, Dmitri Krioukov, and M. Angeles Serrano. «Network Geometry». In: *Nature Reviews Physics* 3.2 (Feb. 2021), pp. 114–135. ISSN: 2522-5820. DOI: 10.1038/s42254-020-00264-4. arXiv: 2001.03241 (cit. on p. 5).

[24] Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguñá. «Hyperbolic Geometry of Complex Networks». In: *Physical Review E* 82.3 (Sept. 2010), p. 036106. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.82.036106 (cit. on pp. 5, 66).

[25] Wilhelmus Hendricus Schikhof. *Ultrametric Calculus: An Introduction to p-Adic Analysis*. Cambridge Studies in Advanced Mathematics 4. Cambridge [Cambridgeshire] ; New York: Cambridge University Press, 1984. ISBN: 978-0-521-24234-9 (cit. on p. 7).

[26] David H. Fremlin. *Measure Theory. 1: The Irreducible Minimum*. 2. ed. Colchester: Torres Fremlin, 2011. ISBN: 978-0-9538129-8-1 (cit. on p. 10).

[27] Frank Jones. *Lebesgue Integration on Euclidean Space*. Rev. ed. Jones and Bartlett Books in Mathematics. Sudbury, Mass: Jones and Bartlett, 2001. ISBN: 978-0-7637-1708-7 (cit. on p. 10).

[28] Pertti Mattila. *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability*. Cambridge Studies in Advanced Mathematics 44. Cambridge [England] ; New York: Cambridge University Press, 1995. ISBN: 978-0-521-46576-2 (cit. on p. 11).

[29] Dmitri Burago, Sergei Ivanov, and Yaroslav Kurylev. «Spectral Stability of Metric-Measure Laplacians». In: *arXiv:1506.06781 [math]* (Aug. 2018). arXiv: 1506.06781 [math] (cit. on p. 26).

[30] A. N. Shiryayev. *Probability*. Vol. 95. Graduate Texts in Mathematics. New York, NY: Springer New York, 1984. ISBN: 978-1-4899-0020-3 978-1-4899-0018-0. DOI: 10.1007/978-1-4899-0018-0 (cit. on p. 28).

[31] Fan R. K. Chung. *Spectral Graph Theory*. Regional Conference Series in Mathematics no. 92. Providence, R.I: Published for the Conference Board of the mathematical sciences by the American Mathematical Society, 1997. ISBN: 978-0-8218-0315-8 (cit. on p. 35).

[32] Henry Kenlay, Dorina Thanou, and Xiaowen Dong. «On the Stability of Polynomial Spectral Graph Filters». In: (), p. 5 (cit. on p. 59).

[33] Henry Kenlay, Dorina Thanou, and Xiaowen Dong. «On the Stability of Graph Convolutional Neural Networks under Edge Rewiring». In: *arXiv:2010.13747 [cs]* (Feb. 2021). arXiv: 2010.13747 [cs] (cit. on p. 59).

[34] Henry Kenlay, Dorina Thanou, and Xiaowen Dong. «Interpretable Stability Bounds for Spectral Graph Filters». In: *arXiv:2102.09587 [cs]* (Feb. 2021). arXiv: 2102.09587 [cs] (cit. on pp. 59, 65).

[35] Nicolas Keriven, Alberto Bietti, and Samuel Vaiter. «Convergence and Stability of Graph Convolutional Networks on Large Random Graphs». In: *arXiv:2006.01868 [cs, stat]* (Oct. 2020). arXiv: 2006.01868 [cs, stat] (cit. on p. 59).

[36] Nicolas Keriven, Alberto Bietti, and Samuel Vaiter. «On the Universality of Graph Neural Networks on Large Random Graphs». In: (), p. 29 (cit. on p. 59).

[37] Fernando Gama, Joan Bruna, and Alejandro Ribeiro. «Stability of Graph Neural Networks to Relative Perturbations». In: *arXiv:1910.09655 [cs, eess, stat]* (Oct. 2019). arXiv: 1910.09655 [cs, eess, stat] (cit. on p. 59).

[38] Fernando Gama, Joan Bruna, and Alejandro Ribeiro. «Stability Properties of Graph Neural Networks». In: *IEEE Transactions on Signal Processing* 68 (2020), pp. 5680–5695. ISSN: 1053-587X, 1941-0476. DOI: 10.1109/TSP.2020.3026980. arXiv: 1905.04497 (cit. on p. 59).

[39] Luana Ruiz, Luiz F. O. Chamon, and Alejandro Ribeiro. «Graphon Signal Processing». In: *arXiv:2003.05030 [eess]* (Mar. 2021). arXiv: 2003.05030 [eess] (cit. on p. 59).

[40] Gilbert Helmberg. *Introduction to Spectral Theory in Hilbert Space*. 2. printing. North-Holland Series in Applied Mathematics and Mechanics 6. Amsterdam: North-Holland Publ, 1975. ISBN: 978-0-444-10822-7 978-0-7204-2356-3 (cit. on p. 66).

[41] Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Philadelphia: Society for Industrial and Applied Mathematics, 2008. ISBN: 978-0-89871-646-7 (cit. on p. 66).