The Cheeger Constant: from Discrete to Continuous

Ery Arias-Castro, Bruno Pelletier, Pierre Pudlo

The Cheeger Constant: from Discrete to Continuous

Ery Arias-Castro*, Bruno Pelletier† and Pierre Pudlo‡

May 12, 2010

Abstract. Let M be a bounded domain of $\mathbb{R}^d$ with smooth boundary. We relate the Cheeger constant of M and the conductance of a neighborhood graph defined on a random sample from M. By restricting the minimization defining the latter over a particular class of subsets, we obtain consistency (after normalization) as the sample size increases, and show that any minimizing sequence of subsets has a subsequence converging to a Cheeger set of M.

Index Terms: Cheeger isoperimetric constant of a manifold, conductance of a graph, neighborhood graph, spectral clustering, U-processes, empirical processes.

AMS 2000 Classification: 62G05, 62G20.

1 Introduction and main results

The Cheeger isoperimetric constant may be defined for a Euclidean domain as well as for a graph. In either case it quantifies how well the set can be bisected or 'cut' into two pieces that are as little connected as possible. Motivated by recent developments in spectral clustering and computational geometry, we relate the Cheeger constant of a neighborhood graph defined on a sample from a domain and the Cheeger constant of the domain itself.

Given a graph G with weights $\{\delta_{ij}\}$, the *normalized cut of* a subset $S \subset G$ is defined as

$$h(S; G) = \frac{\sigma(S)}{\min\{\delta(S), \delta(S^c)\}}, \tag{1.1}$$

where S^c denotes the complement of S in G, and

$$\delta(S) = \sum_{i \in S} \sum_{j \neq i} \delta_{ij}, \quad \sigma(S) = \sum_{i \in S} \sum_{j \in S^c} \delta_{ij}, \tag{1.2}$$

are the discrete volume and perimeter of S. The *Cheeger constant* or *conductance* of the graph G is defined as the value of the optimal normalized cut over all non-empty subsets of G, i.e.

$$H(G) = \min\{h(S; G) : S \subset G, S \neq \emptyset\}. \tag{1.3}$$

A corresponding quantity can be defined for a domain of a Euclidean space. Let M be a bounded domain (i.e. open, connected subset) of $\mathbb{R}^d$ with smooth boundary ∂M. For an integer $1 \leq k \leq d$,

*Department of Mathematics, University of California, San Diego, USA
†Département de Mathématiques, IRMAR – UMR CNRS 6625, Université Rennes II, France
‡Département de Mathématiques, I3M – UMR CNRS 5149, Université Montpellier II, France

let Vol_k denote the k-dimensional volume (Hausdorff measure) in $\mathbb{R}^d$. For an open subset $A \subset \mathbb{R}^d$, define its normalized cut with respect to M by

$$h(A; M) = \frac{\mathrm{Vol}_{d-1}(\partial A \cap M)}{\min\{\mathrm{Vol}_d(A \cap M), \mathrm{Vol}_d(A^c \cap M)\}},$$

where A^c denotes the complement of A in $\mathbb{R}^d$ and with the convention that $0/0 = \infty$. The Cheeger (isoperimetric) constant of M is defined as

$$H(M) = \inf\{h(A; M) : \partial A \cap M \text{ is a smooth submanifold of co-dimension } 1\}.$$

Equivalently, the infimum may be restricted to all open subsets of M. This quantity was introduced by Cheeger [15] in order to bound the eigengap of the spectrum of the Laplacian on a manifold. A Cheeger set is a subset $A \subset M$ such that $h(A; M) = H(M)$; there is always a Cheeger set and it is unique under some conditions on the domain M [12]. For $A \subset M$, we call $\partial A \cap M$ its relative boundary.

1.1 Consistency of the normalized cut

Suppose that we observe an i.i.d. random sample $\mathcal{X}_n = (X_1, \ldots, X_n)$ from the uniform distribution μ on M. For $r > 0$, let $G_{n,r}$ be the graph with nodes the sample points and edge weights $\delta_{ij} = \mathbf{1}\{\|X_i - X_j\| \le r\}$, which is an instance of a random geometric graph [29]. Let ω_d denote the d-volume of the unit d-dimensional ball, and define

$$\gamma = \int_{\mathbb{R}^d} \big(\langle u, z\rangle\big)_+ \mathbf{1}\{\|z\| \le 1\}\, \mathrm{d}z, \tag{1.4}$$

where u is any unit-norm vector of $\mathbb{R}^d$. Actually γ is the average volume of a spherical cap when the height is chosen uniformly at random. We establish the pointwise consistency of the normalized cut, which yields an asymptotic upper bound on the Cheeger constant of the neighborhood graph based on the Cheeger constant of the manifold. This is the first result we know of that relates these two quantities.

Theorem 1. *Let A be a fixed subset of M with smooth relative boundary. Fix a sequence $r_n \to 0$ with $nr_n^{d+1}/\log n \to 0$, and let $S_n = A \cap G_{n,r_n}$. Then with probability one*

$$\frac{\omega_d}{\gamma r_n} h(S_n; G_{n,r_n}) \to h(A; M),$$

and, consequently,

$$\limsup_{n\to\infty} \frac{\omega_d}{\gamma r_n} H(G_{n,r_n}) \le H(M).$$

We do not know whether the Cheeger constant of the neighborhood graph, for an appropriate choice of the connectivity radius and properly normalized, converges to the Cheeger constant of the domain.

1.2 Consistent estimation of the Cheeger constant and Cheeger sets

We obtain a consistent estimator of the Cheeger constant $H(M)$ by restricting the minimization defining the conductance of the neighborhood graph (1.3) to subsets associated with subsets of $\mathbb{R}^d$ with controlled reach. The reach of a subset $S \subset \mathbb{R}^d$ [20], denoted $\mathrm{reach}(S)$, is the supremum over $\eta > 0$ such that, for each x within distance η of S, there is a unique point in S that is closest to x.

Theorem 2. *Let $\{r_n\}$, $\{\rho_n\}$ and $\{\alpha_n\}$ be sequences going to 0 as $n \to \infty$ such that $\rho_n > \alpha_n > 2r_n$. Let $\{\beta_n\}$ be a sequence such that $\beta_n \to \infty$ as $n \to \infty$ and suppose that*

$$\frac{\rho_n \alpha_n^{2d}}{\beta_n r_n} \to \infty \quad \text{and} \quad \frac{n r_n^{d+2}}{\log(n)} \to \infty.$$

Let $\mathcal{R}_n$ be a class of (bounded) open subsets $R \subset \mathbb{R}^d$ such that $\mathrm{Vol}_d(R) \leq \beta_n$ and $\mathrm{reach}(\partial R) \geq \rho_n$. Define the functional $h_n^\ddagger$ over $\mathcal{R}_n$ by

$$h_n^\ddagger(R) = \frac{\omega_d}{\gamma r_n} h\left(R \cap \mathcal{X}_n; G_{n,r_n}\right)$$

if both R and R^c contain a ball of radius α_n centered at a sample point and $h_n^\ddagger(R) = \infty$ otherwise.

(i) With probability one,

$$\min_{R \in \mathcal{R}_n} h_n^\ddagger(R) \to H(M), \quad n \to \infty.$$

(ii) Let $\{R_n\}$ be a sequence satisfying

$$R_n \in \mathcal{R}_n, \qquad h_n^\ddagger(R_n) = \min\{h_n^\ddagger(R) : R \in \mathcal{R}_n\}. \tag{1.5}$$

Then with probability one, $\{R_n \cap M\}$ admits a subsequence converging in the L^1-metric. Moreover, any subsequence of $\{R_n \cap M\}$ converging in the L^1-metric converges to a Cheeger set.

Note that the infimum defining R_n in (1.5) is attained in $\mathcal{R}_n$ since the function $h_n^\ddagger$ takes only a finite number of values. Under our conditions, if we choose r_n as a power of n then, up to $\log(n)$ factors, the smallest choice for r_n is on the order of $n^{-1/(d+2)}$. Similarly, the smallest orders for ρ_n and α_n are $n^{-1/((d+2)(1+2d))}$, up to $\log(n)$ factors. The sequence β_n is used to uniformly bound the perimeters over the class $\mathcal{R}_n$ by β_n/ρ_n, up to a multiple constant. It is only necessary that β_n becomes larger than $\mathrm{Vol}_d(M)$ for all n large enough, so that β_n can be allowed to grow slowly with n.

Part *(ii)* of Theorem 2 would provide a consistent estimate of a Cheeger set of M, if it where for the fact that only $R_n \cap M$ converges, which depends on M. On the other hand reconstructing an unknown set from a random sample of it is an independent problem for which there exists multiple techniques and an important literature (see e.g., [6] and the references therein). In the following Theorem, we construct a random discrete measure which does not require the knowledge of M, and we prove that its accumulation points are the uniform measures supported by a Cheeger set of M.

Theorem 3. *Let $\{R_n\}$ be a sequence as in Theorem 2-(ii), $\{R_{n_k}\}$ a subsequence of $\{R_n\}$ with $R_{n_k} \cap M \to A_\infty$ in L^1. Define the random discrete measure $Q_n = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{R_n}(X_i)\delta_{X_i}$ and the measure $Q = \mathbf{1}_{A_\infty}(.)\mu$. Then, that Q_n converges weakly to Q is an event which holds with probability one.*

As an example of an estimate of a Cheeger set of M, one can consider a union of balls of radius κ_n centered at the observations falling in R_n. Under appropriate conditions, it is known that this estimate converges in L^1; see [6].

Let us mention that with our result, only the "regular" part of a Cheeger set can be reconstructed. Indeed, in dimension $d \geq 8$, the boundary of a Cheeger set is not necessarily regular and may contain parts of codimension greater than 1.

1.3 Connections to the literature

Our results relating the respective Cheeger constants of a domain and of a neighborhood graph defined from a sample from the domain are the first of their kind, as far as we know. The connections to the literature stem from the concept of normalized cut taking a central place in graph partitioning and related methods in clustering; from a recent trend in computational geometry (and topology) aiming at estimating geometrical (and topological) attributes of a set based on a sample; and from the fact that we can use the conductance to bound the mixing time of a random walk on the neighborhood graph.

Clustering. In spectral graph partitioning, the goal is to partition a graph G into subgraphs based on the eigenvalues and eigenvectors of the Laplacian [32, 16]. It arises as a convex relaxation of the combinatorial search of finding an optimal bisection in terms of the normalized cut. Given a set of points $X_1, \ldots, X_n$ and a dissimilarity measure (or kernel) ϕ, spectral clustering applies spectral graph partitioning to the graph with nodes the data points and edge weight $\delta_{ij} = \phi(X_i, X_j)$ between X_i and X_j [34]. For instance, if the points are embedded in a Euclidean space, the kernel ϕ is often of the form $\phi(x, y) = \psi(\|x - y\|/\sigma)$, where σ is a tuning parameter, and ψ is, e.g., the Gaussian kernel $\psi(t) = \exp(-t^2)$ or the simple kernel $\psi(t) = \mathbf{1}_{[0,1]}(t)$ [26, 3]. The consistency of spectral methods has been analyzed in this context [35, 28, 4, 21, 31]. However, there is nothing in the literature about the consistency of the normalized cut. We partially fill that gap by relating the Cheeger constant of the graph to the Cheeger constant of the underlying domain from which the points are sampled.

Computational geometry (and topology). The Cheeger constant $H(M)$, and Cheeger sets, are *bona fide* geometric characteristics of the domain M that we might want to estimate, following a fast developing line of research around the estimation of some geometric and topological characteristics of sets from a sample, e.g., the number of connected components [5], the intrinsic dimensionality [25] and, more generally, the homology [27, 10, 11, 37, 14, 30, 13]; the Minkowski content [17], as well as the perimeter and area (volume) [8].

Random walks. Random geometric graphs are gaining popularity as models for real-life networks. Some protocols for passing information between nodes amounts to performing a random walk and it is important to bound the time it takes for information to spread to the whole network; see [2] and references therein. It is well-known that, given a graph G, a lower bound on $H(G)$ may be used to bound the mixing time the random walk on G. This is the path taken in [7, 2] when M is the unit hypercube and the graph is $G_{r_n,n}$. However, in both papers the authors reduce the setting to that of a regular grid without rigorous justification, leaving the problem unresolved (in our opinion) even in this particular case.

1.4 Discussion

Generalizations. With some additional work, our results and methodology extend to settings where the kernel (here the simple kernel) is fast decaying and where the data points are sampled from a probability distribution on M that has a non-vanishing density with respect to the uniform distribution. It would also be interesting to consider the setting where M is a d-dimensional smooth submanifold embedded in some Euclidean ambient space, using the maps as is done e.g., in [9, Lem. 3.4].

An open problem. Whether the normalized Cheeger constants of some sequence of neighborhood graphs converges to the Cheeger constant of the domain is an intriguing question. To paraphrase the question we leave open, is there a sequence $\{r_n\}$ such that, with probability one,

$$\lim_{n\to\infty} \frac{\omega_d}{\gamma r_n} H(G_{n,r_n}) = H(M)?$$

A positive answer would establish the consistency of the normalized cut criterion for graph partitioning. Also, a lower bound on $H(G_{n,r_n})$ would provide a lower bound on the eigengap between the first and second eigenvalue of the Laplacian, which in turn may be used to bound the mixing time of the random walk on G_{n,r_n}, as done in [7, 2] when M is the unit hypercube.

Consistent estimation in polynomial time. Our estimation procedures, though theoretically valid and consistent, are not practical. It would be interesting to know whether there is a consistent estimator for the Cheeger constant that can be implemented in polynomial-time. Note that computing the Cheeger constant of a graph is NP-hard (which motivates the use of spectral methods), and even the best polynomial-time approximations we are aware of are not precise enough to allow consistency [1].

1.5 Content

The rest of the paper is organized as follows. In Section 3, we establish the convergence of discrete volume and perimeter to continuous volume and perimeter for a subset of M with smooth relative boundary based on Hoeffding's inequality for U-statistics [24] and deduce the lim sup bound given in Theorem 1 via the lower semi-continuity of $h(\cdot; M)$. In Section 4, we prove Theorem 2 and Theorem 3 using results on empirical U-processes [18], and examine convergence of the Cheeger set of the random graph to a Cheeger set of M using some compacity properties of the L^1-metric [23]. Some auxiliary results are gathered in Section 5, and some geometrical results of independent interest are collected in Section 6. In the Appendix we state some results on concentration inequalities for U-statistics and some compacity properties of the L^1-metric.

2 Notation and background

The reach coincides with the condition number introduced in [27] for submanifolds without boundary, and the property $\mathrm{reach}(\partial A) > r$ is equivalent to A and A^c being both r-convex [36], in the sense that a ball of radius r rolls freely inside A and A^c. (We say that a ball of radius r rolls freely in A if, for all $p \in \partial A$, there is $x \in A$ such that $p \in \partial B(x, r)$ and $B(x, r) \subset A$.) It is well-known

that the reach bounds the radius of curvature from below [20, Thm. 4.18]. For $r < \mathrm{reach}(\partial M)$, let M_r denote the subset of M made of points at a distance r or more from ∂M.

The uniform measure on M is denoted μ, so that $\mu(A) = \mathrm{Vol}_d(A \cap M)/\mathrm{Vol}_d(M)$; and the normalized perimeter is denoted $\nu(A) = \mathrm{Vol}_{d-1}(\partial A \cap M)/\mathrm{Vol}_d(M)$. Define

$$\mu_n(A) = \frac{\delta(A \cap \mathcal{X}_n; G_{n,r_n})}{\omega_d n(n-1) r_n^d}, \quad \nu_n(A) = \frac{\sigma(A \cap \mathcal{X}_n; G_{n,r_n})}{\gamma n(n-1) r_n^{d+1}}, \quad h_n(A) = \frac{\nu_n(A)}{\mu_n(A)}$$

where δ, σ are given in (1.2), $\mathcal{X}_n$ is the sample, and G_{n,r_n} the neighborhood graph. Hence, we have

$$h_n(A) = \frac{\omega_d}{\gamma r_n} h(A \cap \mathcal{X}_n; G_{n,r_n})$$

where h is given in (1.1). The volume of a spherical cap at height η is defined as

$$\pi_d(\eta) = \mathrm{Vol}_d\left\{x : \|x\| \leq 1 \text{ and } \langle u, x\rangle \geq \eta\right\},$$

where u is any unit-norm vector of $\mathbb{R}^d$.

For a real-valued, measurable function ϕ and any measure λ, let $\lambda\phi$ denote the integral of ϕ with respect to λ, i.e., $\int \phi(x)\lambda(\mathrm{d}x)$. For $p, \varepsilon > 0$ and a class $\mathcal{H}$ of measurable functions, let $\mathcal{N}_p(\varepsilon, \mathcal{H}, \lambda)$ denote the ε-covering number in the $L^p(\lambda)$ metric, i.e.

$$\mathcal{N}_p(\varepsilon, \mathcal{H}, \lambda) = \min\left\{N : \exists \phi_1, \ldots, \phi_N \text{ such that } \sup_{\phi \in \mathcal{H}} \min_j \lambda(|\phi - \phi_j|^p) \leq \varepsilon^p\right\}.$$

It is classical to bound those covering numbers independently of λ, see Section 5 and then to take $\lambda = P_n$, where P_n is the empirical measure of the sample, given by

$$P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}.$$

We will use the L^1-metric on subsets of $\mathbb{R}^d$, given by $\mathrm{Vol}_d(A \Delta B) = \int |\mathbf{1}_A(x) - \mathbf{1}_B(x)|\, \mathrm{d}x$. The arrow $\xrightarrow{L^1}$ denotes the convergence in this metric. And, finally, if a is a real number, a_+ and a_- denote its positive and negative parts, so that $a = a_+ - a_-$ and $|a| = a_+ + a_-$.

In the rest of the paper, the generic constant C may vary from line to line, except when stated explicitly otherwise.

3 Proof of Theorem 1: Consistency of the normalized cut

To prove Theorem 1, we establish the almost-sure convergence of $\mu_n(A)$ to $\mu(A)$ and $\nu_n(A)$ to $\nu(A)$ for a subset $A \subset M$ with smooth relative boundary. This follows from the following exponential inequalities.

3.1 Exponential inequalities

Proposition 4. *Fix a sequence $r_n \to 0$. Let $A \subset M$ be an arbitrary open subset of M. There exists a constant C depending only on M such that, for any $\varepsilon > 0$, and all n large enough, we have*

$$\mathbb{P}\left(|\mu_n(A) - \mu(A)| \geq \varepsilon\right) \leq 2\exp\left(-\frac{n r_n^d \varepsilon^2}{C(1+\varepsilon)}\right).$$

In particular, if $n r_n^d / \log n \to \infty$, then $\mu_n(A)$ converges almost surely to $\mu(A)$ when $n \to \infty$.

Proof. Define the symmetric kernel

$$\phi_{A,r}(x,y) = \frac{1}{2}\left\{\mathbf{1}_A(x) + \mathbf{1}_A(y)\right\}\mathbf{1}\{\|x-y\| \le r\}, \tag{3.1}$$

so that

$$\mu_n(A) = \frac{1}{\omega_d n(n-1)r_n^d}\sum_{i\neq j}\phi_{A,r_n}(X_i,X_j).$$

By the triangle inequality, we have

$$|\mu_n(A) - \mu(A)| \le |\mu_n(A) - \mathbb{E}\left[\mu_n(A)\right]| + |\mathbb{E}\left[\mu_n(A)\right] - \mu(A)|.$$

For all n large enough such that $r_n \le \mathrm{reach}(\partial M)$, the second term on the right-hand side (the bias term) is bounded by Cr_n with C depending only on M by Lemmas 12 and 15. Assume that n is large enough such that $2Cr_n \le \varepsilon$. We then apply Lemma 22, which is Hoeffding's Inequality for U-statistics [24], to the first term (the deviation term) on the right-hand side with the kernel

$$\phi := \phi_{A,r_n} - \mathbb{E}\left[\phi_{A,r_n}(X_1,X_2)\right]$$

and $t = \omega_d r^d \varepsilon/2$. The kernel satisfies $\|\phi\|_\infty \le 1$, and simple calculations yields

$$\mathrm{Var}(\phi(X_1,X_2)) \le \mathbb{E}\left[\phi_{A,r_n}(X_1,X_2)^2\right] \le \mu(A)\omega_d r_n^d \le \omega_d r_n^d.$$

From this we obtain the large deviation bound. The almost sure convergence is then a simple consequence of the Borel-Cantelli Lemma. □

Proposition 5. *Fix a sequence $r_n \to 0$. Let A be an open subset of M with smooth relative boundary and positive reach. There exists a constant C depending only on M such that, for any $\varepsilon > 0$, and for all n large enough, we have*

$$\mathbb{P}\left(|\nu_n(A) - \nu(A)| \ge \epsilon\right) \le 2\exp\left(-\frac{nr_n^{d+1}\epsilon^2}{C(\nu(A)+\epsilon)}\right).$$

In particular, if $nr_n^{d+1}/\log n \to \infty$, then

$$\nu_n(A) \to \nu(A), \quad n\to\infty, \quad \textit{almost surely.}$$

Proof. Define the symmetric kernel

$$\bar{\phi}_{A,r}(x,y) = \frac{1}{2}\left\{\mathbf{1}_A(x)\mathbf{1}_{A^c}(y) + \mathbf{1}_A(y)\mathbf{1}_{A^c}(x)\right\}\mathbf{1}\{\|x-y\| \le r\}, \tag{3.2}$$

so that

$$\nu_n(A) = \frac{1}{\gamma n(n-1)r_n^{d+1}}\sum_{i\neq j}\bar{\phi}_{A,r_n}(X_i,X_j).$$

By the triangle inequality, we have

$$|\nu_n(A) - \nu(A)| \le |\nu_n(A) - \mathbb{E}\left[\nu_n(A)\right]| + |\mathbb{E}\left[\nu_n(A)\right] - \nu(A)|.$$

By Lemma 13-(i), the second term on the right-hand side goes to 0 as $n \to \infty$. We then apply, for n large enough, Lemma 22 to the first term on the right-hand side with the kernel

$$\phi := \bar{\phi}_{A,r_n} - \mathbb{E}\left[\bar{\phi}_{A,r_n}(X_1, X_2)\right]$$

and $t := \gamma r^{d+1}\nu(A)\epsilon/2$. The kernel satisfies $\|\phi\|_\infty \leq 1$, and using Lemma 13-(i), we have

$$\mathrm{Var}(\phi(X_1, X_2)) \leq \mathbb{E}\left[\bar{\phi}_{A,r_n}(X_1, X_2)^2\right] = \mathbb{E}\left[\bar{\phi}_{A,r_n}(X_1, X_2)\right] \leq 2\gamma\nu(A)r_n^{d+1},$$

where the last inequality follows from Lemma 13-(i) for n large enough. From this we obtain the large deviation bound, and the almost sure convergence is a consequence of the Borel-Cantelli Lemma. □

3.2 Proof of Theorem 1

The first statement of Theorem 1 is an immediate consequence of Propositions 4 and 5. To prove the second statement, under the conditions of Theorem 1, for any subset A with smooth relative boundary, with probability one $\lim_n h_n(A) = h(A; M)$ while $h_n(A) \geq \frac{\omega_d}{\gamma r_n} H(G_{n,r_n})$, so that $\limsup_n \frac{\omega_d}{\gamma r_n} H(G_{n,r_n}) \leq h(A; M)$. Then we obtain the upper bound of Theorem 1 by taking the infimum over all such subsets A.

4 Proof of Theorems 2 and 3: consistent estimation

4.1 Exponential inequalities

Given $\beta > 0$ and $\rho > 0$, with $\beta > \omega_d \rho^d$, let $\mathcal{A}_{\beta,\rho}$ be the class of subsets of M defined by

$$\mathcal{A}_{\beta,\rho} = \{R \cap M \ : \ \mathrm{reach}(\partial R) > \rho, \mathrm{Vol}_d(R) \leq \beta\}. \tag{4.1}$$

Proposition 6. *There exists a constant C depending only on M, such that, for any $\varepsilon > 0$, any $r > 0$, and all n satisfying $n/\log(n) > C/\varepsilon^2$ and $nr^d > C/\varepsilon$, we have*

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}_{\beta,\rho}} |\mu_n(A) - \mathbb{E}\left[\mu_n(A)\right]| \geq \varepsilon\right) \leq C \exp\left(-\frac{n\varepsilon^2}{C}\right) + C \exp\left(-\frac{nr^d\varepsilon}{C}\right). \tag{4.2}$$

Proof. Define the kernel class

$$\mathcal{F} = \{\phi_{A,r} \ : \ A \in \mathcal{A}_{\beta,\rho}\}, \tag{4.3}$$

where $\phi_{A,r}$ is defined in (3.1). And set

$$\mathcal{F}_1 = \left\{x \mapsto \mu\phi(x, .) - \mu^{\otimes 2}\phi \ : \ \phi \in \mathcal{F}\right\}, \tag{4.4}$$

$$\mathcal{F}_2 = \left\{(x, y) \mapsto \phi(x, y) - \mu\phi(x, .) - \mu\phi(y, .) + \mu^{\otimes 2}\phi \ : \ \phi \in \mathcal{F}\right\}. \tag{4.5}$$

Define the following functionals respectively over $\mathcal{F}_1$ and $\mathcal{F}_2$:

$$M_n(\phi) = \frac{1}{n}\sum_{i \neq j} \phi(X_i), \quad U_n(\phi) = \frac{1}{n(n-1)}\sum_{i \neq j} \phi(X_i, X_j).$$

Observe that

$$\sup_{A\in\mathcal{A}_{\beta,\rho}} |\mu_n(A) - \mathbb{E}[\mu_n(A)]| = \frac{1}{\omega_d r^d} \sup_{\phi\in\mathcal{F}} \left|U_n(\phi) - \mu^{\otimes 2}(\phi)\right|,$$

which we bound by the sum of a first-order term and a second-order term:

$$\sup_{\phi\in\mathcal{F}} \left|U_n(\phi) - \mu^{\otimes 2}(\phi)\right| \le \sup_{\phi\in\mathcal{F}_1} |M_n(\phi)| + \sup_{\phi\in\mathcal{F}_2} |U_n(\phi)|,$$

so that

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}_{\beta,\rho}} |\mu_n(A) - \mathbb{E}[\mu_n(A)]| \ge \varepsilon\right) \le \mathbb{P}\left(\sup_{\phi\in\mathcal{F}_1} |M_n(\phi)| > \frac{\omega_d r^d \varepsilon}{2}\right) + \mathbb{P}\left(\sup_{\phi\in\mathcal{F}_2} |U_n(\phi)| > \frac{\omega_d r^d \varepsilon}{2}\right).$$

First order term. Take any $\phi \in \mathcal{F}_1$ of the form $\phi(x) = \mu\phi_{A,r}(x,\cdot) - \mu^{\otimes 2}\phi_{A,r}$. Since for all x and y, $\phi_{A,r}(x,y) \le \mathbf{1}\{\|x-y\| \le r\}$, we have

$$\|\phi\|_\infty := \sup_{x\in M} |\phi(x)| \le \sup_{x\in M} |\mu\phi_{A,r}(x,\cdot)| + \mu^{\otimes 2}\phi_{A,r} \le \omega_d[1 + \mathrm{Vol}_d(M)]r^d =: C_1 r^d. \tag{4.6}$$

Therefore

$$\sup_{\phi\in\mathcal{F}_1} \mathrm{Var}\left(\phi(X)\right) \le C_1^2 r^{2d}, \tag{4.7}$$

and so, for all $n > 32C_1^2/(\omega_d\varepsilon^2)$, we have

$$1 - \frac{4}{n(\omega_d r^d \varepsilon/2)^2} \sup_{\phi\in\mathcal{F}_1} \mathrm{Var}\left(\phi(X)\right) > \frac{1}{2}.$$

Hence, we are in a position of applying the symmetrization inequality for probabilities [33, Lem 2.3.7]. Let $\{\xi_i\}_{i\ge 1}$ be a Rademacher sequence. We have

$$\mathbb{P}\left(\sup_{\phi\in\mathcal{F}_1} |M_n(\phi)| > \frac{\omega_d r^d \varepsilon}{2}\right) \le 4\mathbb{P}\left(\sup_{\phi\in\mathcal{F}_1} \left|\frac{1}{n}\sum_{i=1}^n \xi_i\phi(X_i)\right| > \frac{\omega_d r^d \varepsilon}{8}\right).$$

Let $\mathcal{C}$ be a minimal $(\omega_d r^d\varepsilon/8)$-covering of $\mathcal{F}_1$ with respect to the $L^1(P_n)$ metric. A simple union bound (conditional on $\mathcal{X}$) yields

$$\mathbb{P}\left(\sup_{\phi\in\mathcal{F}_1} |M_n(\phi)| > \frac{\omega_d r^d \varepsilon}{2}\right) \le 4\mathbb{E}\left\{\mathcal{N}_1\left(\frac{\omega_d r^d\varepsilon}{8}, \mathcal{F}_1, P_n\right) \max_{\phi\in\mathcal{C}} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \xi_i\phi(X_i)\right| > \frac{\omega_d r^d\varepsilon}{8}\right)\right\}.$$

By Lemma 10

$$\log \mathcal{N}_1\left(\frac{\omega_d r^d\varepsilon}{8}, \mathcal{F}_1, P_n\right) \le C_2 \log(C_2/(r^d\varepsilon)),$$

for some constant C_2 not depending on n. Since $\|\phi\|_\infty \le C_1 r^d$ for all $\phi \in \mathcal{F}_1$ with (4.6), it follows by Hoeffding's inequality that for all $\phi \in \mathcal{C}$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \xi_i\phi(X_i)\right| > \frac{\omega_d r^d\varepsilon}{8}\right) \quad \le \quad 2\exp\left(-\frac{1}{2}\frac{n^2(\omega_d r^d\varepsilon/8)^2}{n(C_1 r^d)^2}\right) =: 2\exp\left(-C_3 n\varepsilon^2\right).$$

Therefore

$$\mathbb{P}\left(\sup_{\phi\in\mathcal{F}_1}|M_n(\phi)|>\frac{\omega_d r^d\varepsilon}{2}\right)\leq 8\exp\left[-C_3n\varepsilon^2\left(1-\frac{C_2}{C_3n\varepsilon^2}\log\left(\frac{C_2}{r^d\varepsilon}\right)\right)\right],$$

and so, for all n satisfying

$$\frac{n}{\log(n)}>\frac{C_2}{2C_3\varepsilon^2}\quad\text{and}\quad nr^d>\frac{C_2}{\varepsilon},$$

we have

$$\mathbb{P}\left(\sup_{\phi\in\mathcal{F}_1}|M_n(\phi)|>\frac{\omega_d r^d\varepsilon}{2}\right)\leq 8\exp\left(-\frac{C_3n\varepsilon^2}{2}\right). \tag{4.8}$$

Second order term. Using Lemma 10, we have, for all $n\geq 1$, and all $\eta>0$, that

$$\log\mathcal{N}_2(\eta,\mathcal{F}_2,P)\leq C_5\log(C_5/\eta),$$

for some constants C_5 not depending on n, and for all probability measure P on $M\times M$. This, together with the fact that $\|\phi\|_\infty\leq 2$ for all ϕ in $\mathcal{F}_2$, yields

$$\sup_P\int_0^4\log\mathcal{N}_2\left(\eta,\mathcal{F}_2,P\right)\mathrm{d}\eta\leq C_6,$$

for some constant $C_6<\infty$. Then by Lemma 23, there exists a constant C_7 such that, for all n satisfying

$$nr^d>\frac{C_7\log(C_7)C_6}{\omega_d\varepsilon/2},$$

we have

$$\mathbb{P}\left(\sup_{\phi\in\mathcal{F}_2}|U_n(\phi)|>\frac{\omega_d r^d\varepsilon}{2}\right)\leq C_7\exp\left(-\frac{nr^d\varepsilon}{C_7}\right). \tag{4.9}$$

Proposition 6 follows from (4.8) and (4.9). □

Proposition 7. *There exists a constant C depending only on M such that, for all $\varepsilon>0$, all r in $\big(0;\min\{\mathrm{reach}(\partial M),\rho/2\}\big)$, and all n satisfying*

$$\frac{n}{\log(n)}>\frac{C}{r^2\varepsilon^2}\quad,\quad nr^{d+1}>\frac{C}{\varepsilon}\quad,\quad n>\frac{C\beta}{r^2\rho^2\varepsilon^2}$$

we have

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}_{\beta,\rho}}|\nu_n(A)-\mathbb{E}[\nu_n(A)]|\geq\varepsilon\right)\leq C\exp\left(-\frac{nr^2\varepsilon^2}{C}\right)+C\exp\left(-\frac{nr^{d+1}\varepsilon}{C}\right).$$

Proof. The proof follows that of Proposition 6, with the symmetric kernel $\bar{\phi}_{A,n}$ defined in 3.2 and corresponding classes $\bar{\mathcal{F}}$, $\bar{\mathcal{F}}_1$ and $\bar{\mathcal{F}}_2$ defined by (4.3), (4.4), and (4.5), with $\phi_{A,r}$ replaced by $\bar{\phi}_{A,r}$. Observe that

$$|\nu_n(A)-\mathbb{E}\left[\nu_n(A)\right]|=\frac{1}{\gamma r^{d+1}}\sup_{\phi\in\bar{\mathcal{F}}}\left|U_n(\phi)-\mu^{\otimes 2}(\phi)\right|,$$

which we decompose into first-order and second-order terms:

$$\sup_{\phi\in\bar{\mathcal{F}}}\left|U_n(\phi)-\mu^{\otimes 2}(\phi)\right| \le \sup_{\phi\in\bar{\mathcal{F}}_1}|M_n(\phi)| + \sup_{\phi\in\bar{\mathcal{F}}_2}|U_n(\phi)|\,,$$

Therefore

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}_{\beta,\rho}}|\nu_n(A)-\mathbb{E}[\nu_n(A)]|\ge\varepsilon\right) \le \mathbb{P}\left(\sup_{\phi\in\bar{\mathcal{F}}_1}|M_n(\phi)|>\frac{\gamma r^{d+1}\varepsilon}{2}\right)+\mathbb{P}\left(\sup_{\phi\in\bar{\mathcal{F}}_2}|U_n(\phi)|>\frac{\gamma r^{d+1}\varepsilon}{2}\right).$$

First order term. Take any $\phi\in\bar{\mathcal{F}}_1$ of the form $\phi(x)=\left(\mu\bar{\phi}_{A,r}(x,\cdot)-\mu^{\otimes 2}\bar{\phi}_{A,r}\right)$. Denote by D the set

$$D=\{x\in A\,:\,\mathrm{dist}(x,M\setminus A)\le r\}\cup\{x\in M\setminus A\,:\,\mathrm{dist}(x,A)\le r\}.$$

Then for all x in M,

$$\mu\bar{\phi}_{A,r}(x,.)\le\frac{\omega_d r^d}{2}\mathbf{1}_D(x),$$

and by using Lemma 14 and Lemma 16, we have

$$\mathrm{Vol}_d(D)\le\mathrm{Vol}_d\,\mathcal{V}\left(\partial R,r\right)\le\left(1+\frac{r}{\rho}\right)^{d-1}\mathrm{Vol}_{d-1}(\partial R)2r\le 2^d r\frac{d\,\mathrm{Vol}_d(R)}{\rho}\le d2^d\frac{r\beta}{\rho}.$$

Therefore,

$$\sup_{\phi\in\bar{\mathcal{F}}_1}\mathrm{Var}\left(\phi(X)\right)\le\sup_{A\in\mathcal{A}_{\beta,\rho}}\mathbb{E}\left[\left(\mu\bar{\phi}_{A,r}(X,\cdot)\right)^2\right]\le C_1\frac{r^{2d+1}\beta}{\rho},$$

where C_1 is a constant depending only on M (through $\mathrm{Vol}_d(M)$), and so, for all $n>\frac{32C_1}{\gamma^2}\beta/(r\rho\varepsilon^2)$, we have

$$1-\frac{4}{n(\gamma r^{d+1}\varepsilon/2)^2}\sup_{\phi\in\bar{\mathcal{F}}_1}\mathrm{Var}\left(\phi(X)\right)>\frac{1}{2}.$$

We may therefore apply the symmetrization inequality for probabilities [33, Lem 2.3.7] and proceed as before, this time using Lemma 11 to control the (random) entropy term, to deduce that, for all n such that

$$\frac{n}{\log(n)}>\frac{C_2}{r^2\varepsilon^2}\quad\text{and}\quad nr^{d+1}>\frac{C_2}{\varepsilon}$$

we have

$$\mathbb{P}\left(\sup_{\phi\in\bar{\mathcal{F}}_1}|M_n(\phi)|>\frac{\gamma r^{d+1}\varepsilon}{2}\right)\le C_2\exp\left(-\frac{nr^2\varepsilon^2}{C_2}\right),$$

for some constant C_2.

Second order term. We use again using Lemma 11 to control the entropy term, and combine it with Lemma 23 as before, to deduce that,

$$\mathbb{P}\left(\sup_{\phi\in\bar{\mathcal{F}}_2}|U_n(\phi)|>\frac{\gamma r^{d+1}\varepsilon}{4}\right)\le C_3\exp\left(-\frac{nr^{d+1}\varepsilon}{C_3}\right).$$

for all n such that $nr^{d+1}>C_3/\varepsilon$. □

4.2 A uniform control on $h_n(A)$

Define the (random) class $\mathcal{A}_n$ of subsets of M by:

$$\mathcal{A}_n = \{A = R \cap M \,:\, R \in \mathcal{R}_n,\, \exists i, j \,\text{such that}\, B(X_i, \alpha_n) \subset R,\, B(X_j, \alpha_n) \subset R^c\}\,. \tag{4.10}$$

Since, by definition, $h_n^{\ddagger}$ is finite if and only if both R and R^c contain a ball of radius α_n centered at a sample point, we have

$$\min_{R \in \mathcal{R}_n} h_n^{\ddagger}(R) = \min_{A \in \mathcal{A}_n} \frac{\nu_n(A)}{\min\{\mu_n(A), \mu_n(A^c)\}} = \min_{A \in \mathcal{A}_n} \frac{\nu_n(A)}{\mu_n(A)}.$$

Note also that

$$\mathcal{A}_n \subset \mathcal{A}_{\beta_n, \rho_n}, \tag{4.11}$$

where we $\mathcal{A}_{\beta,\rho}$ is defined in (4.1), so that $\mathcal{A}_{\beta_n,\rho_n} = \{R \cap M \,:\, R \in \mathcal{R}_n\}$.

Lemma 8. *We have*

$$\liminf_{n\to\infty} \inf_{A \in \mathcal{A}_n} \left(h_n(A) - \frac{\mathrm{Vol}_{d-1}(\partial A \cap M_{r_n})}{\mathrm{Vol}_d(A \cap M_{r_n})} \right) \geq 0. \tag{4.12}$$

Proof. For any $A \in \mathcal{A}_n$, we denote by $A_n := A \cap M_{r_n}$. It is easy to check that

$$\begin{aligned} \left(h_n(A) - \frac{\nu(A_n)}{\mu(A_n)} \right) \left(\frac{\mu_n(A)}{\mu(A_n)} \right) = \frac{\nu(A_n)}{\mu(A_n)} \times \frac{\nu_n(A) - \nu(A_n)}{\nu(A_n)} \qquad & + \frac{\nu(A_n)}{\mu(A_n)} \times \frac{\mu(A_n) - \mu_n(A)}{\mu(A_n)} \\ =: \zeta_n(A) \qquad & + \xi_n(A). \end{aligned}$$

Now we define the probability event

$$\Omega_n = \left[\sup_{A \in \mathcal{A}_n} \left| 1 - \frac{\mu_n(A)}{\mu(A_n)} \right| \leq \frac{1}{2} \right].$$

on which we have $\frac{1}{2} \leq \mu_n(A)/\mu(A_n) \leq \frac{3}{2}$ for all A in $\mathcal{A}_n$. We deduce that on this event Ω_n,

$$\inf_{A \in \mathcal{A}_n} \left[\left(h_n(A) - \frac{\nu(A_n)}{\mu(A_n)} \right) \frac{\mu_n(A)}{\mu(A_n)} \right] \leq \frac{3}{2} \inf_{A \in \mathcal{A}_n} \left(h_n(A) - \frac{\nu(A_n)}{\mu(A_n)} \right)_+ - \frac{1}{2} \inf_{A \in \mathcal{A}_n} \left(h_n(A) - \frac{\nu(A_n)}{\mu(A_n)} \right)_-.$$

Consequently, for all $\varepsilon > 0$,

$$\mathbb{P}\left(\left[\inf_{A \in \mathcal{A}_n} \zeta_n(A) > -\frac{\varepsilon}{4} \right] \cap \left[\inf_{A \in \mathcal{A}_n} \xi_n(A) > -\frac{\varepsilon}{4} \right] \cap \Omega_n \right) \leq \mathbb{P}\left(\frac{1}{2} \inf_{A \in \mathcal{A}_n} \left(h_n(A) - \frac{\nu(A_n)}{\mu(A_n)} \right) > -\frac{\varepsilon}{2} \right),$$

so that

$$\begin{aligned} \mathbb{P}\left(\inf_{A \in \mathcal{A}_n} \left(h_n(A) - \frac{\nu(A_n)}{\mu(A_n)} \right) < -\varepsilon \right) &\leq \mathbb{P}\left(\inf_{A \in \mathcal{A}_n} \zeta_n(A) < -\frac{\varepsilon}{4} \right) + \mathbb{P}\left(\inf_{A \in \mathcal{A}_n} \xi_n(A) < -\frac{\varepsilon}{4} \right) + \mathbb{P}\left(\Omega_n^c\right) \\ &=: I_1 + I_2 + \mathbb{P}\left(\Omega_n^c\right). \end{aligned} \tag{4.13}$$

Bounding I_1. By Lemma 18, $\mu(A_n) = \mathrm{Vol}_d(A \cap M_{r_n})/\mathrm{Vol}_d(M) \geq C\alpha_n^d$ for all A in $\mathcal{A}_n$, so that

$$\inf_{A\in\mathcal{A}_n} \zeta_n(A) \geq -\frac{C}{\alpha_n^d} \inf_{A\in\mathcal{A}_n} \big(\nu_n(A) - \nu(A_n)\big)_-.$$

Consequently

$$\begin{aligned} I_1 &\leq \mathbb{P}\left[\inf_{A\in\mathcal{A}_n} (\nu_n(A) - \nu(A_n)) < -C\varepsilon\alpha_n^d\right] \\ &\leq \mathbb{P}\left[\inf_{A\in\mathcal{A}_n} \big(\nu_n(A) - \mathbb{E}[\nu_n(A)]\big) + \inf_{A\in\mathcal{A}_n} \big(\mathbb{E}[\nu_n(A)] - \nu(A_n)\big) < -C\varepsilon\alpha_n^d\right]. \end{aligned}$$

Using Lemma 13 together with Lemma 16, we have

$$\inf_{A\in\mathcal{A}_n} \big(\mathbb{E}[\nu_n(A)] - \nu(A_n)\big) \geq -C\frac{\beta_n r_n}{\rho_n^2}.$$

By assumption, $\rho_n\alpha_n^{2d}/(\beta_n r_n) \to \infty$ and $\rho_n > \alpha_n$, so that $\beta_n r_n/(\rho_n^2\alpha_n^d) \to 0$. And $\mathcal{A}_n \subset \mathcal{A}_{\beta_n,\rho_n}$. Hence, for all n large enough,

$$I_1 \leq \mathbb{P}\left[\inf_{A\in\mathcal{A}_n} (\nu_n(A) - \mathbb{E}[\nu_n(A)]) < -C\varepsilon\alpha_n^d/2\right] \leq \mathbb{P}\left[\sup_{A\in\mathcal{A}_{\beta_n,\rho_n}} |\nu_n(A) - \mathbb{E}[\nu_n(A)]| > C\varepsilon\alpha_n^d/2\right].$$

Moreover, $\beta_n r_n/(\rho_n^2\alpha_n^d) \to 0$ implies that $\alpha_n^d \gg r_n$. Consequently, since $nr_n^{d+2}/\log(n) \to \infty$ by assumption, and since $d \geq 2$, it follows that

$$\frac{nr_n^2\alpha_n^{2d}}{\log(n)} \geq \frac{nr_n^4}{\log(n)} \to \infty \quad ; \quad nr_n^{d+1}\alpha_n^d \geq nr_n^{d+2} \to \infty \quad ; \quad \frac{nr_n^2\rho_n^2\alpha_n^{2d}}{\beta_n} = nr_n^3\alpha_n^d \times \frac{\rho_n^2\alpha_n^d}{\beta_n r_n} \to \infty.$$

We may therefore apply Proposition 7 to deduce that

$$I_1 \leq C\exp\left(-\frac{nr_n^2\alpha_n^{2d}}{C}\right) + C\exp\left(-\frac{nr_n^{d+1}\alpha_n^d}{C}\right)$$

for some constant $C > 0$ and all n large enough. Since $nr_n^2\alpha_n^{2d}/\log(n) \to \infty$ and $nr_n^{d+1}\alpha_n^d/\log(n) \geq nr_n^{d+2}/\log(n) \to \infty$, we obtain that, for all $\varepsilon > 0$,

$$\sum_n \mathbb{P}\left[\inf_{A\in\mathcal{A}_n} \zeta_n(A) < -\varepsilon/4\right] < \infty. \tag{4.14}$$

Bounding I_2. We have

$$I_2 = \mathbb{P}\left(\inf_{A\in\mathcal{A}_n} \xi_n(A) < -\varepsilon/4\right) \leq \mathbb{P}\left(\sup_{A\in\mathcal{A}_n} |\xi_n(A)| > \varepsilon/4\right).$$

Using Lemma 16, $\nu(A_n)/\mu(A_n)^2 \leq C\beta_n/(\rho_n\alpha_n^{2d})$ for all A in $\mathcal{A}_n$, so that

$$I_2 \leq \mathbb{P}\left(\sup_{A\in\mathcal{A}_n} |\mu_n(A) - \mu(A_n)| > \frac{\rho_n\alpha_n^{2d}\varepsilon}{4C\beta_n}\right). \tag{4.15}$$

Using Lemma 12 and Lemma 15, for all A in $\mathcal{A}_n$ and all n large enough,

$$\begin{aligned} |\mu_n(A) - \mu(A_n)| &\leq |\mu_n(A) - \mathbb{E}[\mu_n(A)]| + |\mathbb{E}[\mu_n(A)] - \mu(A)| + |\mu(A) - \mu(A_n)| \\ &\leq |\mu_n(A) - \mathbb{E}[\mu_n(A)]| + Cr_n. \end{aligned}$$

Hence, since $\rho_n \alpha_n^{2d}/(\beta_n r_n) \to \infty$ by assumption, we have

$$I_2 \leq \mathbb{P}\left(\sup_{A \in \mathcal{A}_n} |\mu_n(A) - \mathbb{E}[\mu_n(A)]| > \frac{\rho_n \alpha_n^{2d} \varepsilon}{8C\beta_n} \right) \leq \mathbb{P}\left(\sup_{A \in \mathcal{A}_{\beta_n, \rho_n}} |\mu_n(A) - \mathbb{E}[\mu_n(A)]| > \frac{\rho_n \alpha_n^{2d} \varepsilon}{8C\beta_n} \right).$$

Since $\rho_n \alpha_n^{2d}/(\beta_n r_n) \to \infty$ implies that $\alpha_n^{2d} \gg \beta_n r_n/\rho_n$, and since $n r_n^{d+2}/\log(n)$ by assumption, we have

$$\frac{n\rho_n \alpha_n^{4d}}{\beta_n^2 \log(n)} \geq \frac{n r_n^2}{\log(n)} \to \infty \quad ; \quad \frac{n r_n^d \rho_n \alpha_n^{2d}}{\beta_n} \geq \frac{n r_n^d \rho_n \alpha_n^{2d}}{\beta_n \log(n)} \geq \frac{n r_n^{d+1}}{\log(n)} \to \infty.$$

Hence we may apply Proposition 6 to deduce that

$$I_2 \leq C \exp\left(-\frac{n\rho_n^2 \alpha_n^{2d} \varepsilon^2}{\beta_n^2} \right) + C \exp\left(-\frac{n r_n^d \rho_n \alpha_n^{2d} \varepsilon}{\beta_n} \right),$$

for some constant $C > 0$ and all n large enough, which yields

$$\sum_n \mathbb{P}\left(\inf_{A \in \mathcal{A}_n} \xi_n(A) < -\varepsilon/4 \right) < \infty. \tag{4.16}$$

Bounding $\mathbb{P}(\Omega_n^c)$. Since $\mu(A_n) > C\alpha_n^d$ for some C uniformly over $A \in \mathcal{A}_n$ by Lemma 18, we have

$$\begin{aligned} \mathbb{P}(\Omega_n^c) &= \mathbb{P}\left(\sup_{A \in \mathcal{A}_n} \frac{|\mu_n(A) - \mu(A_n)|}{\mu(A_n)} > \frac{1}{2} \right) \\ &\leq \mathbb{P}\left(\sup_{A \in \mathcal{A}_n} |\mu_n(A) - \mu(A_n)| > C\alpha_n^d \right), \end{aligned}$$

which we may bound by the right-hand side of (4.15) for all $\varepsilon > 0$ and all n large enough, so that

$$\sum_n \mathbb{P}(\Omega_n^c) < \infty. \tag{4.17}$$

Conclusion. Reporting (4.14), (4.16) and (4.17) in (4.13), we deduce that for all $\varepsilon > 0$,

$$\sum_n \mathbb{P}\left[\inf_{A \in \mathcal{A}_n} \left(h_n(A) - \frac{\nu(A_n)}{\mu(A_n)} \right) < -\varepsilon \right] < \infty.$$

Consequently, applying the Borel-Cantelli lemma, we conclude the proof. □

4.3 Proof of (i) in Theorem 2

Lower bound. Now let R_n be a sequence in $\mathcal{R}_n$ satisfying

$$h_n^\ddagger(R_n) = \min_{R\in\mathcal{R}_n} h_n^\ddagger(R).$$

Then

$$\begin{aligned} h_n^\ddagger(R_n) - h(M) &= \left[h_n^\ddagger(R_n) - h(R_n; M_{r_n})\right] + \left[h(R_n; M_{r_n}) - H(M_{r_n})\right] + \left[H(M_{r_n}) - H(M)\right]. \\ &\geq \inf_{A\in\mathcal{A}_n} \left(h_n(A) - \frac{\mathrm{Vol}_{d-1}(\partial A \cap M_{r_n})}{\mathrm{Vol}_d(A\cap M_{r_n})}\right) + \left[H(M_{r_n}) - H(M)\right], \end{aligned}$$

since the second term $[h(R_n; M_{r_n}) - H(M_{r_n})]$ is non-negative for all n by definition of $H(M_{r_n})$. By Proposition 21 $H(M_{r_n}) \to H(M)$ as $n\to\infty$ and using this, together with Lemma 8, we obtain that

$$\liminf_{n\to\infty} \min_{R\in\mathcal{R}_n} h_n^\ddagger(R) \geq H(M) \quad \text{a.s.} \tag{4.18}$$

Upper bound. To obtain the matching upper bound, fix a subset $A\subset M$ with smooth relative boundary and such that $0 < \mathrm{Vol}_d(A) \leq \mathrm{Vol}_d(M\backslash A) < \mathrm{Vol}_d(M)$. Then, for n large enough, there exists R_n in $\mathcal{R}_n$ such that $R_n \cap M = A$, implying that

$$\min_{R\in\mathcal{R}_n} h_n^\ddagger(R) \leq h_n(A).$$

By Theorem 1, $h_n(A) \to \nu(A)/\mu(A) = h(A; M)$ almost surely, so that

$$\limsup_{n\to\infty} \min_{R\in\mathcal{R}_n} h_n^\ddagger(R) \leq h(A; M) \quad \text{a.s.}$$

By minimizing over A, we obtain

$$\limsup_{n\to\infty} \min_{R\in\mathcal{R}_n} h_n^\ddagger(R) \leq H(M) \quad \text{a.s.} \tag{4.19}$$

Combining the lower and upper bounds (4.18) and (4.19), we conclude that

$$\lim_{n\to\infty} \min_{R\in\mathcal{R}_n} h_n^\ddagger(R) = H(M) \quad \text{a.s.} \tag{4.20}$$

4.4 Proof of (ii) in Theorem 2

Let R_n be a sequence in $\mathcal{R}_n$ satisfying

$$h_n^\ddagger(R_n) = \min_{R\in\mathcal{R}_n} h_n^\ddagger(R),$$

and set $A_n = R_n \cap M$. Fix a subset $A^0 \subset M$ with smooth relative boundary and such that $h(A^0) < \infty$. Then for n large enough, there exists R in $\mathcal{R}_n$ such that $A^0 = R\cap M$. Hence $h_n(A_n) \leq h_n(A^0)$ and since $h_n(A^0) \to h(A^0)$ by Theorem 1, we have

$$\limsup_{n\to\infty} \mathrm{Vol}_{d-1}(A_n) \leq \limsup_{n\to\infty} h(A_n) \min\{\mathrm{Vol}_d(A_n), \mathrm{Vol}_d(A_n^c \cap M)\} \leq h(A^0)\,\mathrm{Vol}_d(M)/2.$$

Therefore by Proposition 24, with probability one, $\{A_n\}$ admits a subsequence converging in the L^1-metric.

On the one hand,

$$h(A_n; M_{r_n}) - H(M) = [h(A_n; M_{r_n}) - H(M_{r_n})] + [H(M_{r_n}) - H(M)],$$

where the first difference term on the right-hand side is non-negative by definition, while the second difference term tends to zero by Proposition 21, so that with probability one:

$$\liminf_{n\to\infty} h(A_n; M_{r_n}) \geq H(M).$$

On the other hand,

$$\begin{aligned} h(A_n; M_{r_n}) - H(M) &= \left[h(A_n; M_{r_n}) - h_n^{\ddagger}(A_n)\right] + \left[h_n^{\ddagger}(A_n) - H(M)\right] \\ &\leq -\inf_{R\in\mathcal{R}_n} \left(h_n^{\ddagger}(R) - h(R; M_{r_n})\right) + \left[h_n^{\ddagger}(A_n) - H(M)\right] \end{aligned}$$

so that

$$\limsup_{n\to\infty} h(A_n; M_{r_n}) - H(M) \leq -\liminf_{n\to\infty} \inf_{R\in\mathcal{R}_n} \left(h_n^{\ddagger}(R) - h(R; M_{r_n})\right) + \left[h_n^{\ddagger}(A_n) - H(M)\right]$$

which goes to 0 as $n \to \infty$ from (4.12) and (4.20). Hence

$$\lim_{n\to\infty} h(A_n; M_{r_n}) \to H(M) \quad \text{a.s.}$$

Now let f_n denote the bi-Lipschitz function mapping M_{r_n} to M defined in Lemma 20 with r and s replaced by r_n and s_n, where $s_n/r_n \to \infty$. Define $B_n = f_n(A_n \cap M_{r_n})$. By Lemmas 19 and 20, we have

$$h(B_n; M) \leq \left(1 + \frac{2r_n}{s_n - r_n}\right)^{2d} h(A_n; M_{r_n}),$$

so that $h(B_n; M) \to H(M)$ almost surely as $n \to \infty$. Moreover, by Proposition 24, with probability one, there exists a subset B_∞ of M and a subsequence $\{B_{n_k}\}$ such that B_{n_k} converges to B_∞ in the L^1-metric. Since $h(\cdot; M)$ is lower-semi-continuous by Proposition 25, with probability one, $\liminf_{n\to\infty} h(B_n; M) \geq h(B_\infty; M)$. Since we also have $\liminf_{n\to\infty} h(B_n; M) = H(M)$ a.s., it follows that $h(B_\infty; M) = H(M)$ a.s. and so B_∞ is a Cheeger set of M.

Moreover, since f_n leaves M_{s_n} unchanged,

$$\mathrm{Vol}_d(A_n \Delta B_n) \leq \mathrm{Vol}_d(M \backslash M_{s_n}) \to 0 \quad \text{as } n \to \infty.$$

Hence with probability one, $\mathbf{1}_{A_n} - \mathbf{1}_{B_n} \to 0$ in L^1. Consequently, the sequences $\{A_n\}$ and $\{B_n\}$ have the same accumulation points, and so any convergent subsequence of $\{A_n\}$ converges to a Cheeger set of M.

4.5 Proof of Theorem 3

Let $A_n = R_n \cap M$. For all $n \geq 1$, and all f in the class of bounded and continuous functions on M, say $\mathcal{C}_b(M)$, we have

$$\left| Q_n f - \int_M f(x) \mathbf{1}_{R_n}(x) \mu(\mathrm{d}x) \right| \leq \sup_{A \in \mathcal{A}_n} \left| P_n \left(f \mathbf{1}_A \right) - \mu \left(f \mathbf{1}_A \right) \right|,$$

where $\mathcal{A}_n$ is the class of subsets of M defined in (4.10). Using the bound on the covering numbers in Lemma 9, it is a classical exercise to prove that the collection of functions $x \mapsto f(x)\mathbf{1}_A(x)$ where A ranges over $\mathcal{A}_n$ is a Glivenko-Cantelli class, whence

$$\left| Q_n f - \int_M f(x) \mathbf{1}_{R_n}(x) \mu(\mathrm{d}x) \right| \to 0 \quad \text{a.s. as } n \to \infty.$$

Next,

$$\left| \int_M f(x) \mathbf{1}_{R_{n_k}}(x) \mu(\mathrm{d}x) - Qf \right| = \left| \int_M f(x) \mathbf{1}_{A_{n_k}}(x) \mu(\mathrm{d}x) - Qf \right| \leq \|f\|_\infty P \left(A_{n_k} \Delta A_\infty \right),$$

which tends to 0 by definition of A_∞. Thus, we have shown that, for all f in $\mathcal{C}_b(M)$, $\mathbb{P}\left(Q_n f \to Qf \right) = 1$. Using the separability of $\mathcal{C}_b(M)$, see, e.g., [19, p. 131], we deduce that

$$\mathbb{P}\big[\forall f \in \mathcal{C}_b(M),\, Q_n f \to Qf \big] = 1,$$

so that the event "Q_n converge weakly to Q" is of probability 1.

5 Auxiliary results

5.1 Covering numbers

For $\rho > 0$, let $\mathcal{A}^\rho$ be the class of subsets of $A \subset M$ with $\mathrm{reach}(\partial A \cap M) \geq \rho$. Note that $\mathcal{A}_{\beta,\rho}$ given (4.1) is a sub-class of $\mathcal{A}^\rho$ for all β.

Lemma 9. *There exists constant C, v depending only on M such that, for any measure P on M, $p \geq 1$, $\rho > 0$ and $\varepsilon > 0$:*

$$\mathcal{N}_p(\varepsilon, \mathcal{A}^\rho, P) \leq \left(\frac{C}{\varepsilon} \right)^{pv}.$$

Proof. Consider 2^d points of M forming the vertices of a cube and a point x at the center of the cube. Upon choosing the side of the cube small enough, the center point x cannot be isolated by some $A \in \mathcal{A}^\rho$. Indeed, if $A \in \mathcal{A}^\rho$ contains x, because $\mathrm{reach}(\partial A) \geq \rho$, there is $y \in A$ such that $x \in B(y, \rho) \subset A$; and when r is small enough relative to ρ, $B(y, \rho)$ contains at least one vertex of the cube. This shows that $\mathcal{A}^\rho$ has VC-index less than $2^d - 1$ and so is a VC-class. The bound then follows from [33, Thm. 2.6.4]. □

Lemma 10. *Let $\mathcal{F}_{r,\rho} = \{\phi_{A,r} : A \in \mathcal{A}^\rho\}$, where $\phi_{A,r}$ is defined in (3.1), and for a probability measure Q on M, let $Q\mathcal{F}_{r,\rho} = \{x \mapsto Q\phi(x, .) : \phi \in \mathcal{F}_{r,\rho}\}$. There exists constant C, v depending only on M such that, for any probability measures P and Q on M,*

$$\mathcal{N}_1\left(\varepsilon, \mathcal{F}_{r,\rho}, P \otimes Q\right) \leq \left(\frac{C}{\varepsilon} \right)^v, \quad \mathcal{N}_2\left(\varepsilon, \mathcal{F}_{r,\rho}, P \otimes Q\right) \leq \left(\frac{C}{\varepsilon} \right)^v \quad \text{and} \quad \mathcal{N}_1(\varepsilon, Q\mathcal{F}_{r,\rho}, P) \leq \left(\frac{C}{\varepsilon} \right)^v.$$

Proof. By Lemma 9, we have

$$\begin{aligned}\mathcal{N}_1(\varepsilon, \mathcal{F}_{r,\rho}, P \otimes Q) &\leq \mathcal{N}_1\Big(2\varepsilon, \{(x,y) \mapsto \mathbf{1}_A(x) + \mathbf{1}_A(y) \,:\, A \in \mathcal{A}^\rho\}, P \otimes Q\Big) \\ &\leq \mathcal{N}_1\left(\varepsilon, \mathcal{A}^\rho, P\right) \times \mathcal{N}_1\left(\varepsilon, \mathcal{A}^\rho, Q\right) \leq \left(\frac{C}{\varepsilon}\right)^{2v}.\end{aligned}$$

That the L^2-covering number is bounded by the same quantity follows from the fact that functions in $\mathcal{F}_{r,\rho}$ are uniformly bounded by 1, and the bound on $\mathcal{N}_1(\varepsilon, Q\mathcal{F}_{r,\rho}, P)$ follows from the inequality

$$\mathcal{N}_1(\varepsilon, Q\mathcal{F}_{r,\rho}, P) \leq \mathcal{N}_1(\varepsilon, \mathcal{F}_{r,\rho}, P \otimes Q). \qquad \square$$

Lemma 11. *Let $\bar{\mathcal{F}}_{r,\rho} = \{\bar{\phi}_{A,r} : A \in \mathcal{A}^\rho\}$, where $\bar{\phi}_{A,r}$ is defined in* (3.2)*, and for a probability measure Q on M, let $Q\bar{\mathcal{F}}_{r,\rho} = \{x \mapsto Q\bar{\phi}(x,.) \,:\, \bar{\phi} \in \bar{\mathcal{F}}_{r,\rho}\}$. There exists constant C, v depending only on M such that, for any probability measures P and Q on M,*

$$\mathcal{N}_1\left(\varepsilon, \bar{\mathcal{F}}_{r,\rho}, P \otimes Q\right) \leq \left(\frac{C}{\varepsilon}\right)^v, \quad \mathcal{N}_2\left(\varepsilon, \bar{\mathcal{F}}_{r,\rho}, P \otimes Q\right) \leq \left(\frac{C}{\varepsilon}\right)^v \quad \text{and } \mathcal{N}_1(\varepsilon, Q\bar{\mathcal{F}}_{r,\rho}, P) \leq \left(\frac{C}{\varepsilon}\right)^v.$$

Proof. By Lemma 9, we have

$$\begin{aligned}\mathcal{N}_1(\varepsilon, \bar{\mathcal{F}}_n, P \otimes Q) &\leq \mathcal{N}_1\Big(2\varepsilon, \{(x,y) \mapsto \mathbf{1}_A(x)\mathbf{1}_{A^c}(y) + \mathbf{1}_A(y)\mathbf{1}_{A^c}(x) \,:\, A \in \mathcal{A}^\rho\}, P \otimes Q\Big) \\ &\leq \mathcal{N}_1\Big(\varepsilon, \{(x,y) \mapsto \mathbf{1}_A(x)\mathbf{1}_{A^c}(y)\}, P \otimes Q\Big)^2 \leq \mathcal{N}_1\left(\frac{\varepsilon}{2}, \mathcal{A}^\rho, P\right)^4 \leq \left(\frac{2C}{\varepsilon}\right)^{4v}.\end{aligned}$$

The remainder of the proof is similar to the one of Lemma 10. $\square$

5.2 Technical results on bias terms

Lemma 12. *Let $\phi_{A,r}$ be defined as in* (3.1)*. There exists a constant C, depending only on M, such that, for any $A \subset M$ and $r < \mathrm{reach}(\partial M)$,*

$$\left|\frac{1}{\omega_d r^d}\mathbb{E}\left[\phi_{A,r}(X_1, X_2)\right] - \mu(A)\right| \leq 2\mu(A \cap M_r^c).$$

Proof. We first note that

$$\mathbb{E}\left[\phi_{A,r}(X_1, X_2)\right] = \mathbb{E}\left[\mathbf{1}_A(X_1)\mathbf{1}\{\|X_1 - X_2\| \leq r\}\right].$$

We partition A into $A \cap M_r$ and $A \cap M_r^c$. By conditioning on X_1, we have

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}_{A\cap M_r}(X_1)\mathbf{1}\{\|X_1 - X_2\| \leq r\}\right] &= \omega_d r^d \mu(A \cap M_r) = \omega_d r^d \mu(A) - \omega_d r^d \mu(A \cap M_r^c); \\ \mathbb{E}\left[\mathbf{1}_{A\cap M_r^c}(X_1)\mathbf{1}\{\|X_1 - X_2\| \leq r\}\right] &\leq \omega_d r^d \mu(A \cap M_r^c).\end{aligned}$$

Hence the result. $\square$

Lemma 13. *Let $A = R \cap M$, where R is a bounded domain with smooth boundary and* $\mathrm{reach}(\partial R) = \rho > 0$*. Let $\bar{\phi}_{A,r}$ be defined as in* (3.2)*.*

(*i*) *There exists a universal constant C, depending only on d, such that, for any $A \subset M$ and $r < \min\{\rho/2, \mathrm{reach}(\partial M)\}$,*

$$\left| \frac{1}{\gamma r^{d+1}} \mathbb{E}\left[\bar{\phi}_{A,r}(X_1, X_2)\right] - \nu(A) \right| \leq C \left(\mathrm{Vol}_{d-1}(\partial R \cap \mathcal{V}(\partial M, r)) + \mathrm{Vol}_{d-1}(\partial R \cap M) \frac{r}{\rho} \right).$$

(*ii*) *There exists a universal constant C, depending only on d, such that, for any $A \subset M$ and $r < \min\{\rho/2, \mathrm{reach}(\partial M)\}$,*

$$\frac{1}{\gamma r^{d+1}} \mathbb{E}\left[\bar{\phi}_{A,r}(X_1, X_2)\right] - \frac{\mathrm{Vol}_{d-1}(\partial A \cap M_r)}{\mathrm{Vol}_d(M)} \geq -C\nu(A)\frac{r}{\rho}. \tag{5.1}$$

Proof. Assume without loss of generality that $\mathrm{Vol}_d(M) = 1$. Let S denote $\partial R \cap M$. Then

$$\mathbb{E}\left[\bar{\phi}_{A,r}(X_1, X_2)\right] = \mathbb{E}\left[\mathbf{1}_A(X_1)\mathbf{1}_{A^c}(X_2)\mathbf{1}\left\{\|X_1 - X_2\| \leq r\right\}\right] = \int_D \mathrm{Vol}_d\left[B(x,r) \cap A^c\right] \mu(\mathrm{d}x),$$

where

$$D = \left\{x \in A \,:\, \mathrm{dist}(x, \partial R) \leq r\right\}.$$

Since $r < \rho$, the projection on ∂R is well-defined on D, and any x in D can be written as $x = p + te_p$, for $p \in \partial R$, and with e_p the unit normal vector of ∂R at p pointing inwards.

We partition D into $D \cap M_r$ and $D \cap M_r^c$. Denote by S_r the projection of $D \cap M_r$ on S. We have

$$\begin{aligned}
\int_{D \cap M_r} \mathrm{Vol}_d\left[B(x,r) \cap A^c\right] \mathrm{d}x &= \int_{S_r} \int_{-r}^{0} \mathrm{Vol}_d\left[B(p + te_p, r) \cap A^c\right] \vartheta(p,t) \mathrm{d}t \, v_\sigma(\mathrm{d}p) \\
&= r \int_{S_r} \int_0^1 \mathrm{Vol}_d\left[B(p - \eta r e_p, r) \cap A^c\right] \vartheta(p, r\eta) \mathrm{d}\eta \, v_\sigma(\mathrm{d}p).
\end{aligned}$$

Therefore

$$\begin{aligned}
&\left| \frac{1}{r^{d+1}} \int_{D \cap M_r} \mathrm{Vol}_d\left[B(x,r) \cap A^c\right] \mathrm{d}x - \gamma\nu(A) \right| \\
&\quad \leq \frac{1}{r^d} \int_{S_r} \int_0^1 \left| \mathrm{Vol}_d\left[B(p - \eta r e_p, r) \cap A^c\right] - \pi_d(\eta) r^d \right| \vartheta(p, r\eta) \mathrm{d}\eta \, v_\sigma(\mathrm{d}p) \\
&\qquad + \left| \int_{S_r} \int_0^1 \pi_d(\eta) \vartheta(p, r\eta) \mathrm{d}\eta \, v_\sigma(\mathrm{d}p) - \gamma\nu(A) \right|.
\end{aligned} \tag{5.2}$$

Using Lemma 17, and then Lemma 14, the first term on the right-hand side is bounded by

$$2\omega_{d-1}(r/\rho) \int_{S_r} \int_0^1 \vartheta(p, r\eta) \mathrm{d}\eta \, v_\sigma(\mathrm{d}p) \leq 2^d \omega_{d-1}(r/\rho) \, \mathrm{Vol}_{d-1}(S_r).$$

Next, using the expansion $\vartheta(p, r\eta) = 1 + \vartheta'(p, r\xi_\eta) r\eta$ for some $0 < \xi_\eta < 1$, with $|\vartheta'(p,t)| \leq (d-1)2^d/\rho$ for all $t \in (-r, r)$ by Lemma 14, and since $r < \rho/2$, we bound the second term by

$$\begin{aligned}
&\left| \int_{S_r} \int_0^1 \pi_d(\eta) \mathrm{d}\eta \, v_\sigma(\mathrm{d}p) - \gamma\nu(A) \right| + r \int_{S_r} \int_0^1 \eta \pi_d(\eta) |\vartheta'(p, r\xi_\eta)| \mathrm{d}\eta \, v_\sigma(\mathrm{d}p) \\
&\quad \leq \gamma \left| \mathrm{Vol}_{d-1}(S_r) - \mathrm{Vol}_{d-1}(S) \right| + (d-1)2^d \gamma(r/\rho) \, \mathrm{Vol}_{d-1}(S_r) \\
&\quad \leq \gamma \, \mathrm{Vol}_{d-1}(S \cap M_r^c) + (d-1)2^d \gamma(r/\rho) \, \mathrm{Vol}_{d-1}(S_r),
\end{aligned}$$

since $S\backslash S_r \subset M_r^c$ because $S \cap M_r \subset S_r$. Collecting terms, the term in (5.2) is bounded by

$$\gamma \operatorname{Vol}_{d-1}(S \cap M_r^c) + C\frac{r}{\rho}\operatorname{Vol}_{d-1}(S_r),$$

for some constant C independent of M.

For the integral over $D \cap M_r^c$, since D is included in the intersection of tubes of radius r about ∂R and ∂M, i.e., $D \subset \mathcal{V}(\partial R, r) \cap \mathcal{V}(\partial M, r)$, we have

$$\begin{aligned}
\int_{D\cap M_r^c} \operatorname{Vol}_d\left[B(x,r)\cap A^c\right] \mathrm{d}x &\leq \int_{\partial R\cap\mathcal{V}(\partial M,r)} \int_{-r}^{0} \operatorname{Vol}_d\left[B(p+te_p,r)\cap A^c\right]\vartheta(p,t)\mathrm{d}t\, v_\sigma(\mathrm{d}p) \\
&= r\int_{\partial R\cap\mathcal{V}(\partial M,r)} \int_{0}^{1} \operatorname{Vol}_d\left[B(p-\eta r e_p,r)\cap A^c\right]\vartheta(p,r\eta)\mathrm{d}\eta\, v_\sigma(\mathrm{d}p) \\
&\leq 2^{d-1}\omega_d r^{d+1}\operatorname{Vol}_{d-1}(\partial R\cap\mathcal{V}(\partial M,r)),
\end{aligned}$$

where we used Lemma 14 in the last inequality.

Combining the bounds on the integrals over $D \cap M_r$ and $D \cap M_r^c$, we obtain that

$$\begin{aligned}
&\left|\frac{1}{\gamma r^{d+1}}\mathbb{E}\left[\bar{\phi}_{A,r}(X_1,X_2)\right] - \nu(A)\right| \\
&\quad\leq \operatorname{Vol}_{d-1}(S\cap M_r^c) + C\frac{r}{\rho}\operatorname{Vol}_{d-1}(S_r) + 2^{d-1}\omega_d\operatorname{Vol}_{d-1}(\partial R\cap\mathcal{V}(\partial M,r)) \\
&\quad\leq C\left(\operatorname{Vol}_{d-1}(\partial R\cap\mathcal{V}(\partial M,r)) + \operatorname{Vol}_{d-1}(S)\frac{r}{\rho}\right),
\end{aligned}$$

which proves the first bound stated in Lemma 13.

To prove (ii), using the bound on (5.2), we deduce that

$$\begin{aligned}
\frac{1}{\gamma r^{d+1}}\mathbb{E}\left[\bar{\phi}_{A,r}(X_1,X_2)\right] &\geq \frac{1}{\gamma r^{d+1}}\int_{D\cap M_r}\operatorname{Vol}_d\left[B(x,r)\cap A^c\right]\mathrm{d}x \\
&\geq \operatorname{Vol}_{d-1}(S) - \left[\operatorname{Vol}_{d-1}(S\cap M_r^c) + \frac{C}{\gamma}\frac{r}{\rho}\operatorname{Vol}_{d-1}(S_r)\right] \\
&\geq \operatorname{Vol}_{d-1}(S\cap M_r) - C\frac{r}{\rho}\operatorname{Vol}_{d-1}(S_r),
\end{aligned}$$

and since $S_r \subset S$, the result follows. □

6 Geometrical results

6.1 Integration in tubes

We introduce the notion of tubes and some of their properties; see [22] for an extensive treatment. Let S be a submanifold of $\mathbb{R}^d$. The *tubular neighborhood* of radius $r > 0$ about S, denoted $\mathcal{V}(S,r)$, is the set of points x in $\mathbb{R}^d$ for which there exists $s \in S$ with $\|x - s\| < r$ and such that the line joining x and s is orthogonal to S at s. When S is without boundary, $\mathcal{V}(S,r)$ coincides with the set of points x in $\mathbb{R}^d$ at a distance no more than r from S. If S has boundary, then the tube

coincides with the set of points at distance no more than r, with the ends removed, corresponding to the points projecting onto ∂S. Assume S is of codimension 1, and oriented, and define e_p as the (unit) normal vector of S at $p \in S$. When and $r < \mathrm{reach}(S)$, $\mathcal{V}(S,r)$ admits the following parameterization

$$\mathcal{V}(S,r) = \{x = p + te_p : p \in S, -r \le u \le r\}.$$

Denote by $\mathbb{I}\!\mathbb{I}_p$ the second fundamental form of S at $p \in S$. The infinitesimal change of volume function is defined on $S \times (-r;r)$ by $\vartheta(p,t) = \det(I - t\mathbb{I}\!\mathbb{I}_p)$; the dependence of ϑ on S is omitted. Given an integrable function g on $\mathcal{V}(S,r)$, we have:

$$\int_{\mathcal{V}(S,r)} g(x)\mathrm{d}x = \int_S \int_{-r}^{r} g(p,t)\vartheta(p,t)\mathrm{d}t\, v_\sigma(\mathrm{d}p),$$

where v_σ is the Riemannian volume measure on S.

Lemma 14. *Assume S is a submanifold of $\mathbb{R}^d$ of codimension 1, with $\rho := \mathrm{reach}(S) > 0$. Then, for all $r < \rho$,*

$$\sup_{p\in S}\sup_{-r\le t\le r} \vartheta(p,t) \le (1 + r/\rho)^{d-1},$$

and

$$\sup_{p\in S}\sup_{-r\le t\le r} |\vartheta'(p,t)| \le \frac{(d-1)(1+r/\rho)^{d-1}}{\rho - r}.$$

Proof. By [20, Thm. 4.18], the reach bounds the radius of curvature from below so that the principal curvatures $\kappa^{(1)}, \ldots, \kappa^{(d-1)}$ (the eigenvalues of the second fundamental form) are everywhere bounded (in absolute value) from above by $1/\rho$. Therefore, for $r < \rho$ and $-r \le t \le r$,

$$0 \le \vartheta(p,t) = \det(I - t\mathbb{I}\!\mathbb{I}_p) = \prod_{i=1}^{d-1}\left(1 - \kappa_p^{(i)} t\right) \le (1 + r/\rho)^{d-1}.$$

For the derivative of ϑ, we have

$$\frac{\vartheta'(p,t)}{\vartheta(p,t)} = -\sum_{i=1}^{d-1} \frac{\kappa_p^{(i)}}{1 - \kappa_p^{(i)} t}.$$

Hence

$$|\vartheta'(p,t)| \le \vartheta(p,t)(d-1)\frac{1/\rho}{1 - r/\rho} \le \frac{(d-1)(1+r/\rho)^{d-1}}{\rho - r}.$$ □

Lemma 15. *There exists a positive constant C, depending only on M, such that, for all $r <$ $\mathrm{reach}(\partial M)$,*

$$\mu\left[\mathcal{V}(\partial M, r)\right] \le Cr.$$

Proof. Let $\rho = \mathrm{reach}(\partial M)$. By Lemma 14,

$$\begin{aligned}\mu\left[\mathcal{V}(\partial M, r)\right] &= \frac{1}{\mathrm{Vol}_d(M)} \int_{\partial M}\int_{-r}^{r} \vartheta(p,u)\mathrm{d}u\, v_\sigma(\mathrm{d}p) \\ &\le \frac{\mathrm{Vol}_{d-1}(\partial M)}{\mathrm{Vol}_d(M)} 2r(1 + r/\rho)^{d-1} \le \frac{2^d \,\mathrm{Vol}_{d-1}(\partial M)}{\mathrm{Vol}_d(M)} r.\end{aligned}$$ □

6.2 Perimeter bounds

Let E, F be two Borel subsets of $\mathbb{R}^d$. Recall the following isoperimetric inequality in $\mathbb{R}^d$ (see e.g., Evans and Gariepy, 1992):

$$d\omega_d^{1/d}\,\mathrm{Vol}_d(E)^{1-1/d} \le \mathrm{Vol}_{d-1}(\partial E). \tag{6.1}$$

We also have

$$\mathrm{Vol}_{d-1}\left(\partial(E\cup F)\right) + \mathrm{Vol}_{d-1}\left(\partial(E\cap F)\right) \le \mathrm{Vol}_{d-1}(\partial E) + \mathrm{Vol}_{d-1}(\partial F). \tag{6.2}$$

Lemma 16. *Let R be a bounded open subset of $\mathbb{R}^d$ with smooth boundary and $\mathrm{reach}(\partial R) = \rho > 0$. Then,*

$$\mathrm{Vol}_{d-1}(R) \le d\,\mathrm{Vol}_d(R)/\rho.$$

Proof. Since $\mathrm{reach}(\partial R) = \rho > 0$, a ball of radius ρ rolls freely in R. Consequently R can be written as a countable union of balls of radius ρ, i.e.,

$$R = \bigcup_{i=1}^{\infty} B(x_i, \rho).$$

Set $R_n = \cup_{i=1}^n B_i$ where $B_i = B(x_i, \rho)$.

Using the decomposition $R_{n+1} = R_n \cup B_{n+1}$, on the one hand we have

$$\mathrm{Vol}_d(R_{n+1}) = \mathrm{Vol}_d(R_n \cup B_{n+1}) = \mathrm{Vol}_d(R_n) + \omega_d\rho^d - \mathrm{Vol}_d(R_n \cap B_{n+1}),$$

and on the other hand, using inequality (6.2), we have

$$\mathrm{Vol}_{d-1}(\partial R_{n+1}) = \mathrm{Vol}_{d-1}(\partial(R_n \cup B_{n+1})) \le \mathrm{Vol}_{d-1}(\partial R_n) + d\omega_d\rho^{d-1} - \mathrm{Vol}_{d-1}(\partial(R_n \cap B_{n+1})).$$

Consequently

$$\begin{aligned}\mathrm{Vol}_{d-1}(\partial R_{n+1}) - \frac{d}{\rho}\mathrm{Vol}_d(R_{n+1}) \;\le\; & \mathrm{Vol}_{d-1}(\partial R_n) - \frac{d}{\rho}\mathrm{Vol}_d(R_n) \\ & + \left[\frac{d}{\rho}\mathrm{Vol}_d(R_n \cap B_{n+1}) - \mathrm{Vol}_{d-1}(\partial(R_n \cap B_{n+1}))\right].\end{aligned}$$

But, using the isoperimetric inequality (6.1), we may write

$$\begin{aligned}&\frac{d}{\rho}\mathrm{Vol}_d(R_n \cap B_{n+1}) - \mathrm{Vol}_{d-1}\left(\partial(R_n \cap B_{n+1})\right) \\ &\quad\le \frac{d}{\rho}\mathrm{Vol}_d(R_n \cap B_{n+1}) - d\omega_d^{1/d}\Big(\mathrm{Vol}_d(R_n \cap B_{n+1})\Big)^{1-1/d} \\ &\quad\le \Big(\mathrm{Vol}_d(R_n \cap B_{n+1})\Big)^{1-1/d}\left[\frac{d}{\rho}\mathrm{Vol}_d(R_n \cap B_{n+1})^{1/d} - d\omega_d^{1/d}\right] \le 0\end{aligned}$$

since, in the last bracket, $\mathrm{Vol}_d(R_n \cap B_{n+1}) \le \mathrm{Vol}_d(B_{n+1}) = \omega_d\rho^d$. Therefore, for all $n \ge 1$, we have

$$\mathrm{Vol}_{d-1}(\partial R_{n+1}) - \frac{d}{\rho}\mathrm{Vol}_d(R_{n+1}) \le \mathrm{Vol}_{d-1}(\partial R_n) - \frac{d}{\rho}\mathrm{Vol}_d(R_n).$$

But since R_1 is a ball of radius ρ, we have $\mathrm{Vol}_{d-1}(\partial R_1) - d\,\mathrm{Vol}_d(R_1)/\rho = 0$ and so

$$\mathrm{Vol}_{d-1}(\partial R_n) - \frac{d}{\rho}\,\mathrm{Vol}_d(R_n) \leq 0 \quad \text{for all } n \geq 1.$$

Since R_n converges to R in L^1, it follows from the lower semi-continuity of the perimeter, see e.g. [23, Prop. 2.3.6], that $\liminf_n \mathrm{Vol}_{d-1}(\partial R_n) \geq \mathrm{Vol}_{d-1}(\partial R)$. This concludes the proof. □

6.3 Volume bounds

Lemma 17. *Let R be a bounded open subset of $\mathbb{R}^d$ with smooth boundary and $\mathrm{reach}(\partial R) = \rho > 0$. Set $A = R \cap M$. For any $r < \min\{\mathrm{reach}(\partial M); \rho\}$, any $0 \leq \eta \leq 1$, and all p in $\partial A \cap M_r$, we have*

$$\left|\mathrm{vol}_d\left(B(p + \eta r e_p, r) \cap A^c\right) - \pi_d(\eta) r_n^d\right| \leq 2\omega_{d-1} r^{d+1}/\rho,$$

where e_p denotes the unit normal vector at p pointing inward A.

Proof. For ease of notation, set $B = B(p + \eta r e_p, r)$. Let $(\tilde{e}_1, \ldots, \tilde{e}_d)$ be an orthonormal frame at p, with $\tilde{e}_d = e_p$. Denote by $\tilde{x}_1, \ldots, \tilde{x}_d$ the local coordinates in this frame, such that p has coordinates 0. Then $\partial A \cap M$ can be expressed locally as the set of points $\tilde{x}$ such that $\tilde{x}^d = F(\tilde{x}^1, \ldots, \tilde{x}^{d-1})$ for some function F, and, if we set $\tilde{x}^{(d)} = (\tilde{x}^1, \ldots, \tilde{x}^{d-1})$, then

$$\begin{aligned}
\mathrm{Vol}_d(B \cap A^c) &= \int_B \mathbf{1}\{\tilde{x}^d < F(\tilde{x}^{(d)})\}\mathrm{d}\tilde{x} \\
&= \int_B \left[\mathbf{1}\{\tilde{x}^d < F(\tilde{x}^{(d)})\}\mathbf{1}\{\tilde{x}^d < 0\} + \mathbf{1}\{\tilde{x}^d < F(\tilde{x}^{(d)})\}\mathbf{1}\{\tilde{x}^d > 0\}\right] \mathrm{d}\tilde{x}
\end{aligned}$$

Since

$$\pi_d(\eta) r^d = \int_B \mathbf{1}\{\tilde{x}^d < 0\}\mathrm{d}\tilde{x}$$

it follows that

$$\begin{aligned}
\left|\mathrm{vol}_d\left(B_n \cap A^c\right) - \pi_d(\eta) r_n^d\right| &\leq \int_B \left[\mathbf{1}\{\tilde{x}^d > F(\tilde{x}^{(d)})\}\mathbf{1}\{\tilde{x}^d < 0\} + \mathbf{1}\{\tilde{x}^d < F(\tilde{x}^{(d)})\}\mathbf{1}\{\tilde{x}^d > 0\}\right] \mathrm{d}\tilde{x} \\
&\leq \int_{B_n} \mathbf{1}\left\{|\tilde{x}^d| \leq |F(\tilde{x}^{(d)})|\right\} \mathrm{d}\tilde{x} \leq 2 \int_{\{\|\tilde{x}^{(d)}\| \leq r\}} |F(\tilde{x}^{(d)})| \mathrm{d}\tilde{x}^{(d)}.
\end{aligned}$$

Expanding F at 0, we have, for all $\tilde{x}$ with $\|\tilde{x}\| \leq r$,

$$F(\tilde{x}^{(d)}) = \sum_{i,j=1}^{d-1} H_{ij}(\xi)\tilde{x}^i\tilde{x}^j,$$

for some $\xi := \xi(\tilde{x}^{(d)})$. Since the reach bounds the principal curvatures by $1/\rho$ [20], we have $\sup_{p \in \partial A \cap M_r} \|H(p)\| \leq 1/\rho$. Then, using the change of variable $u = r\tilde{x}$, we deduce that

$$\begin{aligned}
\left|\mathrm{vol}_d\left(B(p + \eta r e_p, r) \cap A^c\right) - \pi_d(\eta) r_n^d\right| &\leq 2\omega_{d-1} \sup_{p \in \partial A \cap M} \|H(p)\| r^{d+1} \\
&\leq 2\omega_{d-1} r^{d+1}/\rho.
\end{aligned}$$

□

Lemma 18. *There exists a constant $C > 0$ such that, for all α, r satisfying $0 < 2r \leq \alpha \leq \mathrm{reach}(\partial M)$, and all x in M,*

$$\mathrm{Vol}_d(B(x,\alpha) \cap M_r) \geq C\alpha^d.$$

Proof. The main argument is to include a ball of radius $\alpha/4$ into $B(x,\alpha) \cap M_r$. First, because $\rho := \mathrm{reach}(\partial M) > 0$, for any $x \in M$ there is $y \in M$ such that $x \in B(y,\rho) \subset M$. Second, since $\mathrm{dist}(y, \partial M) \geq \rho$ and $\rho \geq 2r$, we have $y \in M_r$ and $B(y, \rho - r) \subset M_r$. Hence

$$B(x,\alpha) \cap B(y,\rho - r) \subset B(x,\alpha) \cap M_r.$$

If $y = x$, the result is trival. Otherwise, let $z := x + (r + \alpha/4)(y - x)/\|y - x\|$ and note that $B(z, \alpha/4)$ is a ball of radius $\alpha/4$ included in $B(x,\alpha) \cap B(y, \rho - r)$. □

6.4 Some properties of the Cheeger constant and Cheeger sets

In this section, we prove some properties of the Cheeger constant and Cheeger sets. As in all the paper, M denotes a bounded, connected, open subset of $\mathbb{R}^d$ with smooth boundary. To this aim, we will make use a bi-Lipschitz deformation of M. For a Lipschitz map f, let $\|f\|_{\mathrm{Lip}}$ denote its Lipschitz constant. If f is bi-Lipschitz, we define its condition number by $\mathrm{cond}(f) := \|f\|_{\mathrm{Lip}} \|f^{-1}\|_{\mathrm{Lip}}$. We shall need the following two lemmas.

Lemma 19. *Let f be a bi-Lipschitz on M. Then for any $A \subset M$ measurable,*

$$\max\left\{ \frac{h(f(A); f(M))}{h(A; M)}, \frac{h(A; M)}{h(f(A); f(M))} \right\} \leq \mathrm{cond}(f)^d.$$

Proof. For any $A \subset M$, $\partial f(A) = f(\partial A)$ and $f(A)^c \cap f(M) = f(A^c \cap M)$, and if A is measurable, for $k = 1, \ldots, d$,

$$\|f^{-1}\|_{\mathrm{Lip}}^{-k} \mathrm{Vol}_k(A) \leq \mathrm{Vol}_k(f(A)) \leq \|f\|_{\mathrm{Lip}}^k \mathrm{Vol}_k(A).$$

Therefore,

$$\begin{aligned} h(f(A); f(M)) &= \frac{\mathrm{Vol}_{d-1}(f(\partial A \cap M))}{\min\{\mathrm{Vol}_d(f(A)), \mathrm{Vol}_d(f(A^c \cap M))\}} \\ &\leq \frac{\|f\|_{\mathrm{Lip}}^{d-1} \mathrm{Vol}_{d-1}(\partial A \cap M)}{\|f^{-1}\|_{\mathrm{Lip}}^{-d} \min\{\mathrm{Vol}_d(A), \mathrm{Vol}_d(A^c \cap M)\}} \\ &\leq \mathrm{cond}(f)^d \, h(A; M). \end{aligned}$$

And vice-versa. □

Lemma 20. *For any $r < \rho_M$, we denote by M_r the subset of M made of points at a distance r or more from ∂M. Fix $r < s \leq \mathrm{reach}(\partial M)$. Then, there is a bi-Lipschitz map between M_r and M that leaves M_s unchanged, and with condition number at most $(1 + 2r/(s - r))^2$.*

Proof. For x in M such that $\delta(x) := \mathrm{dist}(x, \partial M) < s$, let $\xi(x) \in M$ be its metric projection onto ∂M and u_x be the unit normal vector of M at $\xi(x)$ pointing outwards. We define the map

$$f_r : M_r \mapsto M, \quad f_r(x) = x + \frac{r(s - \delta(x))_+}{s - r} \, u_x,$$

where a_+ denotes the positive part of $a \in \mathbb{R}$. By construction, f is one-to-one, with inverse

$$f_r^{-1} : M \mapsto M_r, \quad f_r^{-1}(x) = x - \frac{r(s - \delta(x))_+}{s}\, u_x.$$

By [20, Thm. 4.8(1)], δ is Lipschitz with constant at most 1, therefore so is $x \mapsto (s - \delta(x))_+$; and since the reach bounds the radius of curvature from below [20, Thm. 4.18], $x \mapsto u_x$ is Lipschitz with constant at most $1/\operatorname{reach}(\partial M)$. Therefore, using the fact that $(s - \delta(x))_+ \le s$ and $\|u_x\| = 1$, f_r and f_r^{-1} are Lipschitz with constants at most $1 + 2r/(s - r)$ and $1 + 2r/s$ respectively. □

Proposition 21. *For $r < \operatorname{reach}(\partial M)$, let M_r denote the subset of M made of points at a distance r or more from ∂M. Then*

$$H(M_r) = (1 + O(r))\, H(M), \quad r \to 0.$$

Proof. From Lemmas 19 and 20, we deduce that

$$\max\left\{ \frac{H(M_r)}{H(M)}, \frac{H(M)}{H(M_r)} \right\} \le (1 + 2r/(\rho_M - r))^{2d},$$

for any $r < \rho_M := \operatorname{reach}(\partial M)$, which immediately yields the desired result. □

The Appendix

A U-statistics

The following is Hoeffding's Inequality for U-statistics [24] and is a special case of [18, Thm. 4.1.8].

Lemma 22. *Let ϕ be a measurable, bounded kernel on $\mathbb{R}^d \times \mathbb{R}^d$ and let $\{X_k : k \in \mathbb{N}\}$ be i.i.d. random vectors in $\mathbb{R}^d$. Assume that $\mathbb{E}[\phi(X_1, X_2)] = 0$ and that $b := \|\phi\|_\infty < \infty$, and let $\sigma^2 = \operatorname{Var}(\phi(X_1, X_2))$. Then, for all $t > 0$,*

$$\mathbb{P}\left(\frac{1}{n(n-1)} \sum_{i \ne j} \phi(X_i, X_j) \ge t \right) \le \exp\left(-\frac{nt^2}{5\sigma^2 + 3bt} \right).$$

The following is a uniform version of Lemma 22 and is a special case of [18, Thm. 5.3.15].

Lemma 23. *Let $\mathcal{H}$ be a class of symmetric kernels ϕ on $\mathbb{R}^d \times \mathbb{R}^d$ such that $b := \sup_{\phi \in \mathcal{H}} \|\phi\|_\infty < \infty$ and*

$$c := \sup_P \int_0^{2b} \log \mathcal{N}_2(\varepsilon, \mathcal{H}, P) \mathrm{d}\varepsilon < \infty,$$

where the supremum is over all the probability measures on $\mathbb{R}^d \times \mathbb{R}^d$. Let $\{X_k : k \in \mathbb{N}\}$ be i.i.d. random vectors in $\mathbb{R}^d$ and assume that $\mathbb{E}[\phi(X_1, X_2)] = 0$ for all $\phi \in \mathcal{H}$. Then there exists a constant C_1, not depending on n or $\mathcal{H}$, such that, for all $\varepsilon > 0$ and all $n \ge \frac{C_1 \log(C_1) c}{\varepsilon}$,

$$\mathbb{P}\left(\sup_{\phi \in \mathcal{H}} \left| \frac{1}{n(n-1)} \sum_{i \ne j} \phi(X_i, X_j) \right| > \varepsilon \right) \le C_1 \exp\left(-\frac{n\varepsilon}{C_1 c} \right).$$

B Convergence of sets in the L^1-metric

The following propositions are adapted from [23, Thm. 2.3.10] and [23, Prop. 2.3.6] respectively.

Proposition 24. *Let E_n be a sequence of measurable subsets of M. Suppose that*

$$\limsup_{n\to\infty} \mathrm{Vol}_{d-1}(\partial E_n \cap M) < \infty.$$

Then (E_n) admits a subsequence converging for the L^1-metric.

Proposition 25. *Let E_n and E be bounded measurable subsets of M such that $\mathbf{1}_{E_n} \xrightarrow{L^1} \mathbf{1}_E$ and $h(E;M) < \infty$. Then*

$$\liminf_n h(E_n;M) \geq h(E;M).$$

Acknowledgments

Ery Arias-Castro was partially supported by a grant from the US National Science Foundation (DMS-06-03890) and wishes to thank Lei Ni for stimulating discussions. Bruno Pelletier and Pierre Pudlo were supported by the French National Research Agency (ANR) under grant ANR-09-BLAN-0051-01. Bruno Pelletier would like to thank Michel Pierre and Jimmy Lamboley for insightful discussions on Cheeger sets.

References

[1] S. Arora, E. Hazan, and S. Kale. $O(\sqrt{\log n})$-approximation to sparsest cut in $\tilde{O}(n^2)$ time. In *FOCS '04: Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, pages 238–247, Washington, DC, USA, 2004. IEEE Computer Society.

[2] C. Avin and G. Ercal. On the cover time and mixing time of random geometric graphs. *Theor. Comput. Sci.*, 380(1-2):2–22, 2007.

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems*, 1:585–592, 2002.

[4] M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. *J. Comput. System Sci.*, 74(8):1289–1308, 2008.

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

[6] G. Biau, B. Cadre, and B. Pelletier. Exact rates in density support estimation. *Journal of Multivariate Analysis*, 99(10):2185–2207, 2008.

[7] S. P. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Mixing times for random walks on geometric random graphs. In C. Demetrescu, R. Sedgewick, and R. Tamassia, editors, *SIAM Workshop on Analytic Algorithmics & Combinatorics (ANALCO)*, pages 240–249. SIAM, 2005.

[8] H. Bräker and T. Hsing. On the area and perimeter of a random convex hull in a bounded convex set. *Probab. Theory Related Fields*, 111(4):517–550, 1998.

[9] P. Buser. A note on the isoperimetric constant. *Ann. Sci. École Norm. Sup. (4)*, 15(2):213–230, 1982.

[10] G. Carlsson. Topology and data. *Bull. Amer. Math. Soc. (N.S.)*, 46(2):255–308, 2009.

[11] G. Carlsson and A. Zomorodian. The theory of multidimensional persistence. *Discrete Comput. Geom.*, 42(1):71–93, 2009.

[12] V. Caselles, A. Chambolle, and M. Novaga. Some remarks on uniqueness and regularity of Cheeger sets. *Rend. Sem. Mat. Univ. Padova*, 2009.

[13] F. Chazal, L. Guibas, S. Oudot, and P. Skraba. Analysis of scalar fields over point cloud data. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1021–1030. Society for Industrial and Applied Mathematics, 2009.

[14] F. Chazal and A. Lieutier. Weak feature size and persistant homology: computing homology of solids in $\mathbb{R}^n$ from noisy data samples. In *Computational geometry (SCG'05)*, pages 255–262. ACM, New York, 2005.

[15] J. Cheeger. A lower bound for the smallest eigenvalue of the Laplacian. In *Problems in analysis (Papers dedicated to Salomon Bochner, 1969)*, pages 195–199. Princeton Univ. Press, Princeton, N. J., 1970.

[16] F. R. K. Chung. *Spectral graph theory*. Number 92 in Regional Conference Series in Mathematics. Amer. Math. Soc., Providence, 1997.

[17] A. Cuevas, R. Fraiman, and A. Rodríguez-Casal. A nonparametric approach to the estimation of lengths and surface areas. *Ann. Statist.*, 35(3):1031–1051, 2007.

[18] V. H. de la Peña and E. Giné. *Decoupling*. Probability and its Applications (New York). Springer-Verlag, New York, 1999. From dependence to independence, Randomly stopped processes. U-statistics and processes. Martingales and beyond.

[19] J. L. Doob. *Measure theory*, volume 143 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1994.

[20] H. Federer. Curvature measures. *Trans. Amer. Math. Soc.*, 93:418–491, 1959.

[21] 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.

[22] A. Gray. *Tubes*, volume 221 of *Progress in Mathematics*. Birkhäuser Verlag, Basel, second edition, 2004. With a preface by Vicente Miquel.

[23] A. Henrot and M. Pierre. *Variation et optimisation de formes*, volume 48 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer, Berlin, 2005. Une analyse géométrique. [A geometric analysis].

[24] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[25] E. Levina and P. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems*, volume 17, pages 777–784. MIT Press, Cambridge, Massachusetts, 2005.

[26] A. Y. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems (NIPS)*, volume 14, pages 849–856, 2001.

[27] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.*, 39(1-3):419–441, 2008.

[28] B. Pelletier and P. Pudlo. Operator norm convergence of spectral clustering on level sets. Available from `http://fr.arXiv.org/abs/1002.2313`, 2009.

[29] M. Penrose. *Random Geometric Graphs*, volume 5 of *Oxford Studies in Probability*. Oxford University Press, Oxford, 2003.

[30] V. Robins. Towards computing homology from finite approximations. In *Proceedings of the 14th Summer Conference on General Topology and its Applications (Brookville, NY, 1999)*, volume 24, pages 503–532, 1999.

[31] A. Singer. From graph to manifold Laplacian: the convergence rate. *Appl. Comput. Harmon. Anal.*, 21(1):128–134, 2006.

[32] D. A. Spielman and S.-H. Teng. Spectral partitioning works: planar graphs and finite element meshes. *Linear Algebra Appl.*, 421(2-3):284–305, 2007.

[33] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[34] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[35] U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, 36(2):555–586, 2008.

[36] G. Walther. Granulometric smoothing. *Ann. Statist.*, 25(6):2273–2299, 1997.

[37] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete Comput. Geom.*, 33(2):249–274, 2005.