Max–Planck–Institut für biologische Kybernetik
Max Planck Institute for Biological Cybernetics

Technical Report No. TR-134

Consistency of Spectral Clustering

Ulrike von Luxburg[1], Mikhail Belkin[2], Olivier Bousquet[3]

December 2004

[1] Department for Empirical Inference, email:ulrike.luxburg@tuebingen.mpg.de
[2] The University of Chicago, Department of Computer Science, 1100 E 58th st., Chicago, USA, email: misha@cs.uchicago.edu
[3] Department for Empirical Inference, email:olivier.bousquet@tuebingen.mpg.de

This report is available in PDF–format via anonymous ftp at ftp://ftp.kyb.tuebingen.mpg.de/pub/mpi-memos/pdf/TRspectral04.pdf.
The complete series of Technical Reports is documented at: http://www.kyb.tuebingen.mpg.de/techreports.html

Consistency of Spectral Clustering

Ulrike von Luxburg, Mikhail Belkin, Olivier Bousquet

Abstract. Consistency is a key property of statistical algorithms, when the data is drawn from some underlying probability distribution. Surprisingly, despite decades of work, little is known about consistency of most clustering algorithms. In this paper we investigate consistency of a popular family of spectral clustering algorithms, which cluster the data with the help of eigenvectors of graph Laplacian matrices. We show that one of the two of major classes of spectral clustering (normalized clustering) converges under some very general conditions, while the other (unnormalized), is only consistent under strong additional assumptions, which, as we demonstrate, are not always satisfied in real data. We conclude that our analysis provides strong evidence for the superiority of normalized spectral clustering in practical applications. We believe that methods used in our analysis will provide a basis for future exploration of Laplacian-based methods in a statistical setting.

---

1 Introduction

Clustering is a popular technique, widely used in statistics, computer science and various data analysis applications. Given a set of data points the goal is to separate the points in several groups based on some notion of similarity. For example, for document retrieval applications it may be useful to organize documents by topic, while an online store may be interested in separating customers in several groups, based on their preference profiles.

Assuming that the data is drawn from an underlying probability distribution, which seems to be a natural setting for many applications of clustering, the overall goal is to find a partition of the data space satisfying certain optimality criteria. To achieve this one has to answer two different questions:

- Assuming that the underlying probability distribution is known, what is a desirable clustering of the data space?
- Given finitely many data points sampled from an unknown probability distribution, how to find a way to approximate that optimal partitioning empirically?

Interestingly, while extensive literature exists on clustering and partitioning (e.g., see Jain et al. (1999) for a review), very few clustering algorithms have been analyzed or shown to converge in the setting where the data is sampled from a continuous probability distribution. Exceptions are the k-means algorithm (Pollard, 1981), the single linkage algorithm (Hartigan, 1981), and the clustering algorithm suggested in Niyogi and Karmarkar (2000). Even these limited results are far from satisfactory. Pollard (1981) shows consistency of the minimizer of the objective function for k-means clustering. However, most commonly used k-means algorithms are local optimization techniques without any global performance or convergence guarantees. Hartigan (1981) demonstrates a weaker notion of consistency, proving that the algorithm will identify certain high-density regions, but does not prove a general consistency result. Finally, the algorithm in Niyogi and Karmarkar (2000) minimizes an unusual objective function, and convergence is shown in the one-dimensional case only. This lack of consistency guarantees is especially striking as many clustering algorithms are widely used in statistics, computer science, and pattern recognition, where they are applied to various tasks of exploratory data analysis.

In this paper we investigate the limit behavior of a class of spectral clustering algorithms. Spectral clustering is a popular technique going back to Donath and Hoffman (1973) and Fiedler (1973). In its simplest form it uses the second eigenvector of the graph Laplacian matrix constructed from the affinity graph between the sample points to obtain a partition of the samples into two groups. The main difference between spectral clustering algorithms is whether they use normalized or unnormalized graph Laplacian. Different versions of spectral clustering have been used for load balancing, parallel computations (Hendrickson and Leland, 1995), VLSI design (Hagen and Kahng, 1992) and sparse matrix partitioning (Pothen et al., 1990). Laplacian-based clustering algorithms also have found

success in applications to image segmentation (Shi and Malik, 2000), text mining (Dhillon, 2001) and as general purpose methods for data analysis and clustering (Alpert and Yao, 1995, Kannan et al., 2000, Ding et al., 2000, Ng et al., 2002). A nice survey on the history of spectral clustering can be found in Spielman and Teng (1996).

We establish consistency results and convergence rates for several versions of spectral clustering. To prove those results, the main step is to establish the convergence of eigenvalues and eigenvectors of random graph Laplace matrices for growing sample size. Interestingly, our analysis shows that while one type of spectral clustering ("normalized") is consistent under very general conditions, another popular version of spectral clustering ("unnormalized") is only consistent under some very specific conditions which do not have to be satisfied in practice. We therefore conclude that the normalized clustering algorithm should be the preferred method in practical applications.

While there has been some work on theoretical properties of spectral clustering on finite point sets (e.g.,Spielman and Teng, 1996, Guattery and Miller, 1998, Kannan et al., 2000), we do not know of any results discussing the limit behavior of spectral clustering for samples drawn from some continuous probability distribution. Related to the question of convergence of graph Laplacians is the question of convergence of similarity matrices constructed from sample points. The convergence of eigenvalues and eigenvectors of positive definite similarity matrices has already attracted some attention in the machine learning community, as can be seen in Shawe-Taylor et al. (2002), Bengio et al. (2003) and Williams and Seeger (2000). The authors build on work of Baker (1977) or Koltchinskii (1998) and Koltchinskii and Giné (2000). However, those results cannot be applied for the case of unnormalized spectral clustering. We note that methods of Baker (1977) are not valid in the case of randomly drawn data points (an issue which has been ignored by the machine learning community so far). They were developed in a deterministic setting, and it is not clear how they can be adapted to random kernel matrices (see Section II.10 of von Luxburg (2004) for details). The results in Koltchinskii (1998) and Koltchinskii and Giné (2000) are very general, and they apply to all reasonable similarity matrices on arbitrary sample spaces. However, for their methods it is of vital importance that the operators under consideration are Hilbert-Schmidt, which turns out not to be the case for the unnormalized Laplacian. In this paper we develop methods which also work for non-compact operators. As a by-product we recover certain results from Koltchinskii (1998) and Koltchinskii and Giné (2000) by using considerably simpler techniques.

There has been some debate on the question of whether normalized or unnormalized spectral clustering should be used. Recent papers using the normalized version include Van Driessche and Roose (1995), Shi and Malik (2000), Kannan et al. (2000), Ng et al. (2002), Meila and Shi (2001), while Barnard et al. (1995) and Guattery and Miller (1998) use the unnormalized clustering. Comparing the empirical behavior of both approaches, Van Driessche and Roose (1995) and Weiss (1999) find some evidence that the normalized version should be preferred. On the other hand, there is a recent study (Higham and Kibble, 2004) which under certain conditions advocates for the unnormalized version. It seems difficult to resolve this question theoretically from purely graph-theoretic considerations as both normalized and unnormalized spectral clustering can be justified by similar graph theoretic principles (see next section). In our work we now obtain the first theoretical results on this question. They clearly show the superiority of normalized spectral clustering over unnormalized spectral clustering from a statistical point of view.

It is interesting to note that several recent methods for semi-supervised and transductive learning are based on eigenvectors of similarity graphs (cf. Chapelle et al., 2003, Belkin and Niyogi, 2004, and closely related Zhou et al., 2004, Zhu et al., 2003). An algorithm for data representation (Laplacian Eigenmaps) based on eigenvectors of the graph Laplacian and its connection to spectral clustering and differential geometry of probability distributions was introduced and discussed in Belkin and Niyogi (2003). We observe that our theoretical framework can also be applied to investigate the consistency of these algorithms with respect to the unlabeled data.

This paper is organized as follows:
1. **Introduction**
2. **Spectral clustering**
we review the problem of graph partitioning and show how spectral can be obtained as a real-valued relaxation of NP-hard discrete-valued graph partitioning problems.

3. **Informal statement of the results**
4. **Prerequisites and notation**
we introduce notations and certain results necessary for stating and proving our results
5. **Convergence of normalized spectral clustering**
6. **Rates of convergence of normalized spectral clustering**
7. **The unnormalized case**
we derive conditions necessary to ensure consistency of unnormalized clustering.
8. **Non-isolated eigenvalues**
we investigate the spectral properties of the limit operators corresponding to normalized and unnormalized spectral clustering, point out some important differences, and show theoretical and practical examples where the convergence conditions in the unnormalized case are violated.
9. **Spectral clustering: from discrete to continuous**
we discuss the problem of clustering in the case of continuous distributions and some directions of future research.
10. **Conclusion**.

2 Spectral clustering

The purpose of this section is to introduce the statistical audience to the potentially unfamiliar problem of graph partitioning and show how spectral clustering emerges as a simple and algorithmically compelling heuristic for approximating various NP-hard bisectioning problems. We will not attempt to make these ideas fully precise or discuss the more involved problem of multiway partitioning, referring the interested reader to the appropriate literature cited in the introduction.

Informally speaking, the problem of graph partitioning is to cut a given graph in two (or more) parts which are "as disjoint as possible". The natural formulation is to try to partition the graph in parts which are of roughly the same size and are connected by as few edges as possible. However, graph cut problems are often NP-hard and therefore not feasible computationally. Even good approximations are difficult to obtain and are sometimes known to be NP-hard as well (see Arora et al. (2004) for some recent work on complexity of approximations). It turns out, however, that a good heuristic can be obtained by writing the cost function as a quadratic form and relaxing the discrete optimization problem of finding the characteristic function of the cut to a real-valued optimization problem. This relaxation leads to an eigenproblem, which can be easily solved using the standard numerical techniques. We formalize these intuitions below.

Let $G = \{V, E\}$ be an undirected graph with vertices V and edges E. We assume that the edges are non-negatively weighted with the weight matrix W. We will use $[i]$ to denote the i'th vertex and $[i] \sim [j]$ when $[i]$ is adjacent to $[j]$. The corresponding edge will be denoted by $[i, j]$. Of course, the numbering of vertices is arbitrary. The weighted *degree* of the vertex $[i]$ is defined to be

$$\deg[i] = \sum_{[i]\sim[j]} w_{ij}.$$

The number of vertices in some subset $S \subset V$ will be denoted by $|S|$. The *volume* of a subset of vertices $S \subset V$ is the total degree of all vertices in the subset

$$\mathrm{vol}(S) = \sum_{[i]\in S} \deg[i].$$

If $S \subset V$ is a set of vertices of G, we define its *edge boundary* δS to be the set of edges connecting S and its complement $\bar{S} = V - S$. Similarly, the volume of δS is defined as

$$\mathrm{vol}(\delta S) = \sum_{[i,j]\in E,\, i\in S, j\in\bar{S}} w_{ij}.$$

This quantity is also known as the *expansion of the cut* given by S, $\bar{S}$.

Perhaps the simplest and the most direct way of partitioning a graph is to consider the minimum cut (mincut). The problem is simply to find a partition $S, \bar{S}$, which minimizes $\mathrm{vol}(\delta S)$. It turns out that efficient algorithms can

be devised to minimize this quantity, e.g., see Stoer and Wagner (1997) and the discussion therein. However, in practice mincut is often not a satisfactory partition. The problem is that no penalty is paid for unbalanced partitions and therefore nothing prevents the algorithm from making single vertices "clusters". An example of such behavior is shown in Figure 1. We see that a more balanced partition is desirable. One standard way to define such partitions

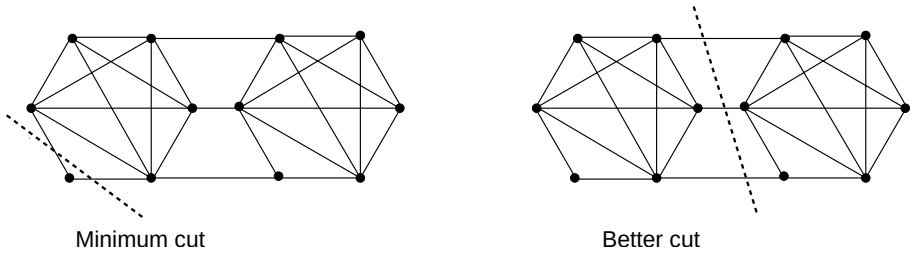

Figure 1: Mincut often splits off single vertices. Instead, one prefers to get more balanced cuts.

is through the Cheeger constant (minimum conductance):

$$c(G) = \operatorname*{argmin}_{S \subset V} \frac{\mathrm{vol}(\delta S)}{\min(\mathrm{vol}(S), \mathrm{vol}(\bar{S}))}.$$

Another notion is the *normalized cut* proposed in Shi and Malik (2000). The authors define normalized cut as

$$n(G) := \operatorname*{argmin}_{S \subset V} \frac{\mathrm{vol}(\delta S)}{\mathrm{vol}(S)} + \frac{\mathrm{vol}(\delta S)}{\mathrm{vol}(\bar{S})}.$$

Observing that for $a, b > 0$, $\min(a, b) \leq \frac{1}{\frac{1}{a}+\frac{1}{b}} \leq 2\min(a, b)$, we see that the Cheeger cut and the normalized cut are closely related. Both problems are NP-hard, and therefore need to be approximated by computationally feasible heuristics in practice.

Another possibility to split the graph is the balanced cut, where the clusters are forced to be of the same size. We will distinguish weighted and unweighted versions of the balanced cut:

$$b_w(G) = \operatorname*{argmin}_{S \subset V,\ \mathrm{vol}(S) = \mathrm{vol}(\bar{S})} \mathrm{vol}(\delta S)$$

$$b_{uw}(G) = \operatorname*{argmin}_{S \subset V,\ |S| = |\bar{S}|} \mathrm{vol}(\delta S).$$

For the purpose of this introduction we simply assume that such balanced partitions exist (e.g., for the unweighted version the number of vertices has to be even). For the general case one can easily modify the definition by requiring the partition to be "almost balanced", and this will lead to the same relaxed optimization problems in the end. We observe that the unweighted version balances the number of vertices in each partition, rather than their volumes. Both of these problems can also be shown to be NP-hard.

We will now define normalized and unnormalized spectral clustering and show how it can be obtained as a relaxation of the weighted or unweighted balanced cut, respectively. By similar arguments, the Cheeger cut and the normalized cut lead to the same relaxed optimization problem as the weighted balanced cut.

Given a graph G with adjacency matrix W, let D be the diagonal matrix with $d_{ii} = \deg[i]$. This matrix is called the *degree matrix* of the graph. We define the *unnormalized graph Laplacian* to be

$$L = D - W$$

and the *normalized graph Laplacian*

$$L' = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}.$$

L and L' are the main objects in spectral graph theory (e.g., Chung, 1997, Mohar, 1991). Various graph invariants can be estimated in terms of eigenvalues of these matrices. Similarly we will see that spectral clustering allows one to replace difficult optimization problems with standard linear algebra. Given a vector $f = (f_1, \ldots, f_n) \in I\!R^n$, the following key identity can be easily verified:

$$fLf^t = 2 \sum_{[i]\sim[j]} w_{ij}(f_i - f_j)^2. \tag{1}$$

Note that this equation shows that L is positive semi-definite, and since $D^{1/2}$ is positive definite the same also holds for L'. We will now see that clustering properties of the graph Laplacian follow directly from this property. Given a subset $S \subset V$, define the column vector $f_S = (f_{S_1}, ..., f_{S_n})' \in I\!R^n$ as follows:

$$f_{S_i} = \begin{cases} 1, & [i] \in S \\ -1, & [i] \in \bar{S} \end{cases}$$

It is immediate that

$$f_S^t L f_S = \sum_{[i]\sim[j]} w_{ij}(f_{Si} - f_{Sj})^2 = 4\,\mathrm{vol}(\delta S),$$

and for all $S \subset V$ we have

$$f_S^t D f_S = \sum_{[i]\in V} w_{ij} = \mathrm{vol}(V).$$

Now consider the case of the weighted balanced cut. We denote by $\mathbf{1}$ the column vector of all ones. We have

$$f_S^t D\mathbf{1} = \sum_{[i]\in S} \deg[i] - \sum_{[i]\in \bar{S}} \deg[i] = \mathrm{vol}\, S - \mathrm{vol}\, \bar{S}.$$

Hence, $f_S^t D\mathbf{1} = 0$ if and only if the cut corresponding to S is volume-balanced, that is $\mathrm{vol}\, S = \mathrm{vol}\, \bar{S}$. Therefore we can reformulate the problem of computing the weighted balanced cut as

$$b_w(G) = \min_{f\in\{-1,1\}^n,\, fD\mathbf{1}=0} f^t L f.$$

Moreover, as $f^t D f$ has the constant value $1/4\,\mathrm{vol}(V)$ for all $f \in \{-1, 1\}^n$, we can rewrite this as

$$b_w(G) = \frac{1}{4}\,\mathrm{vol}(V) \min_{f\in\{-1,1\}^n,\, fD\mathbf{1}=0} \frac{f^t L f}{f^t D f}.$$

Stated in this form, the discrete optimization problem admits a simple relaxation by letting f to take real values instead of $\{-1, 1\}$. Noticing that $L\mathbf{1} = 0$, a standard linear algebra argument shows that

$$\lambda_2 = \min_{f\in I\!R^n,\, fD\mathbf{1}=0} \frac{f^t L f}{f^t D f} \tag{2}$$

where λ_2 is the second smallest eigenvalue of the generalized eigenvector problem $Lf = \lambda Df$. It is clear that the smallest eigenvalue λ_1 of L is 0 and the corresponding eigenvector is $\mathbf{1}$. Moreover, it is easy to show that for a connected graph the second eigenvalue satisfies $\lambda_2 > 0$. Thus, the vector f for which the minimum in Equation (2) is attained is the eigenvector corresponding to λ_2. This line of reasoning leads directly to the following bipartitioning algorithm as relaxation of the weighted balanced cut:

1. *Compute the matrices L and D.*
2. *Find the eigenvector e corresponding to the second smallest eigenvalue of the following generalized eigenvalue problem:*
$$Lf = \lambda Df \tag{3}$$
3. *Obtain the partition: $S = \{[i] : e_i > 0\}$, $\bar{S} = \{[i] : e_i \leq 0\}$.*

This is the basic algorithm for *normalized spectral clustering*.

For *unnormalized spectral clustering*, a similar argument shows that

$$b_{uw}(G) = \frac{1}{4}|V| \min_{f \in \{-1,1\}^n,\, f\mathbf{1}=0} \frac{f^t L f}{f^t f},$$

and by relaxation we obtain an identical algorithm, except that the eigenproblem

$$Lf = \lambda f \tag{4}$$

is solved instead.

Note that in the normalized case, using the second eigenvector of the generalized eigenproblem $Lf = \lambda Df$ is equivalent to using the second eigenvector of the normalized Laplacian $L' = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$. That is easily seen by substituting $v = D^{\frac{1}{2}} f$ into Eq. (3), which can then be rewritten as $L'v = \lambda v$. Since D is a diagonal matrix with positive entries, the partition corresponding to v is the same as the partition corresponding to f. An alternative normalization of the Laplacian which is used sometimes is $L'' = D^{-1}L = I - D^{-1}W$. As for L' it is clear that eigenvectors of L'' are solutions to Eq. (3). While L'' is no longer symmetric, it is easy to see that $D^{-1}W$ is the stochastic matrix corresponding to the random walk on the graph G, where the probability of transition from $[i]$ to $[j]$ is $w_{ij}/\deg[i]$ if the vertices are adjacent and 0 otherwise. $1 - \lambda_2$, the second largest eigenvalue of $D^{-1}W$, controls the *mixing rate* of the random walk (see, e.g. Lovász, 1996). It is possible to derive certain properties of normalized spectral clustering by studying the corresponding random walk on the graph, see for example Meila and Shi (2001).

In most applications we are not given any apriori graph structure and have to construct the graph based on some notion of similarity. This notion is typically represented by a symmetric similarity function $k(x, y)$ on the data space. We assume that $k(x, y)$ is large when x and y are "similar" or "close", and is small or zero otherwise. The graph is then constructed as follows. The vertices of the graph correspond to the data points. Two vertices are connected if the similarity of the corresponding data points is positive, and the edge weights are then given by the similarity. A generic spectral bipartitioning algorithm for data analysis is summarized in the table below:

	Spectral clustering
Input:	n data points $\{X_i\}_{i=1}^n$
Output:	Partition of the data set S, $\bar{S}$.
Step 1	► Choose a symmetric similarity function $k(x, y)$.
Step 2	► Construct a graph with the data points as vertices and edge weights $w_{ij} := k(X_i, X_j)$.
Step 3	► Compute graph Laplacian: $L = D - W.$
Step 4	► Find the eigenvector e corresponding to the second smallest eigenvalue for one of the following problems: $Lf = \lambda Df$ **normalized** $\qquad$ $Lf = \lambda f$ **unnormalized**
Step 5	► Obtain the partition: $S = \{[i] : e_i > 0\}$, $\bar{S} = \{[i] : e_i \leq 0\}$.

Note that this algorithm represents the most basic form of spectral clustering, and the versions used in practice can differ in various details. For example, often the eigenvector e is not thresholded at 0 to obtain the partition (as in step 5), but at some other real value t which depends on the sample. Moreover, in the case when one is interested in obtaining more than two clusters, one typically uses not only the second but also the next few eigenvectors to construct a partition. For details we refer to the literature cited in the introduction.

To summarize, normalized and unnormalized spectral clustering construct a partition of a graph by computing the first few eigenvectors of the normalized or unnormalized Laplacians L and L'. The justification of those algorithms lies in the fact that their solutions approximate the weighted or unweighted balanced cut, respectively, of the graph. In the next sections we investigate behavior of these eigenproblems and the corresponding partitions when points X_i are randomly drawn from some underlying probability distribution.

3 Informal statement of the results

In this section we want to present our main results in a slightly informal but intuitive manner. For the mathematical details and proofs we refer the reader to the following sections. The goal of this article is to study the behavior of normalized and unnormalized spectral clustering on random samples when the sample size n is growing. In Section 2 we have seen that spectral clustering partitions a given sample $X_1, ..., X_n$ according to the coordinates of the first eigenvectors of the (normalized or unnormalized) Laplace matrix. To stress that the Laplace matrices depend on the sample size n, from now on we denote the unnormalized and normalized graph Laplacians by L_n and L'_n (instead of L and L' as in the last section). To investigate whether the various spectral clustering algorithms converge we will have to establish conditions under which the eigenvectors of the Laplace matrices "converge". To see which kind of convergence results we aim at consider the case of the second eigenvector $(v_1, ..., v_n)'$ of L_n. It can be interpreted as a function f_n on the discrete space $\mathcal{X}_n := \{X_1, ..., X_n\}$ by defining $f_n(X_i) := v_i$, and clustering is then performed according to whether f_n is smaller or larger than a certain threshold. It is clear that in the limit for $n \to \infty$, we would like this discrete function f_n to converge to a function f on the whole data space $\mathcal{X}$ such that we can use the values of this function to partition the data space. In our case it will turn out that this space can be chosen as $C(\mathcal{X})$, the space of continuous functions on $\mathcal{X}$. In particular, we will construct a degree function $d \in C(\mathcal{X})$ which will be the "limit" of the discrete degree vector $(d_1, ..., d_n)$. Moreover, we will explicitly construct linear operators U, U', and U'' on $C(\mathcal{X})$ which will be the limit of the discrete operators L_n, L'_n, and L''_n. Certain eigenvectors of the discrete operators are then proved to "converge" (in a certain sense to be explained later) to eigenfunctions of those limit operators. Those eigenfunctions will then be used to construct a partition of the whole data space $\mathcal{X}$.

In the case of normalized spectral clustering it will turn out that this limit process behaves very nicely. We can prove that under certain mild conditions, the partitions constructed on finite samples converge to a sensible partition of the whole data space. In meta-language, this result can be stated as follows:

Result 1 (Convergence of normalized spectral clustering) *Under mild assumptions, if the first r eigenvalues $\lambda_1, ..., \lambda_r$ of the limit operator U' satisfy $\lambda_i \neq 1$ and have multiplicity 1, then the same holds for the first r eigenvalues of L'_n for sufficiently large n. In this case, the first r eigenvalues of L'_n converge to the first r eigenvalues of U' a.s., and the corresponding eigenvectors converge a.s. The clusterings constructed by normalized spectral clustering from the first r eigenvectors on finite samples converge almost surely to a limit clustering of the whole data space.*

In the unnormalized case, the convergence theorem looks quite similar, but there are some subtle differences that will turn out to be important.

Result 2 (Convergence of unnormalized spectral clustering) *Under mild assumptions, if the first r eigenvalues of the limit operator U have multiplicity 1 and do not lie in the range of the degree function d, then the same holds for the first r eigenvalues of $\frac{1}{n}L_n$ for sufficiently large n. In this case, the first r eigenvalues of $\frac{1}{n}L_n$ converge to the first r eigenvalues of U a.s., and the the corresponding eigenvectors converge a.s. The clusterings constructed by unnormalized spectral clustering from the first r eigenvectors on finite samples converge almost surely to a limit clustering of the whole data space.*

On the first glance, both results look very similar: if first eigenvalues are "nice", then spectral clustering converges. However, the difference between Results 1 and 2 is what it means for an eigenvalue to be "nice". For the convergence statements to hold, in Result 1 we only need the condition $\lambda_i \neq 1$, while in Result 2 the condition is $\lambda_i \notin \mathrm{rg}(d)$ has to be satisfied. Both conditions are needed to ensure that the eigenvalue λ_i is isolated in the spectrum of the limit operator, which is a fundamental requirement for applying perturbation theory to the convergence of eigenvectors. We will see that in the normalized case, the limit operator U' has the form $Id - T$ where T is a compact linear operator. As a consequence, the spectrum of U' is very benign, and all eigenvalues $\lambda \neq 1$ are isolated and have finite multiplicity. In the unnormalized case however, the limit operator will have the form $U = M - S$ where M is a multiplication operator and S a compact integral operator. The spectrum of U is not as

nice as the one of U', and in particular it contains the continuous interval $\mathrm{rg}(d)$. Eigenvalues of this operator will only be isolated in the spectrum if they satisfy the condition $\lambda \notin \mathrm{rg}(d)$. As the following proposition shows, this condition has important consequences.

Result 3 (The condition $\lambda \notin \mathrm{rg}(d)$ is necessary)

1. *There exist examples of similarity functions such that there exists no non-zero eigenvalue outside of* $\mathrm{rg}(d)$.
2. *If this is the case, the sequence of second eigenvectors of* $\frac{1}{n}L_n$ *computed by any numerical eigensolver converges to* $\min d(x)$. *The corresponding eigenvectors do not yield a sensible clustering of the data space.*
3. *For a large class of reasonable similarity functions, there are only finitely many eigenvalues (say,* r_0*) inside the interval* $]0, \min d(x)[$. *In this case, the same problems as above occur if the number* r *of eigenvalues used for clustering satisfies* $r > r_0$. *Moreover, we cannot determine* r_0 *from a finite sample.*
4. *The condition* $\lambda \notin \mathrm{rg}(d)$ *refers to the limit case and hence cannot be verified on the finite sample.*

This result complements Result 2. The main message is that firstly, there are many examples where the conditions of Result 2 are not satisfied, secondly, in this case unnormalized spectral clustering fails completely, and thirdly, we cannot detect on a finite sample whether the convergence conditions are satisfied or not.

To further investigate the statistical properties of normalized spectral clustering we compute rates of convergence. Informally, our result is:

Result 4 (Rates of convergence) *The rates of convergence of normalized spectral clustering can be expressed in terms of regularity conditions of the similarity function* k. *For example, for the case of the widely used Gaussian similarity function* $k(x, y) = \exp(-\|x - y\|^2/\sigma^2)$ *on* $\mathbb{R}^d$ *we obtain a rate of* $\mathcal{O}(1/\sqrt{n})$.

Finally, we show how our theoretical results influence the results of spectral clustering in practice. In particular, we demonstrate differences between the behavior of normalized and unnormalized spectral clustering.

Our results show an important difference between normalized and unnormalized spectral clustering: under standard conditions, normalized spectral clustering always converges, while for unnormalized spectral clustering the same is only true if some non-trivial conditions are satisfied. Hence, from a statistical point of view we recommend to use normalized spectral clustering rather than the unnormalized version.

4 Prerequisites and notation

Before we can start we would like to introduce some notation and and summarize some facts from spectral and perturbation theory of bounded linear operators. In the rest of the paper we always make the following **general assumptions:** *The data space* $\mathcal{X}$ *is a compact metric space,* $\mathcal{B}$ *the Borel* σ*-algebra on* $\mathcal{X}$*, and* P *a probability measure on* $(\mathcal{X}, \mathcal{B})$. *Without loss of generality we assume that the support of* P *coincides with* $\mathcal{X}$. *The sample points* $(X_i)_{i \in \mathbb{N}}$ *are drawn independently according to* P. *The similarity function* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is supposed to be symmetric, continuous, and bounded away from 0 by a positive constant, that is there exists a constant* $l > 0$ *such that* $k(x, y) > l$ *for all* $x, y \in \mathcal{X}$.

The affinity graph of a given finite sample $X_1, ..., X_n$ has the sample points as vertices, and the edges $[i, j]$ are weighted by the similarity $k(X_i, X_j)$. As in Section 2 the degree of vertex $[i]$ will be denoted by $deg[i]$. In the following we will denote the degree and the similarity matrices by D_n and K_n, that is D_n is the diagonal matrix with entries $deg[i]$ on the diagonal, K_n the matrix with entries $k(X_i, X_j)$. Similarly we will denote the unnormalized and normalized Laplace matrices by $L_n = D_n - K_n$ and $L'_n = D_n^{-1/2} L_n D_n^{-1/2}$. The eigenvalues of the Laplace matrices $0 = \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$ will always be ordered in increasing order, respecting multiplicities. The term "first eigenvalue" always refers to the trivial eigenvalue $\lambda_1 = 0$. Hence, the interesting eigenvalues are the second, third, and so on.

For a real-valued function f we always denote the range of the function by $\mathrm{rg}(f)$. If $\mathcal{X}$ is connected and f is continuous $\mathrm{rg}(f) := [\inf f(x), \sup f(x)]$. The restriction operator $\rho_n : C(\mathcal{X}) \to \mathbb{R}^n$ denotes the (random)

operator which maps a function to its values on the first n data points, that is $\rho_n f = (f(X_1), ..., f(X_n))$.

Now we want to recall certain facts from spectral and perturbation theory. For more details we refer to Chatelin (1983), Anselone (1971), and Kato (1966). By $\sigma(T)$ we denote the spectrum of a bounded linear operator T on some Banach space E. We define the *discrete spectrum* σ_{d} to be the part of $\sigma(T)$ which consists of all isolated eigenvalues with finite algebraic multiplicity, and the *essential spectrum* $\sigma_{\mathrm{ess}}(T) = \sigma(T) \setminus \sigma_{\mathrm{d}}(T)$. The essential spectrum is always closed, and the discrete spectrum can only have accumulation points on the boundary to the essential spectrum. It is well known (e.g., Theorem IV.5.35 in Kato, 1966) that compact perturbations do not affect the essential spectrum, that is for a bounded operator T and a compact operator V we have $\sigma_{\mathrm{ess}}(T+V) = \sigma_{\mathrm{ess}}(T)$. A subset $\tau \subset \sigma(T)$ is called isolated if there exists an open neighborhood $M \subset \mathbb{C}$ of τ such that $\sigma(T) \cap M = \tau$. For an isolated part $\tau \subset \sigma(T)$ the spectral projection Pr_τ is defined as $\frac{1}{2\pi i} \int_\Gamma (T - \lambda \mathrm{I})^{-1} d\lambda$ where Γ is a closed Jordan curve separating τ from the rest of the spectrum. In particular if $\tau = \{\lambda\}$ for an isolated eigenvalue λ, then Pr_τ coincides with the projection on the invariant subspace related to λ. If λ is a simple eigenvalue (i.e., it has algebraic multiplicity 1), then the spectral projection Pr_τ coincides with the projection on the eigenfunction corresponding to λ.

In the following we will consider different types of convergence of operators:

Definition 5 (Convergence of operators) *Let $(E, \|\cdot\|_E)$ be an arbitrary Banach space, and B its unit ball. Let $(S_n)_n$ be a sequence of bounded linear operators on E.*

- $(S_n)_n$ converges pointwise, *denoted by $S_n \xrightarrow{p} S$, if $\|S_n x - Sx\|_E \to 0$ for all $x \in E$.*
- $(S_n)_n$ converges compactly, *denoted by $S_n \xrightarrow{c} S$, if it converges pointwise and if for every sequence $(x_n)_n$ in B, the sequence $(S - S_n)x_n$ is relatively compact (has compact closure) in $(E, \|\cdot\|_E)$.*
- $(S_n)_n$ converges in operator norm, *denoted by $S_n \xrightarrow{\|\cdot\|} S$, if $\|S_n - S\| \to 0$ where $\|\cdot\|$ denotes the operator norm.*
- $(S_n)_n$ *is called* collectively compact *if the set $\bigcup_n S_n B$ is relatively compact in $(E, \|\cdot\|_E)$.*
- $(S_n)_n$ converges collectively compactly, *denoted by $S_n \xrightarrow{cc} S$, if it converges pointwise and if there exists some $N \in \mathbb{N}$ such that the operators $(S_n - S)_{n>N}$ are collectively compact.*

Both operator norm convergence and collectively compact convergence imply compact convergence. The latter is enough to ensure the convergence of spectral properties in the following sense (cf. Proposition 3.18. and Sections 3.6. and 5.1. in Chatelin, 1983):

Proposition 6 (Perturbation results for compact convergence) *Let $(E, \|\cdot\|_E)$ be an arbitrary Banach space and $(T_n)_n$ and T bounded linear operators on E with $T_n \xrightarrow{c} T$. Then:*

1. Upper semi-continuity: *Let $\tau \subset \sigma(T)$ be an isolated part of $\sigma(T)$ and $\lambda_n \in \sigma(T_n) \cap M$ a converging sequence with limit point λ. Then $\lambda \in \tau$.*
2. Lower semi-continuity: *Let $\lambda \in \sigma(T)$ be an isolated eigenvalue of T with finite algebraic multiplicity. Then there exists some neighborhood $M \subset \mathbb{C}$ of λ such that for large n, $\sigma(T_n) \cap M = \{\lambda_n\}$, and $(\lambda_n)_n$ converges to λ.*
3. Convergence of spectral projections: *Let $\lambda \in \sigma(T)$ an isolated eigenvalue with finite multiplicity and $\lambda_n \in \sigma(T_n)$ a sequence of isolated eigenvalues with finite multiplicity such that $\lambda_n \to \lambda$. Let Pr_n and Pr be the corresponding spectral projections. Then $\mathrm{Pr}_n \xrightarrow{p} \mathrm{Pr}$.*
4. Convergence of eigenvectors: *Under the conditions of Part (3), if λ is a simple eigenvalue, so are λ_n for n large enough. Then the corresponding eigenfunctions f_n converge up to a change of sign (i.e., there exists a sequence $(a_n)_n$ of signs $a_n \in \{-1, +1\}$ such that $a_n f_n$ converges).*

Proof. See Proposition 3.18. and Sections 3.6. and 5.1. in Chatelin (1983). ☺

To prove rates of convergence we will also need some quantitative perturbation theory results for spectral projections. The following theorem can be found in Atkinson (1967):

Theorem 7 (Atkinson) *Let $(E, \|\cdot\|_E)$ be an arbitrary Banach space, B its unit ball. Let $(K_n)_{n\in\mathbb{N}}$ and K be compact linear operators on E such that $K_n \xrightarrow{cc} K$. For a nonzero eigenvalue $\lambda \in \sigma(K)$ denote the corresponding spectral projection by $\Pr_\lambda$. Let $(\lambda_n)_n \in \sigma(K_n)$ a sequence of eigenvalues with $\lambda_n \to \lambda$, and $(\Pr_{\lambda_n})_n$ the corresponding spectral projections. Then there exists a constant $C > 0$ such that for every $x \in \Pr_\lambda E$*

$$\|x - \Pr_{\lambda_n} x\| \leq C\left(\|(K_n - K)x\| + \|x\|\|(K - K_n)K_n\|\right).$$

The constant C is independent of x, but it depends on λ and $\sigma(K)$.

Later we will need some basic tools from empirical process theory to prove certain uniform convergence statements. For a probability measure P and a function $f \in C(\mathcal{X})$ we introduce the abbreviation $Pf := \int f(x)dP(x)$. Let $(X_n)_n$ a sequence of iid random variables drawn according to P, and denote by $P_n := \sum_{i=1}^n \delta_{X_i}$ the corresponding empirical distributions. A set $\mathcal{F} \subset C(\mathcal{X})$ is called a Glivenko-Cantelli class if

$$\sup_{f\in\mathcal{F}} |Pf - P_n f| \to 0 \quad \text{a.s.}$$

Finally, the covering numbers $N(\mathcal{F}, \varepsilon, d)$ of a set totally bounded set $\mathcal{F}$ with metric d are defined as the smallest number n such that $\mathcal{F}$ can be covered with n balls of radius ε.

5 Convergence of normalized spectral clustering

In this section we present our results on the convergence of normalized spectral clustering. We start with an overview over our methods and tools, then proceed to prove several propositions, and finally state and prove our main theorems at the end of this section. The case of unnormalized spectral clustering will be treated in Section 7.

5.1 Overview over the methods

On a high level, the approach to prove convergence of spectral clustering is very similar in both the normalized and unnormalized case. In this section we focus on the normalized case. To study the convergence of spectral clustering we have to investigate whether the eigenvectors of the Laplacians constructed on n sample points "converge" for $n \to \infty$. For simplicity, let us discuss the case of the second eigenvector. For all $n \in \mathbb{N}$, let $v_n \in \mathbb{R}^n$ the second eigenvector of L'_n. The technical difficulty for proving convergence of $(v_n)_{n\in\mathbb{N}}$ is that for different sample sizes n, the vectors v_n live in different spaces. Thus standard notions of convergence cannot be applied. What we want to show instead is that there exists a function $f \in C(\mathcal{X})$ such that the difference between the eigenvector v_n and the restriction of f to the sample converges to 0, that is $\|v_n - \rho_n f\|_\infty \to 0$. Our approach to achieve this takes one more detour. We replace the vector v_n by a function $f_n \in C(\mathcal{X})$ such that $v_n = \rho_n f_n$. This function f_n will be the second eigenfunction of an operator U'_n acting on the space $C(\mathcal{X})$. Then we use the fact that

$$\|v_n - \rho_n f\|_\infty = \|\rho_n f_n - \rho_n f\|_\infty \leq \|f_n - f\|_\infty.$$

Hence, it will be enough to show that $\|f_n - f\|_\infty \to 0$. As the sequence f_n will be random, this convergence will hold almost surely.

Step 1: Relating the matrices L'_n to linear operators U'_n on $C(\mathcal{X})$. First we will construct a family $(U'_n)_{n\in\mathbb{N}}$ of linear operators on $C(\mathcal{X})$ which, if restricted to the sample, "behaves' as $(L'_n)_{n\in\mathbb{N}}$: for all $f \in C(\mathcal{X})$ we will have the relation $\rho_n U'_n f = L'_n \rho_n f$. In the following we will then study the convergence of $(U'_n)_n$ on $C(\mathcal{X})$ instead of the convergence of $(L'_n)_n$.

Step 2: Relation between $\sigma(L'_n)$ and $\sigma(U'_n)$. In Step 1 we replaced the *operators* L'_n by operators U'_n on $C(\mathcal{X})$. But as we are interested in the *eigenvectors* of L'_n we have to check whether they can actually be recovered from the eigenfunctions of U'_n. By elementary linear algebra we can prove that the "interesting" eigenfunctions f_n and eigenvectors v_n of U'_n and L'_n are in a one-to-one relationship and can be computed from each other by the relation $v_n = \rho_n f_n$. As a consequence, if the eigenfunctions f_n of U'_n converge, the same can be concluded for the eigenvectors of L'_n.

Step 3: Convergence of $U'_n \to U'$. In this step we want to prove that certain eigenvalues and eigenfunctions of U'_n converge to the corresponding quantities of some limit operator U'. For this we will have to establish

a rather strong type of convergence of linear operators. Pointwise convergence is in general too weak for this purpose; on the other hand it will turn out that operator norm convergence does not hold in our context. The type of convergence we will consider is compact convergence, which is between pointwise convergence and operator norm convergence and is just strong enough for proving convergence of spectral properties. The notion of compact convergence has originally been developed in the context of (deterministic) numerical approximation of integral operators. We adapt those methods to a framework where the spectrum of a linear operator U' is approximated by the spectra of *random* operators U_n'. Here, a key element is the fact that certain classes of functions are Glivenko-Cantelli classes: the integrals over the functions in those classes can be approximated uniformly by empirical integrals based on the random sample.

Step 4: Assembling all pieces. Now the main work has been done and we just have to put together the different pieces. In Step 1 we replaced the operators L_n' by the operators U_n', which does not affect the spectra according to Step 2. The operators U_n' are shown in Step 3 to converge compactly to some limit operator U'. Under certain conditions on the spectrum of U' this leads to the convergence of the first eigenfunctions of the spectra of U_n', which implies the convergence of spectral clustering.

Now we will implement the program outlined above in detail.

5.2 Step 1: Construction of the operators on $C(\mathcal{X})$

We define the following functions and operators, which are all supposed to act on $C(\mathcal{X})$: The degree functions

$$d_n(x) := \int k(x,y) dP_n(y) \in C(\mathcal{X})$$
$$d(x) := \int k(x,y) dP(y) \in C(\mathcal{X})$$

the multiplication operators,

$$M_{d_n} : C(\mathcal{X}) \to C(\mathcal{X}),\ M_{d_n} f(x) := d_n(x) f(x)$$
$$M_d : C(\mathcal{X}) \to C(\mathcal{X}),\ M_d f(x) := d(x) f(x)$$

integral operators

$$S_n : C(\mathcal{X}) \to C(\mathcal{X}),\ S_n f(x) := \int k(x,y) f(y) dP_n(y)$$
$$S : C(\mathcal{X}) \to C(\mathcal{X}),\ Sf(x) := \int k(x,y) f(y) dP(y)$$

and the corresponding differences

$$U_n : C(\mathcal{X}) \to C(\mathcal{X}),\ U_n f(x) := M_{d_n} f(x) - S_n f(x)$$
$$U : C(\mathcal{X}) \to C(\mathcal{X}),\ Uf(x) := M_d f(x) - Sf(x)$$

The operators U_n and U will be used to deal with the case of unnormalized spectral clustering. For the normalized case we introduce the normalized similarity functions,

$$h_n(x,y) := k(x,y)/\sqrt{d_n(x) d_n(y)}$$
$$h(x,y) := k(x,y)/\sqrt{d(x) d(y)}$$

integral operators

$$T_n : C(\mathcal{X}) \to C(\mathcal{X}),\ T_n f(x) = \int h(x,y) f(y) dP_n(y)$$
$$T_n' : C(\mathcal{X}) \to C(\mathcal{X}),\ T_n' f(x) = \int h_n(x,y) f(y) dP_n(y)$$
$$T : C(\mathcal{X}) \to C(\mathcal{X}),\ Tf(x) = \int h(x,y) f(y) dP(y)$$

and the differences

$$U'_n := \mathrm{I} - T'_n$$
$$U' := \mathrm{I} - T$$

We summarize the properties of those operators in the following proposition. Recall the definition of the restriction operator ρ_n of Section 4.

Proposition 8 (Relations between the operators) *Under the general assumptions, the functions d_n and d are continuous, bounded from below by the constant $l > 0$, and from above by $\|k\|_\infty$. All operators defined above are bounded, and the integral operators are compact. The operator norms of M_{d_n}, M_d, S_n, and S are bounded by $\|k\|_\infty$, the ones of T'_n, T_n, and T by $\|k\|_\infty / l$. Moreover, we have the following relations:*

$$\frac{1}{n} D_n \circ \rho_n = \rho_n \circ M_{d_n} \qquad \frac{1}{n} K_n \circ \rho_n = \rho_n \circ S_n \qquad \frac{1}{n} L_n \circ \rho_n = \rho_n \circ U_n \qquad L'_n \circ \rho_n = \rho_n \circ U'_n$$

Proof. All statements follow directly from the definitions and the general assumptions. Note that in the case of the unnormalized Laplacian L_n we get the scaling factor $1/n$ from the $1/n$-factor hidden in the empirical distribution P_n. In the case of normalized Laplacian, this scaling factor cancels with the scaling factors of the degree functions in the denominators. ☺

The main statement of this proposition is that if restricted to the sample points, U_n "behaves as" $\frac{1}{n} L_n$ and U'_n as L'_n. Moreover, by the law of large numbers it is clear that for fixed $f \in C(\mathcal{X})$ and $x \in \mathcal{X}$ the empirical quantities converge to the corresponding true quantities, in particular $U_n f(x) \to U f(x)$ and $U'_n f(x) \to U' f(x)$. Proving stronger convergence statements for U_n and U'_n will be the main part of Step 3.

5.3 Step 2: Relations between the spectra

The following proposition establishes the connections between the spectra of L'_n and U'_n. We show that that U'_n and L'_n have more or less the same spectrum and that the eigenfunctions f of U'_n and eigenvectors v of L'_n correspond to each other by the relation $v = \rho_n f$.

Proposition 9 (Spectrum of U'_n)

1. *If $f \in C(\mathcal{X})$ is an eigenfunction of U'_n with the eigenvalue λ, then the vector $v = \rho_n f \in \mathbb{R}^n$ is an eigenvector of the matrix L'_n with eigenvalue λ.*
2. *Let $\lambda \neq 1$ be an eigenvalue of U'_n with eigenfunction $f \in C(\mathcal{X})$, and $v := (v_1, ..., v_n) := \rho_n f \in \mathbb{R}^n$. Then f is of the form*
$$f(x) = \frac{\frac{1}{n} \sum_j k(x, X_j) v_j}{1 - \lambda}. \tag{5}$$
3. *If v is an eigenvector of the matrix L'_n with eigenvalue $\lambda \neq 1$, then f defined by equation* (5) *is an eigenfunction of U'_n with eigenvalue λ.*
4. *The spectrum of U'_n consists of at most countably many non-negative eigenvalues with finite multiplicity. The essential spectrum consists of at most one point, namely $\sigma_{ess}(U'_n) = \{1\}$. This is also the only possible accumulation point of $\sigma(U'_n)$. The same statement is true for U'.*

Proof. Part (1): Follows directly from Proposition 8.
Part (2): Follows directly from solving the eigenvalue equation.
Part (3): Define f as in Equation (5). It is well-defined because v is an eigenvector of $\frac{1}{n} L_n$, and f is an eigenfunction of U_n with eigenvalue λ.
Part (4): As T'_n is a compact integral operator according to Proposition 8, its spectrum consists of at most countably many eigenvalues with finite multiplicity, and the only accumulation point is 0. The set $\{0\}$ is the essential spectrum of T'_n. The spectrum $\sigma(U'_n)$ of $U'_n = \mathrm{I} - T'_n$ is given by $1 - \sigma(T'_n)$. The non-negativity of the eigenvalues follows from the non-negativity of the eigenvalues of L'_n and Parts (1)-(3) of the proposition. The analogous statements also hold for U'. ☺

This proposition establishes a one-to-one correspondence between the eigenvalues and eigenvectors of L'_n and U'_n, provided they satisfy $\lambda \neq 1$. The condition $\lambda \neq 1$ needed to ensure that the denominator of Equation (6) does not vanish. As a side remark, note that the set $\{1\}$ is the essential spectrum of U'_n. Thus the condition $\lambda \neq 1$ can also be written as $\lambda \notin \sigma_{\text{ess}}(U'_n)$, which will be analogous to the condition on the eigenvalues in the unnormalized case. This condition ensures that λ is isolated in the spectrum.

5.4 Step 3: Compact convergence

In this section we want to prove that the sequence of random operators U'_n converges compactly to U' almost surely. First we will prove pointwise convergence. Note that on the space $C(\mathcal{X})$, the pointwise convergence of a sequence U'_n of operators is defined as $\|U'_n f - U' f\|_\infty \to 0$, that is for each $f \in C(\mathcal{X})$, the sequence $(U'_n f)_n$ has to converge uniformly over $\mathcal{X}$. To establish this convergence we will need to show that several classes of functions are "not too large", that is they are Glivenko-Cantelli classes. For convenience we introduce the following sets of functions:

$$\mathcal{K} := \{k(x,\cdot);\ x \in \mathcal{X}\} \text{ (where } k \text{ is the given similarity function)}$$
$$\mathcal{H} := \{h(x,\cdot);\ x \in \mathcal{X}\} \text{ (where } h \text{ is the normalized similarity function as defined above)}$$
$$g \cdot \mathcal{H} := \{g(\cdot) \cdot h(x,\cdot);\ x \in \mathcal{X}\} \quad \text{(for some } g \in C(\mathcal{X}))$$
$$\mathcal{H} \cdot \mathcal{H} := \{h(x,\cdot)h(y,\cdot);\ x, y \in \mathcal{X}\}$$

Proposition 10 (Glivenko-Cantelli classes) *Under the general assumptions, the classes $\mathcal{K}$, $\mathcal{H}$, and $g \cdot \mathcal{H}$ (for arbitrary $g \in C(\mathcal{X})$) are Glivenko-Cantelli classes.*

Proof. As k is a continuous function defined on a compact domain, it is uniformly continuous. In this case it is easy to construct, for each $\varepsilon > 0$, a finite ε-cover with respect to $\|\cdot\|_\infty$ of $\mathcal{K}$ from a finite δ-cover of $\mathcal{X}$. Hence $\mathcal{K}$ has finite $\|\cdot\|_\infty$-covering numbers. Then it is easy to see that $\mathcal{K}$ also has finite $\|\cdot\|_{L_1(P)}$-bracketing numbers (cf. van der Vaart and Wellner, 1996, p. 84). Now the statement about the class $\mathcal{K}$ follows from Theorem 2.4.1. of van der Vaart and Wellner (1996). The statements about the classes $\mathcal{H}$ and $g \cdot \mathcal{H}$ can be proved in the same way, hereby observing that h is continuous and bounded as a consequence of the general assumptions. ☺

Note that it is a direct consequence of this proposition that the empirical degree function d_n converges uniformly to the true degree function d, that is

$$\|d_n - d\|_\infty = \sup_{x \in \mathcal{X}} |d_n(x) - d(x)| = \sup_{x \in \mathcal{X}} |P_n k(x,\cdot) - Pk(x,\cdot)| \to 0 \text{ a.s.}$$

Now we can establish the convergence of the integral operators T'_n:

Proposition 11 (T'_n converges pointwise to T a.s.) *Under the general assumptions, $T'_n \xrightarrow{p} T$ almost surely.*

Proof. For arbitrary $f \in C(\mathcal{X})$ we have

$$\|T'_n f - Tf\|_\infty \leq \|T'_n f - T_n f\|_\infty + \|T_n f - Tf\|_\infty.$$

The second term can be written as

$$\|T_n f - Tf\|_\infty = \sup_{x \in \mathcal{X}} |P_n(h(x,\cdot)f(\cdot)) - P(h(x,\cdot)f(\cdot))| = \sup_{g \in f \cdot \mathcal{H}} |P_n g - Pg|$$

which converges to 0 a.s. by Proposition 10. It remains to prove the almost sure convergence of the first term (which we prove in a slightly more technical way than necessary because it can then be used in Section 6):

$$\begin{aligned}
\|T_n f - T'_n f\|_\infty &\leq \|f\|_\infty \sup_{x \in \mathcal{X}} \int |h(x,y) - h_n(x,y)| dP_n(y) \\
&\leq \|f\|_\infty \|k\|_\infty \sup_{x,y \in \mathcal{X}} \left| \frac{1}{\sqrt{d_n(x) d_n(y)}} - \frac{1}{\sqrt{d(x) d(y)}} \right| \\
&\leq \|f\|_\infty \frac{\|k\|_\infty}{l^2} \sup_{x,y \in \mathcal{X}} \left| \sqrt{d_n(x) d_n(y)} - \sqrt{d(x) d(y)} \right|
\end{aligned}$$

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

To bound the last expression we use

$$\begin{aligned}\sup_{x,y\in\mathcal{X}} |d_n(x)d_n(y) - d(x)d(y)| &\leq \sup_{x,y\in\mathcal{X}} |d_n(x)d_n(y) - d_n(x)d(y)| + |d_n(x)d(y) - d(x)d(y)| \\ &\leq \sup_{x,y\in\mathcal{X}} |d_n(x)|\,|d_n(y) - d(y)| + |d(y)|\,|d_n(x) - d(x)| \\ &\leq 2\|k\|_\infty \sup_x |d_n(x) - d(x)| \\ &= 2\|k\|_\infty \sup_{g\in\mathcal{K}} |P_n g - Pg|\end{aligned}$$

Together this leads to

$$\|T_n' f - T_n' f\|_\infty \leq \|f\|_\infty \frac{\|k\|_\infty^2}{l^3} \sup_{g\in\mathcal{K}} |P_n g - Pg|$$

which converges to 0 a.s. by Proposition 10. ☺

Proposition 12 (T_n' converges collectively compactly to T a.s.) *Under the general assumptions, $T_n' \overset{cc}{\to} T$ almost surely.*

Proof. We have already seen the pointwise convergence $T_n' \overset{p}{\to} T$ in Proposition 11. Next we have to prove that for some $N \in \mathbb{N}$, the sequence of operators $(T_n' - T)_{n>N}$ is collectively compact a.s. As T is compact itself, it is enough to show that $(T_n')_{n>N}$ is collectively compact a.s. This will be done using the Arzela-Ascoli theorem. First we fix the random sequence $(X_n)_n$ and hence the random operators $(T_n')_n$. By Proposition 8 we know that $\|T_n'\| \leq \|k\|_\infty / l$ for all $n \in \mathbb{N}$. Hence, the functions in $\bigcup_n T_n' B$ are uniformly bounded by $\sup_{n\in\mathbb{N}, f\in B} \|T_n' f\|_\infty \leq \|k\|_\infty / l$. To prove that the functions in $\bigcup_{n>N} T_n' B$ are equicontinuous we have to bound the expression $|g(x) - g(x')|$ in terms of the distance between x and x', uniformly in $g \in \bigcup_n T_n' B$. For fixed sequence $(X_n)_{n\in\mathbb{N}}$ and all $n \in \mathbb{N}$ we have that for all $x, x' \in \mathcal{X}$,

$$\sup_{f\in B, n\in\mathbb{N}} |T_n' f(x) - T_n' f(x')| = \sup_{f\in B, n\in\mathbb{N}} \left| \int (h_n(x,y) - h_n(x',y)) f(y) dP_n(y) \right|$$
$$\leq \sup_{f\in B, n\in\mathbb{N}} \|f\|_\infty \int |h_n(x,y) - h_n(x',y)|\, dP_n(y) \leq \|h_n(x,\cdot) - h_n(x',\cdot)\|_\infty.$$

Now we have to prove that the right hand side gets small whenever the distance between x and x' gets small.

$$\begin{aligned}\sup_y |h_n(x,y) - h_n(x',y)| &= \sup_y \left| \frac{k(x,y)\sqrt{d_n(x')} - k(x',y)\sqrt{d_n(x)}}{\sqrt{d_n(x)d_n(x')d_n(y)}} \right| \\ &\leq \frac{1}{l^{3/2}} \sup_y \Big(|k(x,y)\sqrt{d_n(x')} - k(x',y)\sqrt{d_n(x')}| + \\ &\qquad + |k(x',y) + \sqrt{d_n(x')} - k(x',y)\sqrt{d_n(x)}| \Big) \\ &\leq \frac{1}{l^{3/2}} \left(\|\sqrt{d_n}\|_\infty \|k(x,\cdot) - k(x',\cdot)\|_\infty + \|k\|_\infty |\sqrt{d_n(x)} - \sqrt{d_n(x')}| \right) \\ &\leq \frac{1}{l^{3/2}} \left(\|k\|_\infty^{1/2} \|k(x,\cdot) - k(x',\cdot)\|_\infty + \frac{\|k\|_\infty}{2l^{1/2}} |d_n(x) - d_n(x')| \right) \\ &\leq \frac{\|k\|_\infty^{1/2}}{l^{3/2}} \|k(x,\cdot) - k(x',\cdot)\|_\infty +\end{aligned}$$

$$\frac{\|k\|_\infty}{2l^2}\Big(|d_n(x)-d(x)|+|d(x)-d(x')|+|d(x')-d_n(x')|\Big)$$
$$\leq C_1\|k(x,\cdot)-k(x',\cdot)\|_\infty + C_2|d(x)-d(x')| + C_3\|d_n-d\|_\infty$$

As $\mathcal{X}$ is a compact space, the continuous functions k (on the compact space $\mathcal{X}\times\mathcal{X}$) and d are in fact uniformly continuous. Thus, the first two (deterministic) terms $\|k(x,\cdot)-k(x',\cdot)\|_\infty$ and $|d(x)-d(x')|$ can be made arbitrarily small for all x, x' whenever the distance between x and x' is small. For the third term $\|d_n-d\|_\infty$, which is a random term, we know by the Glivenko-Cantelli properties of Proposition 10 that it converges to 0 a.s. This means that for each given $\varepsilon>0$ there exists some $N\in\mathbb{N}$ such that for all $n>N$ we have $\|d_n-d\|_\infty\leq\varepsilon$ a.s. (in particular, N can be chosen simultaneously for all *random* operators T_n'). Together, these arguments show that $\bigcup_{n>N} T_n'B$ is equicontinuous a.s. By the Arzela-Ascoli theorem we then know that $\bigcup_{n>N} T_n'B$ is relatively compact a.s., which concludes the proof. ☺

Proposition 13 (U_n' converges compactly to U' a.s.) *Under the general assumptions, $U_n' \xrightarrow{c} U'$ a.s.*

Proof. This follows directly from the facts that collectively compact convergence implies compact convergence, the definitions of U_n' to U', and Proposition 12. ☺

5.5 Step 4: Convergence of normalized spectral clustering

Now we have collected all ingredients to state and prove our convergence result for normalized spectral clustering. The following theorem is the precisely formulated version of the informal Result 1 of the introduction:

Theorem 14 (Convergence of normalized spectral clustering) *Assume that the general assumptions hold. Let $\lambda\neq 1$ be an eigenvalue of U'. Then there exists some $N\in\mathbb{N}$ and some neighborhood $M\subset\mathbb{C}$ of λ such that for $n>N$, $\sigma(L_n')\cap M=\{\lambda_n\}$, and $(\lambda_n)_n$ converges to λ a.s. Moreover, let $\Pr_n'\colon C(\mathcal{X})\to C(\mathcal{X})$ be the spectral projection corresponding to λ_n, and $\Pr$ the one corresponding to $\lambda\in\sigma(U')$. Then $\Pr_n'\xrightarrow{p}\Pr$ a.s. If λ is a simple eigenvalue, then also the eigenvectors converge a.s. up to a change of sign: if v_n is the eigenvector of L_n' with eigenvalue λ_n, $v_{n,i}$ its i-th coordinate, and f the eigenfunction of eigenvalue λ of U', then there exists a sequence $(a_n)_{n\in\mathbb{N}}$ with $a_i\in\{+1,-1\}$ such that $\sup_{i=1,\dots,n}|a_nv_{n,i}-f(X_i)|\to 0$ a.s. In particular, for all $b\in\mathbb{R}$, the sets $\{a_nf_n>b\}$ and $\{f>b\}$ converge, that is their symmetric difference satisfies $P(\{f>b\}\triangle\{a_nf_n>b\})\to 0$.*

Proof. In Proposition 9 we established a one-to-one correspondence between the eigenvalues $\lambda\neq 1$ of L_n' and U_n', and we saw that the eigenvalues λ of U' with $\lambda\neq 1$ are isolated and have finite multiplicity. In Proposition 13 we proved the compact convergence of U_n' to U', which according to Proposition 6 implies the convergence of the spectral projections of isolated eigenvalues with finite multiplicity. For simple eigenvalues, this implies the convergence of the eigenvectors up to a change of sign. The convergence of the sets $\{f_n>b\}$ is a simple consequence of the almost sure convergence of $(a_nf_n)_n$. ☺

6 Rates of convergence of normalized spectral clustering

In this section we want to prove statements about the rates of convergence of normalized spectral clustering. Our main result is the following:

Theorem 15 (Rate of convergence of normalized spectral clustering) *Under the general assumptions, let $\lambda\neq 0$ be a simple eigenvalue of T with eigenfunction u, $(\lambda_n)_n$ a sequence of eigenvalues of T_n' such that $\lambda_n\to\lambda$, and $(u_n)_n$ a corresponding sequence of eigenfunctions. Define $\mathcal{F}=\mathcal{K}\cup u\cdot\mathcal{H}\cup\mathcal{H}\cdot\mathcal{H}$. Then there exists a constant $C'>0$ (which depends on the similarity function k, the spectrum $\sigma(T)$, and the eigenvalue λ) and a sequence $(a_n)_n$ of signs $a_n\in\{+1,-1\}$ such that*

$$\|a_nu_n-u\|_\infty\leq C'\sup_{f\in\mathcal{F}}|P_nf-Pf|.$$

Hence, the speed of convergence of the eigenfunctions is controlled by the speed of convergence of $\sup_{f\in\mathcal{F}}|P_nf-Pf|$.

This theorem shows that the rate of convergence of normalized spectral clustering is at least as good as the rate of convergence of the supremum of the empirical process indexed by $\mathcal{F}$. To determine the latter there exist a variety of tools and techniques from the theory of empirical processes such as covering numbers, VC dimension, Rademacher complexities, see for example van der Vaart and Wellner (1996), Dudley (1999), Mendelson (2003), Pollard (1984). In particular it is the case that "the nicer" the kernel function k is (e.g., k is Lipschitz, or smooth, or positive definite), the faster the rate of convergence on the right hand side will be. As an example we will consider the case of the Gaussian similarity function $k(x,y) = \exp(-\|x-y\|^2/\sigma^2)$, which is widely used in practical applications of spectral clustering.

Example 1 (Rate of convergence for Gaussian kernel) *Let $\mathcal{X}$ be compact subset of $\mathbb{R}^d$ and $k(x,y) = \exp(-\|x-y\|^2/\sigma^2)$. Then the eigenvectors in Theorem 15 converge with rate $\mathcal{O}(1/\sqrt{n})$.*

For the case of unnormalized spectral clustering it is possible to obtain similar results on the speed of convergence, for example by using Proposition 5.3. in Chapter 5 of Chatelin (1983) instead of the results of Atkinson (1967) (note that in the unnormalized case, the assumptions of Theorem 7 are not satisfied as we only have compact convergence instead of collectively compact convergence). As we recommend to use normalized rather than unnormalized spectral clustering anyway we we do not discuss this issue any further. The remaining part of this section is devoted to the proofs of Theorem 15 and Example 1.

6.1 Some technical preparations

Before we can prove Theorem 15 we need to show several technical propositions. With the background from the previous sections, all of them are rather straight forward but a bit lengthy to prove.

Proposition 16 (Bound on $\|T_n' - T_n\|$) *Assume that the general conditions are satisfied. Then:*

$$\|T_n' - T_n\| \leq \frac{\|k\|_\infty^2}{l^3} \sup_{f \in \mathcal{K}} |P_n f - Pf|.$$

Proof. In the proof of Proposition 11 we have already seen that for every $f \in C(\mathcal{X})$,

$$\|T_n f - T_n' f\|_\infty \leq \|f\|_\infty \frac{\|k\|_\infty^2}{l^3} \sup_{g \in \mathcal{K}} |P_n g - Pg|.$$

This also proves Proposition 16. ☺

Proposition 17 (Bound on $\|(T_n - T)g\|_\infty$) *Assume that the general conditions are satisfied. Then for every $g \in C(\mathcal{X})$ we have*

$$\|(T_n - T)g\|_\infty \leq \sup_{f \in g \cdot \mathcal{H}} |P_n f - Pf|.$$

Proof. This proposition follows directly from the definitions:

$$\|(T_n - T)g\|_\infty = \sup_{x \in \mathcal{X}} |\int h(x,y)g(y)dP_n(y) - \int h(x,y)g(y)dP(y)| = \sup_{f \in g \cdot \mathcal{H}} |P_n f - Pf|.$$

☺

Proposition 18 (Bound on $\|(T - T_n)T_n\|$) *Assume that the general conditions are satisfied. Then:*

$$\|(T - T_n)T_n\| \leq \sup_{f \in \mathcal{H} \cdot \mathcal{H}} |P_n f - Pf|.$$

Proof. By Fubini's theorem and the symmetry of h,

$$\begin{aligned}
\|(T-T_n)T_n\| &= \sup_{\|f\|_\infty\le 1}\sup_{x\in\mathcal{X}}|TT_nf(x)-T_nT_nf(x)|\\
&= \sup_{\|f\|_\infty\le 1}\sup_{x\in\mathcal{X}}|\int h(x,z)\int h(z,y)f(y)dP_n(y)dP(z)-\int h(x,z)\int h(z,y)f(y)dP_n(y)dP_n(z)|\\
&= \sup_{\|f\|_\infty\le 1}\sup_{x\in\mathcal{X}}|\int f(y)\left(\int h(x,z)h(z,y)dP(z)-\int h(x,z)h(z,y)dP_n(z)\right)dP_n(y)|\\
&\le \sup_{x,y\in\mathcal{X}}|\int h(x,z)h(z,y)dP(z)-\int h(x,z)h(z,y)dP_n(z)|\\
&= \sup_{f\in\mathcal{H}\cdot\mathcal{H}}|P_nf-Pf|.
\end{aligned}$$

☺

Proposition 19 (Convergence of one-dimensional projections) *Let $(v_n)_n$ be a sequence of vectors in some Banach space $(E,\|\cdot\|)$ with $\|v_n\|=1$, Pr_n the projections on the one-dimensional subspace spanned by v_n, and $v\in E$ with $\|v\|=1$. Then there exists a sequence $(a_n)_n\in\{+1,-1\}$ of signs such that*

$$\|a_nv_n-v\|\le 2\|v-\mathrm{Pr}_n\,v\|.$$

In particular, if $\|v-\mathrm{Pr}_n\,v\|\to 0$ then v_n converges to v up to a change of sign.

Proof. By the definition of Pr_n we know that $\mathrm{Pr}_n\,v=c_nv_n$ for some $c_n\in\mathbb{R}$. Define $a_n:=\mathrm{sgn}(c_n)$. Then

$$|a_n-c_n|=|1-|c_n|\,|=|\,\|v\|-|c_n|\cdot\|v_n\|\,|\le\|v-c_nv_n\|=\|v-\mathrm{Pr}_n\,v\|.$$

From this we can conclude that

$$\|v-a_nv_n\|\le\|v-c_nv_n\|+\|c_nv_n-a_nv_n\|=\|v-c_nv_n\|+|c_n-a_n|\cdot\|v_n\|\le 2\|v-\mathrm{Pr}_n\,v\|.$$

☺

6.2 Proof of Theorem 15

First we fix a realization of the random variables $(X_n)_n$. From the convergence of the spectral projections in Theorem 14 we know that if $\lambda\in\sigma(T)$ is simple, so are $\lambda_n\in\sigma(T_n')$ for large n. Then the eigenfunctions u_n are uniquely determined up to a change of orientation. In Proposition 19 we have seen that the speed of convergence of u_n to u coincides with the speed of convergence of the expression $\|u-\mathrm{Pr}_n\,u\|$ from Theorem 7. As we already know by Section 5, the operators T_n' and T satisfy the assumptions in Theorem 7. Accordingly, $\|u-\mathrm{Pr}_n\,u\|$ can be bounded by the two terms $\|(T_n'-T)u\|$ and $\|(T-T_n')T_n'\|$. It will turn out that both terms are easier to bound if we can replace the operator T_n' by T_n. To accomplish this observe that

$$\begin{aligned}
\|(T-T_n')T_n'\| &\le \|TT_n'-TT_n\|+\|TT_n-T_nT_n\|+\|T_nT_n-T_n'T_n'\|\\
&\le \|T\|\|T_n-T_n'\|+\|(T-T_n)T_n\|+\|T_nT_n-T_nT_n'\|+\|T_nT_n'-T_n'T_n'\|\\
&\le (\|T\|+\|T_n\|+\|T_n'\|)\|T_n-T_n'\|+\|(T-T_n)T_n\|\\
&\le 3\frac{\|k\|_\infty}{l}\|T_n-T_n'\|+\|(T-T_n)T_n\|
\end{aligned}$$

and also

$$\|(T_n'-T)u\|_\infty\le\|(T_n'-T_n)u\|_\infty+\|(T_n-T)u\|_\infty\le\|u\|_\infty\|T_n'-T_n\|+\|(T_n-T)u\|_\infty.$$

At this point, note that T_n does not converge to T in operator norm (cf. page 197 in Section 4.7.4. of Chatelin, 1983). Thus it does not make sense to bound $\|(T_n-T)u\|_\infty$ by $\|T_n-T\|\|u\|_\infty$ or $\|(T-T_n)T_n\|$ by $\|T-T_n\|\|T_n\|$.

Assembling all inequalities, applying Proposition 19 and Theorem 7, and choosing the signs a_n as in the proof of Proposition 19 we obtain

$$\begin{aligned}\|a_n u_n - u\| &\leq 2\|u - \Pr_{\lambda_n} u\| \\ &\leq 2C\left(\|(T_n' - T)u\| + \|(T - T_n')T_n'\|\right) \\ &\leq 2C\left((\frac{3\|k\|_\infty}{l} + 1)\|T_n - T_n'\| + \|(T_n - T)u\|_\infty + \|(T - T_n)T_n\|\right) =: (*)\end{aligned}$$

To bound $(*)$ we now apply Propositions 16, 17, 18 and in the last step merge all occurring constants to one larger constant C' to obtain

$$\begin{aligned}(*) &\leq 2C\left((\frac{3\|k\|_\infty}{l} + 1)\frac{\|k\|_\infty^2}{l^3}\sup_{f\in\mathcal{K}}|P_n f - Pf| + \sup_{f\in u\cdot\mathcal{H}}|P_n f - Pf| + \sup_{f\in\mathcal{H}\cdot\mathcal{H}}|P_n f - Pf|\right) \\ &\leq C'\sup_{f\in\mathcal{K}\cup u\cdot\mathcal{H}\cup\mathcal{H}\cdot\mathcal{H}}|P_n f - Pf|.\end{aligned}$$

As all arguments hold for each fixed realization $(X_n)_n$ of the sample points, they also hold for the random variables themselves almost surely. This concludes the proof of Theorem 15. ☺

6.3 Rate of convergence for the Gaussian kernel

In this subsection we want to prove the convergence rate $\mathcal{O}(1/\sqrt{n})$ stated in Example 1 for the case of a Gaussian kernel function $k(x, y) = \exp(-\|x-y\|^2/\sigma^2)$. In principle, there are many ways to compute rates of convergence for terms of the form $\sup_f |Pf - P_n f|$ (see for example van der Vaart and Wellner, 1996). As discussing those methods is not the main focus of our paper we choose a rather simple covering number approach which suffices for our purposes. We will use the following theorem, which is well known in the empirical process theory (nevertheless we did not find a good reference for it; it can be obtained for example by combining Section 3.4. of Anthony, 2002, and Theorem 2.34 in Mendelson, 2003):

Theorem 20 (Entropy bound) *Let $(\mathcal{X}, \mathcal{A}, P)$ be an arbitrary probability space, $\mathcal{F}$ a class of real-valued functions on $\mathcal{X}$ with $\|f\|_\infty \leq 1$. Let $(X_n)_{n\in\mathbb{N}}$ a sequence of iid random variables drawn according to P, and $(P_n)_{n\in\mathbb{N}}$ the corresponding empirical distributions. Then there exists some constant $c > 0$ such that for all $n \in \mathbb{N}$ with probability at least $1 - \delta$,*

$$\sup_{f\in\mathcal{F}}|P_n f - Pf| \leq \frac{c}{\sqrt{n}}\int_0^\infty \sqrt{\log N(\mathcal{F}, \varepsilon, L_2(P_n))}\, d\varepsilon \quad + \quad \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}.$$

We can see that if $\int_0^\infty \sqrt{\log N(\mathcal{F}, \varepsilon, L_2(P_n))}\ d\varepsilon < \infty$, then the whole expression scales as $\mathcal{O}(1/\sqrt{n})$. As a first step we would like to evaluate this integral for the function class $\mathcal{F} := \mathcal{K}$. As $L_2(P_n)$-covering numbers are difficult to estimate, we replace the $L_2(P_n)$-covering numbers in the integral by the $\|\cdot\|_\infty$-covering numbers. This is valid because we always have $N(\mathcal{K}, \varepsilon, L_2(P_n)) \leq N(\mathcal{K}, \varepsilon, \|\cdot\|_\infty)$. Moreover, we can replace the upper limit ∞ in the integral by 2. The reason is that for the Gaussian kernel k, all functions in $\mathcal{K}$ satisfy $\|k(x, \cdot)\|_\infty \leq 1$ and consequently, $\log N(\mathcal{K}, \varepsilon, \|\cdot\|_\infty) = 0$ for all $\varepsilon \geq 2$. Finally, in case of the Gaussian kernel, tight bounds for the $\|\cdot\|_\infty$-covering numbers of $\mathcal{K}$ have been obtained for example in Zhou (2002). There it was proved that for $\varepsilon < c_0$ for a certain constant $c_0 > 0$ only depending to the kernel width σ, the covering numbers satisfy

$$\log N(\mathcal{K}, \varepsilon, \|\cdot\|_\infty) \leq 32(\log\frac{1}{\varepsilon})^2.$$

(to see this we chose $R = 1$ and $n = 1$ in Proposition 1 of Zhou, 2002). Plugging the covering numbers in our integral we get

$$\begin{aligned}\int_0^\infty \sqrt{\log N(\mathcal{K}, \varepsilon, L_2(P_n))}\, d\varepsilon &\leq \int_0^2 \sqrt{\log N(\mathcal{K}, \varepsilon, \|\cdot\|_\infty)}\, d\varepsilon \\ &\leq \sqrt{32}\int_0^{c_0}\log\frac{1}{\varepsilon}\, d\varepsilon + \int_{c_0}^2 \sqrt{\log N(\mathcal{K}, \varepsilon, \|\cdot\|_\infty)}\, d\varepsilon \\ &\leq \sqrt{32}c_0(1 - \log c_0) + (2 - c_0)\sqrt{\log N(\mathcal{K}, c_0, \|\cdot\|_\infty)} < \infty.\end{aligned}$$

According to Theorem 15, we have to use the entropy bound not only for the function class $\mathcal{F} = \mathcal{K}$, but for the class $\mathcal{F} = \mathcal{K} \cup u \cdot \mathcal{H} \cup \mathcal{H} \cdot \mathcal{H}$. To this end we will bound the $\|\cdot\|_\infty$-covering numbers of $\mathcal{K} \cup u \cdot \mathcal{H} \cup \mathcal{H} \cdot \mathcal{H}$ in terms of the covering numbers of $\mathcal{K}$.

Proposition 21 (Covering numbers of $\mathcal{H}$) *The covering numbers of $\mathcal{H}$ satisfy*

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

where $s = \frac{\|k\|_\infty + 2\sqrt{l\|k\|_\infty}}{2l^2}$.

Proof. The main work consists in bounding the norm between two functions in $\mathcal{H}$ by the one of the corresponding functions in $\mathcal{K}$. This is not difficult but lengthy.

$$\begin{aligned}
&\|h(x,\cdot) - h(y,\cdot)\|_\infty = \sup_{z\in\mathcal{X}} |\frac{k(x,z)}{\sqrt{d(x)d(z)}} - \frac{k(y,z)}{\sqrt{d(y)d(z)}}| \\
&= l^{-3/2}(\sup_{z\in\mathcal{X}} |\sqrt{d(y)}k(x,z) - \sqrt{d(x)}k(y,z)|) \\
&\leq l^{-3/2}(\sup_{z\in\mathcal{X}} |\sqrt{d(y)}k(x,z) - \sqrt{d(x)}k(x,z)| + |\sqrt{d(x)}k(x,z) - \sqrt{d(x)}k(y,z)|) \\
&\leq l^{-3/2}(\|k\|_\infty |\sqrt{d(y)} - \sqrt{d(x)}| + \|k\|_\infty^{1/2} \sup_{z\in\mathcal{X}} |k(x,z) - k(y,z)|) \\
&\leq l^{-3/2}(\|k\|_\infty \frac{|d(y) - d(x)|}{\sqrt{d(y)} + \sqrt{d(x)}} + \|k\|_\infty^{1/2} \sup_{z\in\mathcal{X}} |k(x,z) - k(y,z)|) \\
&\leq l^{-3/2}(\frac{\|k\|_\infty}{2l^{1/2}} \int |k(x,z) - k(y,z)| dP(z) + \|k\|_\infty^{1/2} \sup_{z\in\mathcal{X}} |k(x,z) - k(y,z)|) \\
&\leq l^{-3/2}(\frac{\|k\|_\infty}{2l^{1/2}} + \|k\|_\infty^{1/2}) \sup_{z\in\mathcal{X}} |k(x,z) - k(y,z)| \\
&= s\|k(x,\cdot) - k(y,\cdot)\|_\infty
\end{aligned}$$

Now the statement about the covering numbers is easy to see. ☺

Proposition 22 (Covering numbers of $\mathcal{F}$) *Let $u \in C(\mathcal{X})$. Then the covering numbers for $\mathcal{F} := \mathcal{K} \cup u \cdot \mathcal{H} \cup \mathcal{H} \cdot \mathcal{H}$ satisfy*

$$N(\mathcal{F}, \varepsilon, \|\cdot\|_\infty) \leq 3N(\mathcal{K}, q\varepsilon, \|\cdot\|_\infty)$$

where the constant q is given as $q := \min\{1, \|u\|_\infty s, \frac{\|k\|_\infty}{l} s\}$ with s as in Proposition 21.

Proof. Because of $\|uf_1 - uf_2\|_\infty \leq \|u\|_\infty \|f_1 - f_2\|_\infty$ it is easy to see that

$$N(u \cdot \mathcal{H}, \varepsilon, \|\cdot\|_\infty) \leq N(\mathcal{H}, \|u\|_\infty \varepsilon, \|\cdot\|_\infty).$$

Similarly, because for all $h_1, h_2, h_3 \in \mathcal{H}$ we have

$$\|h_1h_2 - h_1h_3\|_\infty \leq \|h_1\|_\infty \|h_2 - h_3\|_\infty \leq \frac{\|k\|_\infty}{l} \|h_2 - h_3\|_\infty$$

we can conclude that

$$N(\mathcal{H} \cdot \mathcal{H}, \varepsilon, \|\cdot\|_\infty) \leq N(\mathcal{H}, \frac{\|k\|_\infty}{l}\varepsilon, \|\cdot\|_\infty).$$

Together with the result of Proposition 21 and the fact that the covering number of a union of sets is bounded by the sum of the covering numbers we obtain

$$\begin{aligned}
N(\mathcal{F}, \varepsilon, \|\cdot\|_\infty) &\leq N(\mathcal{K}, \varepsilon, \|\cdot\|_\infty) + N(u \cdot \mathcal{H}, \varepsilon, \|\cdot\|_\infty) + N(\mathcal{H} \cdot \mathcal{H}, \varepsilon, \|\cdot\|_\infty) \\
&\leq N(\mathcal{K}, \varepsilon, \|\cdot\|_\infty) + N(\mathcal{H}, \|u\|_\infty\varepsilon, \|\cdot\|_\infty) + N(\mathcal{H}, \frac{\|k\|_\infty}{l}\varepsilon, \|\cdot\|_\infty) \\
&\leq N(\mathcal{K}, \varepsilon, \|\cdot\|_\infty) + N(\mathcal{K}, \|u\|_\infty s\varepsilon, \|\cdot\|_\infty) + N(\mathcal{K}, \frac{\|k\|_\infty}{l} s\varepsilon, \|\cdot\|_\infty) \\
&\leq 3N(\mathcal{K}, q\varepsilon, \|\cdot\|_\infty)
\end{aligned}$$

where the constant q is given as stated in the proposition. ☺

This proposition shows that the covering numbers of $\mathcal{F}$ coincide with the ones of $\mathcal{K}$ up to a multiplicative constant in ε. Finally, to compute the rate of convergence for the Gaussian kernel we evaluate the integral in Theorem 20 for the class $\mathcal{F} = \mathcal{K} \cup u \cdot \mathcal{H} \cup \mathcal{H} \cdot \mathcal{H}$ and obtain

$$\int_0^\infty \sqrt{\log N(\mathcal{F}, \varepsilon, L_2(P_n))}\, d\varepsilon \leq \int_0^\infty \sqrt{\log 3N(\mathcal{K}, q\varepsilon, \|\cdot\|_\infty)}\, d\varepsilon \ < \infty.$$

This shows that the rate of convergence of $\sup_{f\in\mathcal{F}} |P_n f - Pf|$ is $\mathcal{O}(1/\sqrt{n})$, and by Theorem 15 the same now holds for the eigenfunctions of normalized spectral clustering.

While it is not immediately obvious, the fact that the Gaussian kernel k is smooth and positive definite plays an important role in this proof as these properties are of vital importance in the proof of the covering number bounds in Zhou (2002).

7 The unnormalized case

Now we want to turn our attention to the case of unnormalized spectral clustering. It will turn out that this case is not as nice as the normalized case as the convergence results will only hold under strong conditions only. Moreover, those conditions are often violated in practice. In this case, the eigenvectors do not contain any useful information about the clustering of the data space.

7.1 Convergence of unnormalized spectral clustering

The main theorem about convergence of unnormalized spectral clustering (which was informally stated as Result 2 in the introduction) is as follows:

Theorem 23 (Convergence of unnormalized spectral clustering) *Assume that the general assumptions hold. Let $\lambda \notin \mathrm{rg}(d)$ be an eigenvalue of U. Then there exists some $N \in \mathbb{N}$ and some neighborhood $M \subset \mathbb{C}$ of λ such that for $n > N$, $\sigma(\frac{1}{n}L_n) \cap M = \{\lambda_n\}$, and $(\lambda_n)_n$ converges to λ a.s. Moreover, let $\mathrm{Pr}_n : C(\mathcal{X}) \to C(\mathcal{X})$ be the spectral projection corresponding to $\sigma(U_n) \cap M$, and Pr the one corresponding to $\lambda \in \sigma(U)$. Then $\mathrm{Pr}_n \xrightarrow{p} \mathrm{Pr}$ a.s. If λ is a simple eigenvalue, then also the eigenvectors converge a.s. up to a change of sign: if v_n is the eigenvector of $\frac{1}{n}L_n$ with eigenvalue λ_n, $v_{n,i}$ its i-th coordinate, and f the eigenfunction of U with eigenvalue $\lambda \notin \mathrm{rg}(d)$, then there exists a sequence $(a_n)_{n\in\mathbb{N}}$ with $a_i \in \{+1, -1\}$ such that $\sup_{i=1,\ldots,n} |a_n v_{n,i} - f(X_i)| \to 0$ a.s. In particular, for all $b \in \mathbb{R}$ the sets $\{a_n f_n > b\}$ and $\{f > b\}$ converge, that is their symmetric difference satisfies $P(\{f > b\} \triangle \{a_n f_n > b\}) \to 0$.*

This theorem looks very similar to Theorem 14. The only difference is that the condition $\lambda \neq 1$ of Theorem 14 is now replaced by $\lambda \notin \mathrm{rg}(d)$. Note that in both cases, those conditions are equivalent to saying that λ must be an isolated eigenvalue. In the normalized case, this is satisfied for all eigenvalues but $\lambda = 1$, as $U' = Id - T'$ where T' is a compact operator. In the unnormalized case however, this condition can be violated as the spectrum of U contains a large continuous spectrum. Later we will see that this indeed leads to serious problems regarding unnormalized spectral clustering.

The proof of Theorem 7 is very similar to the one we presented in Section 5. The main difference between both cases is the structure of the spectra of U_n and U. The proposition corresponding to Proposition 9 is the following:

Proposition 24 (Spectrum of U_n)

1. *If $f \in C(\mathcal{X})$ is an eigenfunction of U_n with arbitrary eigenvalue λ, then the vector $v = \rho_n f \in \mathbb{R}^n$ is an eigenvector of the matrix $\frac{1}{n}L_n$ with eigenvalue λ.*
2. *Let $\lambda \notin \mathrm{rg}(d_n)$ be an eigenvalue of U_n with eigenfunction $f \in C(\mathcal{X})$, and $v := (v_1, \ldots, v_n) := \rho_n f \in \mathbb{R}^n$. Then f is of the form*

$$f(x) = \frac{\frac{1}{n}\sum_j k(x, X_j) v_j}{d_n(x) - \lambda}. \tag{6}$$

3. *If v is an eigenvector of the matrix $\frac{1}{n}L_n$ with eigenvalue $\lambda \notin \operatorname{rg}(d_n)$, then f defined by Equation* (6) *is an eigenfunction of U_n with eigenvalue λ.*

4. *The essential spectrum of U_n coincides with the range of the degree function, that is $\sigma_{ess}(U_n) = \operatorname{rg}(d_n)$. All eigenvalues of U_n are non-negative and can have accumulation points only in $\operatorname{rg}(d_n)$. The analogous statements also hold for the operator U.*

Proof. The first parts can be proved analogously to Proposition 9. For the last part, remember that the essential spectrum of the multiplication operator M_{d_n} consists of the range of the multiplier function d_n. As S_n is a compact operator, the essential spectrum of $U_n = M_{d_n} - S_n$ coincides with the essential spectrum of M_{d_n} as we have already mentioned in the beginning of Section 4. The accumulation points of the spectrum of a bounded operator always belong to the essential spectrum. Finally, to see the non-negativity of the eigenvalues observe that if we consider the operator U_n as an operator on $L_2(P_n)$ we have

$$\langle U_n f, f\rangle = \int\int (f(x) - f(y))f(x)k(x,y)dP_n(y)dP_n(x) \quad = \frac{1}{2}\int\int (f(x) - f(y))^2 k(x,y)dP_n(y)dP_n(x) \geq 0$$

Thus U is a non-negative operator on $L_2(P_n)$ and as such only has a non-negative eigenvalues. As we have $C(\mathcal{X}) \subset L_2(P)$ by the compactness of $\mathcal{X}$, the same holds for the eigenvalues of U as an operator on $C(\mathcal{X})$.

☺

This proposition establishes a one-to-one relationship between the eigenvalues of U_n and $\frac{1}{n}L_n$, provided the condition $\lambda \notin \operatorname{rg}(d_n)$ is satisfied. Next we need to prove the compact convergence of U_n to U:

Proposition 25 (U_n converges compactly to U a.s.) *Under the general assumptions, $U_n \xrightarrow{c} U$ a.s.*

Proof. We consider the multiplication and integral operator parts of U_n separately. Similarly to Proposition 12 we can prove that the integral operators S_n converge collectively compactly to S a.s., and as a consequence also $S_n \xrightarrow{c} S$ a.s. For the multiplication operators we have operator norm convergence as

$$\|M_{d_n} - M_d\| = \sup_{\|f\|_\infty \leq 1} \|d_n f - df\|_\infty \leq \|d_n - d\|_\infty \to 0 \text{ a.s.}$$

by the Glivenko-Cantelli Proposition 10. As operator norm convergence implies compact convergence we also have $M_{d_n} \xrightarrow{c} M_d$ a.s. Finally, it is easy to see that the sum of two compactly converging operators also converges compactly. Hence, $U_n \xrightarrow{c} U$ a.s. ☺

Now we can prove the convergence Theorem 23 similarly to Theorem 14:

Proof of Theorem 23. In Proposition 24 we established a one-to-one correspondence between the eigenvalues $\lambda \notin \operatorname{rg}(d_n)$ of $\frac{1}{n}L_n$ and U_n, and we saw that the eigenvalues λ of U with $\lambda \notin \operatorname{rg}(d)$ are isolated and have finite multiplicity. In Proposition 25 we proved the compact convergence of U_n to U, which according to Proposition 6 implies the convergence of the spectral projections of isolated eigenvalues with finite multiplicity. For simple eigenvalues, this implies the convergence of the eigenvectors up to a change of sign, and the convergence of the sets $\{f_n > b\}$ is a simple consequence of the almost sure convergence of $(a_n f_n)_n$. ☺

8 Non-isolated eigenvalues

The most important difference between the limit operators of normalized and unnormalized spectral clustering is the condition under which eigenvalues of the limit operator are isolated in the spectrum. In the normalized case this is true for all eigenvalues $\lambda \neq 1$, while in the unnormalized case this is only true for all eigenvalues satisfying $\lambda \notin \operatorname{rg}(d)$. In this section we want to investigate those conditions more closely. We will see that especially in the unnormalized case, this condition can be violated, and that in this case spectral clustering will not yield sensible results. In particular, the condition $\lambda \notin \operatorname{rg}(d)$ is not an artifact of our methods, but plays a fundamental role. It is the main reason why we suggest to use normalized rather than unnormalized spectral clustering.

8.1 Theoretical results

Firstly we will construct an example where all non-trivial eigenvalues $\lambda_2, \lambda_3, ...$ lie inside the range of the degree function.

Example 2 ($\lambda_2 \notin \mathrm{rg}(d)$ violated) *Consider the data space $\mathcal{X} = [1, 2] \subset \mathbb{R}$ and the probability distribution given by a piecewise constant probability density function p on $\mathcal{X}$ with $p(x) = s$ if $4/3 \leq x < 5/3$ and $p(x) = (3-s)/2$ otherwise, for some fixed constant $s \in [0, 3]$ (for example, for $s = 0.3$ this density has two clearly separated high density regions, cf. Figure 2). As similarity function we choose $k(x, y) := xy$. Then the only eigenvalue of U outside of $\mathrm{rg}(d)$ is the trivial eigenvalue 0 with multiplicity one.*

Proof. First note that in this example the general conditions are satisfied: $\mathcal{X}$ is compact, and k is symmetric and ≥ 1 on $\mathcal{X}\times\mathcal{X}$. The degree function in this case is

$$d(x) = \int_1^2 xyp(y)dy = x(\int_1^{4/3} y\frac{3-s}{2}dy + \int_{4/3}^{5/3} ysdy + \int_{5/3}^2 y\frac{3-s}{2}dy) = 1.5x$$

(independently of s) and has range $[1.5, 3]$ on $\mathcal{X}$. A function $f \in C(\mathcal{X})$ is eigenfunction with eigenvalue $\lambda \notin \mathrm{rg}(d)$ of U if the eigenvalue equation is satisfied:

$$Uf(x) = d(x)f(x) - x\int yf(y)p(y)dy \stackrel{!}{=} \lambda f(x). \tag{7}$$

Defining the real number $\beta := \int yf(y)p(y)dy$ we can solve Equation (7) for $f(x)$ to obtain $f(x) = \frac{\beta x}{d(x)-\lambda}$. Plugging this into the definition of β yields the condition

$$1 \stackrel{!}{=} \int \frac{y^2}{d(y)-\lambda}p(y)dy. \tag{8}$$

Hence, λ is an eigenvalue of U if Equation (8) is satisfied. For our simple density function p, the integral in this condition can be solved analytically. It can then been seen that $g(\lambda) := \int \frac{y^2}{d(y)-\lambda}p(y)dy \stackrel{!}{=} 1$ is only satisfied for $\lambda = 0$, hence the only eigenvalue outside of $\mathrm{rg}(d)$ is the trivial eigenvalue 0 with multiplicity one. ☺

In the above example we could see that there indeed exist situations where there the operator U does not possess a non-zero eigenvalue with $\lambda \notin \mathrm{rg}(d)$. The next question is what happens in this situation.

Proposition 26 (Clustering fails if $\lambda_2 \notin \mathrm{rg}(d)$ is violated) *Assume that $\sigma(U) = \{0\} \cup \mathrm{rg}(d)$ with the eigenvalue 0 having multiplicity 1. Assume that the probability distribution P on $\mathcal{X}$ has no point masses. Then the sequence of second eigenvalues of $\frac{1}{n}L_n$ converges to $\min_{x\in\mathcal{X}} d(x)$. The corresponding eigenfunction will approximate the characteristic function of some $x \in \mathcal{X}$ with $d(x) = \min_{x\in\mathcal{X}} d(x)$ or a linear combination of such functions.*

Proof. It is a standard fact (Chatelin, 1983) that for each λ inside the continuous spectrum $\mathrm{rg}(d)$ of U there exists a sequence of functions $(f_n)_n$ with $\|f_n\| = 1$ such that $\|(U - \lambda\,\mathrm{I})f_n\| \to 0$. Hence, for each precision $\varepsilon > 0$ there exists a function f_ε such that $\|(U - \lambda\,\mathrm{I})f_\varepsilon\| < \varepsilon$. This means that for a computer with machine precision ε, the function f_ε appears to be an eigenfunction with eigenvalue λ. Thus, with a finite precision calculation we cannot distinguish between eigenvalues and the continuous spectrum of an operator. Intuitively it seems clear that this also affects the eigenvalues of the empirical approximation U_n of U. To make this precise we want to construct a sequence $(f_n)_n$ such that for all $\varepsilon > 0$ there exists some $J > 0$ such that for all $j > J$ we have $\|(U_n-\lambda\,\mathrm{I})f_n\| \leq \varepsilon$. For given $\lambda \in \mathrm{rg}(d)$ we choose some $x_\lambda \in \mathcal{X}$ with $d(x_\lambda) = \lambda$. Define $B_n := B(x_\lambda, \frac{1}{n})$ as the ball around x_λ with radius $1/n$ (note that B_n does not depend on the sample), and choose some $f_n \in C(\mathcal{X})$ which is constant 1 on B_n and constant 0 outside B_{n-1}. Then $\|f_n\|_\infty = 1$, the sequence $(f_n)_n$ converges pointwise to the characteristic function at x_λ, and $\|(Uf_n - \lambda\,\mathrm{I})\| \to 0$. Now we obtain:

$$\begin{aligned}
\|(U_n - \lambda\,\mathrm{I})f_n\| &\leq \|(M_{d_n} - \lambda\,\mathrm{I})f_n\| + \|S_n f_n\| \\
&\leq \sup_{x\in B_{n-1}} |d_n(x) - d(x_\lambda)| + \sup_{x\in\mathcal{X}} |\int_{B_n} k(x, y)dP_n(y)| \\
&\leq \sup_{x\in B_{n-1}} |d_n(x) - d(x)| + \sup_{x\in B_{n-1}} |d(x) - d(x_\lambda)| + \|k\|_\infty P_n(B_n)
\end{aligned}$$

The term $\sup_{x \in B_{n-1}} |d_n(x) - d(x)|$ converges to 0 a.s. because $\|d_n - d\|_\infty \to 0$ a.s., and the expression $\sup_{x \in B_{n-1}} |d(x) - d(x_\lambda)|$ converges to zero by the continuity of d. It remains to prove that the last term $P_n(B_n)$ converges to 0 a.s. For given $\varepsilon > 0$ fix some $M \in \mathbb{N}$ such that $P(B_M) \leq \varepsilon$ (this is always possible as we assumed that P does not have any point measures). For this fixed set B_M we know that $P_n(B_M) \to P(B_M)$ a.s., that is for each ε and each M there exists some $N \in \mathbb{N}$ such that for all $n > N$ we have $|P_n(B_M) - P(B_M)| \leq \varepsilon$ a.s.. By the choice of M we can conclude that $P_n(B_M) \leq 2\varepsilon$ a.s. for all $n > N$. As $B_m \subset B_M$ for all $m > M$ by construction we get $P_n(B_m) \leq 2\varepsilon$ a.s. for all $n > N$ and $m > M$. In particular, if we set $J := \max\{N, M\}$ then for all $j > J$ we have $P_j(B_j) \leq 2\varepsilon$ a.s. Consequently, $P_n(B_n) \to 0$ a.s..

Now we have seen that for each machine precision ε there exists some $N \in \mathbb{N}$ such that for $n > N$ we have $\|(U_n - \lambda \mathrm{I}) f_n\| \leq \varepsilon$ a.s., and by Proposition 8 we can conclude that then also

$$\|(\frac{1}{n} L_n - \lambda \mathrm{I})(f(X_1), ..., f(X_n))'\| \leq \varepsilon \text{ a.s.}$$

Consequently, if the machine precision of the numerical eigensolver is ε, then this expression cannot be distinguished from 0, and the vector $(f(X_1), ..., f(X_n))'$ appears to be an eigenvector of $\frac{1}{n} L_n$ with eigenvalue λ. As this construction holds for each $\lambda \in \mathrm{rg}(d)$, the smallest non-zero "eigenvalue" discovered by the eigensolver will be $\lambda_2 := \min_{x \in \mathcal{X}} d(x)$. If x_{λ_2} is the unique point in $\mathcal{X}$ with $d(x_{\lambda_2}) = \lambda_2$, then the second eigenvector of $\frac{1}{n} L_n$ will converge to the delta-function at x_{λ_2}. The clustering constructed from the second eigenvector is not a sensible clustering as thresholding the second eigenvector will result in a trivial clustering that only separates a neighborhood of x_{λ_2} from the rest of $\mathcal{X}$. If there are several points $x \in \mathcal{X}$ with $d(x) = \lambda_2$, then the "eigenspace" of λ_2 will be spanned by the delta-functions at all those points. In this case, the eigenvectors of $\frac{1}{n} L_n$ will approximate one of those delta-functions, or a linear combination thereof. ☺

As a side remark, note that as the above construction holds for all elements $\lambda \in \mathrm{rg}(d)$, eventually the whole interval $\mathrm{rg}(d)$ will be covered by eigenvalues of $\frac{1}{n} L_n$. For this it also does not make a difference whether there actually exists some proper eigenvalue inside $\mathrm{rg}(d)$ or not.

So far we have seen that there exist examples where the assumption $\lambda \notin \mathrm{rg}(d)$ in Theorem 23 is violated and that in this case the corresponding eigenfunction does not contain any useful information for clustering. This situation is aggravated by the fact that the condition $\lambda \notin \mathrm{rg}(d)$ can only be verified if the operator U, and hence the probability distribution P on $\mathcal{X}$, is known. As this is not the case in the standard setting of clustering, it is impossible to know whether the condition $\lambda \notin \mathrm{rg}(d)$ is true for the eigenvalues in consideration or not. Consequently, not only spectral clustering can fail in certain situations, but we are unable to check whether this is the case for a given application of clustering or not. The least thing one should do if one really wants to use unnormalized spectral clustering is to estimate the critical region $\mathrm{rg}(d)$ by $[\min_i d_i/n, \max_i d_i/n]$ and check whether the relevant eigenvalues of $\frac{1}{n} L_n$ are inside or close to this interval or not. This observation then gives an indication whether the results obtained can considered to be reliable or not.

Finally we want to show that such problems as described above do not only occur in pathological examples, but they can come up for many similarity functions which are often used in practice.

Proposition 27 (Finite discrete spectrum for analytic similarity functions) *Assume that $\mathcal{X}$ is a compact subset of $\mathbb{R}^n$, and the similarity function k is analytic in a neighborhood of $\mathcal{X} \times \mathcal{X}$. Let P be a probability distribution on $\mathcal{X}$ which has an analytic density function. Assume that the set $\{x^* \in \mathcal{X};\ d(x^*) = \min_{x \in \mathcal{X}} d(x)\}$ is finite. Then $\sigma(U)$ has only finitely many eigenvalues outside $\mathrm{rg}(d)$.*

Proof. This proposition is special case of results on the discrete spectrum of the generalized Friedrichs model which can be found for example in Lakaev (1979), Abdullaev and Lakaev (1991), and Ikromov and Sharipov (1998). In those articles, the authors only consider the case where P is the uniform distribution, but their proofs can be carried over to the case of analytic density functions. ☺

The assumptions in this proposition are for example satisfied in one of the "default" clustering settings where the probability distribution is a (truncated) mixture of Gaussians and the similarity function used is the Gaussian kernel $k(x, y) = \exp(-|x - y|^2/\sigma^2)$. The proposition now tells us that in this case there are only finitely many

eigenvalues below the essential spectrum. We have experimental evidence (see below) that "finitely many" is usually a rather small number, say 2 or 3. If we now use more than those 2 or 3 eigenvectors of the unnormalized Laplacian, we immediately run into the problems we described above: the eigenvectors carry no information about the clustering of the data space, and hence we get misleading results. Moreover, again we have the problem that we do not know how many eigenvalues of U are below $\mathrm{rg}(d)$. Again we suggest that the least thing one should do is to estimate $\mathrm{rg}(d)$ by $\mathrm{rg}(d_n)$ on the sample and use only those eigenvectors whose eigenvalues are not too close to $\mathrm{rg}(d_n)$.

8.2 Empirical results

To illustrate what happens for unnormalized spectral clustering if the condition $\lambda \notin \mathrm{rg}(d)$ is violated, we want to analyze several empirical examples and compare the results of unnormalized and normalized spectral clustering.

We first start with Example 2 of Section 8.1. We draw n=100 sample points according to the piecewise constant density on the space $\mathcal{X} = [1, 2]$ as given in the example (with parameter $s = 0.3$). This density is shown in the left panel of Figure 2. As similarity function we use the linear similarity function $k(x, y) = xy$ as in the example. The middle panel shows all eigenvalues of the unnormalized Laplace matrix $\frac{1}{n}L_n$, ordered according to magnitude (i.e., we plot i versus λ_i). We can see that apart from the trivial eigenvalue 0, all eigenvalues lie inside the range of the empirical degree function d_n (indicated by the dashed lines). The right panel shows the coordinates of the second eigenvector of $\frac{1}{n}L_n$, plotted versus the corresponding data points (i.e., for sample $X_1, ..., X_n$ and eigenvector $v = (v_1, ..., v_n)$ we plot X_i versus v_i). As we predicted in Section 8.1, this eigenvector approximates a Dirac function. Thus it does not contain any information about the clusters in the data space, and hence unnormalized spectral clustering fails.

While the example above is a bit artificial we now want to show that the same problems can occur in situations which are highly relevant to practical applications. As data space we choose $\mathcal{X} = I\!R$ with a density which is a mixture of four Gaussian distributions where all Gaussians have the same standard deviation 0.5 and are well separated from each other (the means are 2,4,6, and 8). A histogram of this distribution is shown in Figure 3. This distribution a prototypical example, and every clustering algorithm should be able to identify the clusters in this toy example. As similarity function we choose the Gaussian kernel function $k(x, y) = \exp(-\|x - y\|^2/\sigma^2)$, which is the similarity function most widely used in applications of spectral clustering. In this situation it is difficult to prove analytically how many eigenvalues will lie below $\mathrm{rg}(d)$; by Proposition 27 we only know that they are finitely many. However, in practice it turns out that "finitely many" often means "very few", for example two or three. In the following we show plots of the eigenvalues and eigenvectors of the normalized and unnormalized Laplacians, for different values of the kernel width parameter σ. We will see in all those plots that in the unnormalized case, as soon as an eigenvalue gets close to the range of the degree function, its shape approximates a Dirac delta function.

We start with the kernel parameter $\sigma = 0.5$ (which is the true width of the Gaussian densities), cf. Figure 4. In the plot of eigenvalues of the unnormalized Laplacian (first row, left panel of Figure 4) we can see that the second, third, and fourth eigenvalues are close to 0, while the fifth eigenvalue already is close to the range of the degree function. The eigenvectors reflect this observation (second row of 4): the second, third, and fourth eigenvectors contain useful information about the clustering, while the fifth only contains few information and the sixth no information at all. All higher eigenvalues are contained in $\mathrm{rg}(d_n)$. A similar situation can be found for the eigenvectors of the normalized Laplacian (third row of 4). Hence, in this example, both normalized and unnormalized spectral clustering work equally good.

Now consider what happens if we increase the kernel width. In general one can observe that all eigenvalues move towards the range of the degree function. Consider the case $\sigma = 2$, cf. Figure 5. We can see that in the unnormalized case, only the second and third eigenvalues are below the range of the degree function, and only the second and third eigenvectors carry information about the clustering of the data space. The remaining eigenvectors approximate Dirac functions. In case of the normalized Laplacian however, all eigenvectors carry information about the clustering. This situation gets even more extreme if we further increase the kernel width to $\sigma = 5$, cf. Figure 6, or even to the ridiculous value $\sigma = 50$ (Figure 7). In the unnormalized case, only the second eigenvalue and eigenvector are informative, and the other eigenvectors are approximately Dirac. In the normalized case we

have at least four informative eigenvectors and can recover all clusters perfectly.

One the one hand, one could explain the failing of unnormalized spectral clustering in the cases where σ is chosen too large by the inappropriately chosen parameter σ. On the other hand, normalized spectral clustering is able to cope with this situation and can recover the correct clusters even if the parameter σ is set to the value of $\sigma = 50$. In practice, one often does not know the number of clusters in the data, and it is not clear how the kernel width σ has to be chosen. As we have illustrated, normalized spectral clustering gives good results for a very wide range of σ, while unnormalized spectral clustering only works for the "correct" value. The failing in the other cases occurs exactly in the way we have proved it in previous sections: the eigenvalues are not isolated in the spectrum and hence the corresponding eigenvectors do not contain information about the clustering. The fact that this can be reproduced in a toy example for clustering where the density is a simple mixture of Gaussians and the similarity function is the Gaussian kernel shows that our findings have to be taken very serious for practical applications.

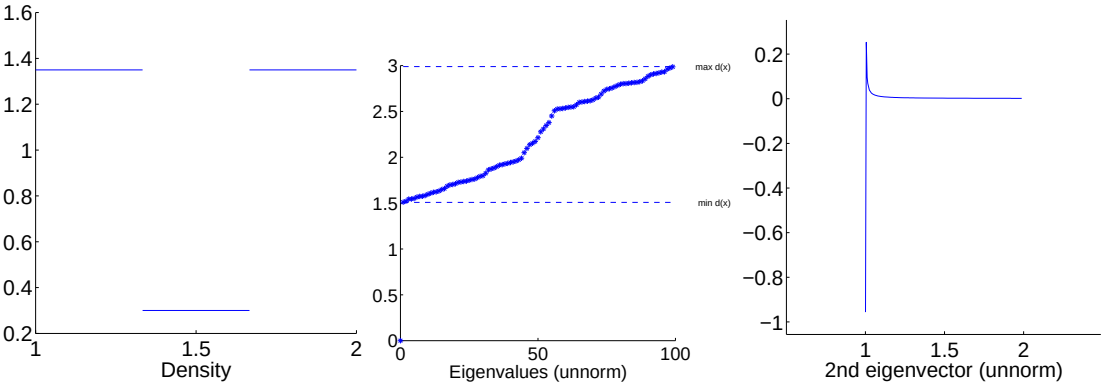

Figure 2: This figure corresponds to Example 2 of Section 8.1. The left panel shows the density function with parameter $s = 0.3$. The middle panel shows the eigenvalues of the unnormalized Laplacian, ordered according to magnitude. The right panel shows the second eigenvector of the unnormalized Laplacian. This eigenvector approximates a Dirac function.

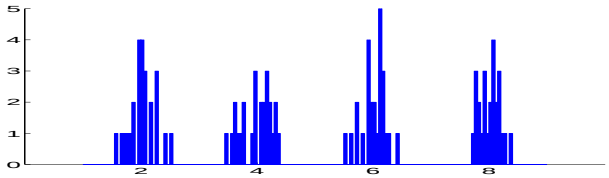

Figure 3: Histogram of a sample drawn according to a mixture of four Gaussians.

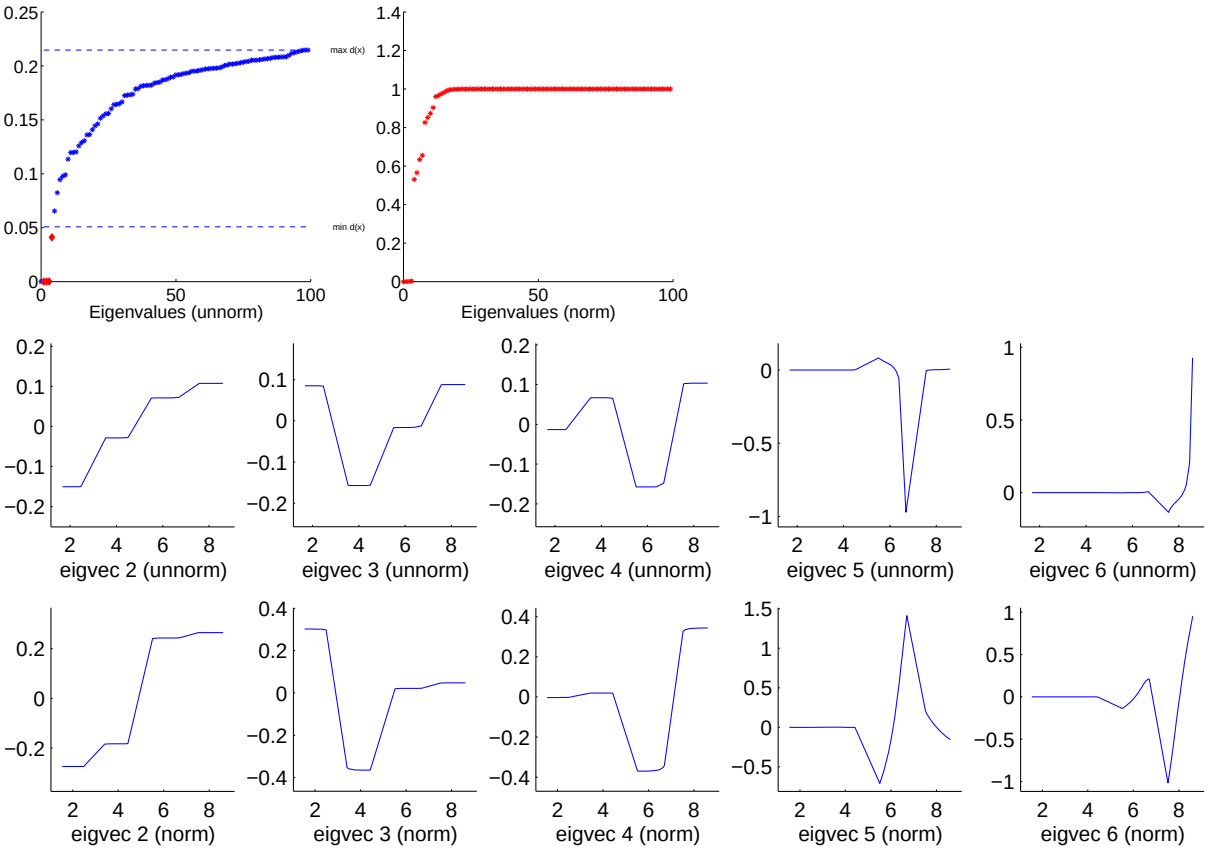

Figure 4: Eigenvalues and eigenvectors of unnormalized and normalized Laplacians for kernel width $\sigma = 0.5$. The first row shows all eigenvalues of the unnormalized (left side) and the normalized (right side) graph Laplacian. The second row shows the first eigenvectors of the unnormalized Laplacian, the third row the eigenvectors of the normalized Laplacian. See text for more details.

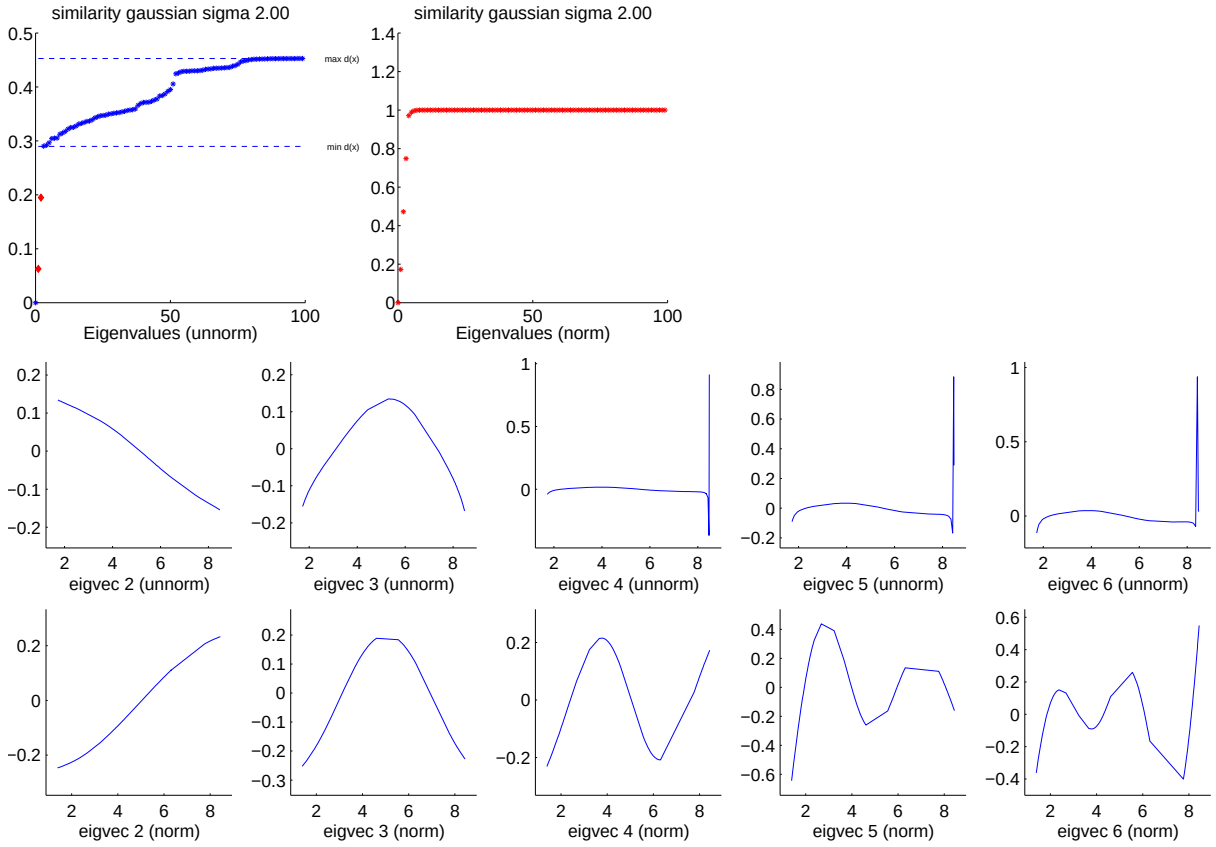

Figure 5: Eigenvalues and eigenvectors of unnormalized and normalized Laplacians for kernel width $\sigma = 2$. See text for more details.

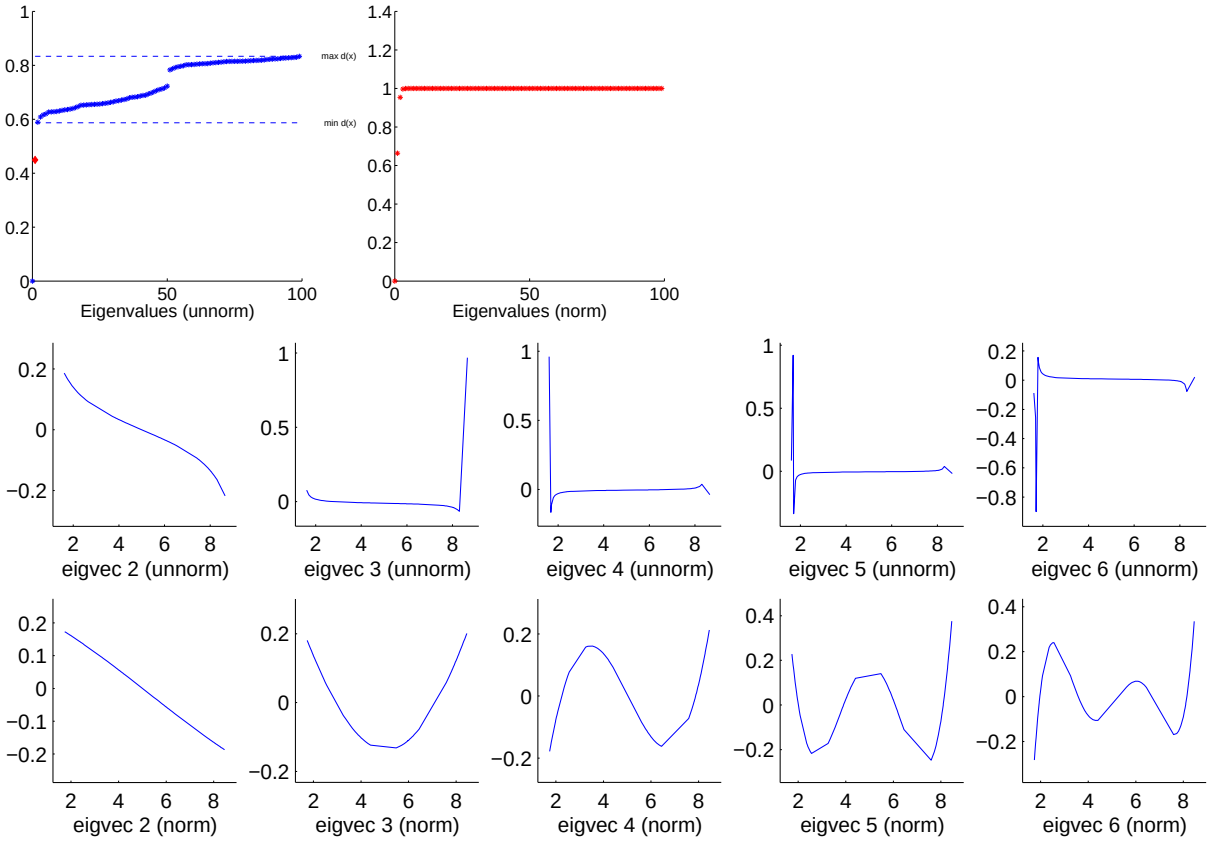

Figure 6: Eigenvalues and eigenvectors of unnormalized and normalized Laplacians for kernel width $\sigma = 5$. See text for more details.

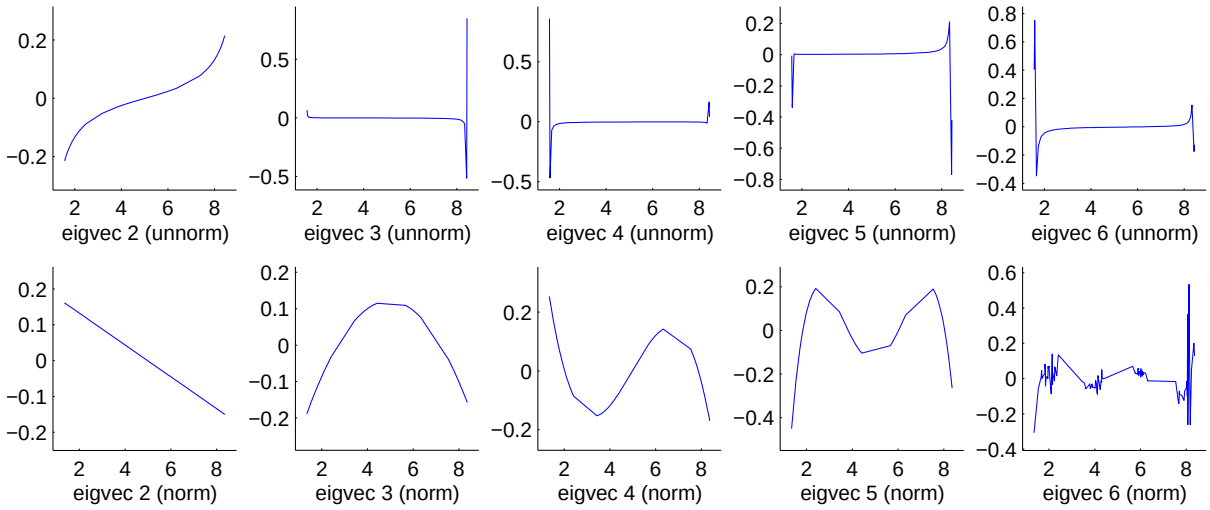

Figure 7: Eigenvectors of unnormalized and normalized Laplacians for kernel width $\sigma = 50$. See text for more details.

9 Spectral clustering: from discrete to continuous

In Section 2 we introduced spectral clustering as an approximation of a balanced graph cut. On discrete sets, this gives a nice and intuitively plausible explanation why spectral clustering constructs desirable clusterings. This justification, however, is no longer satisfactory when we consider the limit clustering on the whole data space. In this section we would like to discuss why also the limit clusterings represent desirable partitions of the probability space. We will suggest a method for bipartitioning a general probability distribution and will outline its connections with spectral clustering. We well keep this section short and informal as a comprehensive discussion of this subject goes beyond the scope of this paper.

Given a probability distribution P supported on some compact manifold or domain $\mathcal{X}$, we would like to partition this domain in two "clusters" satisfying some geometric properties. Similarly to graph partitioning, we want the size of the boundary to be minimized, while keeping the sizes of parts roughly equal.

One natural formulation of this problem when $\mathcal{X}$ is a manifold is due to Cheeger (1970), who introduced the isoperimetric constant (Cheeger constant) for a compact manifold as the following optimization problem. Given a compact n-dimensional Riemannian manifold $\mathcal{M}$ consider a partition $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2$, $\mathcal{M}_2 = \mathcal{M} - \mathcal{M}_1$ into two submanifolds with boundary $\mathcal{B} = \delta\mathcal{M}_1 = \delta\mathcal{M}_2$. The isoperimetric constant was defined by Cheeger as

$$h_{\mathcal{M}} = \inf_{\mathcal{B}=\delta\mathcal{M}_1} \frac{\mathrm{vol}^{n-1}\,\mathcal{B}}{\min\left(\mathrm{vol}^n(\mathcal{M}_1), \mathrm{vol}^n(\mathcal{M}-\mathcal{M}_1)\right)}$$

where $\mathrm{vol}^{n-1}\,\mathcal{B}$ denotes the $n-1$ dimensional volume of the boundary. We will call the partition corresponding to the Cheeger constant (assuming it exists) the Cheeger partition.

Cheeger then observed then that this partition is closely related to properties of the Laplace-Beltrami operator Δ on $\mathcal{M}$. This relation parallels the relation between the Laplacian of a graph and its Cheeger (or balanced) cut. Thus the zero set of the second eigenfunction of Δ is an approximation of the Cheeger cut, which is depicted in Figure 8 (taken from Belkin and Niyogi, 2004).

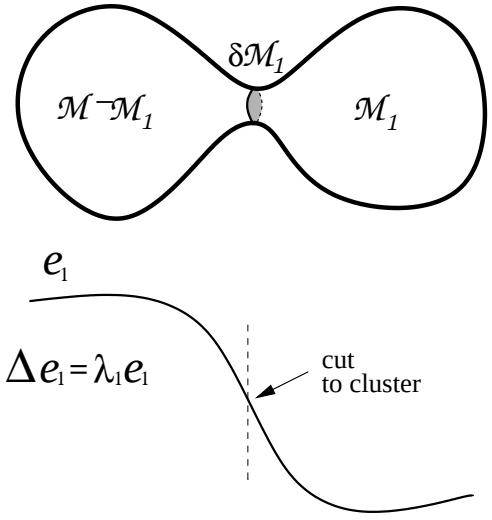

Figure 8: The Cheeger partition of a manifold is induced by the second eigenfunction of the Laplace-Beltrami operator.

What we need, however, is a more general weighted version of the Cheeger constant, when the manifold is considered together with a probability density function $p(x)$. The volumes are then weighted using the $p(x)$ function and the Cheeger partition is appropriately modified, but the conceptual picture does not change. It is now possible to construct a corresponding weighted Laplace-Beltrami operator $\Delta_p = \frac{1}{p}\,\mathrm{div}(p\nabla f)$, where div is the divergence operator on $\mathcal{M}$. Eigenfunctions of Δ_p are analogous to the eigenvectors of the normalized graph

Laplacian and provide a clustering of the probability distribution, approximating the weighted Cheeger partition (cf. Chung et al., 2000).

While much work remains to be done to make these intuitions precise, there are strong indications that these eigenfunction can be approximated from the graph Laplacian with a Gaussian kernel. Results in that direction have recently been obtained in Belkin (2003), where pointwise convergence of the unnormalized graph Laplacian to the Laplace-Beltrami operator is shown for the case of the uniform distribution on an embedded Riemannian manifold, in Bousquet et al. (2004), where it is shown for a probability distribution on a Euclidean domain, and in Lafon (2004), where the general case of an arbitrary probability distribution on a manifold is considered.

We think that those results together with the results obtained in this paper may put us within reach of the general theory of spectral bisectioning for continuous spaces and its empirical approximations.

10 Conclusion

In this article we investigated consistency of spectral clustering algorithm by studying the convergence of eigenvectors of the normalized and unnormalized Laplacian matrices on random samples. We demonstrated that under standard assumptions, the first eigenvectors of the normalized Laplacian converges to eigenfunctions of some limit operator. In the unnormalized case, the same is only true if the eigenvalues of the limit operator satisfy certain properties, namely if these eigenvalues lie below the continuous part of the spectrum. We showed that in many examples this condition is not satisfied. In those cases, the information provided by the corresponding eigenvector is misleading and cannot be used for clustering.

This leads to two main practical conclusions about spectral clustering. First, from a statistical point of view it is clear that normalized rather than unnormalized spectral clustering should be used whenever possible. Second, if for some reason one wants to use unnormalized spectral clustering, one should try to check whether the eigenvalues corresponding to the eigenvectors used by the algorithm lie significantly below the continuous part of the spectrum. If that is not the case, those eigenvectors need to be discarded as they do not provide information about the clustering.

From a mathematical point of view our contribution is to combine different tools to prove results about convergence of spectral properties of random graph Laplacians. While by themselves these tools are not new, they lead to a new method of analysis, which is significantly simpler and easier to understand than previous approaches. While the results obtained in Koltchinskii and Giné (2000) and Koltchinskii (1998) are stronger than our results (for example, they also prove central limit theorems and uniform convergence results), their methods are also much more involved and allow only for analysis of Hilbert-Schmidt operators. That makes them unsuitable for analysis of unnormalized spectral clustering, where the resulting operators are not of Hilbert-Schmidt type.

Finally, our framework can be extended to analysis of various Laplacian based methods other than clustering. We believe that more systematic statistical analysis of various algorithms in computer science may yield new insights into their properties.

References

Z. Abdullaev and S. Lakaev. On the spectral properties of the matrix-valued Friedrichs model. In *Many-particle Hamiltonians: spectra and scattering*, volume 5 of *Adv. Soviet Math.*, pages 1–37. Amer. Math. Soc., Providence, RI, 1991.

Charles J. Alpert and So-Zen Yao. Spectral partitioning: the more eigenvectors, the better. In *Proceedings of the 32nd ACM/IEEE conference on Design automation*, pages 195–200. ACM Press, 1995. ISBN 0-89791-725-1.

P. Anselone. *Collectively compact operator approximation theory*. Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

M. Anthony. Uniform Glivenko-Cantelli theorems and concentration of measure in the mathematical modelling of learning. CDAM Research Report LSE-CDAM-2002-07, 2002.

S. Arora, S. Rao, and U. Vazirani. Expander flows, geometric embeddings and graph partitioning. In *STOC*, pages 222–231, 2004.

K. Atkinson. The numerical solution of the eigenvalue problem for compact integral operators. *TAMS*, 129(3): 458–465, 1967.

C. Baker. *The numerical treatment of integral equations*. Oxford University Press, 1977.

S. Barnard, A. Pothen, and H. Simon. A spectral algorithm for envelope reduction of sparse matrices. *Numerical Linear Algebra with Applications*, 2(4):317–334, 1995.

M. Belkin. *Problems of Learning on Manifolds*. PhD thesis, University of Chicago, 2003.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, June 2003.

M. Belkin and P. Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine Learning*, 56:209–239, 2004. Special Issue on Clustering.

Y. Bengio, P. Vincent, J.-F. Paiement, O. Delalleau, M. Ouimet, and N. Le Roux. Spectral clustering and kernel PCA are learning eigenfunctions. Technical Report TR 1239, University of Montreal, 2003.

O. Bousquet, O. Chapelle, and M. Hein. Measure based regularization. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

O. Chapelle, J. Weston, and B. Schölkopf. Cluster kernels for semi-supervised learning. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15, Cambridge, MA, USA, 2003. MIT Press.

F. Chatelin. *Spectral Approximation of Linear Operators*. Academic Press, New York, 1983.

J. Cheeger. A lower bound for the smallest eigenvalue of the laplacian. In R.C. Gunnings, editor, *Prooblems in analysis*. Princeton University Press, 1970.

F. Chung. *Spectral graph theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC, 1997.

F. Chung, A. Grigor'yan, and S.-T. Yau. Higher eigenvalues and isoperimetric inequalities on riemannian manifolds and graphs. *Communications on Analysis and Geometry*, 8:969–1026, 2000.

Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Knowledge Discovery and Data Mining*, pages 269–274, 2001.

Chris Ding, Xiaofeng He, Hongyuan Zha, Ming Gu, and Horst Simon. A min-max cut algorithm for graph partitioning and data clustering. In *Proceedings of first IEEE International Conference on Data Mining*, pages 107–114, San Jose, CA, 2000.

W. E. Donath and A. J. Hoffman. Lower bounds for the partitioning of graphs. *IBM J. Res. Develop.*, 17:420–425, 1973.

R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge Studies in advanced mathematics. Cambridge University Press, Cambridge, U.K., 1999.

M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Math. J.*, 23:298–305, 1973.

S. Guattery and G. L. Miller. On the quality of spectral separators. *SIAM Journal of Matrix Anal. Appl.*, 19(3), 1998.

L. Hagen and A.B. Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Trans. Computer-Aided Design*, 11(9):1074–1085, 1992.

J. Hartigan. Consistency of single linkage for high-density clusters. *JASA*, 76(374):388–394, 1981.

B. Hendrickson and R. Leland. An improved spectral graph partitioning algorithm for mapping parallel computations. *SIAM J. on Scientific Computing*, 16:452–469, 1995.

D. Higham and M. Kibble. A unified view of spectral clustering. Mathematics Research Report 2, University of Strathclyde, 2004.

I. Ikromov and F. Sharipov. On the discrete spectrum of the nonanalytic matrix-valued Friedrichs model. *Funct. Anal. Appl*, 32(1):49–50, 1998. See also http://www.arxiv.org/abs/funct-an/9502004.

A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, 1999. ISSN 0360-0300.

R. Kannan, S. Vempala, and A. Vetta. On clusterings - good, bad and spectral. Technical report, Computer Science Department, Yale University, 2000.

T. Kato. *Perturbation theory for linear operators*. Springer, Berlin, 1966.

V. Koltchinskii. Asymptotics of spectral projections of some random matrices approximating integral operators. *Progress in Probabilty*, 43, 1998.

V. Koltchinskii and E. Giné. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113 – 167, 2000.

S. Lafon. *Diffusion maps and geometric harmonics*. PhD thesis, Yale University, 2004.

S. N. Lakaev. The discrete spectrum of a generalized Friedrichs model. *Dokl. Akad. Nauk UzSSR*, 4:9–10, 1979.

L. Lovász. Random walks on graphs: a survey. In *Combinatorics, Paul Erdős is eighty, Vol. 2 (Keszthely, 1993)*, volume 2 of *Bolyai Soc. Math. Stud.*, pages 353–397. János Bolyai Math. Soc., Budapest, 1996.

M. Meila and J. Shi. A random walks view of spectral segmentation. In *8th International Workshop on Artificial Intelligence and Statistics*, 2001.

S. Mendelson. A few notes on statistical learning theory. In *Advanced lectures in machine learning*, volume LNCS 2600, pages 1–40. Springer, 2003.

B. Mohar. The Laplacian spectrum of graphs. In *Graph theory, combinatorics, and applications. Vol. 2 (Kalamazoo, MI, 1988)*, Wiley-Intersci. Publ., pages 871–898. Wiley, New York, 1991.

A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*. MIT Press, 2002.

P. Niyogi and N. K. Karmarkar. An approach to data reduction and clustering with theoretical guarantees. In P. Langley, editor, *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, 2000.

D. Pollard. Strong consistency of k-means clustering. *Annals of Statistics*, 9(1):135–140, 1981.

D. Pollard. *Convergence of Stochastic Processes*. Springer, New York, 1984.

A. Pothen, H. D. Simon, and K. P. Liou. Partitioning sparse matrices with egenvectors of graphs. *SIAM Journal of Matrix Anal. Appl.*, 11:430–452, 1990.

J. Shawe-Taylor, C. Williams, N. Cristianini, and J. Kandola. On the eigenspectrum of the Gram matrix and its relationship to the operator eigenspectrum. In N. Cesa-Bianchi, M. Numao, and R. Reischuk, editors, *Proceedings of the 13th International Conference on Algorithmic Learning Theory*. Springer, Heidelberg, 2002.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

D. Spielman and S. Teng. Spectral partitioning works: planar graphs and finite element meshes. In *37th Annual Symposium on Foundations of Computer Science (Burlington, VT, 1996)*, pages 96–105. IEEE Comput. Soc. Press, Los Alamitos, CA, 1996.

Mechthild Stoer and Frank Wagner. A simple min-cut algorithm. *J. ACM*, 44(4):585–591, 1997. ISSN 0004-5411.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

R. Van Driessche and D. Roose. An improved spectral bisection algorithm and its application to dynamic load balancing. *Parallel Comput.*, 21(1), 1995.

U. von Luxburg. *Statistical Learning with Similarity and Dissimilarity Functions*. PhD thesis, Technical University of Berlin, 2004. Submitted.

Y. Weiss. Segmentation using eigenvectors: A unifying view. In *Proceedings of the International Conference on Computer Vision*, pages 975–982, 1999.

C. K. I. Williams and M. Seeger. The effect of the input density distribution on kernel-based classifiers. In P. Langley, editor, *Proceedings of the 17th International Conference on Machine Learning*, pages 1159–1166. Morgan Kaufmann, San Francisco, 2000.

D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In L. Saul S. Thrun and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press,Cambridge, Mass., 2004.

Ding-Xuan Zhou. The covering number in learning theory. *J. Complex.*, 18(3):739–767, 2002.

X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In T. Fawcett and N.Mishra, editors, *Proceedings of the 20th International Conference of Machine Learning*. AAAI Press, 2003.