Journal of Machine Learning Research 12 (2011) 349-380 Submitted 2/10; Revised 4/10; Published 2/11

Operator Norm Convergence of Spectral Clustering on Level Sets

Bruno Pelletier BRUNO.PELLETIER@UNIV-RENNES2.FR
Department of Mathematics
IRMAR – UMR CNRS 6625
Université Rennes II
Place du Recteur Henri Le Moal, CS 24307
35043 Rennes Cedex, France

Pierre Pudlo PIERRE.PUDLO@UNIV-MONTP2.FR
I3M: Institut de Mathématiques et de Modélisation de Montpellier – UMR CNRS 5149
Université Montpellier II, CC 051
Place Eugène Bataillon
34095 Montpellier Cedex 5, France

Editor: Ulrike von Luxburg

Abstract

Following Hartigan (1975), a cluster is defined as a connected component of the t-level set of the underlying density, that is, the set of points for which the density is greater than t. A clustering algorithm which combines a density estimate with spectral clustering techniques is proposed. Our algorithm is composed of two steps. First, a nonparametric density estimate is used to extract the data points for which the estimated density takes a value greater than t. Next, the extracted points are clustered based on the eigenvectors of a graph Laplacian matrix. Under mild assumptions, we prove the almost sure convergence in operator norm of the empirical graph Laplacian operator associated with the algorithm. Furthermore, we give the typical behavior of the representation of the data set into the feature space, which establishes the strong consistency of our proposed algorithm.

Keywords: spectral clustering, graph, unsupervised classification, level sets, connected components

1. Introduction

The aim of data clustering, or unsupervised classification, is to partition a data set into several homogeneous groups relatively separated one from each other with respect to a certain distance or notion of similarity. There exists an extensive literature on clustering methods, and we refer the reader to Anderberg (1973), Hartigan (1975) and McLachlan and Peel (2000), Chapter 10 in Duda et al. (2000), and Chapter 14 in Hastie et al. (2001) for general materials on the subject. In particular, popular clustering algorithms, such as Gaussian mixture models or k-means, have proved useful in a number of applications, yet they suffer from some internal and computational limitations. Indeed, the parametric assumption at the core of mixture models may be too stringent, while the standard k-means algorithm fails at identifying complex shaped, possibly non-convex, clusters.

The class of *spectral clustering* algorithms is presently emerging as a promising alternative, showing improved performance over classical clustering algorithms on several benchmark problems and applications (see, e.g., Ng et al., 2002; von Luxburg, 2007). An overview of spectral clustering algorithms may be found in von Luxburg (2007), and connections with kernel methods are exposed in Fillipone et al. (2008). The spectral clustering algorithm amounts at embedding the data into a feature space by using the eigenvectors of the similarity matrix in such a way that the clusters may be separated using simple rules, for example, a separation by hyperplanes. The core component of the spectral clustering algorithm is therefore the similarity matrix, or certain normalizations of it, generally called graph Laplacian matrices; see Chung (1997). Graph Laplacian matrices may be viewed as discrete versions of bounded operators between functional spaces. The study of these operators has started out recently with the works by Belkin et al. (2004); Belkin and Niyogi (2005), Coifman and Lafon (2006), Nadler et al. (2006), Koltchinskii (1998), Giné and Koltchinskii (2006), Hein et al. (2007) and Rosasco et al. (2010), among others.

In the context of spectral clustering, the convergence of the empirical graph Laplacian operators has been established in von Luxburg et al. (2008). Their results imply the existence of an asymptotic partition of the support of the underlying distribution of the data as the number of samples goes to infinity. However this theoretical partition results from a partition in a feature space, that is, it is the pre-image of a partition of the feature space by the embedding mapping. Therefore interpreting the asymptotic partition with respect to the underlying distribution of the data remains largely an open and challenging question. Similar interpretability questions also arise in the related context of kernel methods where the data is embedded in a feature space. For instance, while it is well-known that the popular k-means clustering algorithm leads to an optimal quantizer of the underlying distribution (MacQueen, 1967; Pollard, 1981; Linder, 2002), "kernelized" versions of the k-means algorithm allow to separate groups using nonlinear decision rules but are more difficult to interpret.

The rich variety of clustering algorithms raises the question of the definition of a cluster, and as pointed out in von Luxburg and Ben-David (2005) and in García-Escudero et al. (2008), there exists many such definitions. Among these, perhaps the most intuitive and precise definition of a cluster is the one introduced by Hartigan (1975). Suppose that the data is drawn from a probability density f on $\mathbb{R}^d$ and let t be a positive number in the range of f. Then a cluster in the sense of Hartigan (1975) is a connected component of the upper t-level set

$$\mathcal{L}(t) = \left\{x \in \mathbb{R}^d \,:\, f(x) \geq t\right\}.$$

This definition has several advantages. First, it is geometrically simple. Second, it offers the possibility of filtering out possibly meaningless clusters by keeping only the observations falling in a region of high density. This proves useful, for instance, in the situation where the data exhibits a cluster structure but is contaminated by a uniform background noise.

In this context, the level t should be considered as a resolution level for the data analysis. For instance, when the threshold t is taken equal to 0, the groups in the sense of Hartigan (1975) are the connected components of the support of the underlying distribution, while as t increases, the clusters concentrate in a neighborhood of the principal modes of the density f. Several clustering algorithms deriving from Hartigan's definition have been introduced building. In Cuevas et al. (2000, 2001), and in the related work by Azzalini and Torelli (2007), clustering is performed by estimating the connected components of $\mathcal{L}(t)$. Hartigan's definition is also used in Biau et al. (2007) to define an estimate of the number of clusters based on an approximation of the level set by a neighborhood graph.

In the present paper, we adopt the definition of a cluster of Hartigan (1975), and we propose and study a spectral clustering algorithm on estimated level sets. The algorithm is composed of two operations. Using the sample $X_1, \ldots, X_n$ of vectors of $\mathbb{R}^d$, we first construct a nonparametric density estimate $\hat{f}_n$ of the unknown density f. Next, given a positive number t, this estimate is used to extract those observations for which the estimated density exceeds the fixed threshold, that is, the observations for which $\hat{f}_n(X_i) \geq t$. In the second step of the algorithm, we perform a spectral clustering of the extracted points. The remaining data points are then left unlabeled.

Our proposal is to study the asymptotic properties of this algorithm. In the whole study, the density estimate $\hat{f}_n$ is arbitrary but supposed consistent, and the threshold t is fixed in advance. For the spectral clustering part of the algorithm, we consider the setting where the kernel function, or similarity function, between any two pairs of observations is non negative and with a compact support of diameter $2h$, for some fixed positive real number h. Our contribution contain two sets of results.

In the first set of results, we establish the almost-sure convergence in operator norm of the empirical graph Laplacian on the estimated level set. In von Luxburg et al. (2008), the authors prove the collectively compact convergence of the empirical operator, acting on the Banach space of continuous functions on some compact set. Finite sample bounds in Hilbert-Schmidt norms on Sobolev spaces are obtained in the paper by Rosasco et al. (2010). In our result, the empirical operator is acting on a Banach subspace of the Holder space $C^{0,1}$ of Lipschitz functions, which we equip with a Sobolev norm. This operator norm convergence is more amenable than the slightly weaker notion of convergence established in von Luxburg et al. (2008), and holds for any value of the scale parameter h, but the functional space that we consider is smaller. As in the related works referenced above, the operator norm convergence is derived using results from the theory of empirical processes to prove that certain classes of functions satisfy a uniform law of large numbers. We also rely on geometrical auxiliary results to obtain the convergence of the preprocessing step of the algorithm. Under mild regularity assumptions, we use the fact that the topology of the level set $\mathcal{L}(t)$ changes only when the threshold t passes a critical value of f. This allow us to define random graph Laplacian operators acting on a fixed space of functions, with large probability.

In the second set of results, we study the convergence of the spectrum of the empirical operator, as a corollary of the operator norm convergence. Depending on the values of the scale parameter h, we characterize the properties of the asymptotic partition induced by the clustering algorithm. First, we assume that h is lower than the minimal distance between any two connected components of the t-level set. Under this condition, we prove that the embedded data points concentrate on several isolated points, each of whose corresponds to a connected component of the level set, that is, observations belonging to the same connected component of the level set are mapped onto the same point in the feature space. As a consequence, in the asymptotic regime, any reasonable clustering algorithm applied on the transformed data partitions the observations according to the connected components of the level set. In this sense, recalling Hartigan's (1975) definition of a cluster, these results imply that the proposed algorithm is strongly consistent and that, asymptotically, observations of $\mathcal{L}(t)$ are assigned to the same cluster if and only if they fall in the same connected component of $\mathcal{L}(t)$. These properties follow from the ones of the continuous (i.e., population version) operator, which we establish by using arguments related to a Markov chain on a general state space. The underlying fact is that the normalized empirical graph Laplacian defines a random walk on the extracted observations, which converges to a random walk on $\mathcal{L}(t)$. Then, asymptotically, when the scale parameter is lower than the minimal distance between the connected components of $\mathcal{L}(t)$, this random walk cannot jump from one connected component to one another. Next, by exploiting the continuity of the operators in the scale parameter h, we obtain similar consistency results when h

is slightly greater than the minimal distance between two connected components of $\mathcal{L}(t)$. In this case, the embedded data points concentrates in several non-overlapping cubes, each of whose corresponds to a connected component of $\mathcal{L}(t)$. This result holds whenever h is smaller than a certain critical value h_{max}, which depends only on the underlying density f.

Finally, let us note that our consistency results hold for any value of the threshold t different from a critical value of the density f, which assume to be twice continuously differentiable. Under the stronger assumption that f is p times continuously differentiable, with $p \geq d$, Sard's lemma imply that the set of critical values of f has Lebesgue measure 0, so that the consistency would hold for almost all t. The special limit case $t = 0$ corresponds to performing a clustering on all the observations, and our results imply the convergence of the clustering to the partition of the support of the density into its connected components, for a suitable choice of the scale parameter. The proofs could be simplified in this setting, though, since no pre-processing step would be needed. Let us mention that this asymptotic partition could also be derived from the results in von Luxburg et al. (2008). At last, we obtain consistency in the sense of Hartigan's definition when the correct number of clusters is requested, which corresponds to the number of connected components of $\mathcal{L}(t)$, and when the similarity function has a compact support . Hence several questions remain largely open which are discussed further in the paper.

The paper is organized as follows. In Section 2, we start by introducing the necessary notations and assumptions. Then we define the spectral clustering algorithm on estimated level sets, and we follow by introducing the functional operators associated with the algorithm. In Section 3, we study the almost-sure convergence in operator norm of the random operators, starting with the unnormalized empirical graph Laplacian operator. The main convergence result of the normalized operator is stated in Theorem 4. Section 4 contains the second set of results on the consistency of the clustering algorithm. We start by studying the properties of the limit operator in the case where the scale parameter h is lower than the minimal distance between two connected components of $\mathcal{L}(t)$. The convergence of the spectrum, and the consistency of the algorithm, is then stated in Theorem 7. This result is extended in Theorem 10 to allow for larger values of h. We conclude this section with a discussion on possible extensions and open problems. The proofs of these theorems rely on several auxiliary technical lemmas which are collected in Sections 5. Finally, to make the paper self contained, materials and some facts from the geometry of level sets, functional analysis, and Markov chains are exposed in Appendices A, B, and C, respectively, at the end of the paper.

2. Spectral Clustering Algorithm

In this section we give a description of the spectral clustering algorithm on level sets that is suitable for our theoretical analysis.

2.1 Mathematical Setting and Assumptions

Let $\{X_i\}_{i\geq 1}$ be a sequence of i.i.d. random vectors in $\mathbb{R}^d$, with common probability measure μ. Suppose that μ admits a density f with respect to the Lebesgue measure on $\mathbb{R}^d$. The *t-level set* of f is denoted by $\mathcal{L}(t)$, that is,

$$\mathcal{L}(t) := \left\{x \in \mathbb{R}^d \,:\, f(x) \geq t\right\},$$

for all positive level t, and given $a \leq b$, $\mathcal{L}_a^b$ denotes the set $\{x \in \mathbb{R}^d : a \leq f(x) \leq b\}$. The differentiation operator with respect to x is denoted by D. We assume that f satisfies the following conditions.

> **Assumption 1.** *(i)* f is of class $\mathcal{C}^2$ on $\mathbb{R}^d$; *(ii)* $\|D_x f\| > 0$ on the set $\{x \in \mathbb{R}^d : f(x) = t\}$; *(iii)* f, Df, and $D^2 f$ are uniformly bounded on $\mathbb{R}^d$.

Note that under Assumption 1, $\mathcal{L}(t)$ is compact whenever t belongs to the interior of the range of f. Moreover, $\mathcal{L}(t)$ has a finite number ℓ of connected components $\mathcal{C}_j$, $j = 1, \ldots, \ell$. For ease of notation, the dependence of $\mathcal{C}_j$ on t is omitted. The minimal distance between the connected components of $\mathcal{L}(t)$ is denoted by d_{min}, that is,

$$d_{min} := \inf_{i \neq j} \operatorname{dist}\big(\mathcal{C}_i, \mathcal{C}_j\big). \tag{1}$$

Let $\widehat{f}_n$ be a consistent density estimate of f based on the random sample $X_1, \ldots, X_n$. The t-level set of $\widehat{f}_n$ is denoted by $\mathcal{L}_n(t)$, that is,

$$\mathcal{L}_n(t) := \big\{x \in \mathbb{R}^d : \widehat{f}_n(x) \geq t\big\}.$$

Let $J(n)$ be the set of integers defined by

$$J(n) := \big\{j \in \{1, \ldots, n\} : \widehat{f}_n(X_j) \geq t\big\}.$$

The cardinality of $J(n)$ is denoted by $j(n)$.

Let $k : \mathbb{R}^d \to \mathbb{R}_+$ be a fixed function. The unit ball of $\mathbb{R}^d$ centered at the origin is denoted by B, and the ball centered at $x \in \mathbb{R}^d$ and of radius r is denoted by $x + rB$. We assume throughout that the function k satisfies the following set of conditions.

> **Assumption 2.** *(i)* k is of class $\mathcal{C}^2$ on $\mathbb{R}^d$; *(ii)* the support of k is B; *(iii)* k is uniformly bounded from below on $B/2$ by some positive number; and *(iv)* $k(-x) = k(x)$ for all $x \in \mathbb{R}^d$.

Let h be a positive number. We denote by $k_h : \mathbb{R}^d \to \mathbb{R}_+$ the map defined by $k_h(u) := k(u/h)$.

2.2 Algorithm

The first ingredient of our algorithm is the *similarity matrix* $\mathbf{K}_{n,h}$ whose elements are given by

$$\mathbf{K}_{n,h}(i, j) := k_h(X_j - X_i),$$

and where the integers i and j range over the random set $J(n)$. Hence $\mathbf{K}_{n,h}$ is a random matrix indexed by $J(n) \times J(n)$, whose values depend on the function k_h, and on the observations X_j lying in the estimated level set $\mathcal{L}_n(t)$. Next, we introduce the diagonal *normalization matrix* $\mathbf{D}_{n,h}$ whose diagonal entries are given by

$$\mathbf{D}_{n,h}(i, i) := \sum_{j \in J(n)} \mathbf{K}_{n,h}(i, j), \quad i \in J(n).$$

Note that the diagonal elements of $\mathbf{D}_{n,h}$ are positive since $\mathbf{K}_{n,h}(i, i) > 0$.

The spectral clustering algorithm is based on the matrix $\mathbf{Q}_{n,h}$ defined by

$$\mathbf{Q}_{n,h} := \mathbf{D}_{n,h}^{-1}\mathbf{K}_{n,h}.$$

Observe that $\mathbf{Q}_{n,h}$ is a random Markovian transition matrix. Note also that the (random) eigenvalues of $\mathbf{Q}_{n,h}$ are real numbers and that $\mathbf{Q}_{n,h}$ is diagonalizable. Indeed the matrix $\mathbf{Q}_{n,h}$ is conjugate to the symmetric matrix $\mathbf{S}_{n,h} := \mathbf{D}_{n,h}^{-1/2}\mathbf{K}_{n,h}\mathbf{D}_{n,h}^{-1/2}$ since we may write

$$\mathbf{Q}_{n,h} = \mathbf{D}_{n,h}^{-1/2}\mathbf{S}_{n,h}\mathbf{D}_{n,h}^{1/2}.$$

Moreover, the inequality $\|\mathbf{Q}_{n,h}\|_\infty \leq 1$ implies that the spectrum $\sigma(\mathbf{Q}_{n,h})$ is a subset of $[-1;+1]$. Let $1 = \lambda_{n,1} \geq \lambda_{n,2} \geq \ldots \geq \lambda_{n,j(n)} \geq -1$ be the eigenvalues of $\mathbf{Q}_{n,h}$, where in this enumeration, an eigenvalue is repeated as many times as its multiplicity.

To implement the spectral clustering algorithm, the data points of the partitioning problem are first embedded into $\mathbb{R}^\ell$ by using the eigenvectors of $\mathbf{Q}_{n,h}$ associated with the ℓ largest eigenvalues, namely $\lambda_{n,1}$, $\lambda_{n,2}$, $\ldots\lambda_{n,\ell}$. More precisely, fix a collection $V_{n,1}$, $V_{n,2}$, $\ldots$, $V_{n,\ell}$ of such eigenvectors with components respectively given by $V_{n,k} = \{V_{n,k,j}\}_{j\in J(n)}$, for $k = 1,\ldots,\ell$. Then the j^{th} data point, for j in $J(n)$, is represented by the vector $\rho_n(X_j)$ of the feature space $\mathbb{R}^\ell$ defined by $\rho_n(X_j) := \{V_{n,k,j}\}_{1\leq k\leq \ell}$. At last, the embedded points are partitioned using a classical clustering method, such as the k-means algorithm for instance.

2.3 Functional Operators Associated With the Matrices of the Algorithm

As exposed in the Introduction, some functional operators are associated with the matrices acting on $\mathbb{C}^{J(n)}$ defined in the previous paragraph. The link between matrices and functional operators is provided by the evaluation map defined in (3) below. As a consequence, asymptotic results on the clustering algorithm may be derived by studying first the limit behavior of these operators.

To this aim, let us first introduce some additional notation. For $\mathcal{D}$ a subset of $\mathbb{R}^d$, let $W(\mathcal{D})$ be the Banach space of complex-valued, bounded, and continuously differentiable functions with bounded gradient, endowed with the norm

$$\|g\|_W := \|g\|_\infty + \|Dg\|_\infty.$$

Consider the non-oriented graph whose vertices are the X_j's for j ranging in $J(n)$. The similarity matrix $\mathbf{K}_{n,h}$ gives random weights to the edges of the graph and the random transition matrix $\mathbf{Q}_{n,h}$ defines a random walk on the vertices of a random graph. Associated with this random walk is the transition operator $Q_{n,h} : W\big(\mathcal{L}_n(t)\big) \to W\big(\mathcal{L}_n(t)\big)$ defined for any function g by

$$Q_{n,h}g(x) := \int_{\mathcal{L}_n(t)} q_{n,h}(x,y)g(y)\mathbb{P}_n^t(dy).$$

In this equation, $\mathbb{P}_n^t$ is the discrete random probability measure given by

$$\mathbb{P}_n^t := \frac{1}{j(n)}\sum_{j\in J(n)}\delta_{X_j},$$

and

$$q_{n,h}(x,y) := \frac{k_h(y-x)}{K_{n,h}(x)}, \quad \text{where } K_{n,h}(x) := \int_{\mathcal{L}_n(t)} k_h(y-x)\mathbb{P}_n^t(dy). \tag{2}$$

In the definition of $q_{n,h}$, we use the convention that $0/0=0$, but this situation does not occur in the proofs of our results.

Given the *evaluation map* $\pi_n : W\big(\mathcal{L}_n(t)\big) \to \mathbb{C}^{j(n)}$ defined by

$$\pi_n(g) := \big\{g(X_j)\big\}_{j\in J(n)}, \tag{3}$$

the matrix $\mathbf{Q}_{n,h}$ and the operator $Q_{n,h}$ are related by $\mathbf{Q}_{n,h}\circ\pi_n = \pi_n \circ Q_{n,h}$. Using this relation, asymptotic properties of the spectral clustering algorithm may be deduced from the limit behavior of the sequence of operators $\{Q_{n,h}\}_n$. The difficulty, though, is that $Q_{n,h}$ acts on $W\big(\mathcal{L}_n(t)\big)$ and $\mathcal{L}_n(t)$ is a random set which varies with the sample. For this reason, we introduce a sequence of operators $\widehat{Q}_{n,h}$ acting on $W\big(\mathcal{L}(t)\big)$ and constructed from $Q_{n,h}$ as follows.

First of all, recall that under Assumption 1, the gradient of f does not vanish on the set $\{x\in\mathbb{R}^d : f(x)=t\}$. Since f is of class $\mathcal{C}^2$, a continuity argument implies that there exists $\varepsilon_0>0$ such that $\mathcal{L}_{t-\varepsilon_0}^{t+\varepsilon_0}$ contains no critical points of f. Under this condition, Lemma 17 states that $\mathcal{L}(t+\varepsilon)$ is diffeomorphic to $\mathcal{L}(t)$ for every ε such that $|\varepsilon|\le\varepsilon_0$. In all of the following, it is assumed that ε_0 is small enough so that

$$\varepsilon_0/\alpha(\varepsilon_0) < h/2, \quad \text{where } \alpha(\varepsilon_0) := \inf\big\{\|Df(x)\|; x\in\mathcal{L}_{t-\varepsilon_0}^{t}\big\}. \tag{4}$$

Let $\{\varepsilon_n\}_n$ be a sequence of positive numbers such that $\varepsilon_n\le\varepsilon_0$ for each n, and $\varepsilon_n\to0$ as $n\to\infty$. In Lemma 17 an explicit diffeomorphism φ_n carrying $\mathcal{L}(t)$ to $\mathcal{L}(t-\varepsilon_n)$ is constructed, that is,

$$\varphi_n : \mathcal{L}(t) \xrightarrow{\cong} \mathcal{L}(t-\varepsilon_n).$$

The diffeomorphism φ_n induces the linear operator $\Phi_n : W\big(\mathcal{L}(t)\big)\to W\big(\mathcal{L}(t-\varepsilon_n)\big)$ defined by $\Phi_n g = g\circ\varphi_n^{-1}$.

Second, let Ω_n be the probability event defined by

$$\Omega_n = \Big[\|\widehat{f}_n - f\|_\infty \le \varepsilon_n\Big] \cap \Big[\inf\Big\{\|D\widehat{f}_n(x)\|, x\in\mathcal{L}_{t-\varepsilon_0}^{t+\varepsilon_0}\Big\} \ge \frac{1}{2}\|Df\|_\infty\Big].$$

Note that on the event Ω_n, the following inclusions hold:

$$\mathcal{L}(t+\varepsilon_n)\subset\mathcal{L}_n(t)\subset\mathcal{L}(t-\varepsilon_n).$$

We assume that the indicator function $\mathbf{1}_{\Omega_n}$ tends to 1 almost surely as $n\to\infty$, which is satisfied by common density estimates $\widehat{f}_n$ under mild assumptions. For instance, consider a kernel density estimate with a Gaussian kernel. It is a classical exercise to prove that $\|\hat{f}_n - \mathbb{E}\hat{f}_n\|_\infty$ converges to 0 almost surely as n goes to infinity (see, e.g., Example 38 in Pollard, 1984, p. 35, or Chapter 3 in Prakasa Rao, 1983) under appropriate conditions on the bandwidth sequence. Moreover, under the conditions on f in Assumption 1, the norm of the gradient of f is uniformly bounded on $\mathbb{R}^d$, so by using a Taylor expansion, it is easy to prove that the bias term $\|\mathbb{E}\hat{f}_n - f\|_\infty\to0$ as well. Hence $\|\hat{f}_n - f\|_\infty\to0$ almost surely. Furthermore, under Assumption 1, $\|D^2f\|$ is uniformly bounded on $\mathbb{R}^d$ so the same reasoning leads to the almost sure convergence to 0 of $\|D\hat{f}_n - Df\|_\infty$. Together, these facts imply that $\mathbf{1}_{\Omega_n}\to1$ almost surely as $n\to\infty$.

We are now in a position to define the operator $\widehat{Q}_{n,h} : W\big(\mathcal{L}(t)\big)\to W\big(\mathcal{L}(t)\big)$. On the event Ω_n, for all function g in $W\big(\mathcal{L}(t)\big)$, we define $\widehat{Q}_{n,h}g$ by the relation

$$\widehat{Q}_{n,h}g(x) = \frac{1}{j(n)}\sum_{j\in J(n)} q_{n,h}(\varphi_n(x),X_j)\,g\big(\varphi_n^{-1}(X_j)\big), \quad \text{for all } x\in\mathcal{L}(t), \tag{5}$$

and we extend the definition of $\widehat{Q}_{n,h}$ to the whole probability space by setting it to the null operator on the complement Ω_n^c of Ω_n, that is, on Ω_n^c, the function $\widehat{Q}_{n,h}g$ is identically zero for each $g \in W\big(\mathcal{L}(t)\big)$. With a slight abuse of notation, we may note that $\widehat{Q}_{n,h} = \Phi_n^{-1} Q_{n,h} \Phi_n$, so that essentially, the operators $\widehat{Q}_{n,h}$ and $Q_{n,h}$ are conjugate and have equal spectra, which are in turn related to the spectrum of the matrix $\mathbf{Q}_{n,h}$. This is made precise in Proposition 1 below.

Proposition 1 *On the event Ω_n, the spectrum of the functional operator is $\widehat{Q}_{n,h}$ is $\sigma(\widehat{Q}_{n,h}) = \{0\} \cup \sigma(\mathbf{Q}_{n,h})$. Moreover, if if $\lambda \neq 0$, the eigenspaces are isomorphic, that is,*

$$\pi_n \Phi_n : N(\widehat{Q}_{n,h} - \lambda) \xrightarrow{\cong} N(\mathbf{Q}_{n,h} - \lambda),$$

where $\pi_n \Phi_n$ acts on $W\big(\mathcal{L}(t)\big)$ as $\phi_n \Phi_n g(x) = g\big(\varphi_n^{-1}(x)\big)$.

Proof From Equation (5), the range $R(\widehat{Q}_{n,h})$ of $\widehat{Q}_{n,h}$ is spanned by the finite collection of functions

$$f_j : \left| \begin{array}{lll} \mathcal{L}(t) & \to & \mathbb{C} \\ x & \mapsto & q_{n,h}(\varphi_n(x), X_j), \end{array} \right.$$

for all $j \in J(n)$. Moreover, these functions form a basis of $R(\widehat{Q}_{n,h})$. To show this, let V be a vector in $\mathbb{C}^{J(n)}$ such that

$$\sum_{j \in J(n)} V_j f_j(x) = 0 \quad \text{for all } x \in \mathcal{L}(t).$$

By definition of $q_{n,h}$, setting $y = \varphi_n(x)$, we have

$$\sum_{j \in J(n)} V_j \frac{k_h(y - X_j)}{K_{n,h}(y)} = 0 \quad \text{for all } y \in \mathcal{L}(t - \varepsilon_n).$$

Since the support of k_h is hB, the support of the function $K_{n,h}$ is equal to $\bigcup_{j \in J(n)} (X_j + hB)$, and since k_h is positive, it follows that $V_j = 0$ for all j in $J(n)$. Hence $\{f_j \,:\, j \in J(n)\}$ is a basis of $R(\widehat{Q}_{n,h})$.

Now let g be an eigenfunction of $\widehat{Q}_{n,h}$ associated with an eigenvalue $\lambda \neq 0$. Then for all x in $\mathcal{L}(t)$

$$\frac{1}{j(n)} \sum_{j \in J(n)} q_{n,h}\big(\varphi_n(x), X_j\big) g\big(\varphi_n^{-1}(X_j)\big) = \lambda g(x). \tag{6}$$

Since we consider a non-zero eigenvalue, g is in the range of $\widehat{Q}_{n,h}$, and since the functions $\{f_j \,:\, j \in J(n)\}$ form a basis of $R(\widehat{Q}_{n,h})$, there exists a unique vector $V = \{V_j\}_{j \in J(n)} \in \mathbb{C}^{j(n)}$ such that

$$g(x) = \frac{1}{\lambda j(n)} \sum_{j \in J(n)} V_j q_{n,h}(\varphi_n(x), X_j), \quad x \in \mathcal{L}(t).$$

Therefore $V_j = g\big(\varphi_n^{-1}(X_j)\big)$ for all j in $J(n)$. Moreover, by evaluating (6) at any $x = \varphi_n^{-1}(X_i)$ with $i \in J(n)$,

$$\sum_{j \in J(n)} q_{n,h}\big(X_i, X_j\big) g\big(\varphi_n^{-1}(X_j)\big) = \lambda g\big(\varphi_n^{-1}(X_i)\big),$$

which implies that $\mathbf{Q}_{n,h} V = \lambda V$. Consequently, V is an eigenvector of $\mathbf{Q}_{n,h}$ associated with the eigenvalue λ. Hence

$$\sigma(\widehat{Q}_{n,h}) \subset \sigma(\mathbf{Q}_{n,h}) \cup \{0\}, \tag{7}$$

and by unicity of V, it follows that the map $\pi_n\Phi_n : N(\widehat{Q}_{n,h} - \lambda) \longrightarrow N(\mathbf{Q}_{n,h} - \lambda)$ is injective.

Conversely, let V be an eigenvector of the matrix $\mathbf{Q}_{n,h}$ associated with a non-zero eigenvalue λ. Consider the function g of $W\big(\mathcal{L}(t)\big)$ defined by

$$g(x) = \frac{1}{\lambda j(n)} \sum_{j\in J(n)} V_j q_{n,h}(\varphi_n(x), X_j), \quad \text{for all } x \in \mathcal{L}(t).$$

Observe that for all j in $J(n)$,

$$\begin{aligned} g\big(\varphi_n^{-1}(X_j)\big) &= \frac{1}{\lambda j(n)} \sum_{j'\in J(n)} q_{n,h}(X_j, X_{j'}) V_{j'} && \text{by definition of } g, \\ &= \frac{1}{\lambda j(n)} \sum_{j'\in J(n)} \frac{j(n)}{\mathbf{K}_{n,h}(j)} k_h(X_{j'} - X_j) V_{j'} && \text{by definition of } \mathbf{K}_{n,h} \text{ and } q_{n,h}, \\ &= \frac{1}{\lambda}\big(\mathbf{Q}_{n,h}V\big)_j = V_j && \text{since } V \text{ is an eigenvector.} \end{aligned}$$

Hence it follows that for all $x \in \mathcal{L}(t)$,

$$\begin{aligned} \widehat{Q}_{n,h} g(x) &= \frac{1}{j(n)} \sum_{j\in J(n)} q_{n,h}(\varphi_n(x), X_j) g\big(\varphi_n^{-1}(X_j)\big) && \text{using (5)} \\ &= \frac{1}{j(n)} \sum_{j\in J(n)} q_{n,h}(\varphi_n(x), X_j) V_j && \text{since } g\big(\varphi_n^{-1}(X_j)\big) = V_j. \\ &= \lambda g(x). \end{aligned}$$

Consequently,

$$\sigma(\mathbf{Q}_{n,h}) \subset \sigma(\widehat{Q}_{n,h}), \tag{8}$$

and the map $\pi_n\Phi_n : N(\widehat{Q}_{n,h} - \lambda) \longrightarrow N(\mathbf{Q}_{n,h} - \lambda)$ is surjective. Combining (7) and (8), and since 0 belongs to $\sigma(\mathbf{Q}_{n,h})$, we obtain the equality

$$\sigma(\widehat{Q}_{n,h}) = \{0\} \cup \sigma(\mathbf{Q}_{n,h}).$$

At last, since $\pi_n\Phi_n$ is both injective and surjective, the subspaces $N(\widehat{Q}_{n,h} - \lambda)$ and $N(\mathbf{Q}_{n,h} - \lambda)$ are isomorphic for any $\lambda \neq 0$. ■

Remark 2 *Albeit the relevant part of $\widehat{Q}_{n,h}$ is defined on Ω_n for technical reasons, this does not bring any difficulty as long as one is concerned with almost sure convergence. To see this, let $(\Omega, \mathcal{A}, P)$ be the probability space on which the X_i's are defined. Denote by Ω_∞ the event on which $\mathbf{1}_{\Omega_n}$ tends to 1, and recall that $P(\Omega_\infty) = 1$ by assumption. Thus, for every $\omega \in \Omega$, there exists a random integer $n_0(\omega)$ such that, for each $n \geq n_0(\omega)$, ω lies in Ω_n. Besides $n_0(\omega)$ is finite on Ω_∞. Hence in particular, if $\{Z_n\}$ is a sequence of random variables such that $Z_n\mathbf{1}_{\Omega_n}$ converges almost surely to some random variable Z_∞, then $Z_n \to Z_\infty$ almost surely.*

3. Operator Norm Convergence

In this section, we start by establishing the uniform convergence of an unnormalized empirical functional operator. The main operator norm convergence result (Theorem 4) is stated in Section 3.2. The proofs of these theorems rely on several auxiliary lemmas which are stated and proved in Section 5.

3.1 Unnormalized Operators

Let $r : \mathcal{L}(t-\varepsilon_0)\times\mathbb{R}^d \to \mathbb{R}$ be a given function. Define the linear operators R_n and R on $W\big(\mathcal{L}(t)\big)$ respectively by

$$R_n g(x) := \int_{\mathcal{L}_n(t)} r\big(\varphi_n(x),y\big)g\big(\varphi_n^{-1}(y)\big)\mathbb{P}_n^t(dy), \qquad \text{and} \qquad Rg(x) := \int_{\mathcal{L}(t)} r(x,y)g(y)\mu^t(dy).$$

Proposition 3 *Assume the following conditions on the function r:*
(i) r is continuously differentiable with compact support ;
(ii) r is uniformly bounded on $\mathcal{L}(t-\varepsilon_0)\times\mathbb{R}^d$, that is, $\|r\|_\infty < \infty$;
(iii) the differential $D_x r$ of the function r with respect to x is uniformly bounded on $\mathcal{L}(t-\varepsilon_0)\times\mathbb{R}^d$, that is, $\|D_x r\|_\infty := \sup\big\{\|D_x r(x,y)\| : (x,y)\in\mathcal{L}(t-\varepsilon_0)\times\mathbb{R}^d\big\} < \infty$.
Then, as $n\to\infty$,

$$\sup\Big\{\big\|R_n g - Rg\big\|_\infty : \|g\|_W \le 1\Big\} \to 0 \quad \textit{almost surely}.$$

The key argument for proving Proposition 3 is that the collection of functions

$$\Big\{y \mapsto r(x,y)g(y)\mathbf{1}_{\mathcal{L}(t)}(y) : x\in\mathcal{L}(t),\ \|g\|_{W(\mathcal{L}(t))} \le 1\Big\}$$

is Glivenko-Cantelli, which is proved in Lemma 13. Let us recall that a collection $\mathcal{F}$ of functions is said to be Glivenko-Cantelli, or to satisfy a uniform law of large number, if

$$\sup_{g\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n g(X_i) - \mathbb{E}[X]\right| \to 0 \quad \text{almost surely,}$$

where $X, X_1, X_2, \dots$ are i.i.d. random variables.

Proof In all this proof, we shall use the following convention: given a function g defined only on some subset $\mathcal{D}$ of $\mathbb{R}^d$, for any subset $\mathcal{A}\subset\mathcal{D}$, and any $x\in\mathbb{R}^d$, the notation $g(x)\mathbf{1}_{\mathcal{A}}(x)$ stands for $g(x)$ is $x\in\mathcal{A}$ and for 0 otherwise. Set

$$\begin{aligned}
S_n g(x) &:= \frac{1}{\mu(\mathcal{L}(t))}\frac{1}{n}\sum_{i=1}^n r\big(\varphi_n(x),X_i\big)\,g\big(\varphi_n^{-1}(X_i)\big)\mathbf{1}_{\mathcal{L}_n(t)}(X_i),\\
T_n g(x) &:= \frac{1}{\mu\big(\mathcal{L}(t)\big)}\frac{1}{n}\sum_{i=1}^n r\big(\varphi_n(x),X_i\big)g(X_i)\mathbf{1}_{\mathcal{L}(t)}(X_i),\\
U_n g(x) &:= \frac{1}{\mu\big(\mathcal{L}(t)\big)}\frac{1}{n}\sum_{i=1}^n r\big(x,X_i\big)g(X_i)\mathbf{1}_{\mathcal{L}(t)}(X_i).
\end{aligned}$$

and consider the inequality

$$\begin{aligned}
\big|R_n g(x) - Rg(x)\big| \le{}& \big|R_n g(x) - S_n g(x)\big| + \big|S_n g(x) - T_n g(x)\big|\\
&+ \big|T_n g(x) - U_n g(x)\big| + \big|U_n g(x) - Rg(x)\big|,
\end{aligned} \tag{9}$$

for all $x \in \mathcal{L}(t)$ and all $g \in W\big(\mathcal{L}(t)\big)$.

The first term in (9) is bounded uniformly by

$$\big|R_n g(x) - S_n g(x)\big| \leq \left|\frac{n}{j(n)} - \frac{1}{\mu\big(\mathcal{L}(t)\big)}\right| \|r\|_\infty \|g\|_\infty$$

and since $j(n)/n$ tends to $\mu(\mathcal{L}(t))$ almost surely as $n \to \infty$, we conclude that

$$\sup\Big\{\big\|R_n g - S_n g\big\|_\infty : \|g\|_W \leq 1\Big\} \to 0 \quad \text{a.s. as } n \to \infty. \tag{10}$$

For the second term in (9), we have

$$\begin{aligned} |S_n g(x) - T_n g(x)| &\leq \frac{\|r\|_\infty}{\mu\big(\mathcal{L}(t)\big)} \frac{1}{n} \sum_{i=1}^{n} \big| g\big(\varphi_n^{-1}(X_i)\big) \mathbf{1}_{\mathcal{L}_n(t)}(X_i) - g(X_i) \mathbf{1}_{\mathcal{L}(t)}(X_i)\big| \\ &= \frac{\|r\|_\infty}{\mu\big(\mathcal{L}(t)\big)} \frac{1}{n} \sum_{i=1}^{n} g_n(X_i), \end{aligned} \tag{11}$$

where g_n is the function defined on the whole space $\mathbb{R}^d$ by

$$g_n(x) = \Big| g\big(\varphi_n^{-1}(x)\big) \mathbf{1}_{\mathcal{L}_n(t)}(x) - g(x) \mathbf{1}_{\mathcal{L}(t)}(x)\Big|.$$

Consider the partition of $\mathbb{R}^d$ given by $\mathbb{R}^d = B_{1,n} \cup B_{2,n} \cup B_{3,n} \cup B_{4,n}$, where

$$\begin{aligned} B_{1,n} &:= \mathcal{L}_n(t) \cap \mathcal{L}(t), & B_{2,n} &:= \mathcal{L}_n(t) \cap \mathcal{L}(t)^c, \\ B_{3,n} &:= \mathcal{L}_n(t)^c \cap \mathcal{L}(t), & B_{4,n} &:= \mathcal{L}_n(t)^c \cap \mathcal{L}(t)^c. \end{aligned}$$

The sum over i in (11) may be split into four parts as

$$\frac{1}{n} \sum_{i=1}^{n} g_n(X_i) = I_1(x,g) + I_2(x,g) + I_3(x,g) + I_4(x,g) \tag{12}$$

where

$$I_k(x,g) := \frac{1}{n} \sum_{i=1}^{n} g_n(X_i) \mathbf{1}\{X_i \in B_{k,n}\}.$$

First, $I_{4,n}(x,g) = 0$ since g_n is identically 0 on $B_{4,n}$. Second,

$$I_2(x,g) + I_3(x,g) \leq \|g\|_\infty \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\mathcal{L}(t)\Delta\mathcal{L}_n(t)}(X_i) \tag{13}$$

Applying Lemma 11 together with the almost sure convergence of $\mathbf{1}_{\Omega_n}$ to 1, we obtain that

$$\frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{\mathcal{L}(t)\Delta\mathcal{L}_n(t)}(X_j) \to 0 \quad \text{almost surely.} \tag{14}$$

Third,

$$\begin{aligned} I_1(x,g) &\leq \sup_{x \in \mathcal{L}(t)} \Big| g\big(\varphi_n^{-1}(x)\big) - g(x)\Big| \leq \|D_x g\|_\infty \sup_{x \in \mathcal{L}(t)} \|\varphi_n^{-1}(x) - x\| \\ &\leq \|D_x g\|_\infty \sup_{x \in \mathcal{L}(t)} \|x - \varphi_n(x)\| \to 0 \end{aligned} \tag{15}$$

as $n \to \infty$ by Lemma 17. Thus, combining (11), (12), (13), (14) and (15) leads to

$$\sup\left\{\left\|S_n g - T_n g\right\|_\infty : \|g\|_W \le 1\right\} \to 0 \quad \text{a.s. as } n \to \infty. \tag{16}$$

For the third term in (9), using the inequality

$$\left|r\big(\varphi_n(x), X_i\big) - r\big(x, X_i\big)\right| \le \|D_x r\|_\infty \sup_{x \in \mathcal{L}(t)} \|\varphi_n(x) - x\|$$

we deduce that

$$\left|T_n g(x) - U_n g(x)\right| \le \frac{1}{\mu\big(\mathcal{L}(t)\big)} \|g\|_\infty \|D_x r\|_\infty \sup_{x \in \mathcal{L}(t)} \|\varphi_n(x) - x\|.$$

and so

$$\sup\left\{\left\|T_n g - U_n g\right\|_\infty : \|g\|_W \le 1\right\} \to 0 \quad \text{a.s. as } n \to \infty, \tag{17}$$

by Lemma 17.

At last, for the fourth term in (9), we conclude by Lemma 13 that

$$\sup\left\{\left\|U_n g - R g\right\|_\infty : \|g\|_W \le 1\right\} \to 0 \quad \text{a.s. as } n \to \infty.$$

Finally, reporting (10), (16) and (17) in (9) yields the desired result. ■

3.2 Normalized Operators

Theorem 4 states that $\widehat{Q}_{n,h}$ converges in operator norm to the limit operator $Q_h : W\big(\mathcal{L}(t)\big) \to W\big(\mathcal{L}(t)\big)$ defined by

$$Q_h g(x) = \int_{\mathcal{L}(t)} q_h(x,y) g(y) \mu^t(dy), \tag{18}$$

where μ^t denotes the conditional distribution of X given the event $\big[X \in \mathcal{L}(t)\big]$, and where

$$q_h(x,y) = \frac{k_h(y-x)}{K_h(x)}, \quad \text{with } K_h(x) = \int_{\mathcal{L}(t)} k_h(y-x)\mu^t(dy). \tag{19}$$

Theorem 4 (Operator Norm Convergence) *Suppose that Assumptions 1 and 2 hold. We have*

$$\left\|\widehat{Q}_{n,h} - Q_h\right\|_W \to 0 \quad \textit{almost surely as } n \to \infty.$$

Proof We will prove that, as $n \to \infty$, almost surely,

$$\sup\left\{\left\|\widehat{Q}_{n,h} g - Q_h g\right\|_\infty : \|g\|_W \le 1\right\} \to 0 \tag{20}$$

and

$$\sup\left\{\left\|D_x\big[\widehat{Q}_{n,h} g\big] - D_x\big[Q_h g\big]\right\|_\infty : \|g\|_W \le 1\right\} \to 0 \tag{21}$$

To this aim, we introduce the operator $\widetilde{Q}_{n,h}$ acting on $W(\mathcal{L}(t))$ as

$$\widetilde{Q}_{n,h} g(x) = \int_{\mathcal{L}_n(t)} q_h(\varphi_n(x), y) g\big(\varphi_n^{-1}(y)\big) \mathbb{P}_n^t(dy).$$

Proof of (20) For all $g \in W\big(\mathcal{L}(t)\big)$, we have

$$\big\|\widehat{Q}_{n,h}g - Q_h g\big\|_\infty \leq \big\|\widehat{Q}_{n,h}g - \widetilde{Q}_{n,h}g\big\|_\infty + \big\|\widetilde{Q}_{n,h}g - Q_h g\big\|_\infty. \tag{22}$$

First, by Lemma 14, the function $r = q_h$ satisfies the condition in Proposition 3, so that

$$\sup\Big\{\|\widetilde{Q}_{n,h}g - Q_h g\|_\infty : \|g\|_W \leq 1\Big\} \to 0 \tag{23}$$

with probability one as $n \to \infty$.

Next, since $\|q_h\|_\infty < \infty$ by Lemma 14, there exists a finite constant C_h such that,

$$\|\widetilde{Q}_{n,h}g\|_\infty \leq C_h \quad \text{for all } n \text{ and all } g \text{ with } \|g\|_W \leq 1. \tag{24}$$

By definition of $q_{n,h}$, for all x, y in the level set $\mathcal{L}(t)$, we have

$$q_{n,h}(x,y) = \frac{K_h(x)}{K_{n,h}(x)} q_h(x,y).$$

So

$$\begin{aligned}\Big|\widehat{Q}_{n,h}g(x) - \widetilde{Q}_{n,h}g(x)\Big| &= \left|\frac{K_n\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)} - 1\right| \Big|\widetilde{Q}_{n,h}g(x)\Big| \\ &\leq C_h \sup_{x \in \mathcal{L}(t)} \left|\frac{K_n\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)} - 1\right|,\end{aligned}$$

where C_h is as in (24). Applying Lemma 16 yields

$$\sup\Big\{\|\widehat{Q}_{n,h}g - \widetilde{Q}_{n,h}g\|_\infty : \|g\|_W \leq 1\Big\} \to 0 \tag{25}$$

with probability one as $n \to \infty$. Reporting (23) and (25) in (22) proves (20).

Proof of (21) We have

$$\left\|D_x\Big[\widehat{Q}_{n,h}g\Big] - D_x\Big[Q_h g\Big]\right\|_\infty \leq \left\|D_x\Big[\widehat{Q}_{n,h}g\Big] - D_x\Big[\widetilde{Q}_h g\Big]\right\|_\infty + \left\|D_x\Big[\widetilde{Q}_{n,h}g\Big] - D_x\Big[Q_h g\Big]\right\|_\infty. \tag{26}$$

The second term in right han side of (26) is bounded by

$$\left\|D_x\Big[\widetilde{Q}_{n,h}g\Big] - D_x\Big[Q_h g\Big]\right\|_\infty \leq \big\|D_x\varphi_n\big\|_\infty \big\|R_n g - Rg\big\|_\infty,$$

where

$$\begin{aligned}R_n g(x) &:= \int_{\mathcal{L}_n(t)} (D_x q_h)(\varphi_n(x), y) g\big(\varphi_n^{-1}(y)\big)\mathbb{P}_n^t(dy) \quad \text{and} \\ Rg(x) &:= \int_{\mathcal{L}(t)} (D_x q_h)(\varphi_n(x), y) g\big(\varphi_n^{-1}(y)\big)\mu^t(dy).\end{aligned}$$

By lemma 17, $x \mapsto D_x\varphi_n(x)$ converges to the identity matrix I_d of $\mathbb{R}^d$, uniformly in x over $\mathcal{L}(t)$. So $\|D_x\varphi_n(x)\|$ is bounded by some finite constant C_φ uniformly over n and $x \in \mathcal{L}(t)$ and

$$\left\|D_x\left[\widetilde{Q}_{n,h}g\right] - D_x\left[Q_h g\right]\right\|_\infty \leq C_\varphi \left\|R_n g - Rg\right\|_\infty.$$

By Lemma 14, the map $r : (x,y) \mapsto D_x q_h(x,y)$ satisfies the conditions in Proposition 3. Thus, $\|R_n g - Rg\|_\infty$ converges to 0 almost surely, uniformly over g in the unit ball of $W(\mathcal{L}(t))$, and we deduce that

$$\sup\left\{\left\|D_x\left[\widetilde{Q}_{n,h}g\right] - D_x\left[Q_h g\right]\right\|_\infty : \|g\|_W \leq 1\right\} \to 0 \quad \text{a.s. as } n \to \infty. \tag{27}$$

For the first term in right hand side of (26), observe first that there exists a constant C'_h such that, for all n and all g in the unit ball of $W\big(\mathcal{L}(t)\big)$,

$$\|R_{n,h}g\|_\infty \leq C'_h, \quad \text{for all } n \text{ and all } g \text{ with } \|g\|_W \leq 1, \tag{28}$$

by Lemma 14.

On the one hand, we have

$$D_x\left[q_{n,h}(\varphi_n(x),y)\right] = \frac{K_h\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)} D_x\varphi_n(x)(D_x q_h)\big(\varphi_n(x),y\big) + D_x\left[\frac{K_h\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)}\right] q_h\big(\varphi_n(x),y\big).$$

Hence,

$$D_x\left[\widehat{Q}_{n,h}g(x)\right] = \frac{K_h\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)} D_x\varphi_n(x) R_n g(x) + D_x\left[\frac{K_h\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)}\right] \widetilde{Q}_{n,h}g(x).$$

On the other hand, since $D_x\left[q_h\big(\varphi_n(x),y\big)\right] = D_x\varphi_n(x)(D_x q_h)\big(\varphi_n(x),y\big)$,

$$D_x\left[\widetilde{Q}_{n,h}g(x)\right] = D_x\varphi_n(x) R_n g(x).$$

Thus,

$$D_x\left[\widehat{Q}_{n,h}g(x)\right] - D_x\left[\widetilde{Q}_h g(x)\right] = D_x\left[\frac{K_h\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)}\right]\widetilde{Q}_{n,h}g(x) + \left(\frac{K_h\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)} - 1\right) D_x\varphi_n(x) R_n g(x).$$

Using the Inequalities (24) and (28), we obtain

$$\left\|D_x\left[\widehat{Q}_{n,h}g\right] - D_x\left[\widetilde{Q}_h g\right]\right\|_\infty \leq C_h \sup_{x\in\mathcal{L}(t)} \left|D_x\left[\frac{K_h\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)}\right]\right| + C'_h C_\varphi \sup_{x\in\mathcal{L}(t)} \left|\frac{K_h\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)} - 1\right|.$$

and by applying Lemma 16, we deduce that

$$\sup\left\{\left\|D_x\left[\widehat{Q}_{n,h}g\right] - D_x\left[\widetilde{Q}_h g\right]\right\|_\infty : \|g\|_W \leq 1\right\} \to 0 \quad \text{a.s. as } n \to \infty. \tag{29}$$

Reporting (27) and (29) in (26) proves (21). ■

4. Consistency of the Algorithm

The consistency of the algorithm relies on the operator norm convergence of $\widehat{Q}_{n,h}$ to the limit operator Q_h (Theorem 4), on the spectral properties of Q_h stated below in Section 4.1, and on the results collected in Appendix B on the perturbation theory of linear operators, The starting point is the fact that, provided that $h < d_{\min}$, the connected components of the level set $\mathcal{L}(t)$ are the recurrent classes of the Markov chain whose transitions are defined by Q_h. Indeed, this process cannot jump from one component to another component. Hence Q_h defines the desired clustering via its eigenspace corresponding to the eigenvalue 1, since this latter is spanned by the characteristic functions of the connected components of $\mathcal{L}(t)$, as stated in Proposition 6 below.

In Section 4.2, the consistency of the clustering is obtained in Theorem 7 in the case where the scale parameter h is lower than d_{min} defined in (1), which is the minimum distance between any two connected components of $\mathcal{L}(t)$. Using the continuity of Q_h in h, we then obtain the main consistency in Theorem 10 of Section 4.3, where h is allowed to be larger than d_{min}, up to a value depending only on the underlying density f.

4.1 Properties of the Limit Operator Q_h When $h < d_{min}$

The transition kernel $q_h(x,dy) := q_h(x,y)\mu^t(dy)$ associated with the operator Q_h defines a Markov chain with state space $\mathcal{L}(t)$, which is not countable. Familiar notions such as irreducibility, aperiodicity, and positive recurrence, which are valid for a Markov chain on a countable state space, may be extended to the non-countable case. The relevant definitions and materials on Markov chains on a general state space are summarized in Appendix C. The properties of the Markov chain with transition kernel $q_h(x,dy)$ are stated in Proposition 5 below.

Recall that $\mathcal{L}(t)$ has ℓ connected components $\mathcal{C}_1,\dots,\mathcal{C}_\ell$ and that d_{min}, defined in (1), is the minimal distance between the connected components of $\mathcal{L}(t)$.

Proposition 5 *Consider the Markov chain with state space $\mathcal{L}(t)$ and transition kernel $q_h(x,dy)$, and assume that $h < d_{min}$.*
1. The chain is Feller and topologically aperiodic.
2. When started at a point x in some connected component of the state space, the chain evolves within this connected component only.
3. When the state space is reduced to some connected component of $\mathcal{L}(t)$, the chain is open set irreducible and positive Harris recurrent.
4. When the state space is reduced to some connected component $\mathcal{C}_k$ of $\mathcal{L}(t)$, the Markov chain has a unique invariant distribution $\nu_k(dy)$ and, for all $x \in \mathcal{C}_k$, the sequence of distributions $\left\{q_h^n(x,dy)\right\}_{n\in\mathbb{N}}$ over $\mathcal{C}_k$ converges in total variation to $\nu_k(dy)$.

Proof Denote by $\{\xi_n\}$ the Markov chain with transition kernel $q_h(x,dy)$. For all $x \in \mathcal{L}(t)$, the distribution $q_h(x,dy) = q_h(x,y)\mu^t(dy)$ is absolutely continuous with respect to the Lebesgue measure, with density $y \mapsto f_h(x,y)$ defined by

$$f_h(x,y) = q_h(x,y)\frac{f(y)}{\int_{y'\in\mathcal{L}(t)} f(y')dy'}\mathbf{1}_{\mathcal{L}(t)}(y).$$

Since the similarity function k_h and the density f are both continuous, the map $(x,y) \mapsto f_h(x,y)$ is continuous.

Now, by induction on n, the distribution of ξ_n conditioned by $\xi_0 = x$, which is $q_h^{n+1}(x,dy)$ is also absolutely continuous with respect to the Lebesgue measure and its density $y \mapsto f_h^n(x,y)$ satisfies

$$f_h^n(x,y) = \int_{z\in\mathcal{L}(t)} f_h^{n-1}(x,z)\, f_h(z,y)dz = \int_{z\in\mathcal{L}(t)} f_h(x,z)\, f_h^{n-1}(z,y)dz, \tag{30}$$

where the last equality follows from the Markov property. Moreover, one easily sees by induction that the map $(x,y) \mapsto f_h^n(x,y)$ is continuous.
1. Since the similarity function k_h is continuous, with compact support hB, the map

$$x \mapsto Q_h g(x) = \int_{\mathcal{L}(t)} q_h(x,dy)g(y)$$

is continuous for every bounded, measurable function g. Hence, the chain is Feller.

Now we have to prove that the chain is topologically aperiodic, that is, that $q_h^n(x, x+\eta B) > 0$ for each $x \in \mathcal{L}(t)$, for all $n \geq 1$ and $\eta > 0$, where $q_h^n(x,\cdot)$ is the distribution of ξ_n conditioned on $\xi_0 = x$. Since the distribution $q_h^n(x,\cdot)$ admits a continuous density $f_h^n(x,\cdot)$, it is enough to prove that $f_h^n(x,x) > 0$. Since k_h is bounded from below on $(h/2)B$ by Assumption 2, the density $f_h(x,y)$ is strictly positive for all $y \in x + hB/2$. By induction over n, using (30), $f_h^n(x,x) > 0$ and the chain is topologically aperiodic.

2. Without loss of generality, since the numbering of the connected components is arbitrary, assume that $x \in \mathcal{C}_1$. Let y be a point of $\mathcal{L}(t)$ which does not belong to $\mathcal{C}_1$. Then $\|y-x\| \geq d_{min} > h$ so that $q_h(x,y) = 0$. Whence,

$$P_x(\xi_1 \in \mathcal{C}_1) = q_h(x,\mathcal{C}_1) = \int_{\mathcal{C}_1} q_h(x,y)\mu^t(dy) = \int_{\mathcal{L}(t)} q_h(x,y)\mu^t(dy) = 1.$$

3. Assume that the state space is reduced to $\mathcal{C}_1$. In order to prove that the chain is open set irreducible, it is enough to prove that, for each $x,y \in \mathcal{C}_1$ and $\eta > 0$, there exists some integer N such that $P_x(\xi_N \in y+\eta B) = q_h^N(x, y+\eta B)$ is positive. Observe that $q_h^n(x,dy)$ is the distribution with density

$$q_h^n(x,y) = \int_{x_1,\ldots,x_{n-1}\in\mathcal{C}_1} q_h(x,x_1)q_h(x_1,x_2)\ldots q_h(x_{n-1},y)dx_1dx_2dx_{n-1}$$

and $(x_1,\ldots,x_{n-1}) \mapsto q_h(x,x_1)q_h(x_1,x_2)\ldots q_h(x_{n-1},y)$ is continuous. Hence, it is enough to prove that there exists some integer N and a finite sequence $x_1,\ldots x_N$ such that

$$q_h(x,x_1)q_h(x_1,x_2)\ldots q_h(x_{N-1},y) > 0.$$

Since $\mathcal{C}_1$ is connected, there exists a finite sequence $x_0, x_1, \ldots x_N$ of points in $\mathcal{C}_1$ such that $x_0 = x$, $x_N = y$, and $\|x_i - x_{i+1}\| \leq h/2$ for each i. Therefore

$$q_h(x,x_1)q_h(x_1,x_2)\ldots q_h(x_{N-1},y) > 0$$

which proves that the chain is open set irreducible.

Since $\mathcal{C}_1$ is compact, the chain is non-evanescent, and so it is Harris recurrent. Recall that $k(x) = k(-x)$ from Assumption 2. Therefore $k_h(y-x) = k_h(x-y)$ which yields

$$K_h(x)q_h(x,dy)\mu^t(dx) = K_h(y)q_h(y,dx)\mu^t(dy).$$

By integrating the previous relation with respect to x over $\mathcal{C}_1$, one may verify that $K_h(x)\mu^t(dx)$ is an invariant measure. At last $\int_{\mathcal{C}_1} K_h(x)\mu^t(dx) < \infty$, which proves that the chain is positive.

4. This ergodic property is a direct application of the last part of Appendix C. ■

Proposition 6 *Assume that $h < d_{min}$. If g is continuous and $Q_h g = g$, then g is constant on the connected components of $\mathcal{L}(t)$.*

Proof The numbering of the connected components is arbitrary. Hence it is enough to prove that g is constant over $\mathcal{C}_1$. For this, fix x in $\mathcal{C}_1$ and note that $g = Q_h g$ implies $g = Q_h^n g$ for any $n \geq 1$. By Proposition 5, the chain is open set irreducible, topologically aperiodic, and positive Harris recurrent on $\mathcal{C}_1$. Moreover, $q_h^n(x, dy)$ converges in total variation norm to $\nu_1(dy)$, where ν_1 is the unique invariant distribution when state space is reduced to $\mathcal{C}_1$. Specifically,

$$Q_h^n g(x) \longrightarrow \int_{\mathcal{C}_1} g(y)\nu_1(dy) \quad \text{as } n \to \infty.$$

Hence, for every x in $\mathcal{C}_1$,

$$g(x) = \int_{\mathcal{C}_1} g(y)\nu_1(dy),$$

and since the last integral does not depend on x, it follows that g is a constant function on $\mathcal{C}_1$. ■

4.2 Spectral Convergence

Theorem 7 states that the representation of the extracted part of the data set into the feature space $\mathbb{R}^\ell$ (see the end of Section 2.2) tends to concentrate around ℓ different centroids. Moreover, each of these centroids corresponds to a cluster, that is, to a connected component of $\mathcal{L}(t)$. As a trivial consequence, any partitioning algorithm (e.g., k-means) applied in the feature space will asymptotically yield the desired clustering. In other words, the clustering algorithm is consistent.

More precisely, using the convergence in operator norm of $\widehat{Q}_{n,h}$ towards Q_h, together with the results of functional analysis given in Appendix B, we obtain the following Theorem which describes the asymptotic behavior of the algorithm. Let us denote by $J(\infty)$ the set of integers j such that X_j is in the level set $\mathcal{L}(t)$. For all $j \in J(\infty)$, define $k(j)$ as the integer such that $X_j \in \mathcal{C}_{k(j)}$.

Theorem 7 *Suppose that Assumptions 1 and 2 hold, and that h is in $(0; d_{min})$.*
1. The first ℓ eigenvalues $\lambda_{n,1}, \lambda_{n,2}, \ldots, \lambda_{n,\ell}$ of $\mathbf{Q}_{n,h}$ converge to 1 almost surely as $n \to \infty$, and there exists $\eta_0 > 0$ such that for all $j > \ell$, $\lambda_{n,j}$ belongs to $\{z : |z-1| \geq \eta_0\}$ for n large enough, with probability one.
2. There exists a sequence $\{\xi_n\}_n$ of invertible linear transformations of $\mathbb{R}^\ell$ such that, for all $j \in J(\infty)$, $\xi_n \rho_n(X_j)$ converges almost surely to $e_{k(j)}$, where $e_{k(j)}$ is the vector of $\mathbb{R}^\ell$ whose components are all 0 except the $k(j)^{\text{th}}$ component equal to 1.

Proof 1. The spectrum of Q_h may be decomposed as $\sigma(Q_h) = \sigma_1(Q_h) \cup \sigma_2(Q_h)$, where $\sigma_1(Q_h) = \{1\}$ and where $\sigma_2(Q_h) = \sigma(Q_h) \setminus \{1\}$. Since 1 is an isolated eigenvalue, there exists η_0 in the open interval $(0;1)$ such that $\sigma(Q_h) \cap \{z \in \mathbb{C} : |z-1| \leq \eta_0\}$ is reduced to the singleton $\{1\}$. Moreover, 1 is an eigenvalue of Q_h of multiplicity ℓ, by Proposition 6. Hence by Theorem 18, $W\big(\mathcal{L}(t)\big)$ decomposes into $W\big(\mathcal{L}(t)\big) = M_1 \oplus M_2$ where $M_1 = N(Q_h - 1)$ and M_2 is mapped into itself by Q_h.

Split the spectrum of $\widehat{Q}_{n,h}$ as $\sigma\big(\widehat{Q}_{n,h}\big) = \sigma_1\big(\widehat{Q}_{n,h}\big) \cup \sigma_2\big(\widehat{Q}_{n,h}\big)$, where

$$\sigma_1\big(\widehat{Q}_{n,h}\big) = \sigma\big(\widehat{Q}_{n,h}\big) \cap \big\{z \in \mathbb{C} : |z-1| < \eta_0\big\}.$$

By Theorem 18, this decomposition of the spectrum of $\widehat{Q}_{n,h}$ yields a decomposition of $W\big(\mathcal{L}(t)\big)$ as $W\big(\mathcal{L}(t)\big) = M_{n,1} \oplus M_{n,2}$, where $M_{n,1}$ and $M_{n,2}$ are stable subspaces under $\widehat{Q}_{n,h}$ and

$$M_{n,1} := \bigoplus_{\lambda \in \sigma_1(\widehat{Q}_{n,h})} N(\widehat{Q}_{n,h} - \lambda).$$

By Proposition 1, $\sigma(\widehat{Q}_{n,h}) = \sigma(\mathbf{Q}_{n,h}) \cup \{0\}$. Statement 6 of Theorem 19 implies that, for all n large enough, the total multiplicity of the eigenvalues in $\sigma_1(\widehat{Q}_{n,h})$ is $\dim(M_1) = \dim(N(Q_h - 1)) = \ell$. Hence, for all $j > \ell$, $\lambda_{n,j}$ belongs to $\{z : |z-1| \geq \eta_0\}$. Moreover, statement 4 of Theorem 19 proves that the first ℓ eigenvalues converges to 1.

2. In addition to the convergence of the eigenvalues of $\mathbf{Q}_{n,h}$, the convergence of the eigenspaces also holds. More precisely, let Π be the projector on $M_1 = N(Q_h - 1)$ along M_2 and Π_n the projector on $M_{n,1}$ along $M_{n,2}$. Statements 2, 3, 5 and 6 of Theorem 19 leads to

$$\|\Pi_n - \Pi\|_W \to 0 \quad a.s. \tag{31}$$

and the dimension of $M_{n,1}$ is equal to ℓ for all n large enough.

Denote by $E_{n,1}$ the subspace of $\mathbb{C}^{j(n)}$ spanned by the eigenvectors of $\mathbf{Q}_{n,h}$ corresponding to the eigenvalues $\lambda_{n,1}, \ldots \lambda_{n,\ell}$. Since

$$M_{n,1} = \bigoplus_{\lambda \in \sigma_1(\widehat{Q}_{n,h})} N(\widehat{Q}_{n,h} - \lambda) \quad \text{and} \quad E_{n,1} = \bigoplus_{\lambda \in \sigma_1(\mathbf{Q}_{n,h})} N(\mathbf{Q}_{n,h} - \lambda),$$

by Proposition 1 the map $\pi_n\Phi_n$ induces an isomorphism between $M_{n,1}$ and $E_{n,1}$. Moreover, Π_n induces a morphism $\tilde{\Pi}_n$ from M_1 to $M_{n,1}$ which converges to the identity map of M_1 in W-norm by (31). Hence, if n is large enough, $\tilde{\Pi}_n$ is invertible and we have the following isomorphisms of vector spaces:

$$\tilde{\Pi}_n : M_1 \xrightarrow{\cong} M_{n,1} \quad \text{and} \quad \pi_n\Phi_n : M_{n,1} \xrightarrow{\cong} E_{n,1}. \tag{32}$$

By Proposition 6, the functions $g_k := \mathbf{1}_{C_k}$, $k = 1, 2 \ldots, \ell$, form a basis of $M_1 = N(Q_h - 1)$. Using the isomorphisms of (32), we may define for all $k \in \{1, \ldots \ell\}$,

$$g_{n,k} := \tilde{\Pi}_n g_k, \qquad \text{and} \qquad \vartheta_{n,k} := \pi_n\Phi_n g_{n,k} = \pi_n\Phi_n\tilde{\Pi}_n g_k.$$

Then the collections $\{g_{n,k}\}_{k=1,\ldots,\ell}$ and $\{\vartheta_{n,k}\}_{k=1,\ldots,\ell}$ are a basis of $M_{n,1}$ and $E_{n,1}$ respectively. Moreover, for all $k \in \{1, \ldots, \ell\}$, $g_{n,k}$ converges to $\mathbf{1}_{C_k}$ in W-norm by (31). And, as $n \to \infty$, if $j \in J(\infty)$,

$$\vartheta_{n,k,j} = \tilde{\Pi}_n(\mathbf{1}_{C_k}) \circ \varphi_n^{-1}(X_j) \to \mathbf{1}_{C_k}(X_j) = \begin{cases} 1 & \text{if } k = k(j), \\ 0 & \text{otherwise.} \end{cases} \tag{33}$$

The eigenvectors $V_{n,1}, \ldots, V_{n,\ell}$ chosen in the algorithm form another basis of $E_{n,1}$. Hence, there exists a matrix ξ_n of dimension $\ell \times \ell$ such that

$$\vartheta_{n,k} = \sum_{i=1}^{\ell} \xi_{n,k,i} V_{n,i}.$$

Hence the j^{th} component of $\vartheta_{n,k}$, for all $j \in J(n)$, may be expressed as

$$\vartheta_{n,k,j} = \sum_{i=1}^{\ell} \xi_{n,k,i} V_{n,i,j}.$$

Since $\rho_n(X_j)$ is the vector of $\mathbb{R}^\ell$ with components $\{V_{n,i,j}\}_{i=1,\ldots,\ell}$, the vector $\vartheta_{n,\bullet,j} = \{\vartheta_{n,k,j}\}_k$ of $\mathbb{R}^\ell$ is related to $\rho_n(X_j)$ by the linear transformation ξ_n, that is,

$$\vartheta_{n,\bullet,j} = \xi_n \rho_n(X_j).$$

The convergence of $\vartheta_{n,\bullet,j}$ to $e_{k(j)}$ then follows from (33) and Theorem 7 is proved. ■

Remark 8 *The last step of the spectral clustering algorithm consists in partitioning the transformed data in the feature space, which can be performed by a standard clustering algorithm, like the k-means algorithm or a hierarchical clustering. Theorem 7 states that there exists a choice for a basis of ℓ eigenvectors such that the transformed data concentrates on the ℓ canonical basis vectors e_k of $\mathbb{R}^\ell$. Consequently, upon choosing a suitable collection $V_{n,1},\ldots,V_{n,\ell}$ of eigenvectors, for any $\varepsilon > 0$, with probability one, for n large enough, the transformed data $\rho_n(X_j)$'s belong to the union of balls centered at $e_1,\ldots,e_\ell$ and of radius ε. Combining this result with known asymptotic properties of the aforementioned clustering algorithms leads to the desired partition.*

For instance, a hierarchical agglomerative method with single linkage allows to separate groups provided that the minimal distance between the groups is larger than the maximal diameter of the groups. In the preceding display, by choosing ε such that $2\varepsilon < \sqrt{2}$, with probability one for n large enough the points belong to ℓ balls of diameter 2ε which are all at a distance strictly larger than 2ε. Consequently, cutting the dendrogram tree of the single linkage hierarchical clustering at a height 2ε will correctly separate the groups, and the algorithm is consistent.

Similarly, for the k-means algorithm, we may note that, upon choosing a suitable basis of eigenvectors, the empirical measure associated with the transformed data converges to a discrete measure supported by the canonical vectors $e_1,\ldots,e_\ell$. Consistency of the grouping then follows from the well-known properties of the vector quantization method; see Pollard (1981).

The existence of an appropriate choice of eigenvectors is guaranteed by Theorem 7. How to choose such a collection of eigenvectors in practice is left for future research. In this direction, we may note that the two clustering methods considered above (i.e., k-means and hierarchical) are invariant by isometries. So the main question concerns the choice of the normalization of an arbitrary collection of eigenvectors.

Remark 9 *Note that if one is only interested in the consistency property, then this result could be obtained through another route. Indeed, it is shown in Biau et al. (2007) that the neighborhood graph with connectivity radius h has asymptotically the same number of connected components as* *the level set. Hence, splitting the graph into its connected components leads to the desired clustering as well. But Theorem 7, by giving the asymptotic representation of the data when embedded in the feature space $\mathbb{R}^\ell$, provides additional insight into spectral clustering algorithms. In particular, Theorem 7 provides a rationale for the heuristic of Zelnik-Manor and Perona (2004) for automatic selection of the number of groups. Their idea is to quantify the amount of concentration of the points embedded in the feature space, and to select the number of groups leading to the maximal concentration. Their method compared favorably with the eigengap heuristic considered in von Luxburg (2007).*

4.3 Further Spectral Convergence

Naturally, the selection of the number of groups is also linked with the choice of the parameter h. In this direction, let us emphasize that the operators $\widehat{Q}_{n,h}$ and Q_h depend continuously on the scale parameter h. Thus, the spectral properties of both operators will be close to the ones stated in Theorem 7 if h is in a neighborhood of the interval $(0;d_{min})$. This follows from the continuity of an isolated set of eigenvalues, as stated in Appendix B. In particular, the sum of the eigenspaces of Q_h associated with the eigenvalues close to 1 is spanned by functions that are close to (in $W(\mathcal{L}(t))$-norm) the characteristic functions of the connected components of $\mathcal{L}(t)$. Hence, the representation of the data set in the feature space $\mathbb{R}^\ell$ still concentrates on some neighborhoods of e_k, $1 \le k \le \ell$ and a simple clustering algorithm such as the k-means algorithm will still lead to the desired partition. This is made precise in the following Theorem.

Theorem 10 *Suppose that assumptions 1 and 2 hold. There exists $h_{max} > d_{min}$ which depends only on the density f, such that, for any $h \in (0;h_{max})$, the event “for all n large enough, the representation of the extracted data set in the feature space, namely $\{\rho_n(X_j)\}_{j\in J(n)}$, concentrates in ℓ cubes of $\mathbb{R}^\ell$ that do not overlap” has probability one. Moreover, on this event of probability one, the ℓ cubes are in one-to-one correspondence with the ℓ connected component of $\mathcal{L}(t)$. Hence, for all n large enough, each $\rho_n(X_j)$ with $j \in J(\infty)$ is in the cube corresponding to the $k(j)^{th}$ cluster for all n large enough.*

This result contrasts with the graph techniques used to recover the connected components, as in, for example, Biau et al. (2007), where an unweighted graph is defined by connecting two observations if and only if their distance is smaller than h. The partition is then obtained by the connected components of the graph. However, when h is taken slightly larger than the critical value d_{min}, at least two connected components cannot be separated using the graph partitioning algorithm.

Proof Let us begin with the following consequence of Proposition 6. For all $h \le d_{min}$ the ℓ largest eigenvalues of Q_h are all equal to 1 and the corresponding eigenspace is spanned by the indicator functions of the connected components of the t-level set. Moreover, 1 is an isolated eigenvalue of $Q_{d_{min}}$, that is, there exists η_0 in the interval $(0;1)$ such that $\sigma(Q_{d_{min}}) \cap \{z \in \mathbb{C} : |z-1| < \eta_0\}$ is the singleton $\{1\}$.

We choose an arbitrary constant C_0 in $(0;1/2)$. Since $h \mapsto Q_h$ is continuous for the topology of the operator norm, Theorem 19 implies that there exists a neighborhood $(h_{min};h_{max})$ of d_{min} such that, for all h in $(h_{min};h_{max})$,

(i) Q_h has exactly ℓ eigenvalues in $\{z \in \mathbb{C} : |z-1| < \eta_0\}$;

(ii) the sum of the corresponding eigenspaces of Q_h is spanned by ℓ functions, say $g_1,\dots,g_\ell$, at distance (in $\|\cdot\|_W$-norm) less than $C_0/2$ from the indicator functions of the connected components

of $\mathcal{L}(t)$:

$$\|g_k - \mathbf{1}_{\mathcal{C}_k}\|_\infty \leq \|g_k - \mathbf{1}_{\mathcal{C}_k}\|_W < C_0/2 \quad \text{for } k = 1, \ldots, \ell. \tag{34}$$

Now, fix h in $(d_{min}; h_{max})$. We follow the arguments leading to Theorem 7. The convergence in (33) becomes

$$\lim_{n\to\infty} \vartheta_{n,k,j} = g_k(X_j) \quad \text{almost surely.}$$

Hence, there exists n_0 such that, for all $n \geq n_0$, $j \in J(n)$ and $k \in \{1, \ldots, \ell\}$, we have $|\vartheta_{n,k,j} - g_k(X_j)| < C_0/2$. With the triangular inequality and (34), we obtain $|\vartheta_{n,k,j} - \mathbf{1}_{\mathcal{C}_k}(X_j)| < C_0$, that is, the representation of the extracted data set in the feature space concentrates in cubes with edge length $2C_0$, centered at e_k, $k = 1, \ldots, \ell$, up to a linear transformation of $\mathbb{R}^\ell$, for all n large enough. Moreover, if X_j with $j \in J(\infty)$ lies in $\mathcal{C}_{k(j)}$, then its representation is in the cube centered at $e_{k(j)}$. Since those cubes have edge length $2C_0 < 1$, they do not overlap. Hence, a classical method such as the k-means algorithm will asymptotically partition the extracted data set as desired. ■

4.4 Generalizations and Open Problems

Our results allow to relate the limit partition of a spectral clustering algorithm with the connected components of either the support of the distribution (case $t = 0$) or of an upper level set of the density (case $t > 0$). This holds for a fixed similarity function with compact support. Interestingly, the scale parameter h of the similarity function may be larger than the minimal distance between two connected components, up to a threshold value h_{max} above which we have no theoretical guarantee that the connected components will be recovered.

Several important questions, though, remain largely open. Among these, interpreting the limit partition of the classical spectral clustering algorithm with the underlying distribution when one asks for more groups than the number of connected components of its support remains largely an unsolved problem. Also in practice, a sequence h_n decreasing to 0 with the number of observations is frequently used for the scale parameter of the similarity function, and to the best of our knowledge, no convergence results have been obtained yet. At last, it would be interesting to alleviate the assumption of compact support on the similarity function. Indeed, a gaussian kernel is a popular choice in practice. In this direction, one possibility would be to consider a sequence of functions with compact support converging towards the gaussian kernel at an appropriate rate.

5. Auxiliary Results for the Operator Norm Convergence

In this section we give technical lemmas that were needed in the proof of our main results. We also recall several facts from empirical process theory in Section 5.2.

5.1 Preliminaries

Let us start with the following simple lemma.

Lemma 11 *Let $\{A_n\}_{n\geq 0}$ be a decreasing sequence of Borel sets in $\mathbb{R}^d$, with limit $A_\infty = \cap_{n\geq 0} A_n$. If $\mu(A_\infty) = 0$, then*

$$\mathbb{P}_n A_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \in A_n\} \to 0 \quad \textit{almost surely as } n \to \infty,$$

where $\mathbb{P}_n$ is the empirical measure associated with the random sample $X_1,\dots,X_n$.

Proof First, note that $\lim_n \mu(A_n) = \mu(A_\infty)$. Next, fix an integer k. For all $n \geq k$, $A_n \subset A_k$ and so $\mathbb{P}_n A_n \leq \mathbb{P}_n A_k$. But $\lim_n \mathbb{P}_n A_k = \mu(A_k)$ almost surely by the law of large numbers. Consequently $\limsup_n \mathbb{P}_n A_n \leq \mu(A_k)$ almost surely. Letting $k \to \infty$ yields

$$\limsup_n \mathbb{P}_n A_n \leq \mu(A_\infty) = 0,$$

which concludes the proof since $\mathbb{P}_n A_n \geq 0$. ■

5.2 Uniform Laws of Large Number and Glivenko-Cantelli Classes

In this paragraph, we prove that some classes of functions satisfy a uniform law of large numbers. We shall use some facts on empirical processes that we briefly summarize below. For materials on the subject, we refer the reader to Chapter 19 in van der Vaart (1998) and the book by van der Vaart and Wellner (2000).

A collection $\mathcal{F}$ of functions is Glivenko-Cantelli if it satisfies a uniform law of large numbers, that is, if

$$\sup_{g\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}[X] \right| \to 0 \quad \text{almost surely,}$$

where $(X_n)_n$ is an i.i.d. sequence of random variables with the same distribution as the random variable X. That a class $\mathcal{F}$ is Glivenko-Cantelli depends on its size. A simple way of measuring the size of $\mathcal{F}$ is in terms of bracketing numbers.

A *bracket* $[f_l, f_u]$ is the set of functions g in $\mathcal{F}$ such that $f_u \leq g \leq f_u$, and an ε-*bracket in* L^p is a bracket $[f_l, f_u]$ such that $\mathbb{E}[(f_u(X) - f_l(X))^p]^{1/p} < \varepsilon$. The *bracketing number* $N_{[]}(\varepsilon, \mathcal{F}, L^p)$ is the minimal number of ε-brackets of size ε in the L^p norm which are needed to cover $\mathcal{F}$. A sufficient condition for a class $\mathcal{F}$ to be Glivenko-Cantelli is that $N_{[]}(\varepsilon, \mathcal{F}, L^1)$ is finite for all $\varepsilon > 0$ (Theorem 2.4.1, van der Vaart and Wellner, 2000, p. 122).

A bound on the L^1-bracketing number of a class $\mathcal{F}$ may be obtained from a bound on its metric entropy in the uniform norm, if appropriate. An ε-*covering of* $\mathcal{F}$ in the supremum norm is a collection of N balls of radius ε and centered at functions $f_1,\dots,f_N$ in $\mathcal{F}$ whose union covers $\mathcal{F}$. For ease of notation, an ε-covering of $\mathcal{F}$ is denoted by the centers of the balls $f_1 \dots, f_N$. The minimal number $\mathcal{N}(\varepsilon, \mathcal{F}, \|.\|_\infty)$ of balls of radius ε in the supremum norm that are needed to cover $\mathcal{F}$ is called the *covering number* of $\mathcal{F}$ in the uniform norm. The *entropy* of the class is the logarithm of the covering number, and $\mathcal{F}$ is said to have *finite entropy* if $\mathcal{N}(\varepsilon, \mathcal{F}, \|.\|_\infty)$ is finite for all ε. If a class $\mathcal{F}$ may be covered by finitely many balls of radius ε in the supremum norm and centered at $f_1,\dots,f_N$, then the brackets $[f_i - \varepsilon; f_i + \varepsilon]$ have size at most 2ε for the L^1 norm and their union covers $\mathcal{F}$. This argument is used to conclude the proof of Lemma 13 below.

Lemma 12 *The two collections of functions*

$$\mathcal{F}_1 := \left\{ y \mapsto k_h(y-x)\mathbf{1}_{\mathcal{L}(t)}(y) \,:\, x \in \mathcal{L}(t-\varepsilon_0) \right\},$$
$$\mathcal{F}_2 := \left\{ y \mapsto D_x k_h(y-x)\mathbf{1}_{\mathcal{L}(t)}(y) \,:\, x \in \mathcal{L}(t-\varepsilon_0) \right\},$$

are Glivenko-Cantelli, where $D_x k_h$ denotes the differential of k_h.

Proof Denote by g_x the functions in $\mathcal{F}_1$, for x ranging in $\mathcal{L}(t-\varepsilon_0)$. We proceed by constructing a covering of $\mathcal{F}_1$ by finitely many L^1-brackets of an arbitrary size, as in, for example, Example 19.8 in van der Vaart (1998). Denote by Q a probability measure on $\mathcal{L}(t)$. Let $\delta > 0$. Since $\mathcal{L}(t-\varepsilon_0)$ is compact, it can be covered by finitely many balls of radius δ, that is, there exists an integer N and points $x_1,\dots,x_N$ in $\mathcal{L}(t-\varepsilon_0)$ such that $\mathcal{L}(t-\varepsilon_0) \subset \bigcup_{i=1}^N B(x_i,\delta)$. Define the functions $g_{i,\delta}^l$ and $g_{i,\delta}^u$ respectively by

$$g_{i,\delta}^l(y) = \inf_{x\in B(x_i,\delta)} g_x(y) \quad \text{and} \quad g_{i,\delta}^u(y) = \sup_{x\in B(x_i,\delta)} g_x(y).$$

Then the union of brackets $[g_{i,\delta}^l, g_{i,\delta}^u]$, for $i = 1,\dots,N$, covers $\mathcal{F}_1$. Observe that $|g_x(y)| \leq \|k_h\|_\infty$ for all $x \in \mathcal{L}(t-\varepsilon_0)$ and all $y \in \mathcal{L}(t)$ since k_h is uniformly bounded, and that for any fixed $y \in \mathcal{L}(t)$, the map $x \mapsto g_x(y)$ is continuous since k is of class $\mathcal{C}^2$ on $\mathbb{R}^d$ under Assumption 2. Therefore the function $g_{i,\delta}^u - g_{i,\delta}^l$ converges pointwise to 0 and $\|g_{i,\delta}^u - g_{i,\delta}^l\|_{L^1(Q)}$ goes to 0 as $\delta \to 0$ by the Lebesgue dominated convergence theorem. Consequently, for any $\varepsilon > 0$, one may choose a finite covering of $\mathcal{L}(t-\varepsilon_0)$ by N balls of radius $\delta > 0$ such that $\max_{i=1,\dots,N} \|g_{i,\delta}^u - g_{i,\delta}^l\|_{L^1(Q)} \leq \varepsilon$. Hence, for all $\varepsilon > 0$ the L^1-bracketing number of $\mathcal{F}_1$ is finite, so $\mathcal{F}_1$ is Glivenko-Cantelli. Since k_h is continuously differentiable, the same arguments apply to each component of $D_x k_h$, and so $\mathcal{F}_2$ is also a Glivenko-Cantelli class. ■

Lemma 13 *Let $r : \mathcal{L}(t) \times \mathbb{R}^d$ be a continuously differentiable function such that*
(i) there exists a compact $\mathcal{K} \subset \mathbb{R}^d$ such that $r(x,y) = 0$ for all $(x,y) \in \mathcal{L}(t) \times \mathcal{K}^c$;
(ii) r is uniformly bounded on $\mathcal{L}(t) \times \mathbb{R}^d$, that is, $\|r\|_\infty < \infty$.
Then the collection of functions

$$\mathcal{F}_3 := \Big\{ y \mapsto r(x,y) g(y) \mathbf{1}_{\mathcal{L}(t)}(y) \,:\, x \in \mathcal{L}(t),\, \|g\|_{W(\mathcal{L}(t))} \leq 1 \Big\}$$

is Glivenko-Cantelli.

Proof Set $\mathcal{R} = \{y \mapsto r(x,y) : x \in \mathcal{L}(t)\}$. Since r is continuous on the compact set $\mathcal{L}(t) \times \mathcal{K}$, it is uniformly continuous. So for any $\varepsilon > 0$, there exists $\delta > 0$ such that $|r(x,y) - r(x',y')| \leq \varepsilon$ whenever the points (x,y) and (x',y') in $\mathcal{L}(t) \times \mathcal{K}$ are at a distance no more than δ. Since $\mathcal{L}(t)$ is compact, it may be covered by finitely many balls of radius δ centered at N points $x_1,\dots,x_N$ of $\mathcal{L}(t)$. Denote by g_i the function in $\mathcal{R}$ defined by $g_i(y) = r(x_i,y)$, and let $\mathcal{R}_i = \{y \mapsto r(x,y) : x \in \mathcal{L}(t)\,,\, \|x - x_i\| \leq \delta\}$. Then the union of the $\mathcal{R}_i$'s cover $\mathcal{R}$, and for any g in $\mathcal{R}_i$, $\|g - g_i\|_\infty \leq \varepsilon$. This shows that $\mathcal{R}$ has finite entropy in the supremum norm, that is, that $\mathcal{N}(\varepsilon, \mathcal{R}, \|.\|_\infty) < \infty$.

Second, consider the unit ball $\mathcal{G}$ in $W(\mathcal{L}(t))$, that is, $\mathcal{G} = \{g : \mathcal{L}(t) \to \mathbb{C} : \|g\|_{W(\mathcal{L}(t))} \leq 1\}$. Denote by $\mathcal{X}$ the convex hull of $\mathcal{L}(t)$, and consider the collection of functions $\tilde{\mathcal{G}} = \{\tilde{g} : \mathcal{X} \to \mathbb{C} : \|\tilde{g}\|_{W(\mathcal{X})} \leq 1\}$. Observe that $\tilde{\mathcal{G}}$ is a subset of the Holder space $C^{0,1}(\mathcal{X})$. It is proved in Theorem 2.7.1, p. 155 in the book by van der Vaart and Wellner (2000) that if $\mathcal{X}$ is a convex bounded subset of $\mathbb{R}^d$, then $C^{0,1}(\mathcal{X})$ has finite entropy in the uniform norm (this theorem was established in van der Vaart (1994) using results of Kolmogorov and Tikhomirov (1961). Consequently, for any $\varepsilon > 0$, there exist N functions $\tilde{g}_1,\dots,\tilde{g}_n$ in $\tilde{\mathcal{G}}$ such that the union of the sets $\{\tilde{g} \in \tilde{\mathcal{G}} : \|\tilde{g} - \tilde{g}_i\|_\infty \leq \varepsilon\}$ covers $\tilde{\mathcal{G}}$. By considering the restrictions g_i of each $\tilde{g}_i$ to $\mathcal{L}$, it follows that the union of the sets $\{g \in \mathcal{G} : \|g - g_i\|_\infty \leq \varepsilon\}$ covers $\mathcal{G}$. So $\mathcal{N}(\varepsilon, \mathcal{G}, \|.\|_\infty) < \infty$ for any $\varepsilon > 0$.

Now fix $\varepsilon > 0$. Let $r_1, \ldots, r_M \in \mathcal{R}$ be an ε-covering of $\mathcal{R}$ in the supremum norm, and let $g_1, \ldots, g_N \in \mathcal{G}$ be an ε-covering of $\mathcal{G}$ in the supremum norm, for some integers M and N. For any function f in $\mathcal{F}_3$ of the form $f(y) = r(x,y)g(y)\mathbf{1}_{\mathcal{L}(t)}$ for some $x \in \mathcal{L}(t)$ and $g \in W(\mathcal{L}(t))$ with $\|g\|_{W(\mathcal{L}(t))} \leq 1$, there exists $1 \leq i \leq M$ and $1 \leq j \leq N$ such that $\|r(x,.) - r_i\|_\infty \leq \varepsilon$ and $\|g - g_j\|_\infty \leq \varepsilon$. Then

$$\begin{aligned}\sup_{y \in \mathbb{R}^d} |f(y) - r_i(y)g_j(y)\mathbf{1}_{\mathcal{L}(t)}(y)| &= \sup_{y \in \mathcal{L}(t)} |r(x,y)g(y) - r_i(y)g_j(y)| \\ &= \sup_{y \in \mathcal{L}(t)} \big|(r(x,y) - r_i(y))g(y) + r_i(y)(g(y) - g_j(y))\big| \\ &\leq \sup_{y \in \mathcal{L}(t)} |r(x,y) - r_i(y)| \, \|g\|_\infty + \|r_i\|_\infty \sup_{y \in \mathcal{L}(t)} \big|g(y) - g_j(y)\big| \\ &\leq \varepsilon + \|r\|_\infty \varepsilon,\end{aligned}$$

since $\|r_i\|_\infty = 1$ for all $i = 1, \ldots, M$ and since $\|g\|_\infty \leq \varepsilon$. So the collection of functions $f_{ij} : y \mapsto r_i(y)g_j(y)\mathbf{1}_{\mathcal{L}(t)}(y)$ form a finite covering of $\mathcal{F}_3$ of size $M \times N$ by balls of radius $(1 + \|r\|_\infty)\varepsilon$ in the supremum norm, and $\mathcal{N}(\varepsilon, \mathcal{F}_3, \|.\|_\infty) < \infty$ for all $\varepsilon > 0$.

To conclude the proof, observe that if $f_1, \ldots, f_N \in \mathcal{F}_3$ is an ε-covering of $\mathcal{F}_3$ in the supremum norm, then the brackets $[f_i - \varepsilon; f_i + \varepsilon]$ have size at most 2ε in the L^1 norm, and their union covers $\mathcal{F}_3$. So for all $\varepsilon > 0$ the L^1-bracketing number of $\mathcal{F}_3$ is finite and $\mathcal{F}_3$ is Glivenko-Cantelli. ■

5.3 Bounds on Kernels

We recall that the limit operator Q_h is given by (18). The following lemma gives useful bounds on K_h and q_h, both defined in (19).

Lemma 14 *1. The function K_h is uniformly bounded from below by some positive number on $\mathcal{L}(t - \varepsilon_0)$, that is,* $\inf\{K_h(x) : x \in \mathcal{L}(t - \varepsilon_0)\} > 0$*;*
2. The kernel q_h is uniformly bounded, that is, $\|q_h\|_\infty < \infty$;
3. The differential of q_h with respect to x is uniformly bounded on $\mathcal{L}(t - \varepsilon_0) \times \mathbb{R}^d$, that is, $\sup\big\{\|D_x q_h(x,y)\| : (x,y) \in \mathcal{L}(t - \varepsilon_0) \times \mathbb{R}^d\big\} < \infty$*;*
4. The Hessian of q_h with respect to x is uniformly bounded on $\mathcal{L}(t - \varepsilon_0) \times \mathbb{R}^d$, that is, $\sup\big\{\|D_x^2 q_h(x,y)\| : (x,y) \in \mathcal{L}(t - \varepsilon_0) \times \mathbb{R}^d\big\} < \infty$.

Proof First observe that the statements 2, 3 and 4 are immediate consequences of statement 1 together with the fact that the function k_h is of class $\mathcal{C}^2$ with compact support, which implies that $k_h(y - x)$, $D_x k_h(y - x)$, and $D_x^2 k_h(y - x)$ are uniformly bounded.

To prove statement 1, note that K_h is continuous and that $K_h(x) > 0$ for all $x \in \mathcal{L}(t)$. Set

$$\alpha(\varepsilon_0) = \inf\big\{\|D_x f(x)\|; x \in \mathcal{L}_{t-\varepsilon_0}^t\big\}.$$

Let $(x,y) \in \mathcal{L}_{t-\varepsilon_0}^t \times \partial\mathcal{L}(t)$. Then

$$\varepsilon_0 \geq f(y) - f(x) \geq \alpha(\varepsilon_0)\|y - x\|.$$

Thus, $\|y - x\| \leq \varepsilon_0 / \alpha(\varepsilon_0)$ and so

$$\operatorname{dist}\big(x, \mathcal{L}(t)\big) \leq \frac{\varepsilon_0}{\alpha(\varepsilon_0)}, \quad \text{for all } x \in \mathcal{L}_{t-\varepsilon_0}^t.$$

Recall from (4) that $h/2 > \varepsilon_0/\alpha(\varepsilon_0)$. Consequently, for all $x \in \mathcal{L}(t-\varepsilon_0)$, the set $(x+hB/2) \cap \mathcal{L}(t)$ contains a non-empty, open set $U(x)$. Moreover k_h is bounded from below by some positive number on $hB/2$ by Assumption 2. Hence $K_h(x) > 0$ for all x in $\mathcal{L}(t-\varepsilon_0)$ and point 1 follows from the continuity of K_h and the compactness of $\mathcal{L}(t-\varepsilon_0)$. ■

In order to prove the convergence of $\widehat{Q}_{n,h}$ to Q_h, we also need to study the uniform convergence of $K_{n,h}$, given in (2). Lemma 15 controls the difference between $K_{n,h}$ and K_h, while Lemma 16 controls the ratio of K_h over $K_{n,h}$.

Lemma 15 *As $n \to \infty$, almost surely,*

1. $\sup_{x \in \mathcal{L}(t-\varepsilon_0)} \left| K_{n,h}(x) - K_h(x) \right| \to 0$ *and*

2. $\sup_{x \in \mathcal{L}(t-\varepsilon_0)} \left| D_x K_{n,h}(x) - D_x K_h(x) \right| \to 0.$

Proof Let

$$K^{\dagger}_{n,h}(x) := \frac{1}{n\mu(\mathcal{L}(t))} \sum_{i=1}^{n} k_h(X_i - x)\mathbf{1}_{\mathcal{L}_n(t)}(X_i), \quad K^{\dagger\dagger}_{n,h}(x) := \frac{1}{n\mu(\mathcal{L}(t))} \sum_{i=1}^{n} k_h(X_i - x)\mathbf{1}_{\mathcal{L}(t)}(X_i).$$

Let us start with the inequality

$$\left| K_{n,h}(x) - K_h(x) \right| \leq \left| K_{n,h}(x) - K^{\dagger}_{n,h}(x) \right| + \left| K^{\dagger}_{n,h}(x) - K_h(x) \right|, \tag{35}$$

for all $x \in \mathcal{L}(t-\varepsilon_0)$. Using the inequality

$$\left| K_{n,h}(x) - K^{\dagger}_{n,h}(x) \right| \leq \left| \frac{n}{j(n)} - \frac{1}{\mu(\mathcal{L}(t))} \right| \|k_h\|_\infty$$

we conclude that the first term in (35) tends to 0 uniformly in x over $\mathcal{L}(t-\varepsilon_0)$ with probability one as $n \to \infty$, since $j(n)/n \to \mu\big(\mathcal{L}(t)\big)$ almost surely, and since k_h is bounded on $\mathbb{R}^d$.

Next, for all $x \in \mathcal{L}(t-\varepsilon_0)$, we have

$$\left| K^{\dagger}_{n,h}(x) - K_h(x) \right| \leq \left| K^{\dagger}_{n,h}(x) - K^{\dagger\dagger}_{n,h}(x) \right| + \left| K^{\dagger\dagger}_{n,h}(x) - K_h(x) \right|. \tag{36}$$

The first term in (36) is bounded by

$$\begin{aligned} \left| K^{\dagger}_{n,h}(x) - K^{\dagger\dagger}_{n,h}(x) \right| &\leq \frac{\|k_h\|_\infty}{\mu\big(\mathcal{L}(t)\big)} \frac{1}{n} \left| \sum_{i=1}^{n} \left\{ \mathbf{1}_{\mathcal{L}_n(t)}(X_i) - \mathbf{1}_{\mathcal{L}(t)}(X_i) \right\} \right| \\ &= \frac{\|k_h\|_\infty}{\mu\big(\mathcal{L}(t)\big)} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\mathcal{L}_n(t)\Delta\mathcal{L}(t)}(X_i), \end{aligned}$$

where $\mathcal{L}_n(t)\Delta\mathcal{L}(t)$ denotes the symmetric difference between $\mathcal{L}_n(t)$ and $\mathcal{L}(t)$. Recall that, on the event Ω_n, $\mathcal{L}(t-\varepsilon_n) \subset \mathcal{L}_n(t) \subset \mathcal{L}(t-\varepsilon_n)$. Therefore $\mathcal{L}_n(t)\Delta\mathcal{L}(t) \subset \mathcal{L}^{t+\varepsilon_n}_{t-\varepsilon_n}$ on Ω_n, and so

$$0 \leq \frac{1}{n} \left| \sum_{i=1}^{n} \left\{ \mathbf{1}_{\mathcal{L}_n(t)}(X_i) - \mathbf{1}_{\mathcal{L}(t)}(X_i) \right\} \right| \mathbf{1}_{\Omega_n} \leq \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{A_n}(X_i),$$

where $A_n = \mathcal{L}_{t-\varepsilon_n}^{t+\varepsilon_n}$. Hence by Lemma 11, and since $\mathbf{1}_{\Omega_n} \to 1$ almost surely as $n \to \infty$, the first term in (36) converges to 0 with probability one as $n \to \infty$.

Next, since the collection $\{y \mapsto k_h(y-x)\mathbf{1}_{\mathcal{L}(t)}(y) \,:\, x \in \mathcal{L}(t-\varepsilon_0)\}$ is Glivenko-Cantelli by Lemma 12, we conclude that

$$\sup_{x\in\mathcal{L}(t-\varepsilon_0)} \left| K_{n,h}^{\dagger\dagger}(x) - K_h(x) \right| \to 0,$$

with probability one as $n \to \infty$. This concludes the proof of the first statement.

The second statement may be proved by developing similar arguments, with k_h replaced by $D_x k_h$, and by noting that the collection of functions $\{y \mapsto D_x k_h(y-x)\mathbf{1}_{\mathcal{L}(t)}(y) \,:\, x \in \mathcal{L}(t-\varepsilon_0)\}$ is also Glivenko-Cantelli by Lemma 12. ■

Lemma 16 *As $n \to \infty$, almost surely,*

$$\sup_{x\in\mathcal{L}(t)} \left| \frac{K_h\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)} - 1 \right| \to 0, \qquad \textit{and} \qquad \sup_{x\in\mathcal{L}(t)} \left\| D_x \left[\frac{K_h\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)} \right] \right\| \to 0.$$

Proof First of all, K_h is uniformly continuous on $\mathcal{L}(t-\varepsilon_0)$ since K_h is continuous and since $\mathcal{L}(t-\varepsilon_0)$ is compact. Moreover, φ_n converges uniformly to the identity map of $\mathcal{L}(t)$ by Lemma 17. Hence

$$\sup_{x\in\mathcal{L}(t)} \left| K_h\big(\varphi_n(x)\big) - K_h(x) \right| \to 0 \quad \text{as } n \to \infty,$$

and since $K_{n,h}$ converges uniformly to K_h with probability one as $n \to \infty$ by Lemma 15, this proves the first convergence result.

We have

$$D_x \left[\frac{K_h\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)} \right] = \Big[K_{n,h}\big(\varphi_n(x)\big) \Big]^{-2} D_x\varphi_n(x) \Big[K_{n,h}\big(\varphi_n(x)\big) D_x K_h\big(\varphi_n(x)\big) - K_h\big(\varphi_n(x)\big) D_x K_{n,h}\big(\varphi_n(x)\big) \Big].$$

Since $D_x\varphi_n(x)$ converges to the identity matrix I_d uniformly over $x \in \mathcal{L}(t)$ by Lemma 17, $\|D_x\varphi_n(x)\|$ is bounded uniformly over n and $x \in \mathcal{L}(t)$ by some positive constant C_φ. Furthermore the map $x \mapsto K_{n,h}(x)$ is bounded from below over $\mathcal{L}(t)$ by some positive constant k_{min} independent of x because i) $\inf_{x\in\mathcal{L}(t-\varepsilon_0)} K_h(x) > 0$ by Lemma 14, and ii) $\sup_{x\in\mathcal{L}(t-\varepsilon_0)} \left| K_{n,h}(x) - K_h(x) \right| \to 0$ by Lemma 15. Hence

$$\left| D_x \left[\frac{K_h\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)} \right] \right| \leq \frac{C_\varphi}{k_{min}^2} \left| K_{n,h}(y) D_x K_h(y) - K_h(y) D_x K_{n,h}(y) \right|,$$

where we have set $y = \varphi_n(x)$ which belongs to $\mathcal{L}(t-\varepsilon_n) \subset \mathcal{L}(t-\varepsilon_0)$. At last, Lemma 15 gives

$$\sup_{y\in\mathcal{L}(t-\varepsilon_0)} \left| K_{n,h}(y) D_x K_h(y) - K_h(y) D_x K_{n,h}(y) \right| \to 0 \quad \text{almost surely},$$

as $n \to \infty$ which proves the second convergence result. ■

Acknowledgments

This work was supported by the French National Research Agency (ANR) under grant ANR-09-BLAN-0051-01. We thank the referees and the Associate Editor for valuable comments and insightful suggestions that led to an improved version of the paper.

Appendix A. Geometry of Level Sets

The proof of the following result is adapted from Theorem 3.1 in (Milnor, 1963, p. 12) and Theorem 5.2.1 in (Jost, 1995, p. 176)

Lemma 17 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function of class $\mathcal{C}^2$. Let $t \in \mathbb{R}$ and suppose that there exists $\varepsilon_0 > 0$ such that $f^{-1}\big([t-\varepsilon_0; t+\varepsilon_0]\big)$ is non empty, compact and contains no critical point of f. Let $\{\varepsilon_n\}_n$ be a sequence of positive numbers such that $\varepsilon_n < \varepsilon_0$ for all n, and $\varepsilon_n \to 0$ as $n \to \infty$. Then there exists a sequence of diffeomorphisms $\varphi_n : \mathcal{L}(t) \to \mathcal{L}(t-\varepsilon_n)$ carrying $\mathcal{L}(t)$ to $\mathcal{L}(t-\varepsilon_n)$ such that:*
1. $\sup_{x \in \mathcal{L}(t)} \|\varphi_n(x) - x\| \to 0$ *and*
2. $\sup_{x \in \mathcal{L}(t)} \|D_x\varphi_n(x) - I_d\| \to 0$*,*
as $n \to \infty$, where $D_x\varphi_n$ denotes the differential of φ_n and where I_d is the identity matrix on $\mathbb{R}^d$.

Proof Recall first that a one-parameter group of diffeomorphisms $\{\varphi_u\}_{u\in\mathbb{R}}$ of $\mathbb{R}^d$ gives rise to a vector field V defined by

$$V_x g = \lim_{u\to 0} \frac{g\big(\varphi_u(x)\big) - g(x)}{u}, \quad x \in \mathbb{R}^d,$$

for all smooth function $g : \mathbb{R}^d \to \mathbb{R}$. Conversely, a smooth vector field which vanishes outside of a compact set generates a unique one-parameter group of diffeomorphisms of $\mathbb{R}^d$; see Lemma 2.4 in (Milnor, 1963, p. 10) and Theorem 1.6.2 in (Jost, 1995, p. 42)

Denote the set $\{x \in \mathbb{R}^d : a \leq f(x) \leq b\}$ by $\mathcal{L}_a^b$, for $a \leq b$. Let $\eta : \mathbb{R}^d \to \mathbb{R}$ be the non-negative differentiable function with compact support defined by

$$\eta(x) = \begin{cases} 1/\|D_x f(x)\|^2 & \text{if } x \in \mathcal{L}_{t-\varepsilon_0}^t, \\ (t+\varepsilon_0 - f(x))/\|D_x f(x)\|^2 & \text{if } x \in \mathcal{L}_t^{t+\varepsilon_0}, \\ 0 & \text{otherwise.} \end{cases}$$

Then the vector field V defined by $V_x = \eta(x) D_x f(x)$ has compact support $\mathcal{L}_{t-\varepsilon_0}^{t+\varepsilon_0}$, so that V generates a one-parameter group of diffeomorphisms

$$\varphi_u : \mathbb{R}^d \to \mathbb{R}^d, \quad u \in \mathbb{R}.$$

We have

$$D_u\left[f\big(\varphi_u(x)\big)\right] = \langle V, D_x f\rangle_{\varphi_u(x)} \geq 0,$$

since η is non-negative. Furthermore,

$$\langle V, D_x f\rangle_{\varphi_u(x)} = 1, \quad \text{if } \varphi_u(x) \in \mathcal{L}_{t-\varepsilon_0}^t$$

Consequently the map $u \mapsto f\big(\varphi_u(x)\big)$ has constant derivative 1 as long as $\varphi_u(x)$ lies in $\mathcal{L}_{t-\varepsilon_0}^{t}$. This proves the existence of the diffeomorphism $\varphi_n := \varphi_{-\varepsilon_n}$ which carries $\mathcal{L}(t)$ to $\mathcal{L}(t-\varepsilon_n)$.

Note that the map $u \in \mathbb{R} \mapsto \varphi_u(x)$ is the integral curve of V with initial condition x. Without loss of generality, suppose that $\varepsilon_n \leq 1$. For all x in $\mathcal{L}_{t-\varepsilon_0}^{t+\varepsilon_0}$, we have

$$\|\varphi_n(x) - x\| \leq \int_{-\varepsilon_n}^{0} \big\|D_u\big(\varphi_u(x)\big)\big\|\, du \leq \varepsilon_n/\beta(\varepsilon_n) \leq \varepsilon_n/\beta(\varepsilon_0)$$

where we have set

$$\beta(\varepsilon) := \inf\big\{\|D_x f(x)\| \,:\, x \in \mathcal{L}_{t-\varepsilon}^{t+\varepsilon}\big\} > 0.$$

This proves the statement 1, since $\varphi_n(x) - x$ is identically 0 on $\mathcal{L}(t+\varepsilon_0)$.

For the statement 2, observe that $\varphi_u(x)$ satisfies the relation

$$\varphi_u(x) - x = \int_0^u D_v\big(\varphi_v(x)\big)\,dv = \int_0^u V\big(\varphi_v(x))\big)\,dv.$$

Differentiating with respect to x yields

$$D_x\varphi_u(x) - I_d = \int_0^u D_x\varphi_v(x) \circ D_x V\big(\varphi_v(x)\big)\,dv.$$

Since f is of class $\mathcal{C}^2$, the two terms inside the integral are uniformly bounded over $\mathcal{L}_{t-\varepsilon_0}^{t+\varepsilon_0}$, so that there exists a constant $C > 0$ such that

$$\|D_x\varphi_n - I\|_x \leq C\varepsilon_n,$$

for all x in $\mathcal{L}_{t-\varepsilon_0}^{t+\varepsilon_0}$. Since $\|D_x\varphi_n - I\|_x$ is identically zero on $\mathcal{L}(t+\varepsilon_0)$, this proves the statement 2. ■

Appendix B. Continuity of an Isolated Finite Set of Eigenvalues

In brief, the spectrum $\sigma(T)$ of a bounded linear operator T on a Banach space is upper semi-continuous in T, but not lower semi-continuous; see Kato (1995), IV§3.1 and IV§3.2. However, an isolated finite set of eigenvalues of T is continuous in T, as stated in Theorem 19 below.

Let T be a bounded operator on the $\mathbb{C}$-Banach space E with spectrum $\sigma(T)$. Let $\sigma_1(T)$ be a finite set of eigenvalues of T. Set $\sigma_2(T) = \sigma(T) \setminus \sigma_1(T)$ and suppose that $\sigma_1(T)$ is separated from $\sigma_2(T)$ by a rectifiable, simple, and closed curve Γ. Assume that a neighborhood of $\sigma_1(T)$ is enclosed in the interior of Γ. Then we have the following theorem; see Kato (1995), III.§6.4 and III.§6.5.

Theorem 18 (Separation of the spectrum) *The Banach space E decomposes into a pair of supplementary subspaces as $E = M_1 \oplus M_2$ such that T maps M_j into M_j ($j = 1,2$) and the spectrum of the operator induced by T on M_j is $\sigma_j(T)$ ($j = 1,2$). If additionally the total multiplicity m of $\sigma_1(T)$ is finite, then* $\dim(M_1) = m$.

Moreover, the following theorem states that a finite system of eigenvalues of T, as well as the decomposition of E of Theorem 18, depends continuously of T, see Kato (1995), IV.§3.5. Let $\{T_n\}_n$ be a sequence of operators which converges to T in norm. Denote by $\sigma_1(T_n)$ the part of the spectrum of T_n enclosed in the interior of the closed curve Γ, and by $\sigma_2(T_n)$ the remainder of the spectrum of T_n.

Theorem 19 (Continuous approximation of the spectral decomposition) *There exists a finite integer n_0 such that the following holds true.*
1. Both $\sigma_1(T_n)$ and $\sigma_2(T_n)$ are nonempty for all $n \geq n_0$ provided this is true for T.
2. For each $n \geq 0$, the Banach space E decomposes into two subspaces as $E = M_{n,1} \oplus M_{n,2}$ in the manner of Theorem 18, that is, T_n maps $M_{n,j}$ into itself and the spectrum of T_n on $M_{n,j}$ is $\sigma_j(T_n)$.
3. For all $n \geq n_0$, $M_{n,j}$ is isomorphic to M_j.
4. If $\sigma_1(T)$ is a singleton $\{\lambda\}$, then every sequence $\{\lambda_n\}_n$ with $\lambda_n \in \sigma_1(T_n)$ for all $n \geq n_0$ converges to λ.
5. If Π is the projector on M_1 along M_2 and Π_n the projector on $M_{n,1}$ along $M_{n,2}$, then Π_n converges in norm to Π.
6. If the total multiplicity m of $\sigma_1(T)$ is finite, then, for all $n \geq n_0$, the total multiplicity of $\sigma_1(T_n)$ is also m and $\dim(M_{n,1}) = m$.

Appendix C. Background Materials on Markov Chains

For the reader not familiar with Markov chains on a general state space, we begin by summarizing the relevant part of the theory.

Let $\{\xi_i\}_{i\geq 0}$ be a Markov chain with state space $\mathcal{S} \subset \mathbb{R}^d$ and transition kernel $q(x,dy)$. We write P_x for the probability measure when the initial state is x and E_x for the expectation with respect to P_x. The Markov chain is called *(strongly) Feller* if the map

$$x \in \mathcal{S} \mapsto Qg(x) := \int_{\mathcal{S}} q(x,dy)g(y) = \mathbb{E}_x f(\xi_1)$$

is continuous for every bounded, measurable function g on $\mathcal{S}$; see (Meyn and Tweedie, 1993, p. 132). This condition ensures that the chain behaves nicely with the topology of the state space $\mathcal{S}$. The notion of irreducibility expresses the idea that, from an arbitrary initial point, each subset of the state space may be reached by the Markov chain with a positive probability. A Feller chain is said *open set irreducible* if, for every points x,y in $\mathcal{S}$, and every $\eta > 0$,

$$\sum_{n\geq 1} q^n(x, y + \eta B) > 0,$$

where $q^n(x,dy)$ stands for the n-step transition kernel; see (Meyn and Tweedie, 1993, p. 135). Even if open set irreducible, a Markov chain may exhibit a periodic behavior, that is, there may exist a partition $\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_N$ of the state space such that, for every initial state $x \in \mathcal{S}_0$,

$$P_x[\xi_1 \in \mathcal{S}_1, \xi_2 \in \mathcal{S}_2, \ldots, \xi_N \in \mathcal{S}_N, \xi_{N+1} \in \mathcal{S}_0, \ldots] = 1.$$

Such a behavior does not occur if the Feller chain is *topologically aperiodic*, that is, if for each initial state x, each $\eta > 0$, there exists n_0 such that $q^n(x, x + \eta B) > 0$ for every $n \geq n_0$; see (Meyn and Tweedie, 1993, p. 479).

Next we come to ergodic properties of the Markov chain. A Borel set A of $\mathcal{S}$ is called *Harris recurrent* if the chain visits A infinitely often with probability 1 when started at any point x of A, that is,

$$P_x\left(\sum_{i=0}^{\infty} \mathbf{1}_A(\xi_i) = \infty\right) = 1$$

for all $x \in A$. The chain is then said to be *Harris recurrent* if every Borel set A with positive Lebesgue measure is Harris recurrent; see (Meyn and Tweedie, 1993, p. 204). At least two types of behavior, called evanescence and non-evanescence, may occur. The event $[\xi_n \to \infty]$ denotes the fact that the sample path visits each compact set only finitely many often, and the Markov chain is called *non-evanescent* if $P_x(\xi_n \to \infty) = 0$ for each initial state $x \in \mathcal{S}$. Specifically, a Feller chain is Harris recurrent if and only if it is non-evanescent; see (Meyn and Tweedie, 1993, p. 122), Theorem 9.2.2.

The ergodic properties exposed above describe the long time behavior of the chain. A measure ν on the state space is said *invariant* if

$$\nu(A) = \int_{\mathcal{S}} q(x,A)\nu(dx)$$

for every Borel set A in $\mathcal{S}$. If the chain is Feller, open set irreducible, topologically aperiodic and Harris recurrent, it admits a unique (up to constant multiples) invariant measure ν; see (Meyn and Tweedie, 1993, p. 235), Theorem 10.0.1. In this case, either $\nu(\mathcal{S}) < \infty$ and the chain is called *positive*, or $\nu(\mathcal{S}) = \infty$ and the chain is called *null*. The following important result provides one with the limit of the distribution of ξ_n when $n \to \infty$, whatever the initial state is. Assuming that the chain is Feller, open set irreducible, topologically aperiodic and positive Harris recurrent, the sequence of distribution $\{q^n(x,dy)\}_{n\geq 1}$ converges in total variation to $\nu(dy)$, the unique invariant probability distribution; see Theorem 13.3.1 of (Meyn and Tweedie, 1993, p. 326). That is to say, for every x in $\mathcal{S}$,

$$\sup_g \left\{ \left| \int_{\mathcal{S}} g(y) q^n(x,dy) - \int_{\mathcal{S}} g(y)\nu(dy) \right| \right\} \to 0 \quad \text{as } n \to \infty,$$

where the supremum is taken over all continuous functions g from $\mathcal{S}$ to $\mathbb{R}$ with $\|g\|_\infty \leq 1$.

References

M.R. Anderberg. *Cluster Analysis for Applications*. Academic Press, New-York, 1973.

A. Azzalini and N. Torelli. Clustering via nonparametric estimation. *Stat. Comput.*, 17:71–80, 2007.

M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. In *Learning Theory*, volume 3559 of *Lecture Notes in Comput. Sci.*, pages 486–500. Springer, Berlin, 2005.

M. Belkin, I. Matveeva, and P. Niyogi. Regularization and semi-supervised learning on large graphs. In *Learning Theory*, volume 3120 of *Lecture Notes in Comput. Sci.*, pages 624–638. Springer, Berlin, 2004.

G. Biau, B. Cadre, and B. Pelletier. A graph-based estimator of the number of clusters. *ESAIM Probab. Stat.*, 11:272–280, 2007.

F.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.

R.R. Coifman and S. Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21:5–30, 2006.

A. Cuevas, M. Febrero, and R. Fraiman. Estimating the number of clusters. *Canadian Journal of Statistics*, 28:367–382, 2000.

A. Cuevas, M. Febrero, and R. Fraiman. Cluster analysis: a further approach based on density estimation. *Comput. Statist. Data Anal.*, 36:441–459, 2001.

R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley Interscience, New-York, 2000.

M. Fillipone, F. Camastra, F. Masulli, and S. Rovetta. A survey of kernel and spectral methods for clustering. *Pattern Recognition*, 41(1):176–190, 2008.

L.A. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar. A general trimming approach to robust cluster analysis. *Ann. Statis.*, 36(3):1324–1345, 2008.

E. Giné and V. Koltchinskii. Empirical graph Laplacian approximation of Laplace-Beltrami operators: large sample results. In *High Dimensional Probability*, volume 51 of *IMS Lecture Notes Monogr. Ser.*, pages 238–259. Inst. Math. Statist., Beachwood, OH, 2006.

J.A. Hartigan. *Clustering Algorithms*. Wiley, New-York, 1975.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New-York, 2001.

M. Hein, J.-Y. Audibert, and U. von Luxburg. Graph Laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 8:1325–1368, 2007.

J. Jost. *Riemannian Geometry and Geometric Analysis*. Universitext. Springer-Verlag, Berlin, 1995.

T. Kato. *Perturbation Theory for Linear Operators*. Classics in Mathematics. Springer-Verlag, Berlin, 1995. Reprint of the 1980 edition.

A.N. Kolmogorov and V.M. Tikhomirov. ε-entropy and ε-capacity of sets in functional space. *Amer. Math. Soc. Transl. (2)*, 17:277–364, 1961.

V.I. Koltchinskii. Asymptotics of spectral projections of some random matrices approximating integral operators. In *High Dimensional Probability (Oberwolfach, 1996)*, volume 43 of *Progr. Probab.*, pages 191–227. Birkhäuser, Basel, 1998.

T. Linder. Learning-theoretic methods in vector quantization. In *Principles of Nonparametric Learning (Udine, 2001)*, volume 434 of *CISM Courses and Lectures*, pages 163–210. Springer, Vienna, 2002.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkely Symp. Math. Statist. Prob.*, volume 1, pages 281–297, 1967.

G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, New-York, 2000.

S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Communications and Control Engineering Series. Springer-Verlag, London, 1993.

J.W. Milnor. *Morse Theory*. Annals of Mathematics Studies, No. 51. Princeton University Press, Princeton, N.J., 1963.

B. Nadler, S. Lafon, R.R. Coifman, and I.G. Kevrekidis. Difusion maps, spectral clustering and reaction coordinates of dynamical systems. *Appl. Comput. Harmon. Anal.*, 21(1):113–127, 2006.

A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, and Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, pages 849–856. MIT Press, 2002.

D. Pollard. Consistency of k-means clustering. *Ann. Statis.*, 9(1):135–140, 1981.

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

B. L. S. Prakasa Rao. *Nonparametric Functional Estimation*. Probability and Mathematical Statistics. Academic Press Inc., New York, 1983.

L. Rosasco, M. Belkin, and E. De Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11:905–934, 2010.

A.W. van der Vaart. Bracketing smooth functions. *Stochastic Process. Appl.*, 52(1):93–105, 1994.

A.W. van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics, Second Edition. Springer, New-York, 2000.

U. von Luxburg. A tutorial on spectral clustering. *Stat. Comput.*, 17(4):395–416, 2007.

U. von Luxburg and S. Ben-David. Towards a statistical theory of clustering. In *PASCAL Workshop on Statistics and Optimization of Clustering*, 2005.

U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *Ann. Statis.*, 36 (2):555–586, 2008.

L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Eighteenth Annual Conference on Neural Information Processing Systems (NIPS)*, 2004.