On the Convergence of Spectral Clustering on Random Samples: the Normalized Case

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

[1] Max Planck Institute for Biological Cybernetics, Tübingen, Germany
{ulrike.luxburg, olivier.bousquet}@tuebingen.mpg.de
[2] The University of Chicago, Department of Computer Science
misha@cs.uchicago.edu

Abstract. Given a set of n randomly drawn sample points, spectral clustering in its simplest form uses the second eigenvector of the graph Laplacian matrix, constructed on the similarity graph between the sample points, to obtain a partition of the sample. We are interested in the question how spectral clustering behaves for growing sample size n. In case one uses the normalized graph Laplacian, we show that spectral clustering usually converges to an intuitively appealing limit partition of the data space. We argue that in case of the unnormalized graph Laplacian, equally strong convergence results are difficult to obtain.

1 Introduction

Clustering is a widely used technique in machine learning. Given a set of data points, one is interested in partitioning the data based on a certain similarity among the data points. If we assume that the data is drawn from some underlying probability distribution, which often seems to be the natural mathematical framework, the goal becomes to partition the probability space into certain regions with high similarity among points. In this setting the problem of clustering is two-fold:

- 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 can we reconstruct that optimal partition empirically on the finite sample?

Interestingly, while extensive literature exists on clustering and partitioning, to the best of our knowledge very few algorithms have been analyzed or shown to converge for increasing sample size. Some exceptions are the k-means algorithm (cf. Pollard, 1981), the single linkage algorithm (cf. Hartigan, 1981), and the clustering algorithm suggested by Niyogi and Karmarkar (2000). The goal of this paper is to 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). It has been used for load balancing (Van Driessche and Roose, 1995), parallel computations (Hendrickson and Leland, 1995), and VLSI design (Hagen and Kahng, 1992). Recently, Laplacian-based clustering algorithms have found success in applications to image segmentation (cf. Shi and Malik, 2000). Methods based on graph Laplacians have also been used for other problems in machine learning, including semi-supervised learning (cf. Belkin and Niyogi, to appear; Zhu et al., 2003). While theoretical properties of spectral clustering have been studied (e.g., Guattery and Miller (1998), Weiss (1999), Kannan et al. (2000), Meila and Shi (2001), also see Chung (1997) for a comprehensive theoretical treatment of the spectral graph theory), we do not know of any results discussing the convergence of spectral clustering or the spectra of graph Laplacians for increasing sample size. However for kernel matrices, the convergence of the eigenvalues and eigenvectors has already attracted some attention (cf. Williams and Seeger, 2000; Shawe-Taylor et al., 2002; Bengio et al., 2003).

2 Background and notations

Let $(\mathcal{X}, \mathrm{dist})$ be a metric space, $\mathcal{B}$ the Borel σ-algebra on $\mathcal{X}$, P a probability measure on $(\mathcal{X}, \mathcal{B})$, and $L_2(P) := L_2(\mathcal{X}, \mathcal{B}, P)$ the space of square-integrable functions. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a measurable, symmetric, non-negative function that computes the similarity between points in $\mathcal{X}$. For given sample points $X_1, ..., X_n$ drawn iid according to the (unknown) distribution P we denote the empirical distribution by P_n. We define the similarity matrix as $K_n := (k(X_i, X_j))_{i,j=1,...,n}$ and the degree matrix D_n as the diagonal matrix with diagonal entries $d_i := \sum_{j=1}^{n} k(X_i, X_j)$. The unnormalized discrete Laplacian matrix is defined as $L_n := D_n - K_n$. For symmetric and non-negative k, L_n is a positive semi-definite linear operator on $\mathbb{R}^n$. Let $a = (a_1, ..., a_n)$ the second eigenvector of L_n. Here, "second eigenvector" refers to the eigenvector belonging to the second smallest eigenvalue, where the eigenvalues $\lambda_1 \leq \lambda_2 ... \leq \lambda_n$ are counted *with multiplicity*. In a nutshell, spectral clustering in its simples form partitions the sample points $(X_i)_i$ into two (or several) groups by thresholding the second eigenvector of L_n: point X_i belongs to cluster 1 if $a_i > b$, and to cluster 2 otherwise, where $b \in \mathbb{R}$ is some appropriate constant. An intuitive explanation of why this works is discussed in Section 4.

Often, spectral clustering is also performed with a normalized version of the matrix L_n. Two common ways of normalizing are $L'_n := D_n^{-1/2} L_n D_n^{-1/2}$ or $L''_n := D_n^{-1} L_n$. The eigenvalues and eigenvectors of both matrices are closely related. Define the normalized similarity matrices $H'_n := D_n^{-1/2} K_n D_n^{-1/2}$ and $H''_n := D_n^{-1} K_n$. It can be seen by multiplying the eigenvalue equation $L'_n v = \lambda v$ from left with $D_n^{-1/2}$ that $v \in \mathbb{R}^n$ is eigenvector of L'_n with eigenvalue λ iff $D_n^{-1/2} v$ is eigenvector of L''_n with eigenvalue λ. Furthermore, rearranging the eigenvalue equations for L'_n and L''_n shows that $v \in \mathbb{R}^n$ is an eigenvector of L'_n with eigenvalue λ iff v is eigenvector of H'_n with eigenvalue $(1 - \lambda)$, and that $v \in \mathbb{R}^n$ is an eigenvector of L''_n with eigenvalue λ iff v is eigenvector of H''_n

with eigenvalue $(1-\lambda)$. Thus, properties about the spectrum of one of the matrices L_n', L_n'', H_n', or H_n'' can be reformulated for the three other matrices as well.

In the following we want to recall some definitions and facts from perturbation theory for bounded operators. The standard reference for general perturbation theory is Kato (1966), for perturbation theory in Hilbert spaces we also recommend Birman and Solomjak (1987) and Weidmann (1980), and Bhatia (1997) for finite-dimensional perturbation theory. We denote by $\sigma(T)$ the spectrum of a linear operator T. Its essential and discrete spectra are denoted by $\sigma_{\mathrm{ess}}(T)$ and $\sigma_{\mathrm{d}}(T)$, respectively.

Proposition 1 (Spectral and perturbation theory).

1. **Spectrum of a compact operator:** *Let T a compact operator on a Banach space. Then $\sigma(T)$ is at most countable and has at most one limit point, namely 0. If $0 \neq \lambda \in \sigma(T)$, then λ is an isolated eigenvalue with finite multiplicity. The spectral projection corresponding to λ coincides with the projection on the corresponding eigenspace.*
2. **Spectrum of a multiplication operator:** *For a bounded function $g \in L_\infty(P)$ consider the multiplication operator $M_g : L_2(P) \to L_2(P),\ f \mapsto gf$. M_g is a bounded linear operator whose spectrum coincides with the essential range of the multiplier g.*
3. **Perturbation of symmetric matrices:** *Let A and B be two symmetric matrices in $\mathbb{R}^{n\times n}$, and denote by $\|\cdot\|$ an operator norm on $\mathbb{R}^{n\times n}$. Then the Hausdorff distance $d(\sigma(A),\sigma(B))$ between the two spectra satisfies $d(\sigma(A),\sigma(B)) \leq \|A-B\|$. Let $\mu_1 > ... > \mu_k$ be the eigenvalues of A counted without multiplicity and $\mathrm{Pr}_1, ..., \mathrm{Pr}_k$ the projections on the corresponding eigenspaces. For $1 \leq r \leq k$ define the numbers*

$$\gamma_r(A) := \min\{|\mu_i - \mu_j|;\ 1 \leq i < j \leq r+1\}.$$

Assume that $\|B\| \leq \varepsilon$. Then for all $1 \leq l \leq r$ we have

$$\|\mathrm{Pr}_l(A+B) - \mathrm{Pr}_l(A)\| \leq 4\frac{\|B\|}{\gamma_r(A)}$$

(cf. Section VI.3 of Bhatia, 1997, Lemma A.1.(iii) of Koltchinskii, 1998, and Lemma 5.2. of Koltchinskii and Giné, 2000).
4. **Perturbation of bounded operators:** *Let $(T_n)_n$ and T be bounded operators on a Banach space E with $T_n \to T$ in operator norm, and λ an isolated eigenvalue of T with finite multiplicity. Then, for n large enough, there exist isolated eigenvalues $\lambda_n \in \sigma(T_n)$ such that $\lambda_n \to \lambda$, and the corresponding spectral projections converge in operator norm. The other way round, for a converging sequence $\lambda_n \in \sigma(T_n)$ of isolated eigenvalues with finite multiplicity, there exists an isolated eigenvalue $\lambda \in \sigma(T)$ with finite multiplicity such that $\lambda_n \to \lambda$ and the corresponding spectral projections converge in operator norm (cf. Theorems 3.16 and 2.23 in Kato, 1966).*

5. **Perturbation of the essential spectrum:** *Let A be a bounded and V a compact operator on some Banach space. Then $\sigma_{ess}(A+V) = \sigma_{ess}(A)$ (cf. Th. 5.35 in Kato, 1966, and Th. 9.1.3 in Birman and Solomjak, 1987).*

Finally we will need the following definition. A set $\mathcal{F}$ of real-valued functions on $\mathcal{X}$ is called a P-Glivenko-Cantelli class if

$$\sup_{f \in \mathcal{F}} \left| \int f dP_n - \int f dP \right| \to 0 \ P\text{-a.s.}$$

3 Convergence of the normalized Laplacian

The goal of this section is to prove that the first eigenvectors of the normalized Laplacian converge to the eigenfunctions of some limit operator on $L_2(P)$.

3.1 Definition of the integral operators

Let $d(x) := \int k(x,y) dP(y)$ the "true degree function" on $\mathcal{X}$, and $d_n(x) := \int k(x,y) dP_n(y)$ the empirical degree function. To ensure that $1/d$ is a bounded function we assume that there exists some constant l such that $d(x) > l > 0$ for all $x \in \mathcal{X}$. We define the normalized similarity functions

$$\begin{aligned} h_n(x,y) &:= k(x,y)/\sqrt{d_n(x) d_n(y)} \\ h(x,y) &:= k(x,y)/\sqrt{d(x) d(y)} \end{aligned} \tag{1}$$

and the operators

$$\begin{aligned} T_n &: L_2(P_n) \to L_2(P_n),\ T_n f(x) = \int h(x,y) f(y) dP_n(y) \\ T_n' &: L_2(P_n) \to L_2(P_n),\ T_n' f(x) = \int h_n(x,y) f(y) dP_n(y) \\ T &: L_2(P) \to L_2(P),\ T f(x) = \int h(x,y) f(y) dP(y). \end{aligned} \tag{2}$$

If k is bounded and $d > l > 0$, then all three operators are bounded, compact integral operators. Note that the scaling factors $1/n$ which are hidden in d_n and P_n cancel. Hence, because of the isomorphism between $L_2(P_n)$ and $\mathbb{R}^n$, the eigenvalues and eigenvectors of T_n' can be identified with the ones of the empirical similarity matrix H_n', and the eigenvectors and values of T_n with those of the matrix $H_n := (h(X_i, X_j))_{ij}$.

Our goal in the following will be to show that the eigenvectors of H_n' converge to those of the integral operator T. The first step will consist in proving that the operators T_n and T_n' converge to each other in operator norm. By perturbation theory results this will allow us to conclude that their spectra also become similar. The second step is to show that the eigenvalues and eigenvectors of T_n converge to those of T. This step uses results obtained in Koltchinskii (1998). Both steps together then will show that the first eigenvectors of the normalized Laplacian matrix converge to the first eigenfunctions of the limit operator T, and hence that spectral clustering converges.

3.2 T_n and T'_n converge to each other

Proposition 2 (d_n converges to d uniformly on the sample). *Let $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be bounded. Then $\max_{i=1,\dots,n} |d_n(X_i) - d(X_i)| \to 0$ a.s. for $n \to \infty$.*

Proof. With $M := \|k\|_\infty < \infty$ we have

$$\begin{aligned}\max_{i=1,\dots,n} |d_n(X_i) - d(X_i)| &= \max_{i=1,\dots,n} |\frac{1}{n}\sum_{j=1}^{n} k(X_i,X_j) - E_X k(X_i,X)| \\ &\le \frac{2M}{n} + \frac{n-1}{n}\max_i |\frac{1}{n-1}\sum_{j\neq i} k(X_i,X_j) - E_X k(X_i,X)|.\end{aligned}$$

For fixed $x \in \mathcal{X}$, the Hoeffding inequality yields

$$P\Big(|\frac{1}{n-1}\sum_{j\neq i} k(x,X_j) - E_X k(x,X)| > \varepsilon\Big) \le \exp(-M(n-1)\varepsilon^2).$$

The same is true conditionally on X_i if we replace x by X_i, because the random variable X_i is independent of X_j for $j \neq i$. Applying the union bound and taking expectations over X_i leads to

$$\begin{aligned}&P\big(\max_{i=1,\dots,n} |\frac{1}{n-1}\sum_{j\neq i} k(X_i,X_j) - E_X k(X_i,X)| > \varepsilon\big) \\ &\le \sum_{i=1}^{n} P\Big(|\frac{1}{n-1}\sum_{j\neq i} k(X_i,X_j) - E_X k(X_i,X)| > \varepsilon \;\Big|\; X_i\Big) \\ &\le n\exp(-M(n-1)\varepsilon^2).\end{aligned}$$

This shows the convergence of $\max_{i=1,\dots,n} |d_n(X_i) - d(X_i)| \to 0$ in probability. As the deviations decrease exponentially, the Borel-Cantelli lemma shows that this convergence also holds almost surely. ☺

Proposition 3 ($\|T'_n - T_n\|_{L_2(P_n)}$ converges to 0). *Let k a bounded similarity function. Assume that there exist constants $u > l > 0$ such that $u \ge d(x) \ge l > 0$ for all $x \in \mathcal{X}$. Then $\|T_n - T'_n\|_{L_2(P_n)} \to 0$ a.s. and $\|H_n - H'_n\|_n \to 0$ a.s., where $\|\cdot\|_n$ denotes the row sum norm for $n\times n$-matrices.*

Proof. By the Cauchy-Schwartz inequality,

$$\begin{aligned}&\|T_n - T'_n\|^2_{L_2(P_n)} = \sup_{\|f\|_{L_2(P_n)}\le 1} \int \left(\int (h_n(x,y) - h(x,y)) f(y) dP_n(y)\right)^2 dP_n(x) \\ &\le \sup_{\|f\|_{L_2(P_n)}\le 1} \int\int (h_n(x,y) - h(x,y))^2 dP_n(y) \int f^2(y) dP_n(y)\;\; dP_n(x) \\ &\le \int\int (h_n(x,y) - h(x,y))^2 dP_n(y) dP_n(x) \\ &\le \max_{i,j=1,\dots,n} |h_n(X_i,X_j) - h(X_i,X_j)|^2\end{aligned}$$

By Proposition 2 we know that for each $\varepsilon > 0$ there exists some N such that for all $n > N$, $|d_n(x) - d(x)| \leq \varepsilon$ for all $x \in \{X_1, ..., X_n\}$. Then

$$|d_n(x)d_n(y) - d(x)d(y)| \leq |d_n(x)d_n(y) - d(x)d_n(y)| + |d(x)d_n(y) - d(x)d(y)| \leq 2u\varepsilon,$$

which implies that $|\sqrt{d_n(x)d_n(y)} - \sqrt{d(x)d(y)}| \leq \sqrt{2u\varepsilon}$. This finally leads to

$$\left| \frac{1}{\sqrt{d_n(x)d_n(y)}} - \frac{1}{\sqrt{d(x)d(y)}} \right| = \left| \frac{\sqrt{d_n(x)d_n(y)} - \sqrt{d(x)d(y)}}{\sqrt{d_n(x)d_n(y)}\sqrt{d(x)d(y)}} \right| \leq \frac{\sqrt{2u\varepsilon}}{l(l - 2u\varepsilon)}$$

for all $x, y \in \{X_1, ..., X_n\}$. This shows that $\|T_n - T_n'\|$ converges to 0 almost surely. The statement for $\|H_n - H_n'\|$ follows by a similar argument. ☺

3.3 Convergence of T_n to T

Now we want to deal with the convergence of T_n to T. By the law of large numbers it is clear that $T_n f(x) \to Tf(x)$ for all $x \in \mathcal{X}$ and $f \in \mathcal{F}$. But this pointwise convergence is not enough to allow any conclusion about the convergence of the eigenvalues, let alone the eigenfunctions of the involved operators. On the other hand, the best convergence statement we can possibly think of would be convergence of T_n to T in operator norm. Here we have the problem that the operators T_n and T are not defined on the same spaces. One way to handle this is to relate the operators T_n, which are currently defined on $L_2(P_n)$, to some operators S_n on the space $L_2(P)$ such that their spectra are preserved. Then we would have to prove that S_n converges to T in operator norm. We believe that such a statement cannot be true in general. Intuitively, the reason for this is the following. Convergence in operator norm means uniform convergence on the unit ball of $L_2(P)$. Independent of the exact definition of S_n, the convergence of S_n to T in operator norm is closely related to the problem

$$\sup_{\|f\| \leq 1} \| \int k(x, y)f(y)dP_n(y) - \int k(x, y)f(y)dP(y) \ \| \stackrel{!}{\to} 0.$$

This statement would be true if the class $\mathcal{G} := \{k(x, \cdot)f(\cdot);\ x \in \mathcal{X}, \|f\| \leq 1\}$ was a P-Glivenko-Cantelli class, which is false in general. This can be made plausible by considering the special case $k \equiv 1$. Then the condition would be that the unit ball of $L_2(P)$ is a Glivenko-Cantelli class, which is clearly not the case for large enough $\mathcal{X}$. As a consequence, we cannot hope to achieve uniform convergence over the unit ball of $L_2(P)$.

A way out of this problem might be not to consider uniform convergence on the whole unit ball, but on a smaller subset of it. Something of a similar flavor has been proved in Koltchinskii (1998). To state his results we first have to introduce some more notation. For a function $f : \mathcal{X} \to \mathbb{R}$ denote its restriction to the sample points by $\tilde{f}$. Let $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a symmetric, measurable similarity function such that $E(h^2(X, Y)) < \infty$. This condition implies that the

integral operator T with kernel h is a Hilbert-Schmidt operator. Let $(\lambda_i)_{i\in I}$ its eigenvalues and $(\Phi_i)_{i\in I}$ a corresponding set of orthonormal eigenfunctions. To measure the distance between two countable sets $A = (a_i)_{i\in\mathbb{N}}$, $B = (b_i)_{i\in\mathbb{N}}$, we introduce the minimal matching distance $\delta(A, B) := \inf_\pi \sum_{i=1}^{\infty} a_i - b_{\pi(i)}$, where the infimum is taken over the set of all permutations π of $\mathbb{N}$. A more general version of the following theorem has been proved in Koltchinskii (1998).

Theorem 4 (Koltchinskii). *Let $(\mathcal{X}, \mathcal{B}, P)$ an arbitrary probability space, $h : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ a symmetric, measurable function such that $E(h^2(X, Y)) < \infty$ and $E(|h(X, X)|) < \infty$, and T_n and T the integral operators as defined in equation (2). Let $(\Phi_i)_{i\in I}$ the eigenfunctions of T, and let $\lambda \neq 0$ the r-th largest eigenvalue of T (counted without multiplicity). Denote by $\Pr$ and $\Pr_n$ the projections on the eigenspaces corresponding to the r-th largest eigenvalues of T and T_n, respectively. Then:*

1. *$\delta(\sigma(T_n), \sigma(T)) \to 0$ a.s.*
2. *Suppose that $\mathcal{G}$ is a class of measurable functions on $\mathcal{X}$ with a square-integrable envelope G with $\|G\|_{L_2(P)} \leq 1$, i.e. $|g(x)| \leq G(x)$ for all $g \in \mathcal{G}$. Moreover, suppose that for all $i \in I$, the set $\mathcal{G}\Phi_i := \{g\Phi_i;\ g \in \mathcal{G}\}$ is a P-Glivenko Cantelli class. Then*

$$\sup_{f,g\in\mathcal{G}} \left| \langle \Pr_n \tilde{f}, \tilde{g}\rangle_{L_2(P_n)} - \langle \Pr f, g\rangle_{L_2(P)} \right| \to 0 \text{ a.s. for } n \to \infty.$$

Coming back to the discussion from above, we can see that this theorem also does not state convergence of the spectral projections uniformly on the whole unit ball of $L_2(P)$, but only on some subset $\mathcal{G}$ of it. The problem that the operators T_n and T are not defined on the same space has been circumvented by considering bilinear forms instead of the operators themselves.

3.4 Convergence of the second eigenvectors

Now we have collected all ingredients to discuss the convergence of the second largest eigenvalue and eigenvector of the normalized Laplacian. To talk about convergence of eigenvectors only makes sense if the eigenspaces of the corresponding eigenvalues are one-dimensional. Otherwise there exist many different eigenvectors for the same eigenvalue. So multiplicity one is the assumption we make in our main result. In order to compare an eigenvector of the discrete operator T_n' and the corresponding eigenfunction of T, we can only measure how distinct they are on the points of the sample, that is by the $L_2(P_n)$-distance. However, as eigenvectors are only unique up to changing their orientations we will compare them only up to a change of sign.

Theorem 5 (Convergence of normalized spectral clustering). *Let $(\mathcal{X}, \mathcal{B}, P)$ a probability space, $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ a symmetric, bounded, measurable function, and $(X_i)_{i\in\mathbb{N}}$ a sequence of data points drawn iid from $\mathcal{X}$ according to*

P. Assume that the degree function satisfies $d(x) > l > 0$ for all $x \in \mathcal{X}$ and some constant $l \in \mathbb{R}$. Denote by $\lambda \neq 0$ the second largest eigenvalue of T (counted with multiplicity), and assume that it has multiplicity one. Let Φ be the corresponding eigenfunction, and Pr *the projection on Φ. Let λ_n, Φ_n and* Pr_n *the same quantities for T_n, and λ'_n, Φ'_n and Pr'_n the same for T'_n. Then there exists a sequence of signs $(a_n)_n$ with $a_n \in \{-1, +1\}$ such that $\|a_n\Phi'_n - \tilde{\Phi}\|_{L_2(P_n)} \to 0$ almost surely.*

Proof. The boundedness of k and $d(x) > l > 0$ imply that the normalized similarity function h is bounded. Hence, the operators T, T_n and T'_n are compact operators. By Proposition 1.1, their non-zero eigenvalues are isolated in their spectra, and their spectral projections correspond to the projections on the eigenspaces. Moreover, the boundedness of h implies $E(h^2(X,Y)) < \infty$ and $E|h(X,X)| < \infty$. Theorem 4 shows $\lambda_n \to \lambda$ for $n \to \infty$, and choosing $\mathcal{F} = \{\Phi\}$ we get

$$\langle \Phi_n, \tilde{\Phi}\rangle^2 = \langle \langle \Phi_n, \tilde{\Phi}\rangle \Phi_n, \tilde{\Phi}\rangle = \langle \mathrm{Pr}_n\, \tilde{\Phi}, \tilde{\Phi}\rangle \to \langle \mathrm{Pr}\, \Phi, \Phi\rangle = \langle \Phi, \Phi\rangle = 1.$$

The eigenfunctions Φ and Φ_n are normalized to 1 in their respective spaces. By the law of large numbers, we also have $\|\tilde{\Phi}\|_{L_2(P_n)} \to 1$ a.s. Hence, $\langle \Phi_n, \tilde{\Phi}\rangle \to 1$ or -1 implies the $L_2(P_n)$-convergence of Φ_n to Φ up to a change of sign.

Now we have to compare λ'_n to λ_n and Φ'_n to Φ_n. In Proposition 3 we showed that $\|T'_n - T_n\| \to 0$ a.s., which according to Proposition 1.3 implies the convergence of $\lambda'_n - \lambda_n$ to zero. Theorem 4 implies the convergence of $\lambda_n - \lambda$ to zero. For the convergence of the eigenfunctions, recall the definition of γ_r in Proposition 1.3. As the eigenvalues of T are isolated we have $\gamma_2(T) > 0$, and by the convergence of the eigenvalues we also get $|\gamma_2(T'_n) - \gamma_2(T)| \to 0$. Hence, $\gamma(T'_n)$ is bounded away from 0 simultaneously for all large n. Moreover, we know by Proposition 3 that $\|T'_n - T_n\| \to 0$ a.s. Proposition 1.3 now shows the convergence of the spectral projections $\|\mathrm{Pr}'_n - \mathrm{Pr}_n\| \to 0$ a.s. This implies in particular that

$$\sup_{\|v\|\leq 1} \langle v, (\mathrm{Pr}_n - \mathrm{Pr}'_n)v\rangle \to 0 \text{ and thus } \sup_{\|v\|\leq 1} |\langle v, \Phi_n\rangle^2 - \langle v, \Phi'_n\rangle^2| \to 0.$$

Since $|a^2 - b^2| = |a - b||a + b|$, we get the convergence of Φ_n to Φ up to a change of sign on the sample, as stated in the theorem. This completes the proof. ☺

Let us briefly discuss the assumptions of Theorem 5. The symmetry of k is a standard requirement in spectral clustering as it ensures that all eigenvalues of the Laplacian are real-valued. The assumption that the degree function is bounded away from 0 prevents the normalized Laplacian from getting unbounded, which is also desirable in practice. This condition will often be trivially satisfied as the second standard assumption of spectral clustering is the non-negativity of k (as it ensures that the eigenvalues of the Laplacian are non-negative). An important assumption in Theorem 5 which is not automatically satisfied is that the second eigenvalue has multiplicity one. But note that if this assumption is not satisfied, spectral clustering will produce more or less arbitrary

results anyway, as the second eigenvector is no longer unique. It then depends on the actual implementation of the algorithm which of the infinitely many eigenvectors corresponding to the second eigenvalue is picked, and the result will often be unsatisfactory. Finally, note that even though Theorem 5 is stated in terms of the second eigenvalue and eigenvector, analogous statements are true for higher eigenvalues, and also for spectral projections on finite dimensional eigenspaces with dimension larger than 1.

To summarize, all assumptions in Theorem 5 are already important for successful applications of spectral clustering on a finite sample. Theorem 5 now shows that with no additional assumptions, the convergence of normalized spectral clustering to a limit clustering on the whole data space is guaranteed.

4 Interpretation of the limit partition

Now we want to investigate whether the limit clustering partitions the data space $\mathcal{X}$ in a desirable way. In this section it will be more convenient to consider the normalized similarity matrix H_n'' instead of H_n' as it is a stochastic matrix. Hence we consider the normalized similarity function $g(x,y) := k(x,y)/d(x)$, its empirical version $g_n(x,y) := k(x,y)/d_n(x)$, and the integral operators

$$R_n'' : L_2(P_n) \to L_2(P_n),\ R_n'' f(x) = \int g_n(x,y) f(y) dP_n(y)$$
$$R : L_2(P) \to L_2(P),\ Rf(x) = \int g(x,y) f(y) dP(y).$$

The spectrum of R_n'' coincides with the spectrum of H_n'', and by the one-to-one relationships between the spectra of H_n'' and H_n' (cf. Section 2), the convergence stated in Theorem 5 for T_n' and T holds analogously for the operators R_n'' and R.

Let us take a step back and reflect what we would like to achieve with spectral clustering. The overall goal in clustering is to find a partition of $\mathcal{X}$ into two (or more) disjoint sets $\mathcal{X}_1$ and $\mathcal{X}_2$ such that the similarity between points from the same set is high while the similarity between points from different sets is low. Assuming that such a partition exists, how does the operator R look like? Let $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ be a partition of the space $\mathcal{X}$ into two disjoint, measurable sets such that $P(\bar{\mathcal{X}}_1 \cap \bar{\mathcal{X}}_2) = 0$. As σ-algebra on $\mathcal{X}_i$ we use the restrictions $\mathcal{B}_i := \{B \cap \mathcal{X}_i;\ B \in \mathcal{B}\}$ of the Borel σ-algebra $\mathcal{B}$ on $\mathcal{X}$. Define the measures P_i as the restrictions of P to $\mathcal{B}_i$. Now we can identify the space $L_2(\mathcal{X}, \mathcal{B}, P)$ with the direct sum $L_2(\mathcal{X}_1, \mathcal{B}_1, P_1) \oplus L_2(\mathcal{X}_2, \mathcal{B}_2, P_2)$. Each function $f \in L_2(\mathcal{X})$ corresponds to a tuple $(f_1, f_2) \in L_2(\mathcal{X}_1) \oplus L_2(\mathcal{X}_2)$, where $f_i : \mathcal{X}_i \to \mathbb{R}$ is the restriction of f to $\mathcal{X}_i$. The operator R can be identified with the matrix $\begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}$ acting on $L_2(\mathcal{X}_1, \mathcal{B}_1, P_1) \oplus L_2(\mathcal{X}_2, \mathcal{B}_2, P_2)$. We denote by d_i the restriction of d to $\mathcal{X}_i$ and by g_{ij} the restriction of g to $\mathcal{X}_i \times \mathcal{X}_j$. With these notations, the operators R_{ij} for

$i, j = 1, 2$ are defined as

$$R_{ij} : L_2(\mathcal{X}_j) \to L_2(\mathcal{X}_i),\ R_{ij} f_j(x) = \int g_{ij}(x, y) f_j(y) dP_j(y).$$

Now assume that our space is ideally clustered, that is the similarity function satisfies $k(x_1, x_2) = 0$ for all $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$, and $k(x_i, x_i') > 0$ for $x_i, x_i' \in \mathcal{X}_1$ or $x_i, x_i' \in \mathcal{X}_2$. Then the operator R has the form $\begin{pmatrix} R_{11} & 0 \\ 0 & R_{22} \end{pmatrix}$. It has eigenvalue 1 with multiplicity 2, and the corresponding eigenspace is spanned by the vectors $(\mathbb{1}, \mathbb{0})$ and $(\mathbb{0}, \mathbb{1})$. Hence, all eigenfunctions corresponding to eigenvalue 1 are piecewise constant on the sets $\mathcal{X}_1, \mathcal{X}_2$, and the eigenfunction orthogonal to the function $(\mathbb{1}, \mathbb{1})$ has opposite sign on both sets. Thresholding the second eigenfunction will recover the true clustering $\mathcal{X}_1 \cup \mathcal{X}_2$. When we interpret the function g as a Markov transition kernel, the operator R describes a Markov diffusion process on $\mathcal{X}$. We see that the clustering constructed by its second eigenfunction partitions the space into two sets such that diffusion takes place within the sets, but not between them.

The same reasoning also applies to the finite sample case, cf. Meila and Shi (2001), Weiss (1999), and Ng et al. (2001). We split the finite sample space $\{X_1, ..., X_n\}$ into the two sets $\mathcal{X}_{i,n} := \{X_1, ..., X_n\} \cap \mathcal{X}_i$, and define

$$R_{ij,n} : L_2(\mathcal{X}_{j,n}) \to L_2(\mathcal{X}_{i,n}),\ R_{ij,n} f_j(x) = \int g_{ij,n}(x, y) f_j(y) dP_{j,n}(y).$$

According to Meila and Shi (2001), spectral clustering tries to find a partition such that the probability of staying within the same cluster is large while the probability of going from one cluster into another one is low (Meila and Shi, 2001). So both in the finite sample case and in the limit case a similar interpretation applies. This shows in particular that the limit clustering accomplishes the goal of clustering to partition the space into sets such that the within similarity is large and the between similarity is low.

In practice, the operator R will usually be irreducible, i.e. there will exist no partition such that the operators R_{12} and R_{21} vanish. Then the goal will be to find a partition such that the norms of R_{12} and R_{21} are as small as possible, while the norms of R_{ii} should be reasonably large. If we find such a partition, then the operators $\begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}$ and $\begin{pmatrix} R_{11} & 0 \\ 0 & R_{22} \end{pmatrix}$ are close in operator norm and according to perturbation theory have a similar spectrum. Then the partition constructed by R will be approximately the same as the one constructed by $\begin{pmatrix} R_{11} & 0 \\ 0 & R_{22} \end{pmatrix}$, which is the partition $\mathcal{X}_1 \cup \mathcal{X}_2$.

The convergence results in Section 3 show that the first eigenspaces of R_n converge to the first eigenspaces of the limit operator R. This statement can be further strengthened by proving that each of the four operators $R_{ij,n}$ converges

to its limit operator R_{ij} compactly, which can be done by methods from von Luxburg et al.. As a consequence, also the eigenvalues and eigenspaces of the single operators $R_{ij,n}$ converge. This statement is even sharper than the convergence statement of R_n to R. It shows that for any fixed partition of $\mathcal{X}$, the *structure* of the operator R_n is preserved when taking the limit. This means that a partition that has been constructed on the finite sample such that the diffusion between the two sets is small also keeps this property when we take the limit.

5 Convergence of the unnormalized Laplacian

So far we always considered the *normalized* Laplacian matrix. The reason is that this case is inherently simpler to treat than the unnormalized case. In the unnormalized case, we have to study the operators

$$U_n f(x) := \int k(x,y)(f(x) - f(y))dP_n(y) = d_n(x)f(x) - \int k(x,y)f(y)dP_n(y)$$
$$U f(x) := \int k(x,y)(f(x) - f(y))dP(y) = d(x)f(x) - \int k(x,y)f(y)dP(y).$$

It is clear that U_n is the operator corresponding to the unnormalized Laplacian $\frac{1}{n}L_n$, and U is its pointwise limit operator for $n \to \infty$. In von Luxburg et al. we show that under mild assumptions, U_n converges to U compactly. Compact convergence is a type of convergence which is a bit weaker than operator norm convergence, but still strong enough to ensure the convergence of eigenvalues and spectral projections (Chatelin, 1983). But there is a big problem related to the structure of the operators U_n and U. Both consist of a difference of two operators, a bounded multiplication operator and a compact integral operator. This is bad news, as multiplication operators are never compact. To the contrary, the spectrum of a multiplication operator consists of the whole range of the multiplier function (cf. Proposition 1.2). Hence, the spectrum of U consists of an essential spectrum which coincides with the range of the degree function, and possibly some discrete spectrum of isolated eigenvalues (cf. Proposition 1.5).

This has the consequence that although we know that U_n converges to U in a strong sense, we are not able to conclude anything about the convergence of the second eigenvectors. The reason is that perturbation theory only allows to state convergence results for *isolated* parts of the spectra. So we get that the essential spectrum of U_n converges to the essential spectrum of U. Moreover, if $\sigma(U)$ has a non-empty discrete spectrum, then we can also state convergence of the eigenvalues and eigenspaces belonging to the discrete spectrum. But unfortunately, it is impossible to conclude anything about the convergence of eigenvalues that lie *inside* the essential spectrum of U. In von Luxburg et al. we actually construct an example of a space $\mathcal{X}$ and a similarity function k such that all non-zero eigenvalues of the unnormalized Laplacian indeed lie inside the essential spectrum of U. Now we have the problem that given a finite sample, we cannot detect whether

the second eigenvalue of the limit operator will lie inside or outside the essential spectrum of U, and hence we cannot guarantee that the second eigenvectors of the unnormalized Laplacian matrices converge. All together this means that although we have strong convergence results for U_n, without further knowledge we are not able to draw any useful conclusion concerning the second eigenvalues.

On the other hand, in case we can guarantee the convergence of unnormalized spectral clustering (i.e., if the second eigenvalue is not inside the essential spectrum), then the limit partition in the unnormalized case can be interpreted similarly to the normalized case by taking into account the form of the operator U on $L_2(\mathcal{X}_1, \mathcal{B}_1, P_1) \oplus L_2(\mathcal{X}_2, \mathcal{B}_2, P_2)$. Similar to above, it is composed of a matrix of four operators $(U_{ij})_{i,j=1,2}$ defined as

$$U_{ii} : L_2(\mathcal{X}_i) \to L_2(\mathcal{X}_i),\ U_{ii} f_i(x) = d_i(x) f_i(x) - \int k_{ii}(x,y) f_i(y) dP_i(y)$$

$$U_{ij} : L_2(\mathcal{X}_j) \to L_2(\mathcal{X}_i),\ U_{ij} f_j(x) = -\int k_{ij}(x,y) f_j(y) dP_j(y) \qquad (\text{ for } i \neq j).$$

We see that the off-diagonal operators U_{ij} for $i \neq j$ only consist of integral operators, whereas the multiplication operators only appear in the diagonal operators U_{ii}. Thus the operators U_{ij} for $i \neq j$ can also be seen as diffusion operators, and the same interpretation as in the normalized case is possible. If there exists a partition such that $k(x_1, x_2) = 0$ for all $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$, then the second eigenfunction is constant on both parts, and thresholding this eigenfunction will recover the "true" partition. Thus, also in the unnormalized case the goal of spectral clustering is to find partitions such that the norms of the off-diagonal operators is small and the norms of the diagonal operators are large. This holds both in the discrete case and in the limit case, but only if the second eigenvalue of U is not inside the range of the degree function.

To summarize, from a technical point of view the eigenvectors of the unnormalized Laplacian are more unpleasant to deal with than the normalized ones, as the limit operator has a large essential spectrum in which the interesting eigenvalues could be contained. But if the second eigenvalue of the limit operator is isolated, some kind of diffusion interpretation is still possible. This means that if unnormalized spectral clustering converges, then it converges to a sensible limit clustering.

6 Discussion

We showed in Theorem 5 that the second eigenvector of the normalized Laplacian matrix converges to the second eigenfunction of some limit operator almost surely. The assumptions in this theorem are usually satisfied in practical applications. This allows to conclude that in the normalized case, spectral clustering converges to some limit partition of the whole space which only depends on the

similarity function k and the probability distribution P. We also gave an explanation of how this partition looks like in terms of a diffusion process on the data space. Intuitively, the limit partition accomplishes the objective of clustering, namely to divide the space into sets such that the similarity within the sets is large and the similarity between the sets is low.

The methods we used to prove the convergence in case of the normalized Laplacian fail in the unnormalized case. The reason is that the limit operator in the unnormalized case is not compact and has a large essential spectrum. Convergence of the second eigenvector in the unnormalized case can be proved with different methods using collectively compact convergence of linear operators, but only under strong assumptions on the spectrum of the limit operator which are not always satisfied in practice (cf. von Luxburg et al.). However, if these assumptions are satisfied, then the limit clustering partitions the data space in a reasonable way. In practice, the fact that the unnormalized case seems much more difficult than the normalized case might serve as an indication that the normalized case of spectral clustering should be preferred.

The observations in Section 4 allow to make some more suggestions for the practical application of spectral clustering. According to the diffusion interpretation, it seems possible to to construct a criterion to evaluate the goodness of the partition achieved by spectral clustering. For a good partition, the off-diagonal operators $R_{12,n}$ and $R_{21,n}$ should have a small norm compared to the norm of the diagonal matrices $R_{11,n}$ and $R_{22,n}$, which is easy to check in practical applications. It will be a topic for future investigations to work out this idea in detail.

There are many open questions related to spectral clustering which have not been addressed in our work so far. The most obvious one is the question about the speed of convergence and the concentration of the limit results. Results in this direction would enable us to make confidence predictions about how close the clustering on the finite sample is to the "true" clustering proposed by the limit operator.

This immediately raises a second question: Which relations are there between the limit clustering and the geometry of the data space? For certain similarity functions such as the Gaussian kernel $k_t(x, y) = \exp(-\|x - y\|^2/t)$, it has been established that there is a relationship between the operator T and the Laplace operator on $\mathbb{R}^n$ (Bousquet et al., 2004) or the Laplace-Beltrami operator on manifolds (Belkin, 2003). Can this relationship also be extended to the eigenvalues and eigenfunctions of the operators?

There are also more technical questions related to our approach. The first one is the question which space of functions is the "natural" space to study spectral clustering. The space $L_2(P)$ is a large space and is likely to contain all eigenfunctions we might be interested in. On the other hand, for "nice" similarity

functions the eigenfunctions are continuous or even differentiable, thus $L_2(P)$ might be too general to discuss relevant properties such as relations to continuous Laplace operators. Moreover, we want to use functions which are pointwise defined, as we are interested in the value of the function at specific data points. But of all spaces, the functions in L_p-spaces do not have this property.

Another question concerns the type of convergence results we should prove. In this work, we fixed the similarity function k and considered the limit for $n \to \infty$. As a next step, the convergence of the limit operators with respect to some kernel parameters, such as the kernel width t for the Gaussian kernel, can be studied as in the works of Bousquet et al. (2004) and Belkin (2003). But it seems more appropriate to take limits in t and n simultaneously. This might reveal other important aspects of spectral clustering, for example how the kernel width should scale with n.

Bibliography

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

M. Belkin and P. Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine Learning*, to appear. Available at http://people.cs.uchicago.edu/~misha.

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.

R. Bhatia. *Matrix Analysis*. Springer, New York, 1997.

M. Birman and M. Solomjak. *Spectral theory of self-adjoint operators in Hilbert space*. Reidel Publishing Company, Dordrecht, 1987.

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.

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

Fan R. K. 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.

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

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.

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.

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

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, 2001.

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.

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.

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, O. Bousquet, and M. Belkin. On the convergence of spectral clustering on random samples: the unnormalized case. to be submitted, available at http://www.kyb.tuebingen.mpg.de/~ule.

J. Weidmann. *Linear Operators in Hilbert spaces*. Springer, New York, 1980.

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.

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.