Master's Thesis

Mathematical Theory of Spectral Clustering

July 2025

Presented by: Vred Rudnick

Student ID: 108022111153

vred.rudnick@ruhr-uni-bochum.de

First Examiner: Prof. Dr. Holger Dette

Second Examiner: Prof. Dr. Axel Bücher

Faculty of Mathematics

Ruhr University Bochum

Contents

1 Introduction

Clustering has generated significant interest in data analysis and machine learning. Its applications range from customer segmentation to social network analysis (Aggarwal and Reddy 2014). Clustering techniques are commonly subdivided into hierarchical and partitional algorithms (Ezugwu et al. 2022). Hierarchical methods follow an iterative structure, which either joins the data points from the bottom up (*agglomerative*) or divides them from the top down (*divisive*) (Saxena et al. 2017). Notable examples include agglomerative techniques such as linkage algorithms (Sibson 1973; Defays 1977; Day and Edelsbrunner 1984), and divisive techniques such as monothetic and polythetic clustering (Chavent 1998; Kim and Billard 2011). Partitional methods, on the other hand, group the data into a predefined number of subsets by optimizing a criterion function on the data (Saxena et al. 2017). These techniques take various forms, with Gaussian mixture models (Banfield and Raftery 1993), k-means clustering (LLoyd 1982; Hartigan and Wong 1979; MacQueen 1967), and fuzzy clustering (Bezdek, Ehrlich, and Full 1984) being notable examples. The focus in this work is on spectral clustering, which represents another example of a partitional algorithm. Spectral clustering has been proposed as an alternative to common techniques such as k-means and others. To group the data into a given number of subsets, the algorithm arranges the data points in a similarity graph and selects the partition that minimizes a cut measure of said graph. This is achieved by computing the eigenvectors to the smallest eigenvalues of a graph Laplacian. The technique offers key advantages, such as the potential to find non-convex data clusters. Notable works on the subject include contributions by Donath and Hoffman (1973), Fiedler (1973), Shi and Malik (2000), Ng, Jordan, and Weiss (2001), and von Luxburg (2007). This work examines the algorithm's ability to recognize inherent cluster structures in data. For this purpose, we study convergence results for eigenvalues and eigenvectors of graph Laplacians for certain similarity graphs, building on the approach provided by von Luxburg, Belkin, and Bousquet (2008).

More specifically, suppose that we draw a sequence of i.i.d. data points $x_1, ..., x_n, ...$ from a compact metric space $\mathcal{X}$ according to a probability distribution P. For $n \in \mathbb{N}$, we interpret the individual elements of the sample $x_1, ..., x_n$ as vertices of an undirected similarity graph. To each pair of points x_i, x_j, we quantify their similarity using a symmetric similarity function

$$\begin{cases} k : \mathcal{X} \times \mathcal{X} \longrightarrow [0, \infty) \\ (x, y) \longmapsto k(x, y) \end{cases}$$

This defines the $(n \times n)$-matrix $K_n = (k(x_i, x_j))_{i,j=1}^n$, which we call the similarity matrix or kernel matrix, and the degree matrix $D_n = \text{diag}[\deg(1), ..., \deg(n)]$, where the degree of a vertex is $\deg(i) = \sum_{i \neq j} w_{ij}$. Assuming we have a given $k < n$, spectral clustering algorithms group the data into k disjoint subsets based on the eigenvectors to the k smallest eigenvalues of the data graph's Laplacian matrices

$$L_n = D_n - K_n$$

and

$$L'_n = I_n - D_n^{-1/2} K_n D_n^{-1/2}$$

where I_n denotes the identity matrix in $\mathbb{R}^{n \times n}$. We call L_n the unnormalized Laplacian and L'_n the normalized Laplacian of the data graph. We refer to clustering based on the eigenvectors of L_n as unnormalized spectral clustering, and to clustering based on those of L'_n as normalized spectral clustering. Our objective throughout the following sections is to study whether there is a sense in which a sequence of k-th eigenvectors (eigenvectors to the k-th smallest eigenvalue) $(v_n)_{n \in \mathbb{N}}$ of $(L'_n)_{n \in \mathbb{N}}$ "converges" for $n \to \infty$. The meaning of this is not straightforward because the dimension of v_n grows with the sample size. In establishing a definition of such "convergence" or "consistency", we follow the approach presented by (von Luxburg, Belkin, and Bousquet 2008) in saying that $(v_n)_{n \in \mathbb{N}}$ converges to a continuous function $f \in C(\mathcal{X})$ if the convergence

$$\max_{i=1,\dots,n} |v_{n,i} - f(x_i)| \longrightarrow 0 \tag{1.1}$$

holds almost surely. Here, the limit function f contains information about the underlying distribution of the data. We will see that, under certain conditions, (1.1) holds true up to changes in orientation of $(v_n)_{n \in \mathbb{N}}$. This, however, requires stricter assumptions for unnormalized compared to normalized spectral clustering. In discussing all of the above, we proceed as follows: in Section 3, we examine spectral clustering in detail, which includes its graph-theoretical foundations, a brief review of the relationship between random walks and spectral clustering, and the stochastic block model. Additionally, we provide an example of practical implementation and view spectral clustering through the lens of Davis-Kahan pertubation theory. Section 4 includes the main results on eigenvector convergence in the sense of (1.1) building on the work presented by von Luxburg, Belkin, and Bousquet (2008): we discuss convergence rates and highlight the differences between normalized and unnormalized spectral clustering. Additionally, we provide simulations to visualize eigenvector convergence of the type (1.1). Before all this, we need to introduce various mathematical concepts that are foundational. First, we need to discuss linear operators on Banach spaces. This is because, to prove convergence of the type (1.1), we construct a sequence $(f_n)_{n \in \mathbb{N}}$ of continuous functions that satisfies

$$v_n = (f_n(x_1), ..., f_n(x_n))^T$$

which allows us to show

$$\|f_n - f\|_\infty \longrightarrow 0 \tag{1.2}$$

instead of (1.1) because

$$\max_{i=1,\dots,n} |v_{n,i} - f(x_i)| \leq \|f_n - f\|_\infty$$

Our construction of $(f_n)_{n \in \mathbb{N}}$ is achieved through defining a sequence of bounded linear operators $(U'_n)_{n \in \mathbb{N}}$ on $(C(\mathcal{X}), \|\cdot\|_\infty)$, of which $(f_n)_{n \in \mathbb{N}}$ are eigenfunctions. We will see that (1.2) is tied to the convergence of $(U'_n)_{n \in \mathbb{N}}$ to a limit operator. This justifies a discussion of operators on Banach spaces, their convergence behavior, and how it relates to spectral

convergence. Some types of operator convergence, including compact convergence, require a review of the Arzela-Ascoli theorem. In addition, addressing empirical process theory and Glivenko-Cantelli theorems is relevant because $(U_n')_{n\in\mathbb{N}}$ are integral operators. We furthermore include a discussion of the Rayleigh-Ritz principal: this is crucial for proving that the classification task in spectral clustering, as presented in Section 3, relates to the smallest Laplacian eigenvalues and their eigenvectors. All of the aforementioned subjects are featured in Section 2, titled *Mathematical Preliminaries*.

2 Mathematical Preliminaries

Discussing the consistency of Spectral Clustering algorithms, as introduced by von Luxburg, Belkin, and Bousquet (2008), requires concepts from various fields of mathematics that we present below. They include: compact subsets of Banach spaces (Section 2.1), compact linear operators on Banach spaces (Section 2.2), a brief presentation of empirical processes, covering numbers and bracketing numbers and Glivenko-Cantelli classes (Section 2.4), and, lastly, a discussion of symmetric matrices and how their eigenvalues relate to Rayleigh quotients. (Section 2.5).

2.1 Topology, Banach Spaces and Compactness

In the following, we discuss foundational topological concepts such as compactness in metric spaces. The theory we discuss is mainly borrowed from Dunford and Schwartz (1958, Chapters I,IV). Suppose that (X, τ) is a topological space (see Heuser (1975, p. 342)). Recall that any subset Y of X is a topological space with respect to the topology

$$\tau_{|Y} \coloneqq \{T \in \tau : T \subset Y\}$$

DEFINITION 2.1. *We call a family $\mathcal{U}$ of subsets of X a cover of Y if and only if*

$$Y \subset \bigcup_{U \in \mathcal{U}} U$$

A subcover of $\mathcal{U}$ is a subset $\mathcal{V}$ of $\mathcal{U}$ which is also a cover of $\mathcal{U}$.

Next, we define compact and relatively compact sets.

DEFINITION 2.2. *A set $K \subset X$ is called compact if and only if every cover of K has a finite subcover. A subset $K \subset X$ is called relatively compact (or conditionally compact) if and only if $\bar{K}$ is compact.*

Continuous functions between topological spaces are those for which the image of any open set is an open set. The image of a compact space under a continuous function is itself a compact space. For any topological space X, we denote the set of all real, continuous functions on X by $C(X)$. Suppose that $f \in C(X)$. We now define uniformly continuous functions, an important subset of continuous functions on metric spaces (see Heuser (1975, p. 252)).

DEFINITION 2.3. *Let (X, d_X) and (Y, d_Y) be metric spaces. A function $f : X \to Y$ is called uniformly continuous if for every $\epsilon > 0$ and for all $x, x' \in X$ there exists a $\delta > 0$ such that*

$$d_X(x, x') < \delta \Longrightarrow d_Y(f(x), f(x')) < \epsilon \tag{2.1}$$

Note that the condition in (2.1) is a globalized version of the local epsilon-delta criterion in the definition of continuous functions between metric spaces. The following statement about uniform continuity is useful and can be proven with a simple argument.

PROPOSITION 2.4. *A continuous function on a compact (metric) domain is uniformly continuous.*

We proceed with an exploration of Hausdorff spaces. Suppose that (X, τ) is any topological space. We begin by defining the concept of a neighborhood.

DEFINITION 2.5. *Suppose $x \in X$. A subset $U \subset X$ is called a neighborhood of x if and only if U contains an open set with respect to the topology τ such that $x \in S$. For a subset $A \subset X$, we call $V \subset X$ a neighborhood of A if it is a neighborhood of each individual element of A.*

The concept of a neighborhood allows us to define isolated points:

DEFINITION 2.6. *Let A be a subset of X. A subset $B \subset A$ is called isolated if there exists a neighborhood U of B such that*

$$U \cap A = B$$

A single point $x \in A$ is called isolated if $\{x\}$ is an isolated subset. If x is not isolated in A it is called an accumulation point of A.

Neighborhoods are also a key concept for the definition of Hausdorff spaces:

DEFINITION 2.7. *(X, τ) is called a Hausdorff space if and only if for every two points $x_1, x_2 \in X$ there exist neighborhoods U_1 and U_2 of x_1, x_2, respectively, such that*

$$U_1 \cap U_2 = \varnothing$$

Suppose that (X, d) is a metric space. Recall that any metric space is a topological space with respect to the topology induced by the metric d. In few steps, we will prove that X is Hausdorff.

PROPOSITION 2.8. *Let (X, d) be a metric space. Then, X is Hausdorff with respect to the topology that d induces.*

Proof. The statement is trivial if x is the only element in X. In case X contains more than one element, we prove that for every distinct pair $x, x' \in X$, there exists an $\epsilon > 0$ such that

$$B_\epsilon(x) \cap B_\epsilon(x') = \varnothing \tag{2.2}$$

Because d is a metric, we have $d(x, x') > 0$. We define $\Delta := d(x, x')$. Choose $\epsilon > 0$ such that $\epsilon < \frac{\Delta}{2}$. We can assume that $B_\epsilon(x)$ and $B_\epsilon(x')$ contain more points besides x, x', respectively. Otherwise the statement is trivial. Let a be in $B_\epsilon(x)$. We show that $a \notin B_\epsilon(x')$. The triangular inequality yields:

$$d(x', a) \geq d(x, x') - d(x, a)$$

$$d(x', a) \geq 2\epsilon - \epsilon$$

which implies $a \notin B_\epsilon(x')$. We have proven that $B_\epsilon(x)$ contains no points that are in $B_\epsilon(x')$, which proves (2.2). □

Suppose that (X, τ) is a compact Hausdorff space. Let $C(X)$ denote the set of all continuous functions from X to $\mathbb{R}$. Let $\|\cdot\|_\infty$ denote the uniform norm on $C(X)$.

DEFINITION 2.9. *$\mathcal{F} \subset C(X)$ is called equicontinuous if and only if for every $x \in X$, and for every $\epsilon > 0$ there exists a neighborhood U_x of x such that*

$$|f(y) - f(x)| < \epsilon$$

for all $y \in U_x$ and all $f \in \mathcal{F}$.

Suppose that (X, τ) is a compact Hausdorff space. With the help of the aforementioned definitions and statements, we present an important statement which characterizes relatively compact families of continuous functions on X.

THEOREM 2.10 (Arzela-Ascoli Theorem)**.** *The relative compactness of any family $\mathcal{F} \subset C(X)$ is equivalent to the following two conditions:*

(i) $\mathcal{F}$ is bounded i.e. $\sup_{f \in \mathcal{F}} \|f\|_\infty < \infty$

(ii) $\mathcal{F}$ is equicontinuous.

For detailed discussions and a proof, we refer to Dunford and Schwartz (1958, p. 266-267).

2.2 Linear Operators and Spectral Convergence

We now turn our attention to bounded linear operators on Banach spaces: we discuss spectral properties, compact operators as well as different types of operator convergence. The theory we present is mainly adapted from Kato (1966) and Dunford and Schwartz (1958). Suppose that $(E, \|\cdot\|_E)$ is a normed vector space over $\mathbb{R}$. Recall that E is called complete with respect to $\|\cdot\|_E$ if and only if every Cauchy sequence in E converges with respect to $\|\cdot\|_E$.

DEFINITION 2.11. *$(E, \|\cdot\|_E)$ is called a Banach space if E is complete with respect to $\|\cdot\|_E$. In this case, we also say that E is a Banach space.*

The most important example of a Banach space for our purposes is the function space $C(K)$ with respect to the uniform norm $\|\cdot\|_\infty$, where K is a compact Hausdorff space. Throughout the following section, let $(E, \|\cdot\|_E)$ denote a Banach space. As for finite-dimensional vector spaces, a map $O : E \to E$ is called a linear map if

$$O(e + f) = Oe + Of$$

and

$$O(\lambda e) = \lambda \cdot Oe$$

hold for all $e, f \in E$ and for all $\lambda \in \mathbb{R}$ (see Heuser (1975, p. 78)).

DEFINITION 2.12. *A linear map $O : E \longrightarrow E$ is called a bounded linear operator if there exists a constant $C > 0$ such that*

$$\|Oe\|_E \leq C\|e\|_E \tag{2.3}$$

for all $e \in E$. We denote the set of all bounded linear operators on E as $\mathbb{B}(E)$.

Note that the set $\mathbb{B}(E)$ is a vector space over $\mathbb{R}$. The function

$$\|\cdot\|_{\text{op}} : \begin{cases} \mathbb{B}(E) \longrightarrow [0,\infty) \\ O \longmapsto \sup\limits_{\|e\|_E=1} \|Oe\|_E \end{cases} \tag{2.4}$$

is called the operator norm on the vector space $\mathbb{B}(E)$. Note that $\|O\|_{\text{op}}$ is a finite number for all $O \in \mathbb{B}(E)$ because of (2.3). $(\mathbb{B}(E), \|\cdot\|_{\text{op}})$ is itself a Banach space (see Kato (1966, Section III.3)). An important subclass of bounded linear operators are compact operators:

DEFINITION 2.13. *An operator $S \in \mathbb{B}(E)$ is called a compact operator if for any bounded sequence $(u_n)_{n\in\mathbb{N}}$ in E, the image $(Su_n)_{n\in\mathbb{N}}$ contains a subsequence that converges (or, equivalently, is a Cauchy sequence).*

Equivalently, compact operators can be characterized as follows:

PROPOSITION 2.14. *Let B_1 be the unit ball of E with respect to $\|\cdot\|_E$. An operator $S \in \mathbb{B}(E)$ is compact if and only if the image of B_1 under S is relatively compact in E.*

This statement can be proven with a simple argument about compact metric spaces (see Heuser (1975, p. 62)). Suppose that X is a compact Hausdorff space. Let $k : X \times X \to \mathbb{R}$ be a symmetric, continuous function. Let μ be a probability measure on $(X, \mathcal{B}(X))$, where $\mathcal{B}(X)$ is the Borel σ-algebra of X. Consider the integral operator

$$\mathcal{I} : C(X) \longrightarrow C(X)$$

$$\mathcal{I}f(x) = \int_X k(x,y) f(y) d\mu(y) \tag{2.5}$$

For integral operators of this type, we have the following result (see Kato (1966, Examples 2.4,4.1).

PROPOSITION 2.15. *The integral operator $\mathcal{I}$ defined in (2.5) is a compact operator.*

Suppose that $S \in \mathbb{B}(E)$. An eigenvalue of the operator S is any number $\lambda \in \mathbb{C}$ for which there exists an element $e \in E$ such that

$$Se = \lambda e$$

holds. Let I be the identity operator. Note that I is bounded because it is isometric. We define the resolvent set, and the spectrum of S.

DEFINITION 2.16. *The resolvent set of $S \in \mathbb{B}(E)$, denoted by $\varrho(S)$ is the set of all $\lambda \in \mathbb{C}$ for which the operator*

$$S - \lambda I$$

is invertible and its inverse is an element of $\mathbb{B}(E)$. In this case, the function

$$R : \begin{cases} \varrho(S) \longrightarrow \mathbb{B}(E) \\ \lambda \longmapsto R(\lambda) = (S - \lambda I)^{-1} \end{cases}$$

is called the resolvent of S. The spectrum $\sigma(S)$ of S is defined as $\varrho(S)^c$.

By definition, $\sigma(S)$ contains all of the eigenvalues of S. However $\sigma(S)$ does not necessarily only include eigenvalues of S as when the dimension of E is infinite. For elements of $\sigma(S)$, we can define isolation and multiplicity. We call a subset of $\sigma(T)$ isolated if it is isolated according to Definition 2.6. As for matrices in $\mathbb{R}^{n\times n}$, we can define the algebraic and geometric multiplicities of an eigenvalue $S \in \mathbb{B}(E)$. Let $\mathrm{N}_\lambda := \ker(S - \lambda I)$ be the nullspace of $S - \lambda I$. If λ is part of the spectrum of S, then N_λ is not $\{0\}$, and subsequently

$$\dim(\mathrm{N}_\lambda) \geq 1$$

DEFINITION 2.17. *For any eigenvalue $\lambda \in \sigma(S)$, we call $\dim(\mathrm{N}_\lambda)$ the geometric multiplicity of λ.*

This concept is analogous to the definition of geometric multiplicity for eigenvalaues of matrices. To define algebraic multiplicity for eigenvalues of an operator $S \in \mathbb{B}(E)$, we need to introduce some definitions.

DEFINITION 2.18. *An operator $P \in \mathbb{B}(E)$ is called idempotent if $P^2 = P$. An idempotent operator is called a projection.*

Recall that $I \in \mathbb{B}(E)$ is the identity operator. Any projection $P \in \mathbb{B}(E)$ defines a decomposition of E in two subspaces:

$$E = M \oplus N$$

where $M = PE$ and $N = (I - P)E$. In the same way, any decomposition $E = M \oplus N$ defines a projection, which we shall refer to as the projection of M along N. Next, we define the decomposition of an operator $S \in \mathbb{B}(E)$.

DEFINITION 2.19. *S is said to be decomposed into the subspaces $M, N \subset E$ if $SM \subset M$ and $SN \subset N$ hold. By S_M and S_N, we denote the restrictions of S on M, N, respectively.*

Suppose that $\sigma(S)$ can be partitioned into σ' and σ'' by a closed Jordan curve $\gamma \subset \varrho(S)$ such that σ' is in the interior, and σ'' a subset of the exterior of γ. For $S \in \mathbb{B}(E)$, the resolvent R is an operator-valued function. It maps $\lambda \in \varrho(S)$ to the inverse operator $(S-\lambda I)^{-1} \in \mathbb{B}(E)$. Recall that $(\mathbb{B}(E), \|\cdot\|_{\mathrm{op}})$ is a Banach space. Thus, the integral

$$\frac{1}{2\pi i}\int_\gamma R(\lambda)d\lambda =: P_{\sigma'} \in \mathbb{B}(E) \tag{2.6}$$

is well-defined (see Heuser (1975, Section XIII.97)). It can be proven that P_σ is idempotent, and thus a projection (see Kato (1966, Section III.4)). It can be shown that E can be decomposed into

$$E = M' \oplus M'' \tag{2.7}$$

where M' and M'' correspond to σ' and σ'', respectively, in the following way: σ' is the spectrum of $T_{M'}$, and σ'' is the spectrum of $T_{M''}$ (see Kato (1966, Chapter III, Theorem 6.17)). Assume that $\lambda \in \sigma(S)$ is an isolated point in the spectrum of S. This implies that

there exists an $\epsilon > 0$ such that the ϵ-sphere $\mathbb{S}_\epsilon(\lambda)$ contains no point of $\sigma(S)$. Then, we can define the projection

$$P_\lambda := \frac{1}{2\pi i} \int_{\mathbb{S}_\epsilon(\lambda)} R(\zeta) d\zeta$$

for the oriented Jordan-curve $\mathbb{S}_\epsilon(\lambda)$. Note that

$$\mathbb{S}_\epsilon(\lambda) := \{z \in \mathbb{C} : |z - \lambda| = \epsilon\}$$

denotes the ϵ-sphere with center λ. Recall that P_λ induces a decomposition of E into $M'_\lambda \oplus M''_\lambda$ in the sense of (2.7). We also call $P_{\sigma'}$ and P_λ the spectral projections of S on σ' and λ, respectively. We now present some additional definitions on the spectrum (Kato 1966; von Luxburg, Belkin, and Bousquet 2008).

DEFINITION 2.20. *If $\lambda \in \sigma(S)$ is also an eigenvalue of S we call the dimension of M'_λ the algebraic multiplicity of λ. Eigenvalues with algebraic multiplicity of 1 we shall refer to as simple eigenvalues. The subset $\sigma_d(S) \subset \sigma(S)$ which contains all isolated eigenvalues with finite algebraic multiplicity is called the discrete spectrum of S. Its complement in $\sigma(S)$ is called the essential spectrum of S, which we denote by $\sigma_{\text{ess}}(S)$.*

THEOREM 2.21 (Kato (1966), Chapter III, Theorem 6.26)**.** *Let $S \in \mathbb{B}(E)$ be a compact operator. Then $\sigma(S)$ consists of countably many, isolated eigenvalues with finite algebraic multiplicity, and 0 is the only accumulation point of $\sigma(S)$.*

DEFINITION 2.22. *$T \in \mathbb{B}(E)$ is called a compact perturbation of the identity operator I if T has the form*

$$T = I - K$$

for a compact operator $K : E \to E$.

Note that a compact perturbation of the identity is itself not a compact operator. However, its spectrum is closely related to the spectrum of K, as the following proposition states.

PROPOSITION 2.23. *Let $T \in \mathbb{B}(E)$ be a compact perturbation of the identity in the sense of Definition 2.22. The spectrum of $\sigma(T)$ is equal to the set*

$$1 - \sigma(K) := \{1 - \nu \mid \nu \in \sigma(K)\}$$

The statement follows directly from the definitions. We next discuss the convergence of bounded linear operator sequences and related spectral behavior. We begin with defining a few types of convergence for sequences in $\mathbb{B}(E)$ (see von Luxburg, Belkin, and Bousquet (2008, Definition 5)):

DEFINITION 2.24. *Let $(S_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{B}(E)$. We say $(S_n)_{n\in\mathbb{N}}$ converges pointwise to $S \in \mathbb{B}(E)$, which we denote by*

$$S_n \xrightarrow{p} S$$

if for all $e \in E$ the the sequence $\|S_n e - Se\|_E$ converges to 0 for $n \to \infty$.

Let $\|\cdot\|_{\text{op}}$ be the operator norm as defined in (2.4). We define convergence in operator norm as follows:

DEFINITION 2.25. *$(S_n)_{n\in\mathbb{N}} \subset \mathbb{B}(E)$ converges in operator norm to $S \in \mathbb{B}(E)$, which we denote by*

$$S_n \xrightarrow{\|\cdot\|_{\text{op}}} S$$

if the sequence $\|S_n - S\|_{\text{op}}$ converges to 0 for $n \to \infty$.

The next two types of operator convergence require some topological notions.

DEFINITION 2.26. *Let $(S_n)_{n\in\mathbb{N}}$ be a sequence of bounded linear operators on E. Let B_1 be the unit ball of E with respect to $\|\cdot\|_E$. We call the sequence $(S_n)_{n\in\mathbb{N}}$ collectively compact if the union $\bigcup_{n\in\mathbb{N}} S_n B_1$ is relatively compact. Additionally, $(S_n)_{n\in\mathbb{N}}$ converges*

(i) ***compactly*** *to $S \in \mathbb{B}(E)$ if it converges pointwise and if the sequence $((S - S_n)e_n)_{n\in\mathbb{N}}$ is relatively compact for any sequence $(e_n)_{n\in\mathbb{N}}$ in B_1. We denote compact convergence as follows:*

$$S_n \xrightarrow{c} S$$

(ii) ***collectively compactly*** *to $S \in \mathbb{B}(E)$ if it converges pointwise and if there exists a number $N \in \mathbb{N}$ such that the sequence $(S - S_n)_{n>N}$ is collectively compact. We denote the convergence by*

$$S_n \xrightarrow{cc} S$$

Suppose that $(S_n)_{n\in\mathbb{N}}$ does not converge compactly to a bounded linear operator S. This implies the existence of a sequence $(e_n)_{n\in\mathbb{N}} \subset B_1$ for which the closure $\overline{((S - S_n)e_n)_{n\in\mathbb{N}}}$ has at least one cover without a finite subcover. Since $(e_n)_{n\in\mathbb{N}}$ is a subset of B_1, it follows that $\bigcup_{n\in\mathbb{N}} (S - S_n) B_1$ has at least one cover without a finite subcover. This contradicts the assumption of collectively compact convergence, which means that $(S_n)_{n\in\mathbb{N}}$ does not converge collectively compactly either. This proves an important implication:

PROPOSITION 2.27. *Collectively compact convergence implies compact convergence.*

Next, we discuss the relationship between compact convergence and spectral convergence (see von Luxburg, Belkin, and Bousquet (2008, Section 4)).

DEFINITION 2.28. *Suppose that $(S_n)_{n\in\mathbb{N}}$ is a sequence in $\mathbb{B}(E)$, and let $S \in \mathbb{B}(E)$. Assume that λ is an isolated, simple eigenvalue of S, and let $M \subset \mathbb{C}$ denote an isolation neighborhood of λ. Let $(\lambda_n)_{n\in\mathbb{N}}$ be a sequence of respective eigenvalues of $(S_n)_{n\in\mathbb{N}}$ such $\lambda_n \in \sigma(S_n) \cap M$. Assume that*

$$\lambda_n \longrightarrow \lambda$$

for $n \to \infty$. Suppose that $(e_n)_{n\in\mathbb{N}}$ is a sequence in E such that for each $n \in \mathbb{N}$, e_n is an eigenvector of S_n to λ_n and suppose that $e \in E$ is an eigenvector of S corresponding to λ. We say that $(e_n)_{n\in\mathbb{N}}$ converges to e if

$$\|e_n - e\|_E \xrightarrow{n\to\infty} 0$$

holds. We say that $(e_n)_{n\in\mathbb{N}}$ *converges to* e *up to a change of sign if there is a sequence of signs* $(a_n)_{n\in\mathbb{N}}$, $a_n \in \{-1, 1\}$ *such that*

$$\|a_n e_n - e\|_E \overset{n\to\infty}{\longrightarrow} 0$$

The following result establishes sufficient conditions for spectral convergence in terms of operator convergence. The version we include stems from von Luxburg, Belkin, and Bousquet (2008), who refer to Chatelin (1983, Sections 3.6, 5.1) and Chatelin (1983, Proposition 3.18) for the underlying theory.

THEOREM 2.29. *Suppose that* $S \in \mathbb{B}(E)$. *Let* $(S_n)_{n\in\mathbb{N}}$ *be a sequence in* $\mathbb{B}(E)$ *such that* $S_n \overset{c}{\to} S$. *Let* $\lambda \in \sigma(S)$ *be an isolated eigenvalue of* S. *Let* M *be a neighborhood of* λ *such that* $\sigma(S) \cap M = \{\lambda\}$, *and let* $(\lambda_n)_{n\in\mathbb{N}}$ *be a sequence of eigenvalues of* $(S_n)_{n\in\mathbb{N}}$ *such that*

$$\lambda_n \in \sigma(S_n) \cap M$$

for $n \in \mathbb{N}$. *By*

$$P_n := P_{\sigma(S_n)\cap M}$$

we denote the spectral projections in the sense of (2.6). Then the following results hold:

(i) $(\lambda_n)_{n\in\mathbb{N}}$ *converges to* λ.

(ii) The spectral projection sequence $(P_n)_{n\in\mathbb{N}}$ *converges to* P_λ *pointwise.*

(iii) Let λ *be a simple eigenvalue of* S *and* $(v_n)_{n\in\mathbb{N}}$ *be a sequence of eigenvectors such that each* v_n *corresponds to* λ_n *for* $n \in \mathbb{N}$. *Then, there exists an eigenfunction* $f \in C(X)$ *of* S *corresponding to* λ *such that* $(v_n)_{n\in\mathbb{N}}$ *converges to* f *up to a change of sign.*

We modify Definition 2.28 for a special case of the above. Suppose that X is a compact metric space. Recall that $C(X)$ is the Banach space of all real, continuous functions on X and that $\|\cdot\|_\infty$ is the uniform norm on $C(X)$. Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of matrices such that for each $n \in \mathbb{N}$, $A_n \in \mathbb{R}^{n\times n}$. Let F be a bounded linear operator on $C(X)$, and let $\lambda \in \sigma(F)$ be a simple, isolated eigenvalue of F. Let $M \subset \mathbb{C}$ be a neighborhood of λ such that $\sigma(F) \cap M = \{\lambda\}$. Let $(\lambda_n)_{n\in\mathbb{N}}$ be a sequence of eigenvalues such that $\lambda_n \in \sigma(A_n) \cap M$ for $n \in \mathbb{N}$. Assume that $\lambda_n \to \lambda$ for $n \to \infty$. Let $(v_n)_{n\in\mathbb{N}}$ be a sequence of eigenvectors such that v_n is an eigenvector of A_n to λ_n for $n \in \mathbb{N}$. For a given sequence $(x_n)_{n\in\mathbb{N}}$, the sequence

$$r_n : \begin{cases} C(X) \longrightarrow \mathbb{R}^n \\ f \longmapsto r_n f := (f(x_1), ..., f(x_n))^T \in \mathbb{R}^n \end{cases}$$

is a sequence of restriction operators: the n-th element of $(r_n)_{n\mathbb{N}}$ restricts any function f to its evaluation on the first n elements of $(x_n)_{n\in\mathbb{N}}$.

DEFINITION 2.30. *Let* F *denote a bounded linear operator on* $(C(X), \|\cdot\|_\infty)$ *We say that* $(v_n)_{n\in\mathbb{N}}$ *converges to an eigenfunction* $f \in C(X)$ *of* F *under the restriction operator* $(r_n)_{n\in\mathbb{N}}$ *if*

$$\|v_n - r_n f\|_{\max,n} \overset{n\to\infty}{\longrightarrow} 0$$

$(v_n)_{n\in\mathbb{N}}$ converges to f under $(r_n)_{n\in\mathbb{N}}$ up to a change of sign if there is a sequence of signs $(a_n)_{n\in\mathbb{N}}$, $a_n \in \{-1, 1\}$ such that

$$\|a_n v_n - r_n f\|_{\max,n} \overset{n\to\infty}{\longrightarrow} 0$$

Here, $\|\cdot\|_{\max,n}$ denotes the maximum norm on $\mathbb{R}^n$.

Next, we introduce a statement which is relevant for convergence rates of the convergence stated in Theorem 2.29. Recall that the operator norm of a bounded linear operator $S \in \mathbb{B}(E)$ is given by

$$\|S\|_{\mathrm{op}} := \sup_{\|e\|_E=1} \|Se\|_E$$

Note that this one of multiple equivalent definitions of the operator norm, which is defined for a broader class of linear operators (see Heuser 1975, Section II.10). The following statement is originally due to Atkinson (1967) and was presented in this form in von Luxburg, Belkin, and Bousquet (2008, Theorem 7).

THEOREM 2.31. *Let $(K_n)_{n\in\mathbb{N}}$ be a sequence of compact linear operators on a Banach space $(E, \|\cdot\|_E)$, and let K be a compact linear operator on E. Assume that $K_n \overset{cc}{\to} K$. Let B_1 denote the unit ball in $(E, \|\cdot\|_E)$. Let $\lambda \neq 0$ be an eigenvalue of K, and $M \in \mathbb{C}$ an open neighborhood of λ such that*

$$\sigma(K) \cap M = \{\lambda\}$$

holds. We denote by P the spectral projection of K on λ in the sense of (2.6). Then, there exists $n_0 \in \mathbb{N}$ such that for all $n > n_0$, the set $\sigma(K_n) \cap M$ is isolated in $\sigma(K_n)$. Additionally, there exists a constant $C > 0$ such that, for all $x \in PE$

$$\|x - P_n x\|_E \leq C\left(\|(K_n - K)x\|_E + \|x\|_E \|(K_n - K)K_n\|\right) \tag{2.8}$$

Here, P_n denotes the spectral projection of K_n on $\sigma(K_n) \cap M$ for $n \in \mathbb{N}$, and the constant C only depends on λ and $\sigma(K)$.

2.3 Hilbert Spaces and L^2-Function Spaces

We briefly turn our attention to Hilbert spaces and L^p spaces. For further reading, see Heuser (1975, Section IV.18) and Rudin (1966, Chapter 3). A dot product on an $\mathbb{R}$-vector space E is a bilinear function

$$\langle\cdot,\cdot\rangle : \begin{cases} E \times E \longrightarrow \mathbb{R} \\ (e_1, e_2) \longmapsto \langle e_1, e_2\rangle \end{cases}$$

that satisfies symmetry and that is positive definite. Note that $\langle\cdot,\cdot\rangle$ induces a norm on E, specifically

$$\|\cdot\|_E : \begin{cases} E \times E \longrightarrow \mathbb{R} \\ e \longmapsto \sqrt{\langle e, e\rangle} \end{cases} \tag{2.9}$$

With this, we present a formal definition of a Hilbert space:

DEFINITION 2.32. *Let $(E, \|\cdot\|_E)$ be a Banach space. E is called a Hilbert space if the norm $\|\cdot\|_E$ is induced by a dot product on E in the sense of (2.9).*

Besides the Euclidean norm and dot product on $\mathbb{R}^d$, $d \in \mathbb{N}$, there is another notable example of this, which is closely linked to our topic: the space of L^2-functions with its respective norm and dot product. Suppose that $(X, \mathcal{M})$ is a measurable space. Let μ be a measure on $(X, \mathcal{M})$. Recall that the vector space $L^2(X)$ is given by

$$L^2(X, \mathcal{M}, \mu) := \mathcal{L}^2(X, \mathcal{M}, \mu)/\sim$$

where $\mathcal{L}^2(X, B(X), \mu)$ is the set of square-integrable functions

$$\mathcal{L}^2(X, \mathcal{M}, \mu) := \left\{ f : X \to \mathbb{R} \text{ measurable} : \int_X f^2 d\mu < \infty \right\}$$

and $\sim$ is an equivalence relation on $\mathcal{L}^2(X, B(X), \mu)$ that is given by

$$f \sim g :\Longleftrightarrow \mu(\{x \in X : f(x) \neq g(x)\}) = 0$$

In the following, we will denote $L^2(X, \mathcal{M}, \mu)$ by $L^2(X, \mu)$ and $L^2(X)$ unless the specific context requires additional detail. It is straightforward to prove that

$$\langle f, g \rangle_{L^2} := \int_X fg \, d\mu$$

defines a dot product on L^2, which induces the L^2-norm

$$\|f\|_{L^2} := \sqrt{\langle f, f \rangle_{L^2}} = \left(\int_X f^2 d\mu \right)^{1/2}$$

The Riesz-Fischer Theorem states that $L^2(X)$ is a complete vector space ((Royden 1963, p. 117, 244)). It thus follows that the space $(L^2(X), \|\cdot\|_{L^2})$ is a Banach space and a Hilbert space. Recall from Section 2.2 that $C(X)$ with the uniform norm is a Banach space. Moreover, under certain conditions, $(C(X), \|\cdot\|_{L^2})$ is a subspace of $(L^2(X), \|\cdot\|_{L^2})$, and thus a Hilbert space.

PROPOSITION 2.33. *Suppose that (X, d_X) is a compact metric space, that $\mathcal{M} = B(X)$, and that μ is a finite measure on $(X, B(X))$. Then, the vector space $(C(X), \|\cdot\|_{L^2})$ is subspace of $(L^2(X), \|\cdot\|_{L^2})$, and in particular, a Hilbert space.*

This statement can be proven using the elementary properties of compactness, continuity, and also the involved norms $\|\cdot\|_\infty$ and $\|\cdot\|_{L^2}$.

2.4 Empirical Processes and Glivenko-Cantelli Theorems

Next, we discuss important results from the theory of empirical processes. Most of the theory we provide can be found in van der Vaart and Wellner (1996). Let $(\mathcal{X}, \mathcal{A})$ be a measurable space, and let $\mathcal{F}$ be a family of measurable functions $f : \mathcal{X} \to \mathbb{R}$. Suppose that $(X_n)_{n \in \mathbb{N}}$ is a sequence of independent random observations in $\mathcal{X}$ with identical probability

distribution $\mathbb{P}$. Recall that the empirical measure on $(\mathcal{X}, \mathcal{A})$ is given by

$$\mathbb{P}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

where for any $x \in \mathbb{R}$, δ_x is the Dirac measure on $(\mathcal{X}, \mathcal{A})$. In the following, let $\mathcal{F}$ denote a set of measurable functions $f : (\mathcal{X}, \mathcal{A}) \to \mathbb{R}$. We use the notations $\mathbb{P}f := \int_{\mathcal{X}} f d\mathbb{P}$ and $\mathbb{P}_n f := \int_{\mathcal{X}} f d\mathbb{P}_n$

DEFINITION 2.34. *The map*

$$\mathbb{G}_n : \begin{cases} \mathcal{F} \longrightarrow \mathbb{R} \\ f \longmapsto \sqrt{n}(\mathbb{P}_n - \mathbb{P})f \end{cases}$$

is called the $\mathcal{F}$-indexed empirical process.

Note that the identity

$$\mathbb{G}_n f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i) - \mathbb{P}f)$$

holds. In the following, we define types of function classes that are particularly relevant for the theory of empirical processes. Note $\mathcal{F}$ satisfies the convergence condition

$$\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - \mathbb{P}f| \longrightarrow 0 \tag{2.10}$$

almost surely for $n \to \infty$, $\mathcal{F}$ is called a Glivenko-Cantelli class. Note that the measurability of sets defined by the expression in (2.10) is not trivial (see van der Vaart and Wellner (1996)). Whether a particular function class $\mathcal{F}$ is Glivenko-Cantelli depends on how "large" the function class is. To conceptualize the size of a collection of functions, we introduce covering numbers and bracketing numbers. For this purpose, suppose that $(\mathcal{G}, \|\cdot\|)$ is a subset of a normed space of real-valued functions on which $\|\cdot\|$ is a norm.

DEFINITION 2.35 (Covering Number)**.** *Let $\|\cdot\|_{L^0}$ denote a norm on the space of measurable functions on $(\mathcal{X}, \mathcal{A})$. For any measurable function $g : \mathcal{X} \to \mathbb{R}$ that satisfies $\|g\|_{L^0} < \infty$, the set*

$$B_\epsilon(g) := \{f : \|g - f\| < \epsilon\}$$

is called the ϵ-ball around g. The minimum number of ϵ-balls $B_\epsilon(g)$ that covers $\mathcal{G}$ is called the ϵ-covering number of $\mathcal{G}$ with respect to $\|\cdot\|_{\mathcal{F}}$, which we denote by $\mathcal{N}(\mathcal{G}, \epsilon, \|\cdot\|)$.

Note that the centers of the cover do not need to belong to $\mathcal{G}$.

DEFINITION 2.36 (Bracketing Number)**.** *For two functions l, u such that, the set*

$$[l, u] := \{f : l \leq f \leq u\}$$

is called an ϵ-bracket if $\|u - l\| < \epsilon$ holds. The ϵ-bracketing number $\mathcal{N}_{[]}(\mathcal{G}, \epsilon, \|\cdot\|)$ is given by the minimum number of ϵ-brackets required to cover $\mathcal{G}$.

We say that the norm $\|\cdot\|$ possesses the Riesz property if for two functions $f, g \in \mathcal{G}$ it

satisfies

$$\|f\| \leq \|g\| \tag{2.11}$$

if $|f| \leq |g|$. Note that for any measure μ on $(\mathcal{X}, \mathcal{A})$, the L^p-norms

$$\|f\|_{L^p} := \left(\int_\Omega |f|^p d\mu \right)^{1/p}$$

possess the Riesz property for $1 \leq p \leq \infty$ (van der Vaart and Wellner (1996, Section 2.1.1)). For such norms, there exists an inequality for covering and bracketing numbers: if $\|\cdot\|$ satisfies the Riesz property, then the inequality

$$\mathcal{N}(\mathcal{G}, \epsilon, \|\cdot\|) \leq \mathcal{N}_{[]}(\mathcal{G}, 2\epsilon, \|\cdot\|) \tag{2.12}$$

holds. Furthermore, in the specific case where $\|\cdot\|$ is the uniform norm $\|\cdot\|_\infty$, the identity

$$\mathcal{N}(\mathcal{G}, \epsilon, \|\cdot\|_\infty) = \mathcal{N}_{[]}(\mathcal{G}, 2\epsilon, \|\cdot\|_\infty) \tag{2.13}$$

holds (see van der Vaart and Wellner (1996, p. 84)).The concept of bracketing numbers allows us to introduce a Glivenko-Cantelli theorem for L^1-norms.

THEOREM 2.37 (van der Vaart and Wellner (1996), Theorem 2.4.1)**.** *A class $\mathcal{F}$ of measurable functions on $(\mathcal{X}, \mathcal{A})$ that satisfies $\mathcal{N}_{[]}(\mathcal{F}, \epsilon, \|\cdot\|) < \infty$ for all $\epsilon > 0$ is Glivenko-Cantelli.*

For a probability measure μ and $1 \leq p \leq \infty$, we denote the space of L^p-functions on $(\mathcal{X}, \mathcal{A}, \mu)$ by $L^p(\mu)$. We denote the associated covering numbers and bracketing numbers by $\mathcal{N}(\mathcal{F}, \epsilon, L^p(\mu))$ and $\mathcal{N}_{[]}(\mathcal{F}, \epsilon, L^p(\mu))$, respectively for a function class $\mathcal{F} \subset L^p(\mu)$. For a probability measure $\mathbb{P}$ on $(\mathcal{X}, \mathcal{A})$, we obtain the inequality

$$\mathcal{N}(\mathcal{F}, \epsilon, L^1(\mathbb{P})) \leq \mathcal{N}(\mathcal{F}, \epsilon, L^\infty(\mathbb{P})) \tag{2.14}$$

because

$$\|f\|_{L^1} = \int_{\mathcal{X}} |f| d\mathbb{P}$$

$$\leq \|f\|_\infty \cdot \mathbb{P}(\mathcal{X}) = \|f\|_\infty$$

The following result is an entropy bound for empirical process convergence. We follow the version presented in von Luxburg, Belkin, and Bousquet (2008, Theorem 19), which integrates results originally due to Anthony (2002) and Mendelson (2003).

PROPOSITION 2.38. *Let $\mathcal{F}$ denote a class of measurable functions on $(\mathcal{X}, \mathcal{A}, \mathbb{P})$ with $\|f\|_\infty$. Let $(Y_n)_{n \in \mathbb{N}}$ be a sequence of i.i.d. random variables with values in $\mathcal{X}$ and probability distribution $\mathbb{P}$. Then there exists a constant $\bar{c} > 0$ such that the following result holds for all $n \in \mathbb{N}$ with a probability of at least $1 - \delta$*

$$\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - \mathbb{P} f| \leq \frac{\bar{c}}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\mathcal{F}, \epsilon, \|\cdot\|_{L^2, n})} d\epsilon + \sqrt{\frac{1}{2n} \log \frac{2}{\delta}}$$

Here, $\|\cdot\|_{L^2,n}$ denotes the L^2-norm on $(\mathcal{X}, \mathcal{A})$ with respect to the empirical measure $\mathbb{P}_n$.

2.5 Rayleigh-Quotients and Eigenvalues of Symmetric Matrices

In the following, we briefly discuss properties of symmetric matrices including the Rayleigh-Ritz principal. Those results are crucial for discussing Laplacian eigenproblems in later sections. The theory we provide closely follows the approach presented by Horn and Johnson (1985). Suppose that $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix. Recall that the spectral decomposition of A is given by

$$A = V\Lambda V^T \tag{2.15}$$

where $V \in \mathbb{R}^{n\times n}$ is an orthogonal matrix, and $\Lambda \in \mathbb{R}^{n\times n}$ is the diagonal matrix

$$\Lambda = \text{diag}[\lambda_1, ..., \lambda_n]$$

where $\lambda_1, ..., \lambda_n \in \mathbb{R}$ are the eigenvalues of A, and $v_1, ..., v_n \in \mathbb{R}^n$, which denote the orthonormal columns of V, the corresponding eigenvectors (Horn and Johnson (1985), Theorem 2.5.4, Definitions 2.2.1, 2.5.3). In the following, we assume that $\lambda_1, ..., \lambda_n$ are ordered in non-decreasing order, i.e. $\lambda_k \leq \lambda_{k+1}$ for $1 \leq k < n$. Given non-zero vector $x \in \mathbb{R}^n$, we call

$$R(A, x) = \frac{x^T Ax}{x^T x} \tag{2.16}$$

the Rayleigh-quotient. We are interested in solving the following minimization and maximization problems:

$$\min_{x\neq 0} R(A, x) \tag{2.17}$$

$$\max_{x\neq 0} R(A, x) \tag{2.18}$$

A short calculation shows that the problems in (2.17) and (2.18) are equivalent to $\min_{\|x\|=1} x^T Ax$ and $\max_{\|x\|=1} x^T Ax$, respectively, where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^n$. The following result establishes a relationship between the two problems and the smallest and largest eigenvalues of A, respectively.

THEOREM 2.39 (Rayleigh-Ritz Theorem; Horn and Johnson (1985), Theorem 4.2.2)**.** *For a symmetric matrix $A \in \mathbb{R}^{n\times n}$ with eigenvalues ordered in non-decreasing order, the identities hold*

$$\lambda_1 = \min_{x\neq 0} R(A, x) \tag{2.19}$$

and

$$\lambda_n = \max_{x\neq 0} R(A, x) \tag{2.20}$$

hold.

For a proof of the result, see Horn and Johnson (1985, p. 176–177). Note that this result has consequences for the remaining eigenvalues: for $1 < k < n$, the eigenvalue λ_k is given by

$$\lambda_k = \min_{x\in\mathcal{V}_{k-1}^{\perp}, x\neq 0} R(A, x) \tag{2.21}$$

where $\mathcal{V}_{k-1} = \text{span}(v_1, ..., v_{k-1})$ and $\mathcal{V}_{k-1}^{\perp}$ denotes the orthogonal complement of $\mathcal{V}_{k-1}^{\perp}$ (see Horn and Johnson (1985, p. 178)). This allows us to derive the following result:

PROPOSITION 2.40. *For any symmetric matrix $A \in \mathbb{R}^{n \times n}$, the solution of the minimization problem*

$$\underset{V \in \mathbb{R}^{n \times k}, V^T V = I}{\text{argmin}} \text{tr}(V^T A V)$$

is given by the eigenvectors $v_1^*, ..., v_k^* \in \mathbb{R}^n$ to the k smallest eigenvalues of A.

Proof. The statement is a direct result of what we discussed: recall that the trace of a matrix is defined as the sum of all its diagonal elements, i.e.

$$\text{tr}(V^T A V) = \sum_{j=1}^{k} v_j^T A v_j$$

Hence, the minimization problem is equivalent to

$$\underset{\substack{v_1, ..., v_k \in \mathbb{R}^n \ \|v_j\|=1 \\ v_i^T v_j = 0, \ i \neq j}}{\text{argmin}} \sum_{j=1}^{k} v_j^T A v_j$$

Because $v_1, ..., v_n$ are unit vectors, each summand is equal to the Rayleigh quotient $R(A, v_j)$, which implies we can rewrite the problem as

$$\underset{\substack{v_1, ..., v_k \in \mathbb{R}^n \ \|v_j\|=1 \\ v_i^T v_j = 0, \ i \neq j}}{\text{argmin}} \sum_{j=1}^{k} R(A, v_j)$$

To obtain the orthonormal basis that minimizes $\sum_{j=1}^{k} R(A, v_j)$, we minimize each summand $R(A, v_j)$. The minimum of each $R(A, v_j)$ is given by λ_j for $j \leq k$ because of both Theorem 2.39 and (2.21). Finally, we obtain

$$\underset{\substack{v_1, ..., v_k \in \mathbb{R}^n \ \|v_j\|=1 \\ v_i^T v_j = 0, \ i \neq j}}{\text{argmin}} \sum_{j=1}^{k} v_j^T A v_j = \{v_1^*, ..., v_k^*\}$$

□

3 Spectral Clustering

In this section, we provide an overview of what spectral clustering is: this includes an exploration of its foundations in graph theory and linear algebra, including the properties of graph Laplacian matrices and the mincut-problem. We briefly discuss an alternative framework of spectral clustering based on random-walks. At the end of the section, we discuss how perturbation theory and the Davis-Kahan Theorem in particular motivate spectral clustering.

3.1 Graphs and Laplacians

In the following, we give a brief introduction to graphs and their most important properties, which builds on the theory presented by Clark and Holton (1991, Chapter 1). Suppose that, for $n \in \mathbb{N}$, $V = \{v_1, ..., v_n\}$ is a non-empty set. Let

$$E = \{\{x_i, x_j\} \mid i, j \in J\}$$

be a selection of pairs from the set V. In this case, $J \subset [n]$ may be chosen arbitrarily. We call the tuple $G = (V, E)$ an undirected graph. An element v_i of V is called a vertex of G, and an element $\{v_i, v_j\}$ of E is called the edge between the vertices v_i and v_j. It is important to note that $J \subsetneq [n]$ may hold, and not all two vertices are connected by an edge.

DEFINITION 3.1. *The matrix*

$$W = (w_{ij})_{i,j=1}^{n} \in \mathbb{R}^{n \times n}$$

is called the weight matrix or adjacency matrix of G. $w_{ij} \geq 0$ is the weight assigned to the edge $\{v_i, v_j\}$ (or (v_i, v_j) in the case of a directed graph) and $w_{ij} = 0$ if v_i and v_j are not connected by an edge. G is called an unweighted graph if

$$w_{ij} \in \{0, 1\}$$

for all $i, j \in [n]$. Otherwise, G is a weighted graph. a weighted graph we denote $G = (V, E, W)$.

Graphs can be directed or undirected. In an undirected graph, each two distinct vertices have two respective edges, one for each direction, i.e., for $I, J \in [n]$:

$$E = \{(v_i, v_j) \mid (i, j) \in I \times J\}$$

For an undirected graph, we introduce the following definition:

DEFINITION 3.2. *An undirected graph $G = (V, E)$ is called complete or fully connected if each pair of distinct vertices v_i, v_j is connected by an edge.*

PROPOSITION 3.3. *If a weighted graph $G = (V, E, W)$ is undirected, its similarity matrix W is symmetric, i.e. $W = W^T$.*

Proof. Because G is undirected, the elements of E have the form $\{v_i, v_j\}$. Let $\{v_i, v_j\}$ be in E. Then,

$$\{v_i, v_j\} = \{v_j, v_i\}$$

holds for all i, j. This implies $w_{ij} = w_{ji}$, which completes the proof. □

In the following, we only discuss weighted, undirected graphs of the form $G = (V, E, W)$. We move on to defining walks and paths in G.

DEFINITION 3.4. *Let $(v_j)_{j=0,...,l}$ for $l \in [n]$ be a subset of V. Assume that $e_1, ..., e_l \in E$ such that for each $j \in [l]$, the edge e_j connects the vertices v_{j-1} and v_j. Then, the finite sequence of vertices and edges*

$$\{v_{j-1} e_j v_j\}_{j=1,...,l}$$

is called a walk in G. More precisely, the above sequence is called the walk from v_0 to v_l, where l is the length of the walk, v_0 its origin, and v_l its terminus.

Note that this definition allows for multiple repetitions of the same edge: suppose, for instance, that V only consists of 0 and 1, and that $\{0, 1\} = \{1, 0\}$ is its only edge. Then, the finite sequence

$$0, \{0, 1\}, 1, \{1, 0\}, 0, \{0, 1\}, 1$$

satisfies Definition 3.4 while it repeats the same edge three times and repeats the segment $0, \{0, 1\}, 1$. To allow for further distinction, we introduce the definition of a path in G.

DEFINITION 3.5. *For $l \in [n]$, the walk*

$$\{v_{j-1} e_j v_j\}_{j=1,...,l}$$

is called a path in G if and only if the vertices are distinct, i.e. $v_i \neq v_j$ holds for $i, j = 0, ..., l$.

The Definitions 3.4 and 3.5 enable us to define graph components. First, we introduce subgraphs and cliques.

DEFINITION 3.6. *The graph $H = (A, E')$ is called a subgraph of G if $A \subset V$ and $E' \subset E$. H is called a clique if it is fully connected. If H is also the largest clique in G, the number of vertices in H is called the clique number of G.*

Suppose that $v_i, v_j \in V$. We say that v_i and v_j are *connected* if there exists a path in the sense of Definition 3.5 with origin v_i and terminus v_j. G is connected if every two vertices in V are connected. Otherwise, we say that G is *disconnected*. Note that every fully-connected graph is connected. This allows for the following definition:

DEFINITION 3.7. *A connected subgraph H of G is called a (connected) component if it is not part of a larger connected subgraph of G.*

We review a few examples to illustrate the above definitions.

EXAMPLE 3.8. *The following examples illustrate the terms introduced in Definition 3.6:*

(i) $G = (V, E)$ *without edges (i.e.* $E = \varnothing$*) is called an empty graph. In such a graph, each vertex defines a component as the subgraph*

$$G_x = (\{x\}, \varnothing)$$

is fully connected and not part of a larger connected subgraph for all $x \in V$*.*

(ii) Let G *be a fully connected graph, i.e. a graph* $G = (V, E, W)$ *for which*

$$w_{ij} > 0$$

holds for all $i, j \in [n]$*. Then* G *itself is the only component of* G*.*

Before we move on to the next subject, we give an important definition.

DEFINITION 3.9. *An undirected graph* $G = (V, E)$ *is called bipartite if there exists a disjoint partition* $X \dot{\cup} Y$ *of* V *such that every edge in* E *connects one node in* X *with one in* Y*. We call the graph non-bipartite if it is not bipartite, i.e. if such partition does not exist.*

We now turn to Laplacian matrices and their relationship with their respective graphs. For our discussion, we follow the approach presented by von Luxburg (2007). Recall that $W = (w_{ij})_{i,j=1}^n$ is the weight matrix of G. We call

$$\deg(i) = \sum_{j=1}^{n} w_{ij}$$

the i-th degree or the degree of the i-th vertex. We call the diagonal matrix

$$D = \operatorname{diag}[\deg(1), ..., \deg(n)] \in \mathbb{R}^n$$

the degree matrix of G. Suppose that A is a non-empty subset of V. We call

$$\operatorname{vol}(A) := \sum_{v_i \in A} \deg(i)$$

the volume of A. This constitutes one way of defining the *size* of A. An alternative is to count the number of vertices in A, i.e.

$$|A| := \sum_{v_i \in V} \mathbb{1}_A(v_i)$$

For another subset $B \neq \varnothing$, we define the weight between A and B.

$$W(A, B) = \sum_{v_i \in A, v_j \in B} w_{ij}$$

Next, we define graph Laplacian matrices.

DEFINITION 3.10. *Let* I_n *be the identity matrix in* $\mathbb{R}^{n \times n}$*. The matrix*

$$L := D - W$$

is called the unnormalized Laplacian (matrix) of G.

Below, we present important results about the unnormalized Laplacian matrix L from von Luxburg (2007, Section 3), beginning with the following statement:

PROPOSITION 3.11. *For every $(u_1, ..., u_n)^T =: u \in \mathbb{R}^n$, the identity*

$$u^T L u = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} (u_i - u_j)^2 \tag{3.1}$$

holds.

Proof. The definition of L, D, and W yield

$$\begin{aligned} u^T L u &= u^T D u - u^T W u \\ &= \sum_{i=1}^{n} \deg(i) u_i^2 - \sum_{i,j=1}^{n} w_{ij} u_i u_j \\ &= \frac{1}{2} \left(\sum_{i=1}^{n} \deg(i) u_i^2 - 2 \sum_{i,j=1}^{n} w_{ij} u_i u_j + \sum_{i=1}^{n} \deg(i) u_i^2 \right) \end{aligned}$$

Plugging in the definition of $\deg(i) = \sum_{j \neq i} w_{ij}$, the statement follows from the binomial theorem (von Luxburg 2007). □

Note that L is symmetric because D and W are symmetric. This is because G is undirected. The following statement is a direct implication of Proposition 3.11.

COROLLARY 3.12. *Let $\mathbf{1}$ denote the vector*

$$\mathbf{1} := (1, ..., 1)^T \in \mathbb{R}^n$$

The unnormalized Laplacian matrix L of G satisfies the following conditions:

(i) L is positive semi-definite, i.e. L satisfies

$$v^T L v \geq 0$$

for all $v \in \mathbb{R}^n$. Equivalenty, $\lambda \in [0, \infty)$ holds for all eigenvalues λ of L.

(ii) 0 is an eigenvalue of L with eigenvector $\mathbf{1}$.

The following statement introduces a relationship between the spectral properties of L and the component structure of G.

THEOREM 3.13 (von Luxburg 2007, Proposition 2)**.** *Consider the undirected graph $G = (V, E, W)$. Let $k \leq n$ be the (geometric) multiplicity of the eigenvalue 0 of the unnormalized Laplacian L. Then, k equals the number of components $A_1, ..., A_k$ of G. Furthermore, the corresponding eigenvectors are $\mathbf{1}_{A_1}, ..., \mathbf{1}_{A_k} \in \mathbb{R}^n$, where, for $j \in [k]$, $\mathbf{1}_{A_j}$ is defined such that*

the i-th component $\mathbf{1}_{A_j}^i$ *of* $\mathbf{1}_{A_j}$ *is given by*

$$\mathbf{1}_{A_j}^i := \begin{cases} 1 & \text{if } x_i \text{ is part of } A_j \\ 0 & \text{else} \end{cases} \tag{3.2}$$

for $i \in [n]$.

Proof. We begin with the case $k = 1$: suppose 0 is an eigenvalue of L. It is easy to verify that this is equivalent to the condition

$$0 = u^T L u$$

for any eigenvector u of L to 0. Because of Proposition 3.11, we obtain

$$0 = \sum_{i,j=1}^{n} w_{ij} \left(u_i - u_j\right)^2$$

which implies that u is an eigenvector if and only if all summands in the term on the right-hand side sum up to zero. Then, it is easy to verify that $u_i = u_j$ holds if and only if x_i and x_j are connected by a path. Thus, it follows that all eigenvectors u are constant within each of the components of G. Since we assumed $k = 1$, there is only one connected component, which is the graph G itself, which yields that $\mathbf{1}$ is an eigenvalue of L to the eigenvalue 0. Next, we apply this argument to the general case of $k \geq 2$. We assume without loss of generality that the vertices of W are numbered in accordance with the components of G, i.e. the vertices in A_1 are numbered $x_1, ..., x_{n_1}$, the vertices in A_2 are $x_{n_1+1}, ..., x_{n_2}$ etc. Note that W is a block diagonal matrix in which the k blocks correspond to the k components of G; by definition, the same holds for L. Note that the blocks $L_1, ..., L_k$ have eigenvalue 0 with constant eigenvector because of what we argued earlier. Then, the block structure yields that 0 is an eigenvalue of L with the k corresponding eigenvectors $\mathbf{1}_{A_1}, ..., \mathbf{1}_{A_k}$. Because $A_1, ..., A_k$ are disjoint subgraphs of G, the vectors $\mathbf{1}_{A_1}, ..., \mathbf{1}_{A_k}$ are linearly independent, which proves the statement (von Luxburg 2007, p.397-398). □

In the following, we introduce normalized Laplacian matrices. We begin with the symmetric normalized Laplacian:

DEFINITION 3.14. *The matrix* $L' \in \mathbb{R}^n$ *is defined as*

$$L' := I - D^{-1/2} W D^{-1/2}$$

and called the symmetric normalized Laplacian matrix of G.

Note that L' is a modification of L since, by definition

$$L' = D^{-1/2}(D - W)D^{-1/2}$$

and thus

$$L = D^{-1/2} L D^{-1/2}$$

holds. Obviously, L' is symmetric by definition. Next, we define another type of normalized Laplacian.

DEFINITION 3.15. *The matrix $L'' \in \mathbb{R}^n$ is defined as*

$$L'' := I - D^{-1}W$$

is called the random-walk normalized Laplacian.

It is important to consider that, as opposed to L and L', L'' is not necessarily symmetric. To see this, suppose that ℓ''_{ij} is a non-diagonal entry of L''. Because of Definition 3.15, we obtain

$$\ell''_{ij} = 1 - \frac{w_{ij}}{\deg(i)}$$

whereas

$$\ell''_{ji} = -\frac{w_{ij}}{\deg(j)}$$

which implies that $\ell''_{ij} = \ell''_{ji}$ only if $\deg(i) = \deg(j)$ for $i \neq j$. In the following, we prove important results for both L' and L''.

PROPOSITION 3.16. *Let $u = (u_1, ..., u_n)^T \in \mathbb{R}^n$. The normalized Laplacian L' satisfies*

$$u^T L' u = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \left(\frac{u_i}{\sqrt{\deg(i)}} - \frac{u_j}{\sqrt{\deg(j)}} \right)^2 \tag{3.3}$$

Proof. The identity (3.3) follows from the definition of D and W, as well as the binomial theorem (von Luxburg 2007, Section 3). □

Note that (3.3) implies that L' is positive semi-definite. L' and L'' are directly related in terms of their spectral properties:

PROPOSITION 3.17. *For the symmetric normalized Laplacian matrix L' and the random-walk normalized Laplacian matrix L'', the following two statements are equivalent*

(i) λ is an eigenvalue of L'' with eigenvector u.

(ii) λ is an eigenvalue of L' with eigenvector $D^{1/2}u$.

Proof. Let λ be an eigenvalue of L' with eigenvector $D^{1/2}u$. This is equivalent to the eigenproblem

$$L' D^{1/2} u = \lambda D^{1/2} u$$

which can be manipulated in the following way using the definitions of L' and L'':

$$L' D^{1/2} u = \lambda D^{1/2} u$$

$$(I - D^{-1/2} W D^{-1/2}) D^{1/2} u = \lambda D^{1/2} u$$

$$D^{1/2} u - D^{-1/2} W u = \lambda D^{1/2} u$$

$$D^{1/2}(u - D^{-1} W u) = D^{1/2}(\lambda u)$$

$$(I - D^{-1}W)u = \lambda u$$

The last identity is equivalent to the eigenproblem $L''u = \lambda u$ because the matrix on the left-hand side is L'' according to Definition 3.15. □

The matrices L' and L'' and their respective spectra are also related to the unnormalized Laplacian matrix L as the following statement shows.

PROPOSITION 3.18. *For the symmetric normalized Laplacian L', the random-walk normalized Laplacian matrix L'', and the unnormalized Laplacian L, the following two statements are equivalent:*

(i) $\lambda \in [0, \infty)$ and $u \in \mathbb{R}^n$ solve the eigenproblem

$$Lu = \lambda Du \tag{3.4}$$

(ii) u is an eigenvector of L'' to the eigenvalue λ.

Proof. The proof follows a structure which is similar to that of Proposition 3.17 using the definitions of L and L''. □

Note that, because of Propositions 3.17 and 3.18, u and λ solve (3.4) if and only if $D^{1/2}u$ is an eigenvector of L' to the eigenvalue λ. We summarize implications of the statements we have proved in a corollary.

COROLLARY 3.19. *L' and L'' satisfy the following conditions:*

(i) 0 is an eigenvalue of L' with eigenvector $D^{1/2}\mathbf{1}$. 0 is an eigenvalue of L'' with eigenvector $\mathbf{1}$.

(ii) L' and L'' have real, non-negative eigenvalues.

Proof. We prove that 0 is an eigenvalue of L'' with eigenvector $\mathbf{1}$:

$$(I - D^{-1}W)\mathbf{1} = \mathbf{1} - D^{-1}W\mathbf{1}$$

Recall that each component of $\mathbf{1}$ equals 1. This implies the following for the i-th component $(D^{-1}W\mathbf{1})_i$ of the vector $D^{-1}W\mathbf{1}$:

$$(D^{-1}W)_i = \sum_{j=1}^{n} \frac{w_{ij}}{\deg(i)}$$

for all $i \in [n]$. For the right-hand side, we obtain

$$\sum_{j=1}^{n} \frac{w_{ij}}{\deg(i)} = \frac{\deg(i)}{\deg(i)} = 1$$

by definition of $\deg(i)$. This implies $D^{-1}W\mathbf{1} = \mathbf{1}$ and hence

$$L''\mathbf{1} = \mathbf{1} - \mathbf{1} = \mathbf{0}$$

where $\mathbf{0} = (0, ..., 0)^T \in \mathbb{R}^n$. This implies $L''\mathbf{1} = 0 \cdot \mathbf{1}$, which proves the first part of (i). Then, $D^{1/2}\mathbf{1}$ is an eigenvector of L' to the eigenvalue 0 because of Proposition 3.17. To prove (ii), recall that L' is positive semi-definite because of (3.3). Equivalently, L' has non-negative, real eigenvalues. The same holds for eigenvalues of L'' because Proposition 3.17 states that every eigenvalue of L' is an eigenvalue of L''. □

We conclude this section with a statement similar to Theorem 3.13: the component structure of G relates to the spectra of its normalized Laplacians.

THEOREM 3.20. *Suppose that the undirected graph $G = (V, E, W)$ has non-negative weights, i.e. there exist $i, j \in [n]$ such that $w_{ij} > 0$. Let $k \in [n]$ be the (geometric) multiplicity of the eigenvalue 0 of L' and L'' (see Corollary 3.19). Then, k is equal to the number of components $A_1, ..., A_k$ of G, and the corresponding eigenvectors are*

$$\mathbf{1}_{A_1}, ..., \mathbf{1}_{A_k} \in \mathbb{R}^n$$

for L'' and

$$D^{1/2}\mathbf{1}_{A_1}, ..., D^{1/2}\mathbf{1}_{A_k} \in \mathbb{R}^n$$

for L', where $\mathbf{1}_{A_j}$ is defined according to (3.2) for $j \in [k]$.

The proof works analogously to the proof of Theorem 3.13, in combination with the results established in this section (see von Luxburg (2007)).

Proof. The proof works analogously to the proof of Theorem 3.13 using the results of Proposition 3.17 and Proposition 3.18. □

3.2 Different Perspectives: Graph Cut and Random Walk

3.2.1 Spectral Clustering from the Graph Cut Perspective

To understand the intuition of Spectral Clustering using the matrices L' and L'', we consider two different perspectives: the graph-cut perspective concerning L and L' and the random-walk perspective concerning L''. Beginning with the former, recall that our goal is to group a given number of data points $x_1, ..., x_n$ into a partition of k subsets for a given number $k \leq n$. For the graph-cut perspective, we assume that $x_1, ..., x_n$ define an undirected graph $G = (V, E, W)$ such that $V = \{x_1, ..., x_n\}$, and suppose that G has non-negative weights, i.e. that

$$w_{ij} > 0$$

for some $i, j \in [n]$, $i \neq j$. For simplicity's sake, we assume $k = 2$ for now – we shall return to the more complicated case at the end of this section. Let, A be a subset of V. Note that A and A^C define a partition of V because $A \cup A^C = V$ and $A \cap A^C = \varnothing$. We also say that A and A^C define a cut in G. Recall that for two subsets $S_1, S_2 \subset V$, we defined

$$W(S_1, S_2) = \sum_{x_i \in S_1, x_j \in S_2} w_{ij}$$

For a partition $S_1, ..., S_k$, $l \in [n]$, we call

$$\text{cut}(S_1, ..., S_k) := \frac{1}{2} \sum_{j=1}^{k} W(S_j, S_j^C)$$

the cut function of $S_1, ..., S_k$. In our specific case of $k = 2$, we have

$$\text{cut}(A, A^C) = W(A, A^C)$$

Recall that the core idea of Spectral Clustering is to minimize the similarity across different clusters. We formalize this as follows:

PROBLEM 3.21 (Min-Cut Problem)**.** *The problem*

$$\underset{A \subset V}{\text{argmin}} \ \text{cut}(A, A^C) \tag{3.5}$$

is called the Min-Cut Problem. The expression in (3.5) is also called the minimum cut of G for $k = 2$.

This, however, does often not lead to satisfying results as it often separates one vertex from the rest of V. To circumvent this, we introduce normalized cuts and ratio cuts. Recall the definitions of $\text{vol}(A)$ and $|A|$.

DEFINITION 3.22 (RatioCut)**.** *For $k \in [n]$ and a cut $S_1, ..., S_k$ of G, the RatioCut function is defined as*

$$\text{RatioCut}(S_1, ..., S_k) = \sum_{j=1}^{k} \frac{\text{cut}(S_j, S_j^C)}{|A_j|}$$

Note that $|S_j|$ in the denominator penalizes size imbalances between the subsets S_j. For $k = 2$, we seek to find the cut A, A^C of G that minimizes $\text{RatioCut}(A, A^C)$, i.e.

$$\underset{A \subset V}{\text{argmin}} \left(\frac{\text{cut}(A, A^C)}{|A|} + \frac{\text{cut}(A, A^C)}{|A^C|} \right) \tag{3.6}$$

Next, we present a statement that establishes a relationship between (3.6) and solving an eigenproblem of L.

THEOREM 3.23. *Let $u := (u_1, ...u_n)^T \in \mathbb{R}^n$ be such that both*

$$u_i = \begin{cases} \sqrt{\frac{|A|}{|A^C|}} & \text{if } x_i \in A \\ -\sqrt{\frac{|A|}{|A^C|}} & \text{if } x_i \in A^C \end{cases} \tag{3.7}$$

for $i \in [n]$, and

$$\langle u, \mathbf{1} \rangle = 0$$

Then, the problem in (3.6) is equivalent to

$$\underset{u}{\text{argmin}} \ u^T L u \tag{3.8}$$

Proof. By definition, $|V| = |A| + |A^C|$, and, thus

$$\text{RatioCut}(A) \cdot |V| = \text{cut}(A) \left(\frac{|A| + |A^C|}{|A|} + \frac{|A| + |A^C|}{|A^C|} \right) \tag{3.9}$$

holds. Hence, solving (3.6) is equivalent to finding $A \subset V$ such that the right-hand side of (3.9) becomes minimal. A simple calculation shows that

$$\frac{|A| + |A^C|}{|A|} + \frac{|A| + |A^C|}{|A^C|} = \frac{|A|}{|A^C|} + \frac{|A^C|}{|A|} + 2$$

Subsequently,we minimize

$$\begin{aligned}
&\text{cut}(A) \left(\frac{|A| + |A^C|}{|A|} + \frac{|A| + |A^C|}{|A^C|} \right) = \text{cut}(A) \left(\frac{|A|}{|A^C|} + \frac{|A^C|}{|A|} + 2 \right) \\
&= \text{cut}(A) \left(\sqrt{\frac{|A|}{|A^C|}} + \sqrt{\frac{|A^C|}{|A|}} \right)^2 \\
&= \frac{1}{2} \left(\sqrt{\frac{|A|}{|A^C|}} + \sqrt{\frac{|A^C|}{|A|}} \right)^2 \left(\sum_{\substack{x_i \in A \\ x_j \in A^C}} w_{ij} + \sum_{\substack{x_i \in A^C \\ x_j \in A}} w_{ij} \right) \\
&= \frac{1}{2} \sum_{\substack{x_i \in A \\ x_j \in A^C}} w_{ij} \left(\sqrt{\frac{|A|}{|A^C|}} + \sqrt{\frac{|A^C|}{|A|}} \right)^2 + \frac{1}{2} \sum_{\substack{x_i \in A^C \\ x_j \in A}} w_{ij} \left(-\sqrt{\frac{|A|}{|A^C|}} - \sqrt{\frac{|A^C|}{|A|}} \right)^2 \\
&= \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \left(u_i - u_j \right)^2
\end{aligned}$$

which proves the result (see von Luxburg (2007, p. 401)). □

Note that problem (3.8) is NP-hard (Section 5). Hence, consider the following relaxation of (3.8):

$$\underset{u \in \mathbb{R}^n, \langle u, \mathbf{1} \rangle = 0}{\text{argmin}} \; u^T L u \tag{3.10}$$

where the condition (3.11) is relaxed to $u \in \mathbb{R}^n$. Note that the expression $\min_{u \in \mathbb{R}^n, \langle u, \mathbf{1} \rangle = 0} u^T L u$ is the Rayleigh quotient in (2.21) for $k = 2$; hence, as a consequence of Theorem 2.39, the eigenvector to the second smallest eigenvalue of L solves (3.10). We shall return to the practical impacts of this relaxation later. The normalized cut function Ncut provides an alternative to RatioCut:

DEFINITION 3.24 (Normalized Cut)**.** *For $k \in [n]$ and a cut $S_1, ..., S_k$ of G, the normalized cut function* $\text{Ncut}(S_1, ..., S_k)$ *is defined as*

$$\text{Ncut}(S_1, ..., S_k) = \sum_{j=1}^{k} \frac{\text{cut}(S_j, S_j^C)}{\text{vol}(S_j)}$$

In the case of minimizing $\text{Ncut}(A, A^C)$ for $k = 2$, we derive a statement similar to that of Theorem (3.23):

THEOREM 3.25. *Define*

$$u_i = \begin{cases} \sqrt{\frac{\text{vol}(A^C)}{\text{vol}(A)}} \text{ if } x_i \in A \\ -\sqrt{\frac{\text{vol}(A)}{\text{vol}(A^C)}} \text{ if } x_i \in A^C \end{cases} \tag{3.11}$$

for $i \in [n]$, *and*

$$\langle v, D^{1/2}\mathbf{1} \rangle = 0$$

hold. Then,

$$\underset{A \subset V}{\text{argmin}} \ \text{Ncut}(A, A^C) \tag{3.12}$$

is equivalent to

$$\underset{A \subset V}{\text{argmin}} \ v^T L' v$$

where $v = D^{1/2}u$.

Similar as in Theorem 3.23, the statement can be proven through a short calculation. Note that solving (3.12) is NP-hard, which is why we turn our attention to the relaxation

$$\underset{v \in \mathbb{R}^n}{\text{argmin}} \ \{v^T L' v \mid \langle v, D^{1/2}\mathbf{1} \rangle = 0\} \tag{3.13}$$

Again, Theorem 2.39 yields that (3.13) is equivalent to finding the second-smallest eigenvalue of the symmetric normalized Laplacian matrix L'. Before return to the practical impacts of relaxations such as in (3.13), let us briefly explore what happens if from the graph-cut perspectives when $k \geq 3$: for RatioCut, the modified problem is the following

$$\underset{A_1, \ldots, A_k}{\text{argmin}} \ \text{RatioCut}(A_1, \ldots, A_k) \tag{3.14}$$

In this case, the associated minimum problem becomes more complex: instead of solving for a vector in $\mathbb{R}^n$, (3.14) corresponds to solving the problem

$$\underset{H}{\text{argmin}} \ \{\text{tr}\left(H^T L H\right) \mid H^T H = I\} \tag{3.15}$$

for a matrix $H \in \mathbb{R}^{n \times k}$ of the form $H = (h_{ij})_{i,j=1}^{n,k}$ such that

$$h_{ij} = \begin{cases} \frac{1}{\sqrt{|A_j|}} \text{ if } x_i \in A_j \\ 0 \text{ otherwise} \end{cases}$$

Similar to the simpler case, solving (3.15) is NP-hard. Therefore, we consider the relaxation

$$\underset{H \in \mathbb{R}^{n \times k}}{\text{argmin}} \ \{\text{tr}\left(H^T L H\right) \mid H^T H = I\} \tag{3.16}$$

where H is an arbitrary matrix in $\mathbb{R}^{n \times k}$. This is equivalent to finding the eigenvectors to the

k smallest non-zero eigenvectors of L according to Proposition 2.40. For Ncut, the situation is similar: the equivalent formulation of minimizing $\text{Ncut}(A_1, ..., A_k)$ is

$$\underset{T}{\text{argmin}} \{\text{tr}\left(T^T L' T\right) \mid T^T T = I\} \tag{3.17}$$

for $T = D^{1/2}H$, $H = (h_{ij})_{i,j=1}^{n,k}$ and

$$h_{ij} = \begin{cases} \frac{1}{\sqrt{\text{vol}(A_j)}} \text{ if } x_i \in A_j \\ 0 \text{ otherwise} \end{cases} \tag{3.18}$$

Again, (3.17) is NP-hard, and the subsequent relaxation

$$\underset{T \in \mathbb{R}^{n \times k}}{\text{argmin}} \{\text{tr}\left(T^T L' T\right) \mid T^T T = I\} \tag{3.19}$$

where $T \in \mathbb{R}^{n \times k}$ is an arbitrary matrix, corresponds to finding the eigenvectors to the k smallest positive eigenvalues of the symmetric normalized Laplacian L' (see Proposition 2.40). Recall that both (3.16) and (3.19) are relaxations of minimizing RatioCut or Ncut. This implies that solutions of the relaxed problems do not correspond to partitions that minimize the cut functions. A vector that solves (3.13) does not necessarily have the form that Theorem 3.25 demands. In fact, only in specific cases, a solution $v \in \mathbb{R}^n$ has such a form, which the following example illustrates. The example is inspired by a discussion in Lee and Wilkinson (2019).

EXAMPLE 3.26 (Stochastic Block Model)**.** *Let $G = (V, E)$ be a an undirected graph, where $V = \{x_1, ..., x_n\}$ denotes the set of vertices and E the set of all edges. For $l = 1, ..., k$, let n_l denote the number of vertices in C_l (i.e. $\sum_l n_l = n$). Let $\mathcal{C} := (C_l)_{l=1}^k$ denote a partition of V. For each $i = 1, ..., n$, the vector $Z_i = (Z_1^{(1)} \cdots Z_1^{(n)})^T \in \mathbb{R}^k$ indicates the group membership of x_i in the following way for $l = 1, ..., k$:*

$$Z_i^{(l)} = \begin{cases} 1 \text{ if } x_i \in C_l \\ 0 \text{ otherwise} \end{cases}$$

By $Z \in \mathbb{R}^{n \times k}$, we denote the matrix $(Z_1 \cdots Z_n)^T$. The edge probability matrix $B \in \mathbb{R}^{k \times k}$ is given by

$$B = (p_{lm})_{l,m=1}^k$$

with $p_{lm} \in [0, 1]$. The similarity matrix $W \in \mathbb{R}^{n \times n}$ of G is given by

$$W = ZBZ^T$$

Note that the rows and columns of B may not sum up to 1 since one node can share edges with other nodes from multiple different subsets in $\mathcal{C}$. For the purpose of a simple illustration of Spectral Clustering, we assume $k = 2$ and $n_1 = n_2 = \frac{n}{2}$. In the following, we compute the eigenvalues and eigenvectors of the normalized Laplacian matrix in this

scenario. To do this, we need to solve

$$L'z = \lambda z$$

Without loss of generality, we assume that the data is indexed clusterwise, which is such that $x_1, ..., x_{n/2} \in C_1$ and $x_{n/2+1}, ..., x_n \in C_2$. From the Propositions 3.17 and 3.18, we know that finding the eigenvalues and eigenvectors of L' is equivalent to solving the generalized eigenproblem (3.4). A short calculation shows that the component-wise version

$$(Lu)_i = \lambda(Du)_i$$

of (3.4) is given by

$$\deg(1) \cdot u_i - \left(\sum_{\substack{j=1 \\ j \neq i}}^{n/2} p_{11} u_j + \sum_{j=n/2+1}^{n} p_{12} u_j \right) = \lambda \cdot \deg(1) \cdot u_i \quad \text{for } x_i \in C_1 \tag{3.20}$$

$$\deg(2) \cdot u_i - \left(\sum_{j=1}^{n/2} p_{12} u_j + \sum_{\substack{j=n/2+1 \\ j \neq i}}^{n} p_{22} u_j \right) = \lambda \cdot \deg(2) \cdot u_i \quad \text{for } x_i \in C_2 \tag{3.21}$$

It is easy to verify that $\lambda = 0$ solves this eigenproblem with $u = \mathbf{1}$. Next, consider the cluster indicator vector $\mathbf{1}_{\mathcal{C}}$, which, since $k = 2$, is given by

$$\mathbf{1}_{\mathcal{C}}^i = \begin{cases} 1 & \text{if } x_i \in C_1 \\ -1 & \text{if } x_i \in C_2 \end{cases}$$

With $u = \mathbf{1}_{\mathcal{C}}$, (3.20) and (3.21) simplify to

$$\deg(1) - \left(\frac{n}{2} - 1\right) p_{11} + \frac{n p_{12}}{2} = \lambda \cdot \deg(1) \quad \text{for } x_i \in C_1$$

$$\deg(2) - \left(\frac{n}{2} - 1\right) p_{22} + \frac{n p_{12}}{2} = \lambda \cdot \deg(2) \quad \text{for } x_i \in C_2$$

which yields

$$\lambda = 1 + \frac{\frac{n}{2} p_{12} - \left(\frac{n}{2} - 1\right) p_{11}}{\frac{n}{2} p_{12} + \left(\frac{n}{2} - 1\right) p_{11}}$$

and

$$\lambda = 1 + \frac{\frac{n}{2} p_{12} - \left(\frac{n}{2} - 1\right) p_{22}}{\frac{n}{2} p_{12} + \left(\frac{n}{2} - 1\right) p_{22}}$$

This implies that $\mathbf{1}_{\mathcal{C}}$ is an eigenvector of L' if and only if the condition

$$\frac{\frac{n}{2} p_{12} - \left(\frac{n}{2} - 1\right) p_{11}}{\frac{n}{2} p_{12} + \left(\frac{n}{2} - 1\right) p_{11}} = \frac{\frac{n}{2} p_{12} - \left(\frac{n}{2} - 1\right) p_{22}}{\frac{n}{2} p_{12} + \left(\frac{n}{2} - 1\right) p_{22}} \tag{3.22}$$

holds. Notably, (3.22) is satisfied if C_1 and C_2 are connected components: in this case,

$p_{12} = 0$ by definition, and (3.22) simplifies to

$$-\frac{\left(\frac{n}{2}-1\right)p_{11}}{\left(\frac{n}{2}-1\right)p_{11}} = -\frac{\left(\frac{n}{2}-1\right)p_{22}}{\left(\frac{n}{2}-1\right)p_{22}} = -1$$

which yields

$$\lambda = 0$$

which aligns with Theorem 3.20.

3.2.2 Spectral Clustering from the Random Walk Perspective

Alternatively, we may view spectral clustering from a random-walk perspective (von Luxburg 2007, Section 6): for this purpose, it is necessary to introduce a few basic results and definitions about Markov chains, stationary distributions (Durrett 1999, p.2-26), and random walks on graphs (see Aldous and Fill (2002) for details). Consider the stochastic process $(X_m)_{m\in\mathbb{N}_0}$ with values in a finite state space $\{1, ..., n\}$. The sequence $(X_m)_{m\in\mathbb{N}_0}$ is called a Markov chain if it satisfies the property

$$\mathbb{P}(X_{m+1} = j \mid X_m = i_m, ..., X_0 = i_0) = \mathbb{P}(X_{m+1} = j \mid X_m = i_m) \tag{3.23}$$

i.e. the probability of a state at time $m+1$ only depends on the probability of the previous state. In the following, we focus on temporally homogeneous Markov chains, as presented in the following definition

DEFINITION 3.27. *A Markov chain $(X_m)_{m\in\mathbb{N}_0}$ is called temporally homogeneous if the transition probability*

$$\mathbb{P}(X_{m+1} = j \mid X_m = i)$$

does not depend on m.

Assuming that $(X_m)_{m\in\mathbb{N}_0}$ is a temporally homogeneous Markov chain, we denote by

$$p_{ij} := \mathbb{P}(X_{m+1} = j \mid X_m = i)$$

the probability of reaching the j-th state from the i-th state in a single step. The transition probability matrix P of $(X_m)_{m\in\mathbb{N}_0}$ is given by $P = (p_{ij})_{i,j=1}^n$. Obviously, P and L'' are directly related. We continue with the following definition:

DEFINITION 3.28. *Let $(X_m)_{m\in\mathbb{N}_0}$ denote a temporally homogeneous Markov chain. A row vector $\pi = (\pi_1 \cdots \pi_n) \in \mathbb{R}^{1\times n}$ is called a stationary distribution of $(X_m)_{m\in\mathbb{N}_0}$ if the conditions*

$$\sum_{i=1}^{n} \pi_i = 1$$

and

$$\pi P = \pi$$

hold.

Note that the multistep transition probability matrix $P^s = (p_{ij}^s)_{i,j=1}^n$ of $(X_m)_{m\in\mathbb{N}_0}$ is defined entry-wise:

$$p_{ij}^s := \mathbb{P}(X_s = j \mid X_0 = i) \tag{3.24}$$

i.e. each p_{ij}^s is given by the probability of reaching state j from state i in exactly s steps. The likelihood of reaching a certain state converges to a stationary probability, as specified in the following result (see Krengel (1988, p. 205f.) for a proof).

PROPOSITION 3.29 (Krengel (1988), Theorem 16.1)**.** *Let $(X_m)_{m\in\mathbb{N}_0}$ denote the temporally homogeneous Markov chain with multistep transition probability matrix P^s. The convergence*

$$(p_{i1}^s \cdots p_{in}^s) \xrightarrow{s\to\infty} (\pi_1 \cdots \pi_n)$$

holds true, where $\pi = (\pi_1 \cdots \pi_n) \in \mathbb{R}^{1\times n}$ denotes the unique stationary distribution of $(X_m)_{m\in\mathbb{N}_0}$ if there exists an $L \in \mathbb{N}$ such that $p_{ij}^L > 0$ for all $i, j \in \{1, ..., n\}$.

We return to our graph $G = (V, E, W)$: let $(Y_m)_{m\in\mathbb{N}_0}$ denote a stochastic process on the finite state space $\{1, ..., n\}$, where i represents the i-th vertex x_i in the vertex set V, such that the transition probabilities are given by

$$\mathbb{P}(Y_{m+1} = j \mid Y_m = i) = \frac{w_{ij}}{\deg(\mathrm{i})} \tag{3.25}$$

for all $m \in \mathbb{N}_0$. Based on this, it is easy to verify that $(Y_m)_{m\in\mathbb{N}_0}$ is a temporally homogeneous Markov chain, meaning we can denote $\mathbb{P}(Y_{m+1} = j \mid Y_m = i)$ by p_{ij}, and that the transition probability matrix P is unique and well-defined. We also refer to $(Y_m)_{m\in\mathbb{N}_0}$ as a random walk on G. We define the row vector $\pi = (\pi_1 \cdots \pi_n) \in \mathbb{R}^{1\times n}$ such that

$$\pi_i = \frac{\deg(i)}{\mathrm{vol}(V)}$$

Then, a short calculation shows that π defines a stationary distribution of $(Y_m)_{m\in\mathbb{N}_0}$. Additionally, the multistep transition probabilities p_{ij}^s of $(Y_m)_{m\in\mathbb{N}_0}$ converge to π_j if G is non-bipartite and connected. This can be proven easily as non-bipartiteness and connectedness are equivalent to $(Y_m)_{n\in\mathbb{N}_0}$ satisfying the condition in Proposition 3.29. This leads us to the following result, which follows from follows from a short calculation using the definitions of the conditional probabilities, Ncut, and the stationary distribution of (Y_m).

THEOREM 3.30 (von Luxburg (2007), Proposition 5)**.** *Assume that $G = (V, E, W)$ is connected and non-bipartite. Let $(Y_m)_{n\in\mathbb{N}_0}$ be the random walk on G defined by (3.25). For $A, B \subset V$, define*

$$\mathbb{P}(B \mid A) := \mathbb{P}(Y_1 \in B \mid Y_0 \in A)$$

Let $A, A^C \subset V$ be a partition of V. Then, the identity

$$\mathrm{Ncut}(A, A^C) = \mathbb{P}(A^C \mid A) + \mathbb{P}(A \mid A^C) \tag{3.26}$$

holds.

In other words, Theorem 3.30 states that solving (3.12) corresponds to finding the cut of

G that minimizes the probability of the random walk (Y_N) "switching" between A and A^C. In the remainder, we mainly approach spectral clustering from the graph cut perspective.

3.3 Spectral Clustering Algorithms

In the following, we present specific algorithms meant to implement spectral clustering (von Luxburg 2007, Section 4). For n data points $x_1, ..., x_n$, let $G = (V, E, W)$ be such that

$$V = \{x_1, ..., x_n\}$$

Recall (3.17) for a given $k \leq n$: in general, solving (3.17) is NP-hard, thus we consider instead the relaxation (3.19). We saw previously that the solution of (3.19) is given by the eigenvectors to the k smallest eigenvalues of L'. We present concrete algorithms for this computation task in the following. It is important to note that their viability depends on the size of the k-th eigengap $|\lambda_k - \lambda_{k-1}|$, as we explore in more detail in Section 3.4. The following two algorithms were presented in von Luxburg (2007, Section 4), with the first being due to Shi and Malik (2000) and the second to von Luxburg (2007). The first algorithm we present concerns normalized spectral clustering. For more details on the k-means clustering algorithm, we refer to Hartigan and Wong (1979) and LLoyd (1982).

ALGORITHM 3.31 (Normalized Spectral Clustering)**.** *Input: data sample $V = x_1, ...x_n$, the undirected graph $G = (V, E)$*

(1) Compute the similarity matrix $W \in \mathbb{R}^{n \times n}$ based on a similarity function, the degree matrix D and the unnormalized Laplacian matrix $L = D - W$.

(2) Solve the generalized eigenproblem (3.4) for the eigenvectors to the k smallest general eigenvalues. Arrange them in the matrix $T = (u_1 \cdots u_k) \in \mathbb{R}^{n \times k}$.

(3) Group the rows $\psi_i \in \mathbb{R}^k$ of T using the k-means algorithm and assign the i-th vertex according to the cluster membership of ψ_i.

Output: a partition $A_1, ..., A_k$ of the vertex set V.

Note that in the case of $k = 2$, we can omit the k-means step, and cluster the input data based on a threshold: more specifically, we assign the i-th vertex to A if $v_2^i \geq 0$, and to A^C if $v_2^i < 0$. This intuition is motivated by Theorem 3.25, where we assumed $k = 2$, and should become clearer from our subsequent discussions in Section 3.4. An alternative algorithm for normalized spectral clustering has been proposed by Ng, Jordan, and Weiss 2001 (also see von Luxburg 2007). Next, we present an algorithm for unnormalized spectral clustering based on an algorithm presented by von Luxburg 2007.

ALGORITHM 3.32 (Unnormalized Spectral Clustering (von Luxburg 2007))**.** *Input: data sample $V = x_1, ...x_n$, the undirected graph $G = (V, E)$*

(1) Compute the similarity matrix $W \in \mathbb{R}^{n \times n}$ based on a similarity function, the

degree matrix D and the unnormalized Laplacian matrix $L = D - W$.

(2) Compute the eigenvectors $v_1, ..., v_k$ to the k smallest positive eigenvalues of L. Arrange them in the matrix $V \in \mathbb{R}^{n \times k}$.

(3) Group the rows $\psi_i \in \mathbb{R}^k$ of T using the k-means algorithm and assign the i-th vertex according to the cluster membership of ψ_i.

Output: a partition $A_1, ..., A_k$ of the vertex set V.

3.4 Laplacian Eigengaps and the Davis-Kahan Theorem

The viability of procedures such as Algorithm 3.31 depends on the size of the eigengap $|\lambda_{k+1} - \lambda_k|$. Building on the approaches introduced by von Luxburg (2007) and Ng, Jordan, and Weiss (2001), we discuss the reasons for this in the following. The reasons for the eigengaps's relevance appear clearest when considering unnormalized spectral clustering: assume the ideal scenario of no across-cluster similarity, i.e. for a partition $\bigcup_{l=1}^{k} C_l = V$, the similarity matrix $W = (w_{ij})_{i,j=1}^{n}$ satisfies

$$w_{ij} = 0 \iff x_i \in C_l \text{ and } x_j \in C_m \text{ for } m \neq l$$

for any two distinct vertices $x_i, x_j \in V$. Thus, $C_1, ..., C_k$ match the components of G exactly. According to Theorem 3.13, k is equal to the geometric multiplicity of the eigenvalue 0 of L, and the corresponding eigenvectors are $\mathbf{1}_{C_1}, ..., \mathbf{1}_{C_k}$. In this ideal case, the matrix $T \in \mathbb{R}^{n \times k}$ obtained through the second step of Algorithm 3.32 is given by

$$T = (\mathbf{1}_{C_1} \cdots \mathbf{1}_{C_k})$$

and its i-th row is

$$\psi_i = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{pmatrix} \in \mathbb{R}^k$$

where the non-zero entry indicates the cluster membership of the i-th vertex. Note that this creates k groups of identical vectors in $\mathbb{R}^k$: for $l = 1, ..., k$, for all, ψ_i has its only non-zero component at the l-th position. Hence, the output of the k-means algorithm in Algorithm 3.32 is trivial: for $l = 1, ..., k$, all ψ_i for which the entry 1 is at the l-th position (i.e. $\psi_i^{(l)} = 1$), the k-means algorithm assigns ψ_i to the l-th group of vectors, and thus assigns the associated vertices x_i to A_l, which produces the output

$$A_l = C_l$$

for $l = 1, ..., k$ which is identical to the given components. We now consider a small perturbation of the ideal scenario we discussed. Then, the inter-cluster similarities may be small, but not exactly equal to 0. Intuitively, we expect the k clusters to be associated with the k smallest positive eigenvalues of L, and the corresponding eigenvectors to be sufficiently close to $\mathbf{1}_{C_1}, ..., \mathbf{1}_{C_k}$ such that the k-means algorithm will group all ψ_i based on the true clusters. Note that this intuitive reasoning is not as straightforward when considering normalized

Spectral Clustering. This is due to the fact that here, in the "ideal" case, for all x_i within a cluster C_l, the rows ψ_i are not identical: Theorem 3.20 states that the eigenvectors of L' to the eigenvalue 0 are given by $D^{1/2}\mathbf{1}_{C_1}, ..., D^{1/2}\mathbf{1}_{C_k}$, which implies that

$$\psi_i = \begin{pmatrix} 0 & 0 & \cdots & 0 & \sqrt{\deg(i)} & 0 & \cdots & 0 \end{pmatrix} \in \mathbb{R}^k$$

However, it is still reasonable to assume that k-means will group ψ_i based on cluster membership of x_i: recall that

$$\deg(i) = \sum_{\substack{j=1 \\ j \neq i}}^{n} w_{ij}$$

Because across-cluster similarity is zero, we obtain

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} w_{ij} = \sum_{\substack{x_j \in C_l \\ j \neq i}} w_{ij}$$

which suggests that vertices within the same cluster are more likely to have similar degrees than vertices from different clusters, in particular for large datasets. The Davis-Kahan theorem helps us formalize the perturbation argument we outlined. Before presenting the result, we need to provide necessary definitions (see Björck and Golub (1973) and Golub and Van Loan (1989, Section 2.3)).

DEFINITION 3.33 (Frobenius Norm)**.** *The Frobenius norm $\|\cdot\|_F$ of a matrix $M \in \mathbb{R}^{d \times p}$ is defined as*

$$\|M\|_F := \sqrt{\operatorname{tr}(M^T M)}$$

Let $\mathcal{U}$ and $\mathcal{V}$ denote two subspaces of $\mathbb{R}^d$ such that

$$q = \dim(\mathcal{V}) \leq \dim(\mathcal{U}) = p$$

holds for $p \leq d$. We define the first principal angle between $\mathcal{U}$ and $\mathcal{V}$ as the minimum angle between vectors in $\mathcal{U}$ and $\mathcal{V}$. In particular, the first principal angle θ_1 is given by the equation

$$\cos(\theta_1) = \max_{u \in \mathcal{U}} \max_{v \in \mathcal{V}} u^T v \text{ for } \|u\| = \|v\| = 1 \tag{3.27}$$

This allows us to define the j-th principal angle between $\mathcal{U}$ and $\mathcal{V}$ in a recursive way (see Björck and Golub 1973 for more detail).

DEFINITION 3.34 (Principal Angles between Subspaces of $\mathbb{R}^d$)**.** *For $j = 1, ..., q$, the j-th principal angle θ_j between $\mathcal{U}$ and $\mathcal{V}$ is defined as*

$$\cos(\theta_j) = \max_{u \in \mathcal{U}} \max_{v \in \mathcal{V}} u^T v \text{ for } \|u\| = \|v\| = 1$$

under the constraint

$$u_i^T u = 0 \text{ and } v_i^T v = 0 \tag{3.28}$$

for $i = 1, ..., j-1$.

Note that if $\mathcal{V}$ and $\widetilde{\mathcal{V}}$ are the column spaces of two orthogonal matrices $M, \widetilde{M} \in \mathbb{R}^{d \times p}$, we

obtain the relationship

$$\sigma_j = \cos(\theta_j),\ j = 1, ..., p \tag{3.29}$$

between the principal angles $\theta_1, ..., \theta_p$ of $\mathcal{V}$ and $\widetilde{\mathcal{V}}$ on one hand, and the singular values $\sigma_p \leq ... \leq \sigma_1$ of $M^T M$ on the other (see Yu, Wang, and Samworth (2014) and Knyazev and Zhu (2012)). By

$$\Theta(\mathcal{V}, \widetilde{\mathcal{V}}) := \text{diag}[\theta_1, ..., \theta_p] \in \mathbb{R}^{p \times p}$$

we denote the diagonal matrix of the principal angles of the column spaces $\mathcal{V}$ and $\widetilde{\mathcal{V}}$. Additionally, the sine matrix $\sin\Theta(M, \widetilde{M})$ is defined by the componentwise definition (von Luxburg 2007, p. 406)

$$\sin\Theta(\mathcal{V}, \widetilde{\mathcal{V}}) := \text{diag}[\sin(\theta_1), ..., \sin(\theta_p)] \in \mathbb{R}^{p \times p}$$

Note that the function $\sin(\theta)$ is increasing on the interval $[0, \pi/2]$. This motivates using $\|\sin(\mathcal{V}, \widetilde{\mathcal{V}})\|_F$ as a measure of the distance between $\mathcal{V}$ and $\widetilde{\mathcal{V}}$. The following version of the Davis-Kahan Theorem has been presented in von Luxburg (2007, Theorem 7), who reference Stewart and Sun (1990, Section V.3) for the underlying theory.

THEOREM 3.35. *Let $A, H \in \mathbb{R}^{d \times d}$ be symmetric matrices, and let $\widetilde{A}$ denote the perturbation*

$$\widetilde{A} := A + H$$

For an interval $J \subset \mathbb{R}$, let $\sigma_J(A)$ and $\sigma_J(\widetilde{A})$ denote the respective portions of $\sigma(A)$ and $\sigma(\widetilde{A})$ that are contained in J. $\mathcal{V}$ and $\widetilde{\mathcal{V}}$ denote the images of the spectral projections $\sigma_J(A)$ and $\sigma_J(\widetilde{A})$ that induce. The quantity

$$\delta = \min_{\lambda \in \sigma(A) \setminus J} \{|\lambda - s| \mid s \in J\}$$

is the minimum distance between $\sigma_J(A)$ and the rest of the spectrum of A. Then, the inequality

$$\|\sin\Theta(\mathcal{V}, \widetilde{\mathcal{V}})\|_F \leq \frac{\|H\|_F}{\delta} \tag{3.30}$$

holds.

Note that this result directly applies to Spectral Clustering: choose $J = [0, \lambda_k]$, and let $A = L$ and $\widetilde{A} = \widetilde{L}$ denote two Laplacians such that $\widetilde{L} = L + H$. When we compare the subspaces

$$\mathcal{V} = \text{span}\{v_1, ..., v_k\}$$

and

$$\widetilde{\mathcal{V}} = \text{span}\{\widetilde{v}_1, ..., \widetilde{v}_k\}$$

then the minimum distance δ is given by

$$\delta = \min_{l=k+1,...,n} \{|\lambda_l - s| s \in J\}$$

$$= |\lambda_{k+1} - \lambda_k|$$

which follows from the non-decreasing order of the eigenvalues and the definition of the interval J. Here, $v_1, ..., v_k$ and $\widetilde{v}_1, ..., \widetilde{v}_k$ denote the eigenvectors to the k smallest eigenvalues of L and $\widetilde{L}$, respectively. We observe from (3.30) that the size of the eigengap $|\lambda_{k+1} - \lambda_k|$ directly impacts the distance between the eigenspaces of L and $\widetilde{L}$ as measured by the sine of their principal angles: for δ approaching zero, the upper bound $\frac{\|H\|_F}{\delta}$ grows infinitely large.

3.5 Similarity Functions

Recall that spectral clustering algorithms such as Algorithm 3.31 compute the similarity matrix W based on a similarity function that is defined on the data. We now address these functions specifically:

DEFINITION 3.36. *Let (X, d) be a metric space. A function $k : X \times X \longrightarrow \mathbb{R}$ is called a similarity function if both of the following conditions are satisfied*

(i) The function is non-negative, i.e. $k(x, y) \geq 0$ for all $x, y \in X$.

(ii) The function is symmetric, i.e. $k(x, y) = k(y, x)$ for all $x, y \in X$.

Subsequently, we address two relevant similarity functions. For additional examples of similarity graphs and functions, we refer to von Luxburg (2007, Section 2).

3.5.1 Gaussian Similarity Kernels and Width Selection

We introduce Gaussian similarity kernels first, as this type is the most relevant to our subsequent discussions (von Luxburg, Belkin, and Bousquet 2008, p. 573). Suppose that we are given a data set $x_1, ..., x_n \in \mathbb{R}^d$ for $d \in \mathbb{N}$, and let $G = (V, E)$ denote an undirected graph such that

$$V = \{x_1, ..., x_n\}$$

Then, the function

$$k : \begin{cases} \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R} \\ (x, y) \longmapsto \exp\left(-\frac{\|x-y\|^2}{\tau^2}\right) \end{cases} \tag{3.31}$$

is called the Gaussian similarity kernel with kernel width $\tau > 0$. It is easy to verify that the graph defined by the matrix

$$W = (k(x_i, x_j))_{i,j=1}^n$$

is weighted and fully connected. The key challenge when working with this function is the choice of τ. To address this in the practical examples that follow, we borrow an approach from the theory of bandwidth selection for kernel density estimation: assuming that $x_1, ..., x_n \in \mathbb{R}$, we choose the bandwidth

$$\tau = 1.06\sigma n^{-1/5} \tag{3.32}$$

This selection process is called Scott's rule (see Silverman (1986, p. 45)). Note that besides the obvious similarity of Gaussian KDE kernels to the function defined in (3.31), our choice is motivated by its robust performance in the examples discussed later. We also emphasize

that this rule only applies to univariate data. For further reading on KDE in general and bandwidth selection in particular, we refer to Silverman (1986).

3.5.2 The ϵ-Neighborhood Graph

We briefly present an alternative to the Gaussian similarity kernel: the ϵ-neighborhood similarity function (von Luxburg 2007, Section 2.2). Assume that the undirected graph $G = (V, E)$ is given by the data points $x_1, ..., x_n \in X$ such that

$$V = \{x_1, ..., x_n\}$$

Here, X is a metric space with a distance metric d. First, we introduce the ϵ-neighborhood similarity function for $\epsilon > 0$

$$k_\epsilon : \begin{cases} X \times X \longrightarrow \mathbb{R} \\ (x, y) \longmapsto \begin{cases} 1 & \text{if } d(x, y) < \epsilon \\ 0 & \text{otherwise} \end{cases} \end{cases}$$

It is easy to verify that this function is symmetric by construction. For a given $\epsilon > 0$, the graph $G = (V, E, W)$ with the weight matrix

$$W = (k_\epsilon(x_i, x_j))_{i,j=1}^n$$

is called the ϵ-neighborhood graph. G is an unweighted graph because $k(x_i, x_j) \in \{0, 1\}$ for all tuples of vertices.

3.6 Real-World Applications of Spectral Clustering

The example we discuss next illustrates the performance of spectral clustering on real-world data: we consider list of $n = 50$ countries with the value that agriculture, fishing and forestry added to their economies in the year 2023, measured in percent of their respective GDP (see World Bank (2025)). To keep things concise, we will refer to this metric simply as *the (economic) indicator* or *the (economic) metric*. Of the countries included, 25 rank "low" or "medium" on the 2023 Human Development Index, whereas the other 25 rank "very high" (see United Nations Development Programme (2025, p. 274-278)). Table 3.1 displays the countries, subdivided into the aforementioned categories. The distribution of the indicator, as shown in Figure 3.1, reveals a sizable gap in the data: very few countries fall into the 5-15 percent range. From Figure 3.2, it appears that the indicator is elevated for countries with "low" or "medium" HDI, and low for countries with "very high" HDI.

We are interested in whether spectral clustering is able to detect these same patterns: we choose a Gaussian similarity function and determine the kernel width using Scott's rule (see Subsection 3.5.1). It is important to note that in this instance, our algorithm uses the

"Low" or "Medium" HDI	"Very High" HDI
Senegal	Norway
Congo, Dem. Rep.	Netherlands
Liberia	Germany
Ethiopia	Australia
Pakistan	Sweden
Gambia	Denmark
Benin	Iceland
Malawi	Switzerland
Guinea-Bissau	Hong Kong SAR, China
Djibouti	Belgium
Sudan	Ireland
Guinea	Finland
Afghanistan	Singapore
Mozambique	United Arab Emirates
Madagascar	United Kingdom
Sierra Leone	United States
Burkina Faso	Korea, Rep.
Burundi	Slovenia
Mali	Austria
Niger	Japan
Chad	Malta
Central African Republic	Luxembourg
Lesotho	France
Haiti	Israel
Tanzania	Spain

Table 3.1: The division is based on the ranking at United Nations Development Programme (2025, p. 274-278).

Gaussian similarity kernel

$$k : \begin{cases} \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R} \\ (x, y) \longmapsto \exp\left(-\frac{\|x-y\|^2}{2\tau^2}\right) \end{cases}$$

instead of the one in (3.31). However, the width heuristic established in Subsection 3.5.1 still applies, as evidenced by the performance of the algorithm below. We compute the similarity matrix and the normalized graph Laplacian with its eigenvalues (see Figure 3.3). We select $k = 2$ to find out if spectral clustering can distinguish the highly developed countries from the others based on the indicator (it is important to note that other eigengaps are larger than the second one, so other choices of k may be feasible). A normalized spectral clustering algorithm yields the output shown in Figure 3.4 and Table 3.2. We observe that the result is close to the division according to HDI, with the exception of Djibouti and Lesotho. This coincides with the fact these two countries are outliers in terms of their economies' reliance on agriculture, forestry, and fishing, which is weaker than for others in the "low or medium HDI" category (see Figure 3.2 and United Nations Development Programme (2025)).

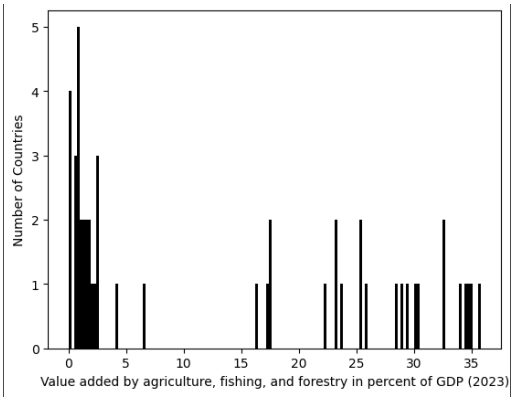


Figure 3.1: Histogram displaying the distribution of the economic metric of across all 50 countries

3.7 The Number of Clusters

We have seen that spectral clustering algorithm requires the number of clusters to be predefined. This raises the challenge of determining the best possible k. In the case of the example discussed in Section 3.6, we were able to select the number of clusters based on prior knowledge about the data. Absent such information, there exist various approaches for choosing k, many of which are listed and referenced in von Luxburg (2007, Section 8.3). One such technique is called the *eigengap heuristic*: In Section 3.4, we saw that the size of the k-th eigengap impacts the feasibility of spectral clustering for a given k. In particular, Theorem 3.35 states that the distance between the subspace associated with the k smallest eigenvalues of a graph Laplacian L' and the corresponding subspace of a perturbation matrix $L' + H$ is bounded in the way

$$\|\sin\Theta(\mathcal{V}, \widetilde{\mathcal{V}})\|_F \leq \frac{\|H\|_F}{|\lambda_k - \lambda_{k+1}|}$$

This suggests choosing k such that the eigengap $|\lambda_k - \lambda_{k+1}|$ is maximal. Another technique for determining the number of clusters has been proposed by Mur et al. (2016). The method, called *Spectral Global Silhouette method*, applies to a range of partitional clustering algorithms and utilizes a framework that is partially grounded in spectral clustering itself.

index	Country Name	2023 [YR2023]
0	Senegal	17.4112512141863
1	Norway	2.08014844089816
2	Netherlands	1.72255687762915
3	Congo, Dem. Rep.	17.4398231813584
4	Liberia	34.5558086560364
5	Ethiopia	35.7864807848152
6	Germany	0.841992091839782
7	Australia	2.57350206211225
8	Sweden	0.992411153445759
9	Denmark	0.756712776534002
10	Pakistan	23.3342653700076
11	Gambia, The	23.2246283367096
12	Benin	25.4019723667563
13	Malawi	30.3792338833699
14	Guinea-Bissau	34.0155209925726
15	Djibouti	2.49007578360367
16	Sudan	30.2756760074835
17	Guinea	29.4714397653024
18	Afghanistan	34.7432471445174
19	Mozambique	25.9114511595886
20	Madagascar	22.4047388943658
21	Sierra Leone	29.0697647996639
22	Burkina Faso	16.3344550497717
23	Burundi	25.341874228406
24	Mali	32.4904123927446
25	Niger	32.5058218469012
26	Chad	35.0417759929849
27	Central African Republic	28.6110555239693
28	Iceland	4.28312373737575
29	Switzerland	0.624912461460987
30	Hong Kong SAR, China	0.037907340516847
31	Belgium	0.769218254127329
32	Ireland	0.875702556096056
33	Finland	2.30587062196186
34	Singapore	0.0288498020075513
35	United Arab Emirates	0.701206777795139
36	United Kingdom	0.575689228141938
37	United States	0.989069218972718
38	Korea, Rep.	1.59560572785029
39	Slovenia	1.51932161939818
40	Austria	1.29697030335468
41	Japan	0.937452748505902
42	Malta	0.0038945813159872
43	Lesotho	6.63922933135679
44	Haiti	17.4571299152426
45	Tanzania	23.6871473778631
46	Luxembourg	0.212423661674197
47	France	1.74050275246234
48	Israel	1.28614593929466
49	Spain	2.4997263609206

Figure 3.2: List of included countries with the economic indicator (two pages)

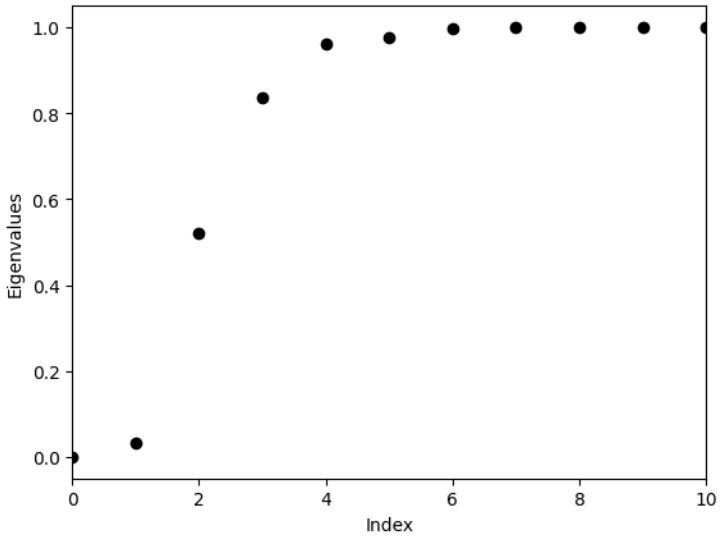


Figure 3.3: Eigenvalues of the normalized Laplacian of the economic indicator data using a Gaussian kernel with Scott's rule for kernel width selection

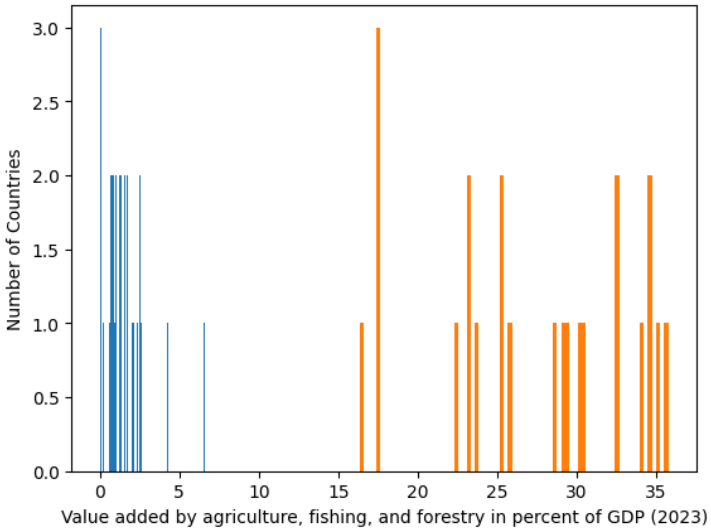


Figure 3.4: Histogram displaying the distribution of the economic metric of across all 50 countries, coloring based on the spectral clustering output for $k = 2$

"Low" or "Medium" HDI	"Very High" HDI
Senegal	Norway
Congo, Dem. Rep.	Netherlands
Liberia	Germany
Ethiopia	Australia
Pakistan	Sweden
Gambia	Denmark
Benin	Iceland
Malawi	Switzerland
Guinea-Bissau	Hong Kong SAR, China
Djibouti	Belgium
Sudan	Ireland
Guinea	Finland
Afghanistan	Singapore
Mozambique	United Arab Emirates
Madagascar	United Kingdom
Sierra Leone	United States
Burkina Faso	Korea, Rep.
Burundi	Slovenia
Mali	Austria
Niger	Japan
Chad	Malta
Central African Republic	Luxembourg
Lesotho	France
Haiti	Israel
Tanzania	Spain

Table 3.2: Clusters of the 50 countries, based on the economic indicator, $k = 2$, indicated by highlighting

4 Constistency of Spectral Clustering

This section contains the main part of our discussion: we explore the conditions under which spectral clustering provides sensible results. For this purpose, we introduce a concept of consistency of spectral clustering that is based on the convergence of Laplacian eigenvectors. We prove sufficient conditions for such convergence, and discuss associated convergence rates. We also highlight crucial distinctions between normalized and unnormalized spectral clustering. All of the above closely follows the approach presented by von Luxburg, Belkin, and Bousquet (2008). Throughout the section, we also provide a simple simulation to illustrate the convergence behavior of Laplacian eigenvectors.

4.1 Motivation and General Assumptions

As we saw in Subsection 3.2.1, minimizing (3.17) is equivalent to finding the partition that minimizes $\mathrm{Ncut}(A_1, ..., A_k)$. However, this does not necessarily hold for the relaxation (3.19). Recall that the solution of the relaxed problem

$$\underset{T \in \mathbb{R}^{n \times k}}{\operatorname{argmin}} \{\operatorname{tr}\left(T^T L' T\right) \mid T^T T = I\}$$

is given by

$$T = (v_1 \cdots v_k) \in \mathbb{R}^{n \times k}$$

Here, $v_1, ..., v_k \in \mathbb{R}^n$ denote the eigenvectors to the k smallest non-zero eigenvalues $\lambda_1 \leq ... \leq \lambda_k$ of L', are the column vectors of T. This raises the question of under what conditions spectral clustering provides sensible results. One such case is when the eigenvalue matrix T equals $D^{1/2}H$, where H is of the form detailed in (3.18). Our next goal is to investigate the question under more general assumptions. We formalize them in the following remark.

REMARK 4.1 (General Assumptions)**.** *Suppose that $\mathcal{X}$ is a compact metric space, $B(\mathcal{X})$ its Borel σ-algebra and $\mathbb{P}$ a probability measure on $(\mathcal{X}, B(\mathcal{X}))$. Let $(X_n)_{n \in \mathbb{N}}$ be a sequence of independent random variables with values in $\mathcal{X}$ and distributed according to the identical distribution $\mathbb{P}$. We further assume that the similarity function*

$$k : \begin{cases} \mathcal{X} \times \mathcal{X} \longrightarrow [0, \infty) \\ (x, y) \longmapsto k(x, y) \end{cases}$$

is continuous and that there exists a constant $c > 0$ such that $k(x, y) > c$ for all $x, y \in \mathcal{X}$. By G_n, we denote the graph with the vertices $X_1, ..., X_n$ and the edge weights $(k(X_i, X_j))$ for $i, j = 1, ..., n$.

Subsequently, G_n has the similarity matrix

$$K_n = (k(X_i, X_j))_{i,j=1}^n \in \mathbb{R}^{n \times n}$$

Note that K_n corresponds to W from Section 3 the difference being that K_n depends on n

and is determined by k and $X_1, ..., X_n$. For $n \in \mathbb{N}$ and $i = 1, ..., n$, we define

$$\deg_n(i) := \sum_{j=1, j\neq i} k(X_i, X_j)$$

as the degree of the vertex X_i. With this notation, we can define the matrix sequence $(D_n)_{n\in\mathbb{N}}$ of the degree matrices of G_n by

$$D_n := \operatorname{diag}[\deg_n(1), \ldots, \deg_n(n)] \in \mathbb{R}^{n\times n}$$

Furthermore, we define the sequence of unnormalized Laplacians $(L_n)_{n\in\mathbb{N}}$ by

$$L_n = D_n - K_n$$

and the sequence $(L'_n)_{n\in\mathbb{N}}$ of normalized Laplacian matrices of G_n

$$L'_n = I_n - D_n^{-1/2} K_n D_n^{-1/2}$$

Recall that

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

is the empirical measure on $(\mathcal{X}, B(\mathcal{X}))$ induced by $X_1, ..., X_n$, where δ_X denotes the Dirac measure on $(\mathcal{X}, B(\mathcal{X}))$ for $x \in \mathcal{X}$. We use this to extend the definition of $\deg_n(i)$ to the whole data space $\mathcal{X}$.

DEFINITION 4.2. *For a continuous similarity function k on $\mathcal{X}$, we define the n-th degree function, for $n \in \mathbb{N}$, with respect to the measure $\mathbb{P}_n$*

$$d_n : \mathcal{X} \longrightarrow \mathbb{R}$$

$$d_n(x) := \int_{\mathcal{X}} k(x, y) d\mathbb{P}_n(y)$$

We also define the function $d : \mathcal{X} \longrightarrow \mathbb{R}$

$$d(x) := \int_{\mathcal{X}} k(x, y) d\mathbb{P}(y)$$

and refer to it as the limit degree function on $\mathcal{X}$.

The functions d_n and d are continuous: this is a consequence of the continuity of k, the compactness of $\mathcal{X}$, and the dominant convergence theorem (Klenke 2006, Corollary 6.26).We will later see that d is indeed the uniform limit of the sequence $(d_n)_{n\in\mathbb{N}}$ under appropriate assumption on k. Next, we prove some properties of the degree function and the limit degree function.

PROPOSITION 4.3. *For any given k, the functions d_n and d have the following properties:*

(i) c bounds d and d_n from below. In particular, d and d_n are strictly positive.

(ii) d and d_n are bounded from above by $\|k\|_\infty$

(iii) For $n \in \mathbb{N}$ and $i = 1, \dots, n$, we have the identity

$$\frac{\deg_n(i)}{n} = d_n(X_i) \tag{4.1}$$

Proof. To prove (i), we consider the definition of d_n and rewrite it as follows

$$\int_{\mathcal{X}} k(x,y) d\mathbb{P}_n(y) = \frac{1}{n} \sum_{i=1}^{n} k(X_i, x) \tag{4.2}$$

For each individual summand on the right-hand side of (4.2) is bounded from below by $c > 0$. It follows that

$$\sum_{i=1}^{n} k(X_i, x) > \frac{nc}{n} = c$$

To prove the same for d, recall that k is a strictly positive function. Therefore, we can bound d as follows:

$$\int_{\mathcal{X}} k(x,y) d\mathbb{P}(y) \geq c \cdot \mathbb{P}(\mathcal{X}) = c \tag{4.3}$$

which proves (i). To prove (ii), recall that for every probability measure μ on $\mathcal{X}$:

$$\int_{\mathcal{X}} k(x,y) d\mu(y) \leq \|k\|_\infty$$

Since both $\mathbb{P}$ and $\mathbb{P}_n$ are probability measures on $\mathcal{X}$, d and d_n are bounded by $\|k\|_\infty$ from above, which proves (ii). To prove (iii), we use the identity in (4.2) to obtain

$$n \cdot d_n(X_i) = \sum_{j=1}^{n} k(X_i, X_j)$$

The right-hand side equals the definition of $\deg_n(i)$, which proves (iii). □

In the remainder of this section, we investigate the question of when spectral clustering, under the assumptions established in Remark 4.1, delivers sensible clustering results. Recall the spectral clustering algorithms we discussed in Subsection 3.3: partitioning the data into $k < n$ disjoint subsets is achieved by retrieving the eigenvectors to the k smallest positive eigenvalues of L'_n (or L_n). For this reason, we study the behavior of such Laplacian eigenvectors for $n \to \infty$. We restrict our investigation to normalized spectral clustering for now – we shall return to the unnormalized case later. Before we proceed, keep in mind that studying the eigenvectors of L'_n and L''_n is equivalent because of Proposition 3.17.

DEFINITION 4.4. *A sequence $(\lambda_n)_{n\in\mathbb{N}}$ is called a sequence of k-th eigenvalues of $(L'_n)_{n\in\mathbb{N}}$ if for all $n \in \mathbb{N}$, λ_n is a k-th smallest positive eigenvalue of L'_n, i.e. L'_n has exactly $k-1$ eigenvalues that are less than or equal to λ_n. In this case, a sequence $(v_n)_{n\in\mathbb{N}}$ is called a sequence of k-th eigenvectors of $(L'_n)_{n\in\mathbb{N}}$ if for all $n \in \mathbb{N}$, $v_n \in \mathbb{R}^n$ is an eigenvector of L'_n to the eigenvalue λ_n.*

Suppose that $(v_n)_{n\in\mathbb{N}}$ is a sequence of k-th eigenvectors of $(L'_n)_{n\in\mathbb{N}}$. Let $(\lambda_n)_{n\in\mathbb{N}}$ be the corresponding sequence of k-th eigenvalues. In the following, we study the convergence behavior of $(v_n)_{n\in\mathbb{N}}$ in the sense of Defintion 2.30. For this purpose, we define the sequence

$(\rho_n)_{n\in\mathbb{N}}$ of restriction operators on $C(\mathcal{X})$ as follows:

$$\rho_n : \begin{cases} C(\mathcal{X}) \longrightarrow \mathbb{R}^n \\ f \mapsto \rho_n f = (f(X_1), ..., f(X_n))^T \end{cases} \tag{4.4}$$

Note that ρ_n maps any f to its evaluations on the first n elements of the random data sequence $(X_n)_{n\in\mathbb{N}}$. As we will see below, this allows us to define convergence of $(v_n)_{n\in\mathbb{N}}$ under $(\rho_n)_{n\in\mathbb{N}}$ in the sense of Definition 2.30. Before presenting the main convergence result, we introduce necessary preliminary results: first, we define the operator

$$T : \begin{cases} C(\mathcal{X}) \longrightarrow C(\mathcal{X}) \\ f \longmapsto \int_{\mathcal{X}} \frac{k(\cdot,y)}{\sqrt{d(\cdot)d(y)}} f(y) d\mathbb{P}(y) \end{cases}$$

and the operator sequence

$$\widehat{T}_n : \begin{cases} C(\mathcal{X}) \longrightarrow C(\mathcal{X}) \\ f \longmapsto \int_{\mathcal{X}} \frac{k(\cdot,y)}{\sqrt{d_n(\cdot)d_n(y)}} f(y) d\mathbb{P}_n(y) \end{cases}$$

which are bounded from above as we prove in the following:

PROPOSITION 4.5. *For $n \in \mathbb{N}$, the operator T and the operator $\widehat{T}_n$ are bounded from above, i.e.*

$$\|Tf\|_\infty \leq \frac{\|k\|_\infty}{c} \|f\|_\infty \tag{4.5}$$

and

$$\|\widehat{T}_n f\|_\infty \leq \frac{\|k\|_\infty}{c} \|f\|_\infty \tag{4.6}$$

for all $f \in C(\mathcal{X})$. In particular, T and $\widehat{T}_n$ are bounded linear operators on $(C(\mathcal{X}), \|\cdot\|_\infty)$ (see Definition 2.12).

Proof. By definition

$$\|Tf\|_\infty = \sup_{x\in\mathcal{X}} |Tf(x)|$$

$$= \sup_{x\in\mathcal{X}} \left| \int_{\mathcal{X}} \frac{k(x,y)}{\sqrt{d(x)d(y)}} f(y) d\mathbb{P}(y) \right|$$

$$\leq \sup_{x\in\mathcal{X}} \int_{\mathcal{X}} \left| \frac{k(x,y)}{\sqrt{d(x)d(y)}} f(y) \right| d\mathbb{P}(y)$$

$$\leq \frac{\|k\|_\infty}{c} \sup_{x\in\mathcal{X}} \int_{\mathcal{X}} |f(y)| d\mathbb{P}(y)$$

$$\leq \frac{\|k\|_\infty}{c} \|f\|_\infty$$

The last inequality follows from $\|f\|_{L^1} \leq \|f\|_\infty$, which holds because $\mathbb{P}$ is a probability measure. This proves the bound in (4.5). The proof of (4.6) works analogously. □

Let I denote the identity operator in $\mathbb{B}(C(\mathcal{X}))$. By U', we denote the integral operator

$$U' : \begin{cases} C(\mathcal{X}) \longrightarrow C(\mathcal{X}) \\ f \longmapsto (I - T)f \end{cases}$$

and by $(U'_n)_{n\in\mathbb{N}}$ the operator sequence

$$U'_n : \begin{cases} C(\mathcal{X}) \longrightarrow C(\mathcal{X}) \\ f \longmapsto (I - \widehat{T}_n)f \end{cases}$$

Note that for $n \in \mathbb{N}$, $\hat{T}_n$ and T are compact operators, which proves useful when studying their spectral properties:

PROPOSITION 4.6. *For $n \in \mathbb{N}$, T and $\widehat{T}_n$ are compact linear operators.*

Proof. Both are integral operators of the form (2.5), hence the result is a consequence of Proposition 2.15. □

This implies that U' and U'_n are compact perturbations of the identity operator, and thus their respective spectra have the following properties:

COROLLARY 4.7. *For $n \in \mathbb{N}$, the operators U' and U'_n have the following spectral properties:*

(i) The spectrum $\sigma(U')$ consists of countably many, non-negative eigenvalues with finite multiplicity. The essential spectrum $\sigma_{\text{ess}}(U')$ of U' is a subset of $\{1\}$.

(ii) The spectrum $\sigma(U'_n)$ consists of a finite number of non-negative eigenvalues with finite multiplicity. The essential spectrum $\sigma_{\text{ess}}(U'_n)$ of U' is a subset of $\{1\}$.

Proof. By definition, U' is a compact perturbations of the identity operator. Hence, (i) follows from Theorem 2.21 and Proposition 2.23. The first part of (ii) is a consequence of Proposition 4.8, as we will see below. The statement about the essential spectrum $\sigma(U'_n)$ also follows because U'_n is a compact perturbation of the identity. □

In the following, we study the relationship between the operator U'_n and the normalized Laplacian L'_n for a given $n \in \mathbb{N}$. Recall that ρ_n is the restriction operator that we defined in (4.4). For any function $f \in C(\mathcal{X})$, we obtain

$$\begin{aligned} \rho_n U'_n f &= \rho_n f - \rho_n \left[\int_{\mathcal{X}} \frac{k(\cdot, y)}{\sqrt{d_n(\cdot) d_n(y)}} f(y) d\mathbb{P}_n(y) \right] \\ &= \rho_n f - \rho_n \left[\frac{1}{n} \sum_{j=1}^{n} \frac{k(\cdot, X_j)}{\sqrt{d_n(\cdot) d_n(X_j)}} f(X_j) \right] \\ &= \rho_n f - \underbrace{\left(\sum_{j=1}^{n} \frac{k(X_1, X_j)}{\sqrt{\deg_n(1)\deg_n(j)}} f(X_j), \ldots, \sum_{j=1}^{n} \frac{k(X_n, X_j)}{\sqrt{\deg_n(n)\deg_n(j)}} f(X_j) \right)^T}_{D_n^{-1/2} K_n D_n^{-1/2} \cdot (f(X_1), \cdots, f(X_n))^T} \end{aligned}$$

where the last identity is a consequence of Proposition 4.3. This last identity further implies that

$$\rho_n U_n' f = \rho_n f - D_n^{-1/2} K_n D_n^{-1/2} (\rho_n f)$$

which implies the crucial result

$$\rho_n U_n' f = L_n' \rho_n f \tag{4.7}$$

This leads us to the following proposition, which establishes a one-to-one relationship between the respective spectra of L_n' and U_n'.

PROPOSITION 4.8. *For $n \in \mathbb{N}$, let $\lambda \in \sigma(U_n')$ be an eigenvalue of U_n', and let $g \in C(\mathcal{X})$ denote a corresponding eigenfunction. Then, λ is also an eigenvalue of L_n', and the vector $v := (v_1, ..., v_n)^T := \rho_n g$ is an eigenvector to λ. Additionally, if $\lambda \neq 1$, g has the form*

$$g(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{k(x, X_j) v_j}{1 - \lambda} \tag{4.8}$$

Conversely, if $\lambda \neq 1$ is an eigenvalue of L_n', and v an eigenvector to λ, then λ is an eigenvalue of U_n', and the function of the form (4.8) is an eigenfunction of U_n' to the eigenvalue λ.

Proof. Let λ be an eigenvalue of U_n' such that $\lambda \neq 1$ holds, i.e.

$$U_n' g = \lambda g$$

which is equivalent to

$$U_n' g(x) = \lambda g(x)$$

for all $x \in \mathcal{X}$, which yields

$$\rho_n U_n' g = \lambda \rho_n g$$

Because of (4.7), we obtain

$$L_n' \rho_n g = \lambda \rho_n g$$

which proves the first statement. We can prove that g is of the form (4.8) by solving the eigenproblem

$$\lambda g = U_n' g$$

for g (see von Luxburg, Belkin, and Bousquet (2008, p.568-569)). Lastly, let $\lambda \neq 1$ be an eigenvalue of L_n', and let $v \in \mathbb{R}^n$ be an eigenvector to λ. Let $g \in C(\mathcal{X})$ be a function of the form (4.8). By definition, we have

$$L_n' v = \lambda v$$

note that g is well defined because v is assumed to be an eigenvector. We can then show that $U_n' g(x) = \lambda g(x)$ holds (see von Luxburg, Belkin, and Bousquet (2008, p. 568–569).) □

This relationship between $(U_n')_{n \in \mathbb{N}}$ and $(L_n')_{n \in \mathbb{N}}$ is essential for proving the following theorem: a result that states that the spectral properties of $(L_n')_{n \in \mathbb{N}}$ converge under certain conditions. Recall that we operate under the assumptions established in Remark 4.1.

THEOREM 4.9 (Convergence of Normalized Laplacian Eigenvectors, Theorem 15 in von Luxburg, Belkin, and Bousquet 2008)**.** *Let $\lambda \neq 1$ be an eigenvalue of U'. Let $M \subset \mathbb{C}$ be a*

neighborhood of λ such that

$$\sigma(U') \cap M = \{\lambda\} \tag{4.9}$$

Let $(L'_n)_{n\in\mathbb{N}}$ be the sequence of normalized Laplacian matrices of the graph sequence $(G_n)_{n\in\mathbb{N}}$. Then, the following statemets are true:

(i) If $(\lambda_n)_{n\in\mathbb{N}}$ is a sequence of eigenvalues such that $\lambda_n \in \sigma(L'_n) \cap M$ for $n \in \mathbb{N}$, then

$$\lambda_n \longrightarrow \lambda \tag{4.10}$$

almost surely.

(ii) If λ is a simple eigenvalue, and $(\lambda_n)_{n\in\mathbb{N}}$ satisfies (4.10), then any sequence of eigenvectors $(v_n)_{n\in\mathbb{N}}$ of $(L'_n)_{n\in\mathbb{N}}$ that corresponds to $(\lambda_n)_{n\in\mathbb{N}}$ converges to an eigenfunction f of U' with eigenvalue λ under $(\rho_n)_{n\in\mathbb{N}}$ up to a change of sign almost surely (see Definition 2.30).

Note that the term "almost surely" refers to the selection of the random sequence $(X_n)_{n\in\mathbb{N}}$ (see Remark 4.1). Some of the arguments that we employ below require us to fix a realization of $(X_n)_{n\in\mathbb{N}}$, in which case we omit the term "almost surely".

4.2 Practical Implications of Theorem 4.9

Recall from Section 3.3 that the implementation of spectral clustering algorithms for a given $k \leq n$ relies on obtaining the eigenvectors to the k smallest eigenvalues of the matrix L'. For this reason, our objective is to retrieve a convergence result for the eigenvector sequence $(v_n)_{n\in\mathbb{N}}$. In light of this, we discuss the practical meaning of Theorem 4.9: we begin with (i), which states that any eigenvalue sequence $(\lambda_n)_{n\in\mathbb{N}}$ of $(L'_n)_{n\in\mathbb{N}}$ converges if it is contained in a sufficiently small interval. This is mostly of theoretical importance as it ensures that the limit function f in (ii) is well-defined. Note that from the perspective of practical implementation, the convergence of eigenvalues is less relevant because the stability of the output eigenvectors depends on the size of the k-th eigengap (see Section 3.4).

The result in (ii) is the actual consistency result – it states that $(v_n)_{n\in\mathbb{N}}$ converges *a.s.* to f under ρ_n up to a change of sign (see Definition 2.30). The result provides insights into the viability of spectral clustering algorithms such as Algorithm 3.31. To see this, suppose that the data points $x_1, ..., x_n$ are sampled from a k-modal distribution $\mathbb{P}$ on $(\mathcal{X}, B(\mathcal{X}))$. This suggests we may apply Algorithm 3.31 or a similar program to identify k distinct clusters if we choose the similarity kernel carefully. According to (ii), the sequences $(v_{1,n})_{n\in\mathbb{N}}, ..., (v_{k,n})_{n\in\mathbb{N}}$ of the first k eigenvectors – the eigenvectors to the k smallest positive eigenvalues – converge to $f_1, ..., f_k \in C(\mathcal{X})$ under $(\rho_n)_{n\in\mathbb{N}}$ *a.s.* up to a change of sign, respectively. In this case, $f_1, ..., f_k$ indicate the inherent partition of the data. We can illustrate this more easily for the case $k = 2$ (see our discussion in Section 3.3 for reference): we assign the i-th vertex to a cluster based on whether the condition $v_n^i \geq 0$ holds for the i-th component of v_n. As we see below, the limit function f indicates cluster membership for $x \in \mathcal{X}$ based on the condition $f(x) \geq 0$. To visualize the convergence of the eigenvector sequence, we present a simulation of the following example:

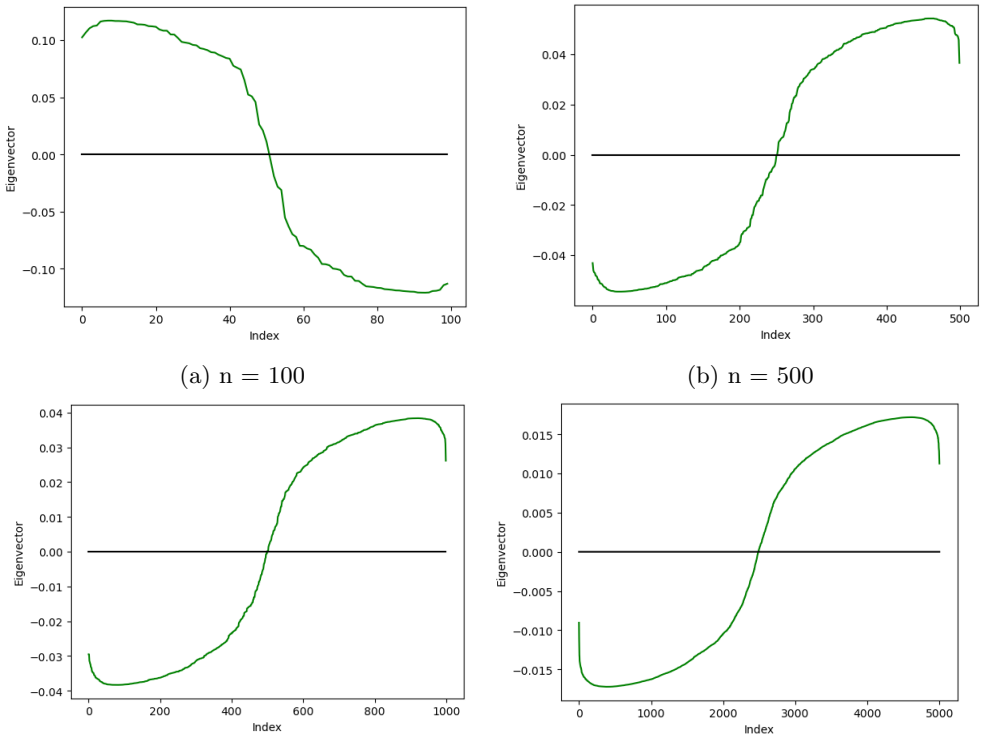
EXAMPLE 4.10. *Let $x_1, ..., x_n$ denote the first n elements of a sequence of real-valued random variables. Suppose, further, that every x_i is drawn independently from the same distribution P, which is given by*

$$P = \frac{1}{2} \cdot \mathcal{N}(0.3, 0.1) + \frac{1}{2} \cdot \mathcal{N}(0.7, 0.1)$$

The similarity function is the Gaussian kernel

$$k : \begin{cases} \mathbb{R} \times \mathbb{R} \longrightarrow (0, \infty) \\ (x, y) \longmapsto \exp\left(-\frac{\|x-y\|^2}{h^2}\right) \end{cases}$$

$h > 0$ denotes the width of the kernel function, which we obtain from applying Scott's KDE rule for $n = 500$ (see (3.32)). The plots in Figure 4.1 display the second eigenvectors of L'_n for $n = 100, 500, 1000, 5000$. Note that for visualization purposes, we sorted the samples x_i and their respective indices in non-decreasing order (we can do this because the convergence in Theorem 4.9 is with respect to the maximum norm, which does not depend on the sample order). We can observe that the curves of the ordered components of the eigenvectors appear



(a) n = 100 (b) n = 500

(c) n = 1000 (d) n = 5000

Figure 4.1: Second Eigenvectors of the normalized Laplacian L'_n for different sample sizes n

smoother as n increases. Additionally, the sign of the eigenvectors changes from $n = 100$ to $n = 500$. This illustrates the convergence to a smooth function up to a change in sign, and aligns with the theory discussed previously: Theorem 4.9 guarantees that under reasonable assumptions about the similarity function and the number of clusters, normalized spectral

lustering provides stable results that align with the inherent structure of the data.

4.3 Proof of Theorem 4.9

We now prove our main result, building on the proof presented by von Luxburg, Belkin, and Bousquet (2008). Recall that $(\lambda_n)_{n\in\mathbb{N}}$ is a sequence of eigenvalues of $(L'_n)_{n\in\mathbb{N}}$ for which the conditions of Theorem 4.9 hold. Because of Proposition 4.8, $(\lambda_n)_{n\in\mathbb{N}}$ is also a sequence of eigenvalues of $(U'_n)_{n\in\mathbb{N}}$. This means that, as a result of Theorem 2.29, (i) is a direct implication of the following Lemma, which we prove further below.

LEMMA 4.11. *The operator sequence $(U'_n)_{n\in\mathbb{N}}$ converges compactly to the operator U'.*

To prove (ii), let $v_{n,i}$ denote the i-th component of the eigenvector v_n to λ_n for $n \in \mathbb{N}$. Suppose that there exists a sequence $(f_n)_{n\in\mathbb{N}}$ in $C(\mathcal{X})$ such that

$$\rho_n f_n = v_n \tag{4.11}$$

for all $n \in \mathbb{N}$. Then, we obtain

$$\|a_n \rho_n f_n - \rho_n f\|_{\max,n} \leq \|a_n f_n - f\|_\infty \tag{4.12}$$

which means it suffices to prove that the right-hand side approaches 0 for $n \to \infty$. Note that, because of (i), $\lambda_n \neq 1$ holds for large enough n. Because of this, we obtain from Proposition 4.8 that there indeed exists a function sequence that satisfies (4.11): the eigenfunction sequence (f_n) given by

$$f_n(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{k(x, X_j) v_{n,j}}{1 - \lambda} \tag{4.13}$$

for all $n \in \mathbb{N}$ that are large enough to satisfy $\lambda_n \neq 1$. As a consequence of (i) and Theorem 2.29, the right-hand side of (4.12) converges to 0 because of Lemma 4.11.

4.4 Proof of Lemma 4.11

We prove a stronger version of Lemma 4.11, which is that $(U'_n)_{n\in\mathbb{N}}$ converges collectively compactly to U'. Recall Definition 2.26: to prove collectively compact convergence, we need to show two properties. First, we need to show that $(\hat{T}_n)_{n\in\mathbb{N}}$ converges to T pointwise, i.e. that for all $f \in C(\mathcal{X})$

$$\|\hat{T}_n f - T f\|_\infty \overset{n\to\infty}{\longrightarrow} 0 \tag{4.14}$$

almost surely. Secondly, we need to prove that the operator sequence $(\hat{T}_n - T)_{n\in\mathbb{N}}$ is collectively compact. We prove (4.14) first. To do this, note that $(\hat{T}_n)_{n\in\mathbb{N}}$ defines a sequence of empirical integrals of a function of the form $\frac{k(x,\cdot)}{\sqrt{d_n(x) d_n(\cdot)}} f(\cdot)$. This suggests we can use Theorem 2.37 to prove (4.14) To prepare this step, we introduce the following function classes:

$$\mathcal{K} := \{k(x, \cdot) | x \in \mathcal{X}\} \tag{4.15}$$

$$\mathcal{H} \coloneqq \left\{ \frac{k(x,\cdot)}{\sqrt{d(x)d(\cdot)}} \middle| x \in \mathcal{X} \right\} \tag{4.16}$$

$$\mathcal{H} \cdot \mathcal{H} \coloneqq \left\{ \frac{k(x,\cdot)}{\sqrt{d(x)d(\cdot)}} \frac{k(\cdot,y)}{\sqrt{d(\cdot)d(y)}} \middle| x, y \in \mathcal{X} \right\} \tag{4.17}$$

$$g \cdot \mathcal{H} \coloneqq \left\{ g(\cdot) \cdot \frac{k(x,\cdot)}{\sqrt{d(x)d(\cdot)}} \middle| x \in \mathcal{X} \right\} \tag{4.18}$$

for any function $g \in C(\mathcal{X})$. Next, we prove that the function classes we introduced are Glivenko-Cantelli classes:

LEMMA 4.12. *The function classes $\mathcal{K}$, $\mathcal{H}$, $g \cdot \mathcal{H}$, $\mathcal{H} \cdot \mathcal{H}$ are Glivenko-Cantelli classes for every $g \in C(\mathcal{X})$.*

Proof. According to Theorem 2.37, a class $\mathcal{G}$ of measurable functions is Glivenko-Cantelli, it suffices to prove that its L^1-bracketing number $N_{[]}(\epsilon, \mathcal{G}, L^1(P))$ is finite for every $\epsilon > 0$ with respect to a probability measure P. Subsequently, we show this for the function class $\mathcal{K}$: $\mathcal{K}$ is defined as

$$\mathcal{K} \coloneqq \{k(x,\cdot) | x \in \mathcal{X}\}$$

Recall that k is a continuous function on the compact domain $\mathcal{X} \times \mathcal{X}$. Hence, k is uniformly continuous. Because $\mathcal{X}$ is compact, we can find a finite δ-cover of $\mathcal{X}$. Let y be in $\mathcal{X}$. Then, the uniform continuity of k implies that, for every $x \in \mathcal{X}$, and for every $\epsilon > 0$

$$x' \in B_\delta(x) \Longrightarrow |k(x,y) - k(x',y)| < \epsilon$$

Because this does not depend on the choice of y, we obtain that

$$x' \in B_\delta(x) \Longrightarrow \sup_{y \in \mathcal{X}} |k(x,y) - k(x',y)| < \epsilon \tag{4.19}$$

Since the function class $\mathcal{K}$ is indexed by $x \in \mathcal{X}$, and (4.19) holds for any $x, x' \in \mathcal{X}$, we can construct a finite ϵ-cover of $\mathcal{X}$ with respect to $\|\cdot\|_\infty$. Therefore, the fact that $\mathcal{K}$ has finite covering numbers with respect to $\|\cdot\|_\infty$, $\mathcal{K}$ has finite ϵ-bracketing numbers with respect to $\|\cdot\|_\infty$. This is because the uniform norm satisfies

$$N(\epsilon, \mathcal{K}, \|\cdot\|_\infty) = N_{[]}(2\epsilon, \mathcal{K}, \|\cdot\|_\infty)$$

for all $\epsilon > 0$ (see (2.13)). Since $\mathbb{P}$ is a probability measure on $\mathcal{X}$, we know, for any measurable function $f : \mathcal{X}, \to \mathbb{R}$, that

$$\|f\|_{L^1} < \|f\|_\infty$$

which implies that

$$N_{[]}(\epsilon, \mathcal{G}, \|\cdot\|_{L^1}) \leq N_{[]}(\epsilon, \mathcal{G}, \|\cdot\|_\infty)$$

for all $\epsilon > 0$. This proves that $\mathcal{K}$ has finite L^1-bracketing number with respect to $\mathbb{P}$. Hence, it follows from Theorem 2.37 that $\mathcal{K}$ is a Glivenko-Cantelli class. The proofs for the other

function classes work analogously because the functions in $\mathcal{H}$ are continuous due to the fact d is continuous (see Section 4.1 and also von Luxburg, Belkin, and Bousquet (2008, Section 5)). □

Suppose that $f \in C(\mathcal{X})$. Recall that we need to prove that the uniform distance $\|\hat{T}_n f - Tf\|_\infty$ converges to 0. We introduce the bounded linear operator

$$T_n : \begin{cases} C(\mathcal{X}) \longrightarrow C(\mathcal{X}) \\ f(x) \longmapsto \int_{\mathcal{X}} \frac{k(x,y)}{\sqrt{d(x)d(y)}} f(y) d\mathbb{P}_n(y) \end{cases} \tag{4.20}$$

Following a similar argument as applied in the proof of Proposition 4.5, it can be proven that T_n is bounded for all $n \in \mathbb{N}$. T_n is also a compact operator because of Proposition 2.15. We use $(T_n)_{n\in\mathbb{N}}$ to bound the sequence in (4.14) as follows:

$$\|\widehat{T}_n f - Tf\|_\infty \leq \underbrace{\|\widehat{T}_n f - T_n f\|_\infty}_{:=\mathrm{I}} + \underbrace{\|T_n f - Tf\|_\infty}_{:=\mathrm{II}} \tag{4.21}$$

Next, we bound I and II from above by $\sup_{g\in\mathcal{K}} |\mathbb{P}_n g - \mathbb{P} g|$ and $\sup_{g\in f\cdot\mathcal{H}} |\mathbb{P}_n g - \mathbb{P} g|$, respectively. We do this in order to utilize the Glivenko-Cantelli property of these function classes. We begin with II: recall that the functions $h \in \mathcal{H}$ have the form

$$h(x,\cdot) = \frac{k(x,\cdot)}{\sqrt{d(x)d(\cdot)}}$$

for a given $x \in \mathcal{X}$. Using the definitions of T and T_n, we derive

$$\|T_n f - Tf\|_\infty = \sup_{x\in\mathcal{X}} |\mathbb{P}_n \left(h(x,\cdot) f(\cdot)\right) - \mathbb{P}\left(h(x,\cdot) f(\cdot)\right)| \tag{4.22}$$

$$= \sup_{g\in f\cdot\mathcal{H}} |\mathbb{P}_n g - \mathbb{P} g| \tag{4.23}$$

which converges to 0 according to Lemma 4.12. Next, we bound I: the definitions of $\widehat{T}_n$ and T_n and basic calculations yield

$$\|\widehat{T}_n f - T_n f\|_\infty \leq \|f\|_\infty \|k\|_\infty \sup_{x,y\in\mathcal{X}} \left| \frac{1}{\sqrt{d_n(x)d_n(y)}} - \frac{1}{\sqrt{d(x)d(y)}} \right| \tag{4.24}$$

We bound the right-hand side of this using the fact that d_n and d are bounded from below by c

$$\|f\|_\infty \|k\|_\infty \sup_{x,y\in\mathcal{X}} \left| \frac{1}{\sqrt{d_n(x)d_n(y)}} - \frac{1}{\sqrt{d(x)d(y)}} \right|$$

$$= \|f\|_\infty \|k\|_\infty \sup_{x,y\in\mathcal{X}} \left| \frac{\sqrt{d_n(x)d_n(y)} + \sqrt{d(x)d(y)}}{\sqrt{d_n(x)d_n(y)} + \sqrt{d(x)d(y)}} \cdot \frac{\sqrt{d(x)d(y)} - \sqrt{d_n(x)d_n(y)}}{\sqrt{d_n(x)d_n(y)}\sqrt{d(x)d(y)}} \right|$$

$$= \|f\|_\infty \|k\|_\infty \sup_{x,y\in\mathcal{X}} \left| \frac{1}{\sqrt{d_n(x)d_n(y)}\sqrt{d(x)d(y)}} \cdot \frac{d_n(x)d_n(y) - d(x)d(y)}{\sqrt{d_n(x)d_n(y)} + \sqrt{d(x)d(y)}} \right|$$

$$\leq \|f\|_\infty \frac{\|k\|_\infty}{c^2} \sup_{x,y\in\mathcal{X}} \left| \frac{d_n(x)d_n(y) - d(x)d(y)}{\sqrt{d_n(x)d_n(y)} + \sqrt{d(x)d(y)}} \right|$$

and thus we obtain

$$\|\widehat{T}_n f - T_n f\|_\infty \leq \|f\|_\infty \frac{\|k\|_\infty}{c^2} \sup_{x,y\in\mathcal{X}} \frac{|d_n(x)d_n(y) - d(x)d(y)|}{\sqrt{d_n(x)d_n(y)} + \sqrt{d(x)d(y)}}$$

We bound the right-hand side using the results $d \geq c$ and $d_n \geq c$ from Proposition 4.3; the subsequent inequality

$$\sqrt{d_n(x)d_n(y)} + \sqrt{d(x)d(y)} \geq 2c$$

yields

$$\|f\|_\infty \frac{\|k\|_\infty}{c^2} \sup_{x,y\in\mathcal{X}} \frac{|d_n(x)d_n(y) - d(x)d(y)|}{\sqrt{d_n(x)d_n(y)} + \sqrt{d(x)d(y)}} \leq \|f\|_\infty \frac{\|k\|_\infty}{2c^3} \sup_{x,y\in\mathcal{X}} |d_n(x)d_n(y) - d(x)d(y)|$$

and thus

$$\|\hat{T}_n f - T_n f\|_\infty \leq \|f\|_\infty \frac{\|k\|_\infty}{2c^3} \sup_{x,y\in\mathcal{X}} |d_n(x)d_n(y) - d(x)d(y)|$$

A short calculation shows that the inequality

$$\sup_{x,y\in\mathcal{X}} |d_n(x)d_n(y) - d(x)d(y)| \leq 2\|k\|_\infty \sup_{x\in\mathcal{X}} |d_n(x) - d(x)|$$

holds. Applying the definitions of the function class $\mathcal{K}$ and the functions d_n and d, we obtain

$$\sup_{x\in\mathcal{X}} |d_n(x) - d(x)| = \sup_{g\in\mathcal{K}} |\mathbb{P}_n g - \mathbb{P} g|$$

which implies

$$\|\hat{T}_n f - T_n f\|_\infty \leq \|f\|_\infty \frac{\|k\|_\infty}{c^3} \sup_{g\in\mathcal{K}} |\mathbb{P}_n g - \mathbb{P} g| \tag{4.25}$$

which shows that I converges to 0 almost surely. Hence, the right-hand side of (4.21) tends to 0 almost surely. This proves the almost sure, pointwise convergence of $\widehat{T}_n$ to the operator T. Besides pointwise convergence, we also need to prove that the sequence $\left(\hat{T}_n - T\right)_{n\in\mathbb{N}}$ is collectively compact, i.e. that $\bigcup_{n\in\mathbb{N}} \left(\hat{T}_n - T\right) B_1$ has compact closure. Because T is a compact operator, it suffices to show that $(\hat{T}_n)_{n\in\mathbb{N}}$ is collectively compact. We prove this in two steps using Theorem 2.10. For this purpose, we fix a realization of our sample sequence $(X_n)_{n\in\mathbb{N}}$. We show first that $\bigcup_{n\in\mathbb{N}} \hat{T}_n B_1$ is bounded. This holds because

$$\sup_{n\in\mathbb{N}, f\in B_1} \|\widehat{T}_n f\|_\infty \leq \frac{\|k\|_\infty}{c}$$

which follows from Proposition 4.3 and the definition of B_1. Next, we need to prove that $\bigcup_{n\in\mathbb{N}} \widehat{T}_n B_1$ is equicontinuous. This means we have to demonstrate that for all $x \in \mathcal{X}$, and

for all $\epsilon > 0$, there exists a neighborhood U_x of x such that

$$|g(x) - g(x')| < \epsilon \tag{4.26}$$

for all $g \in \bigcup_{n\in\mathbb{N}} \widehat{T}_n B_1$, and all $x' \in U_x$. We thus prove that

$$\sup_{g\in\bigcup_{n\in\mathbb{N}} \widehat{T}_n B_1} |g(x) - g(x')| < \epsilon \tag{4.27}$$

which is equivalent to

$$\sup_{n\in\mathbb{N}, f\in B_1} |\widehat{T}_n f(x) - \widehat{T}_n f(x')| < \epsilon \tag{4.28}$$

for all $x' \in U_x$. We bound the left-hand side of (4.28):

$$\begin{aligned}
&\sup_{n\in\mathbb{N}, f\in B_1} |\widehat{T}_n f(x) - \widehat{T}_n f(x')| \\
&= \sup_{n\in\mathbb{N}, f\in B_1} \left| \int_{\mathcal{X}} \frac{k(x,y)}{\sqrt{d_n(x)d_n(y)}} f(y) d\mathbb{P}_n - \int_{\mathcal{X}} \frac{k(x',y)}{\sqrt{d_n(x')d_n(y)}} f(y) d\mathbb{P}_n(y) \right| \\
&= \sup_{n\in\mathbb{N}, f\in B_1} \left| \int_{\mathcal{X}} f(y) \left\{ \frac{k(x,y)}{\sqrt{d_n(x)d_n(y)}} - \frac{k(x',y)}{\sqrt{d_n(x')d_n(y)}} \right\} d\mathbb{P}_n(y) \right| \\
&\leq \sup_{n\in\mathbb{N}, f\in B_1} \int_{\mathcal{X}} |f(y)| \left| \left\{ \frac{k(x,y)}{\sqrt{d_n(x)d_n(y)}} - \frac{k(x',y)}{\sqrt{d_n(x')d_n(y)}} \right\} \right| d\mathbb{P}_n(y) \\
&\leq \|f\|_\infty \sup_{n\in\mathbb{N}, f\in B_1} \int_{\mathcal{X}} \left| \frac{k(x,y)}{\sqrt{d_n(x)d_n(y)}} - \frac{k(x',y)}{\sqrt{d_n(x')d_n(y)}} \right| d\mathbb{P}_n(y) \\
&\leq \sup_{n\in\mathbb{N}, f\in B_1} \int_{\mathcal{X}} \left| \frac{k(x,y)}{\sqrt{d_n(x)d_n(y)}} - \frac{k(x',y)}{\sqrt{d_n(x')d_n(y)}} \right| d\mathbb{P}_n(y)
\end{aligned}$$

Note that the last inequality holds because $\|f\|_\infty = 1$ for $f \in B_1$ by definition. Because $\mathbb{P}_n$ is a probability measure, we obtain

$$\sup_{g\in\bigcup_{n\in\mathbb{N}} \hat{T}_n B_1} |g(x) - g(x')| \leq \left\| \frac{k(x,\cdot)}{\sqrt{d_n(x)d_n(\cdot)}} - \frac{k(x',\cdot)}{\sqrt{d_n(x')d_n(\cdot)}} \right\|_\infty \tag{4.29}$$

A short calculation shows that there exist positive constants C_1, C_2, C_3, independent of x, x', and n, such that the right-hand side of (4.29) is bounded from above in the following way (see von Luxburg, Belkin, and Bousquet (2008, p. 571)):

$$\left\| \frac{k(x,\cdot)}{\sqrt{d_n(x)d_n(\cdot)}} - \frac{k(x',\cdot)}{\sqrt{d_n(x')d_n(\cdot)}} \right\|_\infty \leq C_1 \|k(x,\cdot) - k(x',\cdot)\|_\infty + C_2 |d(x) - d(x')| + C_3 \|d_n - d\|_\infty$$

We bound the summands individually: $\|k(x,\cdot) - k(x',\cdot)\|_\infty$ is infinitesimally small if x and

x' are close, because k is uniformly continuous. The same holds for $|d(x) - d(x')|$ because

$$\begin{aligned} |d(x) - d(x')| &= \left| \int_{\mathcal{X}} k(x,\cdot) d\mathbb{P} - \int_{\mathcal{X}} k(x',\cdot) d\mathbb{P} \right| \\ &= \left| \int_{\mathcal{X}} (k(x,\cdot) - k(x',\cdot)) d\mathbb{P} \right| \\ &\leq \int_{\mathcal{X}} \|k(x,\cdot) - k(x',\cdot)\|_\infty d\mathbb{P} \end{aligned}$$

and because k is uniformly continuous. Recall that d and d_n are defined as

$$d(x) = \int_{\mathcal{X}} k(x,y) d\mathbb{P}(y)$$

and

$$d_n(x) = \int_{\mathcal{X}} k(x,y) d\mathbb{P}_n(y)$$

respectively. Thus, the term

$$\|d_n - d\|_\infty = \|\mathbb{P}_n k(x,\cdot) - \mathbb{P} k(x,\cdot)\|_\infty$$

converges to 0 because $\mathcal{K}$ is a Glivenko-Cantelli class according to Lemma 4.12. Therefore, $\bigcup_{n\in\mathbb{N}} \widehat{T}_n B_1$ is equicontinuous. It follows from Theorem 2.10 that $\bigcup_{n\in\mathbb{N}} \hat{T}_n B_1$ is relatively compact. By definition, this means that the sequence $(\widehat{T}_n)_{n\in\mathbb{N}}$ is collectively compact. Because T is a compact operator, it follows that $(\widehat{T}_n - T)_{n\in\mathbb{N}}$ also is collectively compact. This, in combination with (4.14), proves that $(\widehat{T}_n)_{n\in\mathbb{N}}$ converges collectively compactly to T, which proves Lemma 4.11. We conclude the section with the following statement:

COROLLARY 4.13. *If the assumptions of Theorem 4.9 hold, then, for all $b \in \mathbb{R}$, the set $\{a_n f_n > b\}$ converges to the set $\{f > b\}$, i.e. their symmetric difference satisfies*

$$\mathbb{P}(\{a_n f_n > b\} \triangle \{f > b\}) \longrightarrow 0 \tag{4.30}$$

for $n \to \infty$.

This result is a consequence of the eigenvector convergence in Theorem 4.9.

4.5 Convergence Rates for Normalized Spectral Clustering

According to Theorem 4.9, a sequence of eigenvectors $(v_n)_{n\in\mathbb{N}}$ of $(L'_n)_{n\in\mathbb{N}}$ converges to an eigenfunction f of U' almost surely up to a change of sign under certain assumptions. In particular,

$$\sup_{i=1,\ldots,n} |a_n v_{n,i} - f(X_i)| \longrightarrow 0 \quad \text{a.s.} \tag{4.31}$$

for $n \to \infty$, where $(a_n)_{n\in\mathbb{N}} \in \{-1,1\}^{\mathbb{N}}$ and $v_{n,i}$ denotes the i-th component of v_n. In the following, we discuss the rate of that convergence. Our subsequent discussions, including results we present, follow the approach of von Luxburg, Belkin, and Bousquet (2008). In

(4.12), we showed that for $n \in \mathbb{N}$, the sequence in (4.31) is bounded from above such that

$$\sup_{i=1,\dots,n} |a_n v_{n,i} - f(X_i)| \leq \|a_n f_n - f\|_\infty$$

for eigenfunction sequences $(f_n)_{n\in\mathbb{N}} \subset C(\mathcal{X})$ of $(U_n')_{n\in\mathbb{N}}$. Through the following result, we obtain an upper bound for $\|a_n f_n - f\|_\infty$.

THEOREM 4.14. *Let $\lambda \neq 0$ be a simple eigenvalue of T with eigenfunction $u \in C(\mathcal{X})$. Let $\mathcal{F}$ denote the function class*

$$\mathcal{F} := \mathcal{K} \cup (u \cdot \mathcal{H}) \cup (\mathcal{H} \cdot \mathcal{H})$$

where the function classes are those defined in (4.15)-(4.18). Let $(\lambda_n)_{n\in\mathbb{N}}$ denote sequence of eigenvalues of $(\widehat{T}_n)_{n\in\mathbb{N}}$ and $(u_n)_{n\in\mathbb{N}}$ an associated sequence of eigenfunctions. Then, the result

$$\|a_n u_n - u\|_\infty \leq C' \sup_{g\in\mathcal{F}} |\mathbb{P}_n g - \mathbb{P} g| \tag{4.32}$$

holds almost surely for a constant C' that only depends on k, $\sigma(U')$, and λ.

We introduce the two following lemmas for the proof of 4.14:

LEMMA 4.15 (von Luxburg, Belkin, and Bousquet (2008), Proposition 17)**.** *Suppose that $g \in C(\mathcal{X})$. Then, the following inequalities hold:*

$$\|\hat{T}_n - T_n\|_{\text{op}} \leq \frac{\|k\|_\infty}{l^3} \sup_{f\in\mathcal{K}} |\mathbb{P}_n f - \mathbb{P} f| \tag{4.33}$$

$$\|(T_n - T)g\|_\infty \leq \sup_{f\in g\cdot\mathcal{H}} |\mathbb{P}_n f - \mathbb{P} f| \tag{4.34}$$

$$\|(T_n - T)T_n\|_{\text{op}} \leq \sup_{f\in\mathcal{H}\cdot\mathcal{H}} |\mathbb{P}_n f - \mathbb{P} f| \tag{4.35}$$

Proof. Note that (4.33) is a consequence of (4.25), which we have proven earlier. (4.34) follows from (4.22). To prove (4.35), we write

$$(T_n - T)T_n g(x) = \int_{\mathcal{X}} \frac{k(x,y)}{\sqrt{d(x)d(y)}} T_n g(y) d(\mathbb{P}_n - \mathbb{P})(y) \tag{4.36}$$

$$= \int_{\mathcal{X}} \frac{k(x,y)}{\sqrt{d(x)d(y)}} \int_{\mathcal{X}} \frac{k(y,z)}{\sqrt{d(y)d(z)}} g(z) d\mathbb{P}_n(z) d(\mathbb{P}_n - \mathbb{P})(y) \tag{4.37}$$

$$= \int_{\mathcal{X}} \frac{k(x,y)}{\sqrt{d(x)d(y)}} \int_{\mathcal{X}} \frac{k(y,z)}{\sqrt{d(y)d(z)}} g(z) d\mathbb{P}_n(z) d\mathbb{P}_n(y)$$

$$- \int_{\mathcal{X}} \frac{k(x,y)}{\sqrt{d(x)d(y)}} \int_{\mathcal{X}} \frac{k(y,z)}{\sqrt{d(y)d(z)}} g(z) d\mathbb{P}_n(z) d\mathbb{P}(y) \tag{4.38}$$

for any $g \in C(\mathcal{X})$. Note that $\mathbb{P}$ and $\mathbb{P}_n$ are probability measures and therefore σ-finite. It is obvious that the function

$$F(y,z) = \frac{k(x,y)}{\sqrt{d(x)d(y)}} \cdot \frac{k(y,z)}{\sqrt{d(y)d(z)}} \cdot g(z)$$

is measurable with respect to the Borel product σ-algebra on $\mathcal{X}^2$. We can thus swap the integrals in both terms in (4.38) by applying Fubini's theorem (see Klenke (2006, p. 12, 285)). This yields

$$(T_n - T)T_n g(x) = \int_{\mathcal{X}} g(z) \int_{\mathcal{X}} \frac{k(x,y)}{\sqrt{d(x)d(y)}} \cdot \frac{k(y,z)}{\sqrt{d(y)d(z)}} \cdot d(\mathbb{P}_n - \mathbb{P})(y) d\mathbb{P}_n(z)$$

and therefore

$$\|(T_n - T)T_n\|_{\mathrm{op}} = \sup_{\|g\|_\infty = 1} \left\| \int_{\mathcal{X}} g(z) \int_{\mathcal{X}} \frac{k(\cdot,y)}{\sqrt{d(\cdot)d(y)}} \cdot \frac{k(y,z)}{\sqrt{d(y)d(z)}} \cdot d(\mathbb{P}_n - \mathbb{P})(y) d\mathbb{P}_n(z) \right\|_\infty$$

$$\leq \sup_{f \in \mathcal{H} \cdot \mathcal{H}} |\mathbb{P}_n f - \mathbb{P} f|$$

following the definition of $\mathcal{H} \cdot \mathcal{H}$ (see (4.17)). □

LEMMA 4.16. *Let $(e_n)_{n \in \mathbb{N}}$ be a sequence of vectors in $(E, \|\cdot\|_E)$ such that $\|e_n\|_E = 1$ for all $n \in \mathbb{N}$, and define $e \in E$ such that $\|e\|_E = 1$. For $n \in \mathbb{N}$, let P_n be the projection on the one-dimensional subspace spanned by e_n. Then there exists a sequence of signs $(a_n)_{n \in \mathbb{N}}$ such that*

$$\|a_n e_n - e\| \leq 2\|e - P_n e\| \tag{4.39}$$

Proof. P_n is a projection on the one-dimensional subspace spanned by e_n. Thus, $P_n e$ is of the form $P_n e = c_n e_n$ for a number $c_n \in \mathbb{R}$. Let a_n denote the sign of c_n. Then, it is straightforward to verify that

$$\begin{aligned} |a_n - c_n| &= |1 - |c_n|| \\ &= |\|e\|_E - |c_n| \|e_n\|_E| \\ &\leq \|e - c_n e_n\|_E = \|e - P_n e\|_E \end{aligned}$$

holds. Using the triangular inequality, we argue that

$$\begin{aligned} \|a_n e_n - e\|_E &\leq \|a_n e_n - c_n e_n\|_E + \|c_n e_n - e\|_E \\ &\leq 2\|e - P_n e\|_E \end{aligned}$$

which proves the statement. □

The following result, which is a direct consequence of Theorem 2.29, helps us prove Theorem 4.14:

COROLLARY 4.17. *Assume the assumptions of Theorem 4.9 hold, and $(\lambda_n)_{n \in \mathbb{N}}$ converges to λ almost surely. For $n \in \mathbb{N}$, let P_{λ_n} denote the spectral projection onto λ_n. Then, the sequence $(P_{\lambda_n})_{n \in \mathbb{N}}$ converges pointwise to the spectral projection P_λ onto the eigenvalue $\lambda \in \sigma(T)$ almost surely.*

Proof of Theorem 4.14 Consider a fixed realization of our data sequence $(X_n)_{n\in\mathbb{N}}$. Without loss of generality, assume that $\|u\|_\infty = 1$ and $\|u_n\|_\infty = 1$ for $n \in \mathbb{N}$. We need to show that $\|a_n u_n - u\|_\infty$ is bounded from above by the term $C' \sup_{g\in\mathcal{F}} |\mathbb{P}g - \mathbb{P}_n g|$. Because the spectral projections of $(\widehat{T}_n)_{n\in\mathbb{N}}$ converge, and since λ is a simple eigenvalue, it follows that λ_n is a simple eigenvalue of T for sufficiently large n. Hence, it follows that for large enough n, the spectral projection P_{λ_n} maps to a one-dimensional subspace of $C(\mathcal{X})$. By Lemma 4.16, it follows that

$$\|a_n u_n - u\|_\infty \leq 2\|u - P_{\lambda_n} u\|_\infty \tag{4.40}$$

holds. We bound the right-hand side of (4.40) using Atkinson's Theorem (Theorem 2.31). Note that the conditions in Theorem 2.31 are satisfied: $(\widehat{T}_n)_{n\in\mathbb{N}}$ is a sequence of compact linear operators that converges to the compact operator T as we have shown in the sections 4.1 and 4.4. $\lambda \neq 0$ is satisfied as well. Since $u \in P_\lambda(C(\mathcal{X}))$ by definition, we obtain

$$2\|u - P_{\lambda_n} u\|_\infty \leq 2C(\underbrace{\|(\widehat{T}_n - T)u\|_\infty}_{:=\mathrm{I}} + \|u\|_\infty \underbrace{\|(T - \widehat{T}_n)\widehat{T}_n\|_{\mathrm{op}}}_{:=\mathrm{II}})$$

from Theorem 2.31. We bound I and II individually starting with I. The triangular inequality for $\|\cdot\|_\infty$ and the definition of the operator norm yield that

$$\|(\hat{T}_n - T)u\|_\infty \leq \|u\|_\infty \|\hat{T}_n - T_n\|_{\mathrm{op}} + \|(T_n - T)u\|_\infty \tag{4.41}$$

holds. For term II, a short calculation shows that

$$\begin{aligned}
&\|(T - \widehat{T}_n)\widehat{T}_n\|_{\mathrm{op}} \\
\leq\ & \|T\|_{\mathrm{op}} \|T_n - \widehat{T}_n\|_{\mathrm{op}} + \|(T - T_n)\,\widehat{T}_n\|_{\mathrm{op}} + \|T_n T_n - T_n \widehat{T}_n\|_{\mathrm{op}} + \|T_n \widehat{T}_n - \widehat{T}_n \widehat{T}_n\|_{\mathrm{op}} \\
\leq\ & \frac{3\|k\|_\infty}{c} \|T_n - \widehat{T}_n\|_{\mathrm{op}} + \|(T - T_n)T_n\|_{\mathrm{op}}
\end{aligned}$$

Combining the bounds we obtained for I and II with Lemma 4.15, we find that

$$\begin{aligned}
&\|a_n u_n - u\|_\infty \leq 2\|u - P_n u\|_\infty \\
\leq\ & 2C\left(\|u\|_\infty \|\hat{T}_n - T_n\|_{\mathrm{op}} + \|(T_n - T)u\|_\infty + \|u\|_\infty \left(\frac{3\|k\|_\infty}{c}\|T_n - \widehat{T}_n\|_{\mathrm{op}} + \|(T - T_n)T_n\|_{\mathrm{op}}\right)\right) \\
\leq\ & 2C\left(\left(3\frac{\|k\|_\infty}{c} + 1\right)\|T_n - \hat{T}_n\|_{\mathrm{op}} + \|(T_n - T)u\|_\infty + \|(T - T_n)T_n\|_{\mathrm{op}}\right) \\
\leq\ & C' \sup_{g\in\mathcal{F}} |\mathbb{P}g - \mathbb{P}_n g|
\end{aligned}$$

which follows from the definition of the operators and the fact that $\|u\|_\infty = 1$ and $\|u_n\|_\infty = 1$. □

Note that the inequality (4.32) in Theorem 4.14 states that the convergence of normalized Laplacian eigenvectors in the sense of Theorem 4.9 is at least as fast as the empirical process

indexed by $\mathcal{F}$. Recall the entropy bound in Proposition 2.38: it is useful to bound the right-hand side of the inequality (4.40). The following result helps us apply that proposition specifically to the function class $\mathcal{F} = \mathcal{K} \cup (f \cdot \mathcal{H}) \cup (\mathcal{H} \cdot \mathcal{H})$.

PROPOSITION 4.18 (von Luxburg, Belkin, and Bousquet (2008), Proposition 20)**.** *Define*

$$s = \frac{\|k\|_\infty + 2\sqrt{c\|k\|_\infty}}{2c^2}$$

and

$$q = \min\left\{1, s\|f\|_\infty, \frac{\|k\|_\infty s}{c}\right\}$$

for an arbitrary $f \in C(\mathcal{X})$. Then, the inequalities

$$N(\mathcal{H}, \epsilon, \|\cdot\|_\infty) \leq N(\mathcal{K}, s\epsilon, \|\cdot\|_\infty) \tag{4.42}$$

and

$$N(\mathcal{K} \cup (f \cdot \mathcal{H} \cup (\mathcal{H} \cdot \mathcal{H})), \epsilon, \|\cdot\|_\infty) \leq 3N(\mathcal{K}, q\epsilon, \|\cdot\|_\infty) \tag{4.43}$$

hold.

The proof is elementary and includes calculations similar to the ones used in previous sections (see von Luxburg, Belkin, and Bousquet (2008, p. 576)). The aforementioned results allow us to discuss a relevant example case:

EXAMPLE 4.19. *Suppose that our data space $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$, and that the similarity function k is given by the Gaussian kernel*

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{\sigma^2}\right)$$

Let $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^d$. Then, the rate at which the eigenfunctions in Theorem 4.14 converge is at least as fast as order $\mathcal{O}(n^{-1/2})$.

Proof. We need to prove that the right-hand side in Theorem 4.14 converges to 0 as fast as $n^{-1/2}$. Proposition 2.38 provides an upper bound for such expression. To apply this result, we need to demonstrate that the function class $\mathcal{F}$ satisfies $\|f\|_\infty \leq 1$. It is clear that $\mathcal{K}$ satisfies the condition because all functions in $\mathcal{K}$ are constructed from Gaussian kernel functions. To obtain an upper bound such as in Proposition 2.38 for the entire class $\mathcal{F}$, recall that $\mathcal{F}$ contains $\mathcal{H} \cdot \mathcal{H}$ and $g \cdot \mathcal{H}$ for an arbitrary $g \in C(\mathcal{X})$. However, because of Proposition 4.18, it suffices to evaluate the upper bound in 2.38 on the function class $\mathcal{K}$: up to a constant, we can bound covering numbers of both $\mathcal{H} \cdot \mathcal{H}$ and $g \cdot \mathcal{H}$ with covering numbers of $\mathcal{K}$. We use the following inequality, as detailed in von Luxburg (2007, Section 6.3), and originally due to Zhou (2002, Proposition 1): if $\epsilon < c_0$ for a positive constant c_0, then

$$\log N(\mathcal{K}, \epsilon, \|\cdot\|_\infty) \leq \tilde{C}\left(\log\frac{1}{\epsilon}\right)^2 \tag{4.44}$$

where $\widetilde{C}$ is a constant that depends on the width σ of the kernel k. Then, a short calculation yields that the convergence rate of $\sup_{g \in \mathcal{F}} |\mathbb{P}_n g - \mathbb{P} g|$ is of the order $\mathcal{O}(n^{-1/2})$. □

4.6 Consistency Results for Unnormalized Spectral Clustering

Recall that Theorem 4.9 guarantees convergence of eigenvector sequences of $(L'_n)_{n\in\mathbb{N}}$ under certain conditions. In the following, we discuss a similar result concerning unnormalized spectral clustering. For $n \in \mathbb{N}$, the unnormalized Laplacian matrix of G_n is given by

$$L_n = D_n - K_n$$

Recall that for normalized spectral clustering, we discussed a result concerning the convergence of eigenvalues and eigenvectors of $(L'_n)_{n\in\mathbb{N}}$. In order to control the size of the eigenvalues, we study the scaled sequence $(\frac{1}{n}L_n)_{n\in\mathbb{N}}$. Discussing the eigenvector convergence of $(\frac{1}{n}L_n)_{n\in\mathbb{N}}$ in a way similar to Theorem 4.9 requires the definition of the multiplication operator

$$M : \begin{cases} C(\mathcal{X}) \longrightarrow C(\mathcal{X}) \\ f(x) \longmapsto d(x)f(x) \end{cases}$$

for all $x \in \mathcal{X}$, and the integral operator

$$S : \begin{cases} C(\mathcal{X}) \longrightarrow C(\mathcal{X}) \\ f(x) \longmapsto \int_{\mathcal{X}} k(x,y)f(y)d\mathbb{P}(y) \end{cases}$$

for all $x \in \mathcal{X}$. Note that S is a compact operator because of (2.5). We define the operator U as

$$U := M - S$$

Additionally, we define the operator sequence

$$U_n : \begin{cases} C(\mathcal{X}) \longrightarrow C(\mathcal{X}) \\ f(x) \longmapsto d_n(x)f(x) - \int_{\mathcal{X}} k(x,y)f(y)d\mathbb{P}_n(y) \end{cases}$$

for $n \in \mathbb{N}$. Note that U and $(U_n)_{n\in\mathbb{N}}$ are compact operators because of Proposition 2.15. For a domain D, let $\mathrm{rg}(f)$ denote the range of a function $f : D \to \mathbb{R}$. The range of f is defined as

$$\mathrm{rg}(f) := \{f(x) \mid x \in D\}$$

Note that if the domain M of f is a compact metric space, the range of f satisfies

$$\mathrm{rg}(f) = \left[\min_{x\in D} f(x), \max_{x\in D} f(x)\right]$$

The following statement summarizes important results about the spectra of U and U_n. It contains the unnormalized versions of the results we discussed in Section 4.1, in particular Corollary 4.7 and Proposition 4.8.

PROPOSITION 4.20 (von Luxburg, Belkin, and Bousquet (2008), Proposition 22)**.** *Suppose that $n \in \mathbb{N}$.*

(i) Let $g \in C(\mathcal{X})$ be an eigenfunction of U_n to the eigenvalue λ. Then, $\rho_n g \in \mathbb{R}^n$ is an eigenvector of $\frac{1}{n}L_n$ to the eigenvalue λ.

(ii) Let $\lambda \notin \operatorname{rg}(d_n)$ is an eigenvalue of U_n with eigenfunction $g \in C(\mathcal{X})$, and define $v := (v_1, ..., v_n)^T := \rho_n f$. Then, g has the explicit form

$$g(x) = \frac{\frac{1}{n}\sum_{j=1}^{n} k(x, X_j) v_j}{d_n(x) - \lambda} \tag{4.45}$$

Conversely, if $g \in C(\mathcal{X})$ is of the form (4.45), and v is an eigenvector of $\frac{1}{n}L_n$ to the eigenvalue λ, then g is an eigenfunction of U to the eigenvalue λ.

(iii) The essential spectra of U_n and U satisfy $\sigma_{\text{ess}}(U_n) = \operatorname{rg}(d_n)$ and $\sigma_{\text{ess}}(U) = \operatorname{rg}(d)$. For both operators, their respective eigenvalues are non-negative, and accumulation points occur only in $\operatorname{rg}(d_n)$ and $\operatorname{rg}(d)$, respectively.

Note that the proof of this statement is similar in structure to the proof of Proposition 4.8, e.g. the proof of (i) is a direct implication of the following Lemma, which can be demonstrated in a way similar to how we proved (4.7). To prove the non-negativity of the eigenvalues in (iii) with an red argument about Hilbert spaces: first, we use Proposition 2.33 to show that $(C(\mathcal{X}))$ is an Hilbert space, and then show that $\langle U_n f, f\rangle_{L^2} \geq 0$.

LEMMA 4.21. *For $n \in \mathbb{N}$, the identity*

$$\frac{1}{n} L_n \rho_n = \rho_n U_n$$

holds.

The proof of this Lemma works similarly to how we derived (4.7). This allows us to introduce a result for unnormalized spectral clustering resembling Theorem 4.9:

THEOREM 4.22 (Consistency of Unnormalized Spectral Clustering)**.** *Let $\lambda \notin \operatorname{rg}(d)$ be an eigenvalue of U. Let $M \subset \mathbb{C}$ be a neighborhood of λ such that $\sigma(U) \cap M = \{\lambda\}$. Then, the following convergence results hold:*

(i) Any sequence $(\lambda_n)_{n\in\mathbb{N}}$ of eigenvalues of $(\frac{1}{n}L_n)_{n\in\mathbb{N}}$ such that for all $n \in \mathbb{N}$

$$\lambda_n \in \sigma\left(\frac{1}{n} L_n\right) \cap M$$

converges to λ almost surely, i.e.

$$\lambda_n \longrightarrow \lambda$$

almost surely.

(ii) If λ is a simple eigenvalue, and $(\lambda_n)_{n\in\mathbb{N}}$ satisfies (i), then any sequence of eigenvectors $(v_n)_{n\in\mathbb{N}}$ of $(\frac{1}{n}L_n)_{n\in\mathbb{N}}$ that corresponds to $(\lambda_n)_{n\in\mathbb{N}}$ converges to an eigenfunction f of U corresponding to λ up to a change of sign in the sense of Definition 2.30.

Note that result for spectral projections, similar to Corollary 4.17, also holds (von Luxburg, Belkin, and Bousquet 2008, Theorem 21). At first glance, Theorem 4.22 appears similar to Theorem 4.9: both results state that under certain assumptions, eigenvalues and eigenvectors of the respective matrix sequences converge in the sense of Definition 2.30.

The difference between lies in specific conditions under which convergence holds, in particular with regard to the eigenvalue λ of the limit operator which defines the regions of eigenvalue convergence. While in Theorem 4.9, the condition $\lambda \neq 1$ suffices, Theorem 4.22 requires $\lambda \notin \operatorname{rg}(d)$ to hold. This indicates that unnormalized spectral clustering is only consistent under stricter conditions than normalized spectral clustering. The following example illustrates that there may exist cases in which the condition $\lambda \notin \operatorname{rg}(d)$ is not satisfied.

EXAMPLE 4.23 (von Luxburg, Belkin, and Bousquet 2008, Example 2)**.** *Consider the data space $\mathcal{X} = [1,2]$. For $s \in [0,3]$, the distribution $\mathbb{P}$ on $\mathcal{X}$ is given by the probability density function (see Figure 4.2)*

$$p(x) = \begin{cases} \frac{3-s}{2} \text{ for } x \in \left[1, \frac{4}{3}\right) \\ s \text{ for } x \in \left[\frac{4}{3}, \frac{5}{3}\right) \\ \frac{3-s}{2} \text{ for } x \in \left[\frac{5}{3}, 2\right) \end{cases}$$

For $n \in \mathbb{N}$, we construct the affinity matrix of G_n using the similarity function

$$k : \begin{cases} \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R} \\ (x,y) \longmapsto xy \end{cases}$$

Then, all eigenvalues of U with the exception of 0 are contained within $\operatorname{rg}(d)$.

Note that, as we can observe in Figure 4.2, the distribution $\mathbb{P}$ has two distinct high-density regions for $s = 0.3$ (von Luxburg, Belkin, and Bousquet 2008, Section 8.1). This suggests Example 4.23 is suitable for our topic.

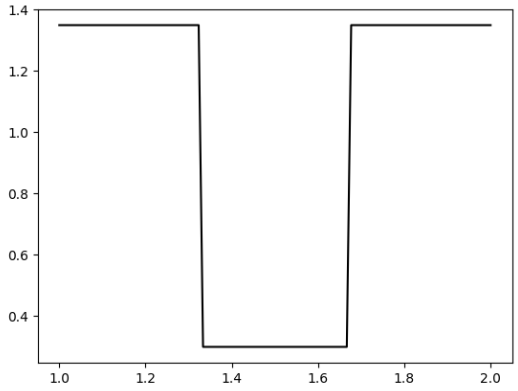


Figure 4.2: Probability Density Function p from Example 4.23 for $s = 0.3$

Proof of Example 4.23 To prove that all positive eigenvalues of U are inside $\operatorname{rg}(d)$, we solve the eigenproblem

$$Uf(x) = \lambda f(x)$$

which is equivalent to the equation

$$d(x)f(x) - x\int_{\mathcal{X}} yf(y)p(y)dy = \lambda f(x) \tag{4.46}$$

We define $\beta := \int_{\mathcal{X}} yf(y)p(y)dy$ and find that eigenfunctions of U have the form

$$f(x) = \frac{\beta x}{d(x) - \lambda} \tag{4.47}$$

By combining (4.47) and the equation that defines β, we obtain

$$\beta = \int_{\mathcal{X}} \frac{\beta y^2}{d(y) - \lambda} p(y)dy \tag{4.48}$$

$$1 = \int_{\mathcal{X}} \frac{y^2}{d(y) - \lambda} p(y)dy \tag{4.49}$$

Hence, λ is an eigenvalue of U if and only if it satisfies (4.49). To find out for which λ this is the case, we define the function

$$g(\lambda) = \int_{\mathcal{X}} \frac{y^2}{d(y) - \lambda} p(y)dy$$

A short calculation shows that the degree function is given by $d(x) = \frac{3}{2}x$. Plugging in the definition of p, we obtain

$$g(\lambda) = \frac{3-s}{2} \int_1^{4/3} \frac{y^2}{\frac{3}{2}y - \lambda} dy \; + s \int_{4/3}^{5/3} \frac{y^2}{\frac{3}{2}y - \lambda} dy \; + \frac{3-s}{2} \int_{5/3}^{2} \frac{y^2}{\frac{3}{2}y - \lambda} dy$$

By computing the integral $\int \frac{y^2}{\frac{3}{2}y - \lambda} dy$, we can solve (4.49) and observe that the condition is equivalent to $\lambda = 0$. This proves that the only part of the $\sigma(U)$ that lies outside essential spectrum $\mathrm{rg}(d)$ is the trivial eigenvalue 0. □

Building on the work of von Luxburg, Belkin, and Bousquet (2008, Section 8), we turn to the reasons behind the more restrictive condition of $\lambda \notin \mathrm{rg}(d)$ in the case of unnormalized spectral clustering. Suppose that U is such that $\sigma(U) = \mathrm{rg}(d) \; \cup \{0\}$ and the eigenvalue 0 of U has multiplicity 1. Also assume that the probability measure $\mathbb{P}$ has no single point masses, i.e. that $\mathbb{P}(\{x\}) = 0$ holds for all $x \in \mathcal{X}$. For a given $\lambda \in \mathrm{rg}(d)$, we choose x_λ such that $d(x_\lambda) = \lambda$. For $n \in \mathbb{N}$, let B_n denote the ball

$$B_n = B_{1/n}(x_\lambda)$$

It is important to note that B_n does not depend on the sample sequence $(X_n)_{n\in\mathbb{N}}$, but only on n itself. Consider a function sequence $(f_n)_{n\in\mathbb{N}}$ in $C(\mathcal{X})$ which satisfies both of the conditions

$$f_n(x) = 1 \quad \text{for } x \in B_n \tag{4.50}$$

$$f_n(x) = 0 \quad \text{for } x \in \mathcal{X} \setminus (B_n \cup B_{n-1}) \tag{4.51}$$

For such function sequences, we find that for $n \to \infty$

$$\|(U_n - \lambda I)f_n\|_\infty \longrightarrow 0 \tag{4.52}$$

holds. This can be proven easily: plugging in the definition U_n, we obtain

$$\|(U_n - \lambda I) f_n\|_\infty = \sup_{x \in \mathcal{X}} \left| d_n(x) f_n(x) - \int_{\mathcal{X}} k(x,y) f_n(y) d\mathbb{P}_n - \lambda f_n(x) \right|$$

We observe that the integral term on the right-hand side vanishes as n increases, which is due to the definition of $(f_n)_{n \in \mathbb{N}}$, the assumption that $\mathbb{P}$ has no point masses, and the fact that $\mathbb{P}_n \to \mathbb{P}$ for $n \to \infty$, as proven earlier. It thus suffices to show that

$$\sup_{x \in \mathcal{X}} |d_n(x) f_n(x) - \lambda f_n(x)| \tag{4.53}$$

converges to 0. To calculate the expression in 4.53, we regard separate cases: if $x \in \mathcal{X} \setminus B_{n-1}$, the expression is equal to zero because $f_n(x) = 0$ by definition. In the other case, we have $x \in B_{n-1}$, thus $f_n(x) = 0$ does not hold necessarily. For large n, we obtain

$$|(d_n(x) - \lambda) f_n(x)| \approx |(d(x) - \lambda) f_n(x)|$$

because d_n converges to d, which follows from the Glivenko-Cantelli property of the function class $\mathcal{K}$. Note that for large n, x needs to be infinitesimally close to x_λ in order for $f_n(x)$ to be positive. Because d is continuous, it follows that $d(x)$ is approximately $d(x_\lambda) = \lambda$. This proves the pointwise convergence

$$|(U_n - \lambda I) f_n(x)| \longrightarrow 0$$

for all $x \in \mathcal{X}$, which proves pointwise convergence. Hence, uniform convergence follows from the Arzela-Ascoli theorem (Theorem 2.10): we demonstrate equicontinuity in a way that is similar to the proof of (4.27). It follows by definition that for every $\epsilon > 0$, there exists an n_0 in $\mathbb{N}$ such that

$$\|(U_n - \lambda I) f_n\|_\infty < \epsilon$$

for all $n \geq n_0$. This fact, according to Lemma 4.21, implies that

$$\left| \left(\frac{1}{n} L_n - \lambda I \right) (f(X_1), ..., f(X_n))^T \right| < \epsilon \tag{4.54}$$

holds. Consider the implications of this to the practical application of spectral clustering: for any machine precision ϵ, any λ in the essential spectrum of U is not distinguishable from an eigenvalue of U for large n. In particular, the outputs for the smallest positive eigenvalue $\lambda_2(\frac{1}{n} L_n)$ obtained through eigensolvers will approach $\min_{x \in \mathcal{X}} d(x)$. Note that the associated eigenvectors converge to the indicator function

$$\mathbb{1}_{\{y \in \mathcal{X} : y = \text{argmin } d(x)\}} : \mathcal{X} \longrightarrow \{0, 1\}$$

This is a consequence of the definition of the sequence $(f_n)_{n \in \mathbb{N}}$. Note that the indicator $\mathbb{1}_{\{y \in \mathcal{X} : y = \text{argmin } d(x)\}}$ is not continuous, so the convergence result in Theorem 4.22 does not hold when $\lambda \in \text{rg}(d)$. Eigenvectors that converge to such an indicator do not lead to sensilble partitions, as we illustrate in a simulation later. It is also important to note that the property

(4.52) mirrors a similar property of U, which holds more generally for the essential spectrum of certain operators (for detail, see von Luxburg, Belkin, and Bousquet (2008) and Chatelin (1983)). The following propositions characterizes a class of cases which violate the condition $\lambda \notin \operatorname{rg}(d)$ (von Luxburg, Belkin, and Bousquet 2008, Proposition 25). For its proof, von Luxburg, Belkin, and Bousquet (2008) refer to literature that touches on more general statements including Lakaev (1989) and Ikmorov and Shapirov (1998).

PROPOSITION 4.24. *If $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$ and $d \in \mathbb{N}$, U only has finitely many eigenvalues outside of $\operatorname{rg}(d)$ if the following conditions hold:*

(1) The similarity function k is analytic in a neighborhood of $\mathcal{X} \times \mathcal{X}$.

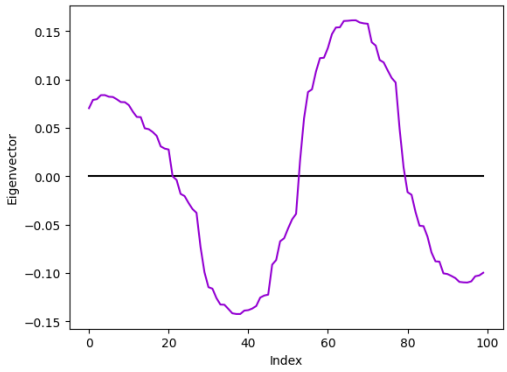
(2) The distribution $\mathbb{P}$ has an analytic density function.

(3) The set $\{y \in \mathcal{X} : y = \operatorname{argmin} d(x)\}$ contains a finite number of elements.

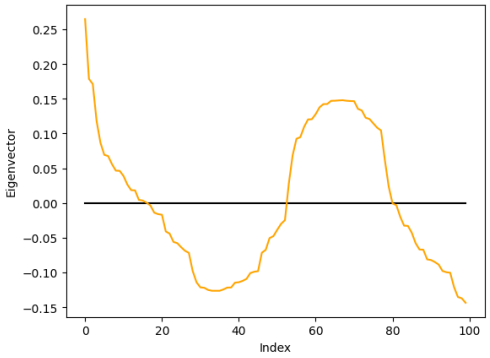
Lastly, we present a simple simulation which illustrates the weaknesses of unnormalized spectral clustering discussed in the previous section. The example we provide draws from von Luxburg, Belkin, and Bousquet (2008, Section 8.2). Suppose that for $\mathcal{X} = \mathbb{R}$, we draw n numbers of a random sequence from the Gaussian mixture distribution

$$P = \frac{1}{4} \cdot \mathcal{N}(2,1) + \frac{1}{4} \cdot \mathcal{N}(5,1) + \frac{1}{4} \cdot \mathcal{N}(8,1) + \frac{1}{4} \cdot \mathcal{N}(11,1)$$

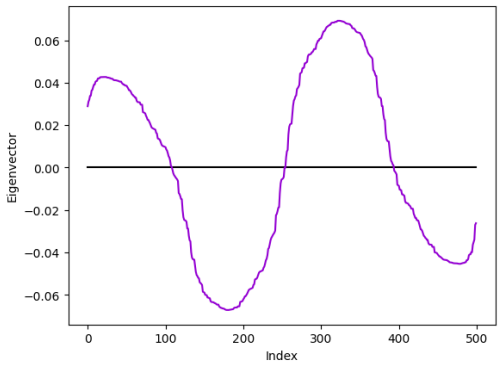
and use the Gaussian similarity function $k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2h^2}\right)$. As in Example 4.10, we select the kernel width using Scott's rule $h = 1.06 \cdot \sigma \cdot n^{-1/5}$ for a random sample of the size $n = 500$ (see Subsection 3.31). To observe the difference between normalized and unnormalized spectral clustering, we compute the fourth eigenvectors of both L'_n and L_n for different values of n, respectively. The plot displays of the eigenvectors in Figure 4.3 indicates what we found in this section: eigenvector sequences of $(L_n)_{n\in\mathbb{N}}$ converge to degenerate functions under certain assumptions. This aligns with the different conditions established in Theorem 4.9 and Theorem 4.22: the condition required for eigenvector sequences of $(L'_n)_{n\in\mathbb{N}}$ to converge is significantly milder than it is in the case of $(L_n)_{n\in\mathbb{N}}$.



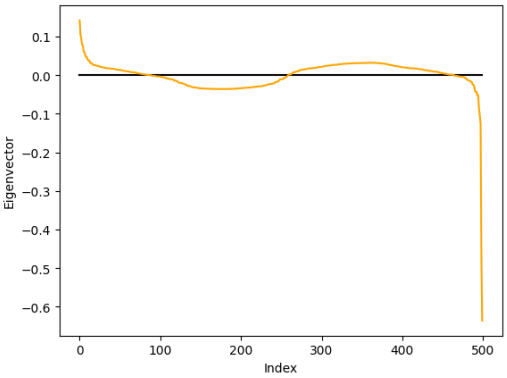
(a) normalized, $n = 100$



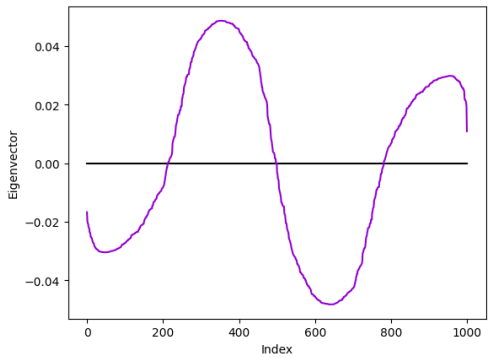
(b) unnormalized, $n = 100$



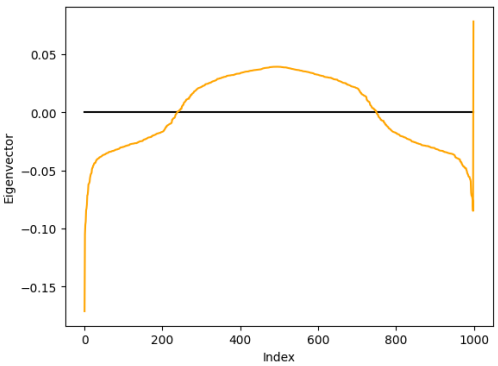
(c) normalized, $n = 500$



(d) unnormalized, $n = 500$



(e) normalized, $n = 1000$



(f) unnormalized, $n = 1000$

Figure 4.3: Comparison of the fourth eigenvectors L'_n and L_n for $n = 100, 500, 1000$ (indexes are ordered with respect to a non-incremental ordering of the sample)

A Appendix: Python Commands and Results

Throughout all the snippets we provide here, we have employed the Python libraries *numpy*, *pandas*, and *matplotlib* (NumPy 2025; pandas 2025; matplotlib 2025). Therefore, they are not mentioned specifically ahead of every individual snippet. Additional packages that we used are listed specifically.

A.1 Section 3

Converting Data, Histogram in Figure 3.1, List in Figure 3.2

```
# data processing
df_visual = pd.read_csv("ag_test_data4.csv", usecols=['Country
    Name','2023 [YR2023]'] ).iloc[0:50]
df_1d = pd.read_csv("ag_test_data4.csv", usecols=['2023
    [YR2023]']).iloc[0:50]
tech_data = df_1d.to_numpy()
td_flattened = tech_data.flatten()
df_visual

# Histogram
plt.hist(td_flattened, color='black', bins = 150)
plt.xlabel("Value added by agriculture, fishing, and forestry in percent
    of GDP (2023)")
plt.ylabel("Number of Countries")
plt.show()
```

Spectral Clustering

For (normalized) spectral clustering, we used *SpectralClustering* from the *clustering* module of the *scikit-learn* library (scikit-learn 2025). For the eigenvalues and eigenvectors, we employed the *scipy.linalg* module from the *SciPy* library (SciPy 2025).

```
# compute and set Gaussian kernel width
n2 = len(td_flattened)
emp_std = np.std(td_flattened)
emp_width = 1.06 * emp_std * n2**(-1/5)
gamma = 1/(2*emp_width**2)

# main algorithm
clustering = SpectralClustering(n_clusters=2, gamma=0.011).fit(tech_data)
aff = clustering.affinity_matrix_
labels = clustering.labels_

# Histogram displaying the clusters
for label in np.unique(labels):
    plt.hist(td_flattened[labels==label], bins=75)

plt.xlabel("Value added by agriculture, fishing, and forestry in percent
    of GDP (2023)")
plt.ylabel("Number of Countries")
plt.show()
```

```
# Compute matrices
D = np.zeros((n2, n2))
for i in range(n2):
    D[i, i] = np.sum(aff[i, :])
L = D - aff
D_inv_sqrt = np.zeros((n2, n2))
for i in range(n2):
    D_inv_sqrt[i, i] = 1 / np.sqrt(D[i, i])
L_prime = D_inv_sqrt @ L @ D_inv_sqrt

# eigenvalues and eigengaps
lambdas, v = eigh(L_prime)
gap2 = lambdas[1] - lambdas[0]
gap3 = lambdas[2] - lambdas[1] # analogous for all other eigengaps
plt.scatter(indices, lambdas, marker = 'o', color = 'black')
plt.xlabel('Index')
plt.ylabel('Eigenvalues')
plt.xlim(0, 10)
plt.show()
```

A.2 Section 4

Simulation in Section 4.2

Again, we used the *scipy.linalg* module from *SciPy* for the eigenvector computations.

```
# generate the data
np.random.seed(1962)
s1 = np.random.normal(0.3, 0.1, 250) # select the third parameter, in
    this case 250 for width selection, as half the desired sample size
np.random.seed(1962)
s2 = np.random.normal(0.7, 0.1, 250)
s = np.concatenate([s1, s2])
sorted_s = np.sort(s)

# determine the kernel width
s_len = len(s) # for len(s)=250
s_std = np.std(s)
s_width = 1.06 * s_std * s_len**(-1/5)

# compute the similarity matrix
def euclidean(u: np.ndarray, v: np.ndarray):
    if u.shape != v.shape:
        raise ValueError("Please choose the correct format.")
    return np.sqrt(np.sum((u-v)**2))
def gaussian_kernel(u: np.ndarray, v: np.ndarray, sigma: float):
    return np.exp(- (euclidean(u, v))**2 / sigma)
def similarity_matrix_1D(d: np.ndarray, sigma: float):
    lend = len(d)
    M = np.zeros((lend, lend))
    for i in range(lend):
        for j in range(lend):
            M[i, j] = gaussian_kernel(d[i], d[j], sigma)
    return M
```

```
K = similarity_matrix_1D(data, 0.0687) # rounded width (Scott's rule)

# compute the Laplacian matrices
D = np.zeros((n, n))
for i in range(n):
    D[i, i] = np.sum(K[i, :])
L = D - K
D_inv_sqrt = np.zeros((n, n))
for i in range(n):
    D_inv_sqrt[i, i] = 1 / np.sqrt(D[i, i])
L_prime = D_inv_sqrt @ L @ D_inv_sqrt

# compute and plot the eigenvectors
lambdas, v = eigh(L_prime)
constant = np.zeros(n)
plt.plot(indices, v[:, 1], color = "green")
plt.title("Second Eigenvector")
plt.plot(indices, constant, color = 'black')
plt.xlabel("Index")
plt.ylabel("Eigenvector")
plt.show()
```

Density Plot in Figure 4.2, Section 4.6

```
S = 0.3
x = np.linspace(1, 2, 100)
y = np.piecewise(x, [(x <= 4/3), (x>4/3) & (x <= 5/3), (x > 5/3)]
    ,[(3-S)/2, S, (3-S)/2])
plt.plot(x, y, color = 'black')
plt.show()
```

Simulation in Section 4.6

We employed the *scipy.linalg* module from *SciPy* for the eigenvector computations.

```
# generate the data
np.random.seed(1962)
x1 = np.random.normal(2, 1.0, 1250) # for sample size n, set third
    parameter to n/4 for x1, x2, x3, x4
np.random.seed(1962)
x2 = np.random.normal(5, 1.0, 1250)
np.random.seed(1962)
x3 = np.random.normal(8, 1.0, 1250)
np.random.seed(1962)
x4 = np.random.normal(11, 1.0, 1250)
x = np.concatenate([x1, x2, x3, x4])
sorted_x = np.sort(x)

# determine the kernel width
x_len = len(x)
x_std = np.std(x)
x_width = 1.06 * x_std * x_len**(-1/5)

# compute the similarity matrix
```

```
def euclidean(u: np.ndarray, v: np.ndarray):
    if u.shape != v.shape:
        raise ValueError("Please choose the correct format.")
    return np.sqrt(np.sum((u-v)**2))
def gaussian_kernel(u: np.ndarray, v: np.ndarray, sigma: float):
    return np.exp(- (euclidean(u, v))**2 / sigma)
def similarity_matrix_1D(d: np.ndarray, sigma: float):
lend = len(d)
M = np.zeros((lend, lend))
for i in range(lend):
    for j in range(lend):
        M[i, j] = gaussian_kernel(d[i], d[j], sigma)
    return M
data = sorted_x
n = len(data)
K = similarity_matrix_1D(data, 1.0671) # rounded width (Scott's rule)

# compute the Laplacian matrices
D = np.zeros((n, n))
for i in range(n):
    D[i, i] = np.sum(K[i, :])
L = D - K
D_inv_sqrt = np.zeros((n, n))
for i in range(n):
    D_inv_sqrt[i, i] = 1 / np.sqrt(D[i, i])
L_prime = D_inv_sqrt @ L @ D_inv_sqrt

# compute and plot the fourth eigenvector
lambdas, v = eigh(L) # unnormalized; replace L with L_prime for
    normalized spectrum
indices = np.arange(n)
plt.plot(indices, constant, color = 'black')
plt.xlabel("Index")
plt.ylabel("Eigenvector")
plt.plot(indices, v[:, 3], color = "orange") # unnormalized; replace
    "orange" with "darkviolet" for the normalized plot
plt.show()
```

References

Aggarwal, Charu C., and Chandan K. Reddy. 2014. *Data Clustering. Algorithms and Applications.* EBook pdf version. Boca Ranton, FL, USA: CRC Press. ISBN: 978-1-4665-5822-9.

Aldous, David, and James Allen Fill. 2002. *Reversible Markov Chains and Random Walks on Graphs.* Unfinished monograph, recompiled 2014 version. https://www.stat.berkeley.edu/users/aldous/RWG/book.pdf.

Anthony, Martin. 2002. *Uniform Glivenko-Cantelli Theorems and Concentration of Measure in the Mathematical Modelling of Learning.* Research Report LSE-CDAM-2002-07. London School of Economics. http://www.cdam.lse.ac.uk/Reports/Files/cdam-2002-07.pdf.

Atkinson, Kendall E. 1967. "The Numerical Solution of the Eigenvalue Problem for Compact Integral Operators." *Transactions of the American Mathematical Society* 129:458–465.

Banfield, Jeffrey D., and Adrian E. Raftery. 1993. "Model-Based Gaussian and Non-Gaussian Clustering." *Biometrics* 49 (3): 803–821. https://sites.stat.washington.edu/raftery/Research/PDF/banfield1993.pdf.

Bezdek, James C., Robert Ehrlich, and William Full. 1984. "FCM: The Fuzzy c-Means Clustering Algorithm." *Computers & Geosciences* 10 (2-3): 191–203. https://doi.org/https://doi.org/10.1016/0098-3004(84)90020-7.

Björck, Åke, and Gene H. Golub. 1973. "Numerical Methods for Computing Angles between Subspaces." *Mathematics of Computation* 27 (123): 579–594. https://doi.org/10.2307/2005662.

Chatelin, Francoise. 1983. *Spectral Approximation of Linear Operators.* Computer Science and Applied Mathematics. New York: Academic Press. ISBN: 0-12-170620-6.

Chavent, Marie. 1998. "A Monothetic Clustering Method." *Pattern Recognition Letters* 19 (11): 989–996. https://doi.org/https://doi.org/10.1016/S0167-8655(98)00087-7.

Clark, John, and Derek Allan Holton. 1991. *A First Look at Graph Theory.* Singapore: World Scientific. ISBN: 9810204906.

Day, William H. E., and Herbert Edelsbrunner. 1984. "Efficient Algorithms for Agglomerative Hierarchical Clustering Methods." *The Computer Journal* 1:7–24. https://doi.org/https://doi.org/10.1007/BF01890115.

Defays, D. 1977. "An Efficient Algorithm for a Complete Link Method." *The Computer Journal* 20 (4): 364–366. https://academic.oup.com/comjnl/article/20/4/364/393966.

Donath, W. E., and A. J. Hoffman. 1973. "Lower Bounds for the Partitioning of Graphs." *IBM Journal of Research and Development* 17 (5): 420–425. https://doi.org/10.1147/rd.175.0420.

Dunford, Nelson, and Jacob T. Schwartz. 1958. *Linear Operators. Part I: General Theory.* Wiley Classics Library Edition Published 1988. New York, NY, USA: Wiley.

Durrett, Richard. 1999. *Essentials of Stochastic Processes.* Springer Texts in Statistics. Third Edition. Switzerland: Springer Cham. https://doi.org/10.1007/978-3-319-45614-0.

Ezugwu, Absalom E., Abiodun M. Ikotun, Olaide O. Oyelade, Laith Abualigah, Jeffery O. Agushaka, Christopher I. Eke, and Andronicus A. Akinyelu. 2022. "A Comprehensive Survey of Clustering Algorithms: State-of-the-Art Machine Learning Applications, Taxonomy, Challenges, and Future Research Prospects." *Engineering Applications of Artificial Intelligence* 110. https://doi.org/10.1016/j.engappai.2022.104743.

Fiedler, Miroslav. 1973. "Algebraic Connectivity of Graphs." *Czechoslovak Mathematical Journal* 23 (2): 298–305. https://eudml.org/doc/12723.

Golub, Gene H., and Charles F. Van Loan. 1989. *Matrix Computations.* Third printing, 1991. Baltimore, MD, USA: The Johns Hopkins University Press. ISBN: 0-8018-3772-3.

Hartigan, J.A., and M.A. Wong. 1979. "A K-Means Clustering Algorithm." *Journal of the Royal Statistical Society,* C, 28 (1): 100–108. https://doi.org/10.2307/2346830.

Heuser, Harro. 1975. *Funktionalanalysis.* 4., durchgesehene Auflage 2006. Wiesbaden: Teubner. ISBN: 9783835100268.

Horn, Roger A., and Charles R. Johnson. 1985. *Matrix Analysis.* Online publication 2012. Cambridge University Press. https://doi.org/10.1017/CBO9780511810817.

Ikmorov, A., and F. Shapirov. 1998. "On the Discrete Spectrum of the Nonanalytic Matrix-Valued Friedrichs Model." *Functional Analysis and its Applications* 32:49–51. https://doi.org/https://link.springer.com/article/10.1007/BF02465757.

Kato, Tosio. 1966. *Perturbation Theory for Linear Operators.* Corrected Printing of the Second Edition. 1980. Berlin: Springer. ISBN: 3-540-07558-5.

Kim, Jaejik, and L. Billard. 2011. "A Polythetic Clustering Process and Cluster Validity Indexes for Histogram-Valued Objects." *Computational Statistics & Data Analysis* 55 (7): 2250–2262. https://doi.org/https://doi.org/10.1016/j.csda.2011.01.011.

Klenke, Achim. 2006. *Wahrscheinlichkeitstheorie.* 4., überarbeitete und ergänzte Auflage 2020. Berlin: Springer Spektrum. https://doi.org/10.1007/978-3-662-62089-2.

Knyazev, A.V., and P. Zhu. 2012. *Principal Angles Between Subspaces and Their Tangents.* Technical Report TR2012-058. Mitsubishi Electric Research Laboratories. https://merl.com/publications/docs/TR2012-058.pdf.

Krengel, Ulrich. 1988. *Einführung in die Wahrscheinlichkeitstheorie und Statistik.* Vieweg Studium: Aufbaukurs Mathematik. 8., erweiterte Auflage 2005. Wiesbaden: Vieweg. ISBN: 3834800635.

Lakaev, S.N. 1989. "Some Spectral Properties of the Generalized Friedrichs Model." Translated from Trudy Seminara imeni I. G. Petrovskogo, No. 11, pp. 210–238, 1986. *Journal of Mathematical Sciences* 45:1540–1563. https://doi.org/10.1007/BF01097277.

Lee, Clement, and Darren Wilkinson. 2019. "A Review of Stochastic Block Models and Extensions for Graph Clustering." *Applied Network Science* 4 (122). https://appliedn etsci.springeropen.com/articles/10.1007/s41109-019-0232-2.

LLoyd, S. 1982. "Least Squares Quantization in PCM." *IEEE Transactions on Information Theory* 28 (2): 129–137. https://doi.org/10.1109/TIT.1982.1056489.

MacQueen, J. 1967. "Some Methods for Classification and Analysis of Multivariate Observations." *Berkeley Symposium on Mathematicas Statistics and Probability* 5 (1): 281–297. https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992.

matplotlib. 2025. "matplotlib." Accessed: July 23, 2025. https://matplotlib.org/.

Mendelson, Shahar. 2003. "A Few Notes on Statistical Learning Theory." In *Advanced Lectures on Machine Learning. Lecture Notes in Computer Science.* Springer. https://doi.org/10.1007/3-540-36434-X_1.

Mur, Angel, Raquel Dormido, Natividad Duro, Sebastian Dormido-Canto, and Jesús Vega. 2016. "Determination of the Optimal Number of Clusters Using a Spectral Clustering Optimization." *Expert Systems with Applications* 65:304–314. https://doi.org/10.1016/j.eswa.2016.08.059.

Ng, Andrew, Michael Jordan, and Yair Weiss. 2001. "On Spectral Clustering: Analysis and an Algorithm." *Advances in Neural Information Processing Systems* 14. https://papers.nips.cc/paper_files/paper/2001/hash/801272ee79cfde7fa5960571fee36b9b-Abstract.html.

NumPy. 2025. "NumPy." Accessed: July 23, 2025. https://numpy.org/.

pandas. 2025. "pandas." Accessed: July 23, 2025. https://pandas.pydata.org/.

Royden, H. L. 1963. *Real Analysis.* Second Edition 1968. Library of Congress catalog card number: 68-10518. New York: Macmillan.

Rudin, Walter. 1966. *Real and Complex Analysis.* McGraw-Hill Series in Higher Mathematics. Second Edition, 1974. New York, NY, USA: McGraw-Hill. ISBN: 0-07-054233-3.

Saxena, Amit, Mukesh Prasad, Akshansh Gupta, Neha Bharill, Om Prakash Patel, Aruna Tiwari, Meng Joo Er, Weiping Ding, and Chin-Teng Lin. 2017. "A Review of Clustering Techniques and Developments." *Neurocomputing* 267:664–681. https://doi.org/10.1016/j.neucom.2017.06.053.

scikit-learn. 2025. "Spectral Clustering." Accessed: July 20, 2025. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html.

SciPy. 2025. "scipy.linalg.eigh." Accessed: July 21, 2025. https://docs.scipy.org/doc/scipy/reference/generated/scipy.linalg.eigh.html.

Shi, Jianbo, and J. Malik. 2000. "Normalized Cuts and Image Segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (8): 888–905. https://doi.org/10.1109/34.868688.

Sibson, R. 1973. "SLINK: An Optimally Efficient Algorithm for the Single-Link Cluster Method." *The Computer Journal* 16 (1): 30–34. https://doi.org/10.1093/comjnl/16.1.30.

Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis.* Vol. 26. Monographs on Statistics and Applied Probability. Printed in 1998. Boca Ranton, FL, USA: Chapman & Hall. ISBN: 978-0-412-24620-3.

Stewart, G.W., and Ji-guang Sun. 1990. *Matrix Perturbation Theory.* United Kingdom Edition published by Academic Press Limited. San Diego, California: Academic Press, Inc. ISBN: 0-12-670230-6.

United Nations Development Programme. 2025. *Human Devlopment Report 2025: A Matter of Choice: People and Possibilities in the Age of AI.* Report. English version. Accessed on July 18, 2025. https://hdr.undp.org/content/human-development-report-2025.

van der Vaart, A.W., and Jon A. Wellner. 1996. *Weak Convergence and Empirical Processes.* Second edition 2023. Switzerland: Springer Cham. https://doi.org/10.1007/978-3-031-29040-4.

von Luxburg, Ulrike. 2007. "A Tutorial on Spectral Clustering." *Statistics and Computing* 17:395–416. https://doi.org/10.1007/s11222-007-9033-z.

von Luxburg, Ulrike, Mikhail Belkin, and Olivier Bousquet. 2008. "Consistency of Spectral Clustering." *Annals of Statistics* 36 (2): 555–586. 10.1214/009053607000000640.

World Bank. 2025. *Agriculture, Forestry, and Fishing, Value Added (% of GDP).* World Bank Development Indicators. Data for the year 2023. Accessed on July 16, 2025. The data was manually selected and downloaded on July 16, 2025. https://databank.worldbank.org/source/world-development-indicators#.

Yu, Yi, Tengyao Wang, and Richard J. Samworth. 2014. *A Useful Version of the Davis-Kahan Theorem for Statisticians.* ArXiv:1405.0680. https://doi.org/https://doi.org/10.48550/arXiv.1405.0680.

Zhou, Ding-Xuan. 2002. "The Covering Number in Learning Theory." *Journal of Complexity* 18 (3): 739–767. https://doi.org/10.1006/jcom.2002.0635.

Declaration of Authorship

German Version (Eigenständigkeitserklärung)

Hiermit versichere ich, dass ich die hier vorliegende Masterarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, alle Ausführungen, die anderen Schriften wörtlich oder sinngemäß entnommen wurden, kenntlich gemacht sind und die Arbeit in gleicher oder ähnlicher Fassung noch nicht Bestandteil einer Studien- oder Prüfungsleistung war.

English Version

I hereby declare that I have written the present master's thesis independently and have used no sources or aids other than those indicated. All passages taken verbatim or in substance from other works have been clearly marked as such. Furthermore, this thesis has not previously been submitted, in whole or in part, for any other academic examination or degree.

Bochum, July 28, 2025

Vred Rudnick