THE UNIVERSITY
of EDINBURGH

Random Geometric Graphs and Data Structures: Discrete Estimations of the Continuous

Henry-Louis de Kergorlay

A thesis presented for the degree of
Doctor of Philosophy

School of Mathematics
University of Edinburgh
2019

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification.

September 10th, 2019, *Henry-Louis de Kergorlay*

Acknowledgements

I would like to thank, following a chronological order, the many people who played an important role in my mathematical development, and without which I would not have written this thesis.

To the exceptional teachers that I was fortunate to have from an early age in France, who transmitted to me their passion and who deeply influenced my understanding of mathematics from an early stage. I am particularly grateful to Marianne Rolland-Billecart for her invaluable support and passion, and her extraordinary pedagogical methods. I am also very grateful to Marielle Say and to Lionel Ponton from whom I learned so much.

My thoughts then flow across the Atlantic, in order to thank the wonderful faculty members which I met at Wesleyan University, during my undergraduate studies. In particular, I am extremely grateful to my supervisor there, Prof. W.K. Chan, who accompanied my time at Wesleyan with invaluable advise and one-on-one tutorials. I was also particularly marked by Prof. Karen Collins and Prof. Adam Fieldsteel, both for their methods and enthusiasm.

During this time, I came to be exposed to my first research experiences. I would like to thank in particular Augustin Huret and Liljana Babinkostova. I am also very grateful to Prof. Dennis Shasha who allowed me to work on a wonderful research project during my last summer in the US, at NYU.

My mathematical journey accross the English Channel starts with the Part III course, at Cambridge. I was particularly marked by two people there: Prof. T. Gowers, for his course on Techniques in Combinatorics and with whom I also wrote my Masters dissertation; and Prof. B. Bollobas for his course Probabilistic and Topological Combinatorics. I feel honoured to have had the privilege to learn from these exceptional mathematicians.

I then moved up North, accross Hadrian's Wall, where my mathematical journey finally entered its PhD phase, here in Edinburgh. First of all, I would like to thank my supervisor, Prof. Higham. He took me onboard towards the end of my PhD, in a difficult situation. Yet, I have learned a lot under his serene and experienced guidance and mentorship. I am grateful for his immense support and his numerous and insightful advise and suggestions. I am also very grateful to Prof. Carbery for his constant support and kindness throughout my time in Edinburgh, and for his mentorship during a truly inspiring first year project on Calderon-Zygmund theory. I would also like to thank Prof. Wright, who I keep meeting in various cafes, for many warm conversations.

I am also grateful to the people I have met during the year I spent at the Alan Turing Institute in London. Ulrike Tillmann and Oliver Vipond with whom I could frequently discuss on ideas related to random topology and from whom I learned a lot. Mihai Cucuringu and Hemant Tyagi who introduced me to exciting problems and who were very available for discussion.

I would also like to thank several friends from Edinburgh and London. William for hosting me during the last couple of months of my time here, and for reading through some bits of the thesis. Alvin for some great conversations and references on the ∞-Wasserstein distance last year, when I started to get interested in it. Alvin also for his shared passion for wine and the great wines from Central Europe and New Zealand that he showed me. Leonardo and Xiling for way too many gatherings, with fun chats and songs and often accompanied with a much recommended combination of whisky and prosciutto. Florentin for introducing me to the wonderful world of Belgium beers. Roland for delightful tea-Thor Sunday afternoons. Roland and Eva for great escapades in London and Edinburgh, here and there.

Last but not least, my thought goes to my family.

A bon-papa, dont le sens de l'honneur et la combativité ont toujours eu un grand écho dans mon coeur. Son exemple saura toujours m'insuffler un esprit guerrier dans les moments les plus âpres.

A maman, pour son aide constante, pour son allant et sa combativité, et pour ses appels quasi-quotidiens qui sont une preuve du grand amour qu'elle me voue. A papa, pour ses conseils, pour sa grande liberté en toute chose, pour l'authenticité de ses nombreux engagements, et pour ces délicieux moments passés tous les deux à Wesleyan et Edimbourg. A Marie-Victoire, ma soeur chérie, si sensible, qui est venue me voir plusieurs fois à Edimbourg et qui pense souvent à moi avec plein de bonnes intentions. A Aimée et Raphaël, dont j'ai reçu l'immense honneur d'être le parrain. A Elisabeth et Gabriel, mes chers cousins de Londres. Gabriel dont je suis aussi le parrain de confirmation. A Emmanuel, Anne-Laure, Aurélia, Edouard, Chantal et bonne-maman, et à mes nombreux petits cousins qui font eux aussi, partie de la fière cousinade. A Arnaud, Gérald et Brigitte, Laurence. A Geoffroy. A Christelle. A Lorry.

Abstract

In recent years, there have been significant efforts to develop rigourous geometric and topological methodologies to better understand data sets, with the view of performing machine learning tasks such as spectral clustering, non-linear dimensionality reduction and Topological Data Analysis. These methodologies have various objectives, use different techniques and are fundamentally different. For instance, the techniques used to tackle the consistency of spectral clustering or Cheeger consistency which are also discussed in this thesis, follow the so-called *variational approach* (proposed by Trillos and Slepcev in a series of works, initiated in [63]). On the other hand, techniques used to investigate topological features of random geometric complexes usually require knowledge of different fields such as Morse theory. While these methodologies are fundamentally different, they may be related through the concept of random geometric graphs. On the one hand, problems such as spectral clustering and Cheeger consistency can be seen as optimisation problems on functionals defined on random geometric graphs. On the other hand, random geometric complexes can be thought as generalisations of random geometric graphs, where we not only take into consideration vertices and edges, but also higher order simplices. Another common point is that all these methodologies are attempts to capture various geometric invariants of a manifold from an underlying sampled set of i.i.d. points (e.g., Betti numbers, Cheeger constants, spectrum of the Laplacian).

While these methodologies have been applied successfully, there is still a lack of results providing theoretical guarantees and rigorously explaining the extent to which these various frameworks effectively work, under various settings. The aim of this thesis is to contribute to the development of such theoretical guarantees, building on existing results and methods.

Lay summary

The aim of this thesis is to investigate theoretical aspects of mathematical problems motivated by applications in machine learning and data science. One type of result we will be concerned with deals with the key task of clustering. This consists in grouping data points according to some affinities, in order to extract structure from a given data set. As such, it is a very common and central task in machine learning and it is desirable to understand it better from a theoretical point of view. Another type of result we shall investigate is on problems motivated by Topological Data Analysis. This field is, loosely speaking, concerned with extracting geometric shapes and patterns from data sets. It finds applications in various fields, such as medicine and neuroscience. In such applications, a key challenge is to design computer algorithms that help us to summarize, visualize or find patterns in large and complex data sets.

Contents

Chapter 1

Introduction

A recurring objective of topological and geometric methodologies in machine learning is to infer geometric properties/invariants of a manifold from a discrete sampled set of points. Many of those methodologies start by connecting nearby points with respect to an underlying metric, thus forming a so-called *random geometric graph*. Random geometric graphs have been studied extensively by Penrose in [55] and they contiune to be actively researched.

A classic problem, then, consists of comparing the asymptotic evolution of geometric graph operators, functionals or quantities, with their continuous counterparts. There are various graph quantities of interest to investigate as the number of vertices $n \to \infty$, or better, with an error estimate given as a function of n.

For instance, one may want to investigate the homology induced from a random geometric graph and ask whether it recovers the homology of the underlying manifold (from which the vertices of the graph are sampled) with high probability (w.h.p.) (e.g., [20, 41, 9, 11, 10, 42]). Given a simplicial complex built from a discrete set of points X_n with bandwidth parameter r_n (two classic choices are the Čech or the Vietoris-Rips complex), one may be interested to know whether the homology groups of different degrees are isomorphic to those of the underlying manifold M from which the points are sampled. In this case, the discrete quantities of interests are the Betti numbers β_k, and we ask under which conditions it holds that $\beta_k($ Čech or Vietoris-Rips complex $) \to \beta_k(M)$ with high probability (w.h.p.). Besides being an interesting problem in its own right, such a problem also finds practical applications with views towards Topological Data Analysis and statistical Persistent Homology. There, it is of interest to estimate threshold values for the bandwidth parameter of a simplicial complex (built from a sampled set), beyond which the induced homology recovers exactly that of the manifold (from which the set was sampled) with probability tending to 1 as $n \to \infty$. See [50, 51, 31, 32, 30, 22, 17, 4] for early works in this direction.

While this set up yields some generalisations of random geometric graphs (see Chapter 5 on *random geometric complexes*), problems about recovering the homology of a manifold will only be addressed later in the thesis, in Chapters 6 and 7, as they use fundamentally different approaches from the other consistency problems which we introduce below.

In another setting, geometric graph objects of interest may arise as operators or functionals defined on the graph, e.g., the graph Laplacian, the Cheeger constant and the minimal bisection functional. Analogous operators or functionals can be defined in the continuous regime, and one may ask under which conditions the graph operator or functional approximates its continuous counterpart.

A very popular operator to study is the graph Laplacian, which is known to incorporate valuable information about the graph and is commonly used to perform machine learning tasks such as spectral clustering (using the spectrum of the graph Laplacian, or of a normalised version). As before, one may be interested in showing spectral clustering consistency by establishing pointwise, or better, spectral convergence of the graph Laplacian to the continuous Laplacian on the underlying domain (the Laplace-Beltrami operator if the domain is a Riemannian manifold).

When a given geometric graph quantity is shown to approximate its continuous counterpart under certain parameters constraints, it is said to be *consistent*. A well known necessary condition for consistency results to hold is generally given by the connectivity threshold value of the bandwidth parameter of a random geometric graph r_n, i.e., the distance under which we connect two random points. For bounded domains, more generally on compact Riemannian manifolds, this threshold value is known to be $r_n \sim (\log n)^{1/d} n^{-1/d}$ ([54]). Indeed, it is straightforward to see that many geometric graph quantities of interest will fail to be consistent if the graph is not connected.

The first four chapters of the thesis deal with consitency results on graphs, such as spectral clustering (via spectral convergence of the graph to the continuous Laplacian) ([63, 64, 61]) or Cheeger consistency ([66, 49]). The common point between these results is that in all cases, the quantities of interest (e.g., eigenvalues of the graph Laplacian, Cheeger constant) arise as minimizers of graph functionals of the form

$$\sum_{x \in X_n} \sum_{y \in X_n} \eta_r(x-y)(u(x)-u(y))^\alpha,$$

where $u \in L^\alpha$, $\eta : \mathbb{R}^d \to \mathbb{R}_+$ is a kernel function (rapidly decaying to 0), and $\eta_r(z) := r^{-d}\eta(z/r)$. As such, the problems of spectral clustering, of Cheeger consistency or of the consistency of the minimal bisection functionals, can all be seen as optimisation problems on graphs. A successful method to tackle such problems was first proposed by Trillos and Slepcev in [63]: what they called

the *variational approach*, using tools from the calculus of variations (e.g., Γ-convergence, defined below).

In Chapters $5, 6, 7$, we will change our paradigm and present some results related to *random geometric complexes* (which can be thought as generalisations of random geometric graphs).

In Chapter 2, we recall some of the successful arguments establishing consistency of various geometric graph quantities, using a variational approach as first introduced in [63].

In Chapter 3, we show that many consistency results on random geometric graphs (spectral clustering, Cheeger constants, etc) hold in the case of k-NN graph constructions (which are well known sparse graph constructions, used in practice). In particular we obtain conditions on k as a function of n for spectral convergence of the graph Laplacian sparisifed via a k-NN construction to the continuous Laplacian (see Theorem 3.3.4, to be compared with Theorem 1.2 in [64]).

In Chapter 4, we show how some of the key results of the variational approach extend to $\mathbb{R}^d$ and discuss some of the current limitations. In particular we show some discrepancy-type results (see Theorems 4.6.1 and 4.6.2) and we discuss the extensions some properties of some functionals used in the bounded domain setting on $\mathbb{R}^d$ (Γ-convergence and compactness property) (see Theorem 4.7.1).

In Chapter 5, we recall some basic definitions and concepts for geometric complexes.

In Chapter 6, we show how well-chosen variable bandwidth constructions can allow us to improve some already known results on vanishing of homology for Čech complexes on $\mathbb{R}^d$, what we call *decrackling the noise* (Theorems 6.3.5 and 6.3.7).

In Chapter 7 finally, we study homology of a Čech complex on a compact Riemannian manifold with smooth non-empty boundary. We emphasise that the content of this chapter will be similar to that of the paper [42], a joint work with Ulrike Tillmann and Oliver Vipond (Theorem 7.1.1).

Chapter 2

Consistency results on random geometric graphs: a variational approach

2.1 Introduction and some notation

In this chapter we mention some of the consistency results obtained on bounded domains, following a variational approach (e.g., [63, 64, 62, 49]), which serve as motivations and set up for later results. We note to the reader that this chapter serves the purpose of introducing several important concepts and results which are be useful for later chapters. In particular, this chapter does not contain original results.

We begin with some notation and some definitions, which we shall refer to throughout the thesis. Let $\mathbb{R}_+ := [0, \infty)$ and let $\mathbb{R}_+^* := \mathbb{R}_+ \setminus \{0\}$. Likewise, we define $\mathbb{N}^* := \mathbb{N} \setminus \{0\}$. Let $d \geq 2$ and let $D \subset \mathbb{R}^d$ be a bounded, connected, open set with Lipschitz boundary. For $x \in \mathbb{R}_+^*$, let $[x] := [1, x] \cap \mathbb{N}$. Given a sequence of nonzero real numbers $(a_n)_{n\in\mathbb{N}}$ and of positive reals $(b_n)_{n\in\mathbb{N}}$, we say that

$$a_n = o(b_n)$$

if

$$\lim_{n\to\infty} \frac{a_n}{b_n} = 0;$$

we say that

$$a_n = O(b_n)$$

if

$$\lim_{n\to\infty} \frac{|a_n|}{b_n} < \infty.$$

We say that $a_n = \omega(b_n)$ if

$$\lim_{n\to\infty} \frac{|a_n|}{b_n} = \infty,$$

and that $a_n = \Omega(b_n)$ if

$$\lim_{n\to\infty} \frac{a_n}{b_n} > 0.$$

We say that $a_n = \Theta(b_n)$ if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$.

Definition 1. *We say that a random variable X is Poisson distributed with parameter $\lambda \in \mathbb{R}^*$, and we write $X \sim Po(\lambda)$, if X has discrete law given by*

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \ \forall\ k \in \mathbb{N}.$$

Definition 2. *Let ν be a probability measure and let $(A_n)_{n\in\mathbb{N}}$ be sequence of ν-measurable events. We say that $(A_n)_{n\in\mathbb{N}}$ is true **with high probability**, abbreviated **w.h.p.**, if*

$$\lim_{n\to\infty} \mathbb{P}(A_n) = 1.$$

By a classic abuse of notation, we will say that A_n is true w.h.p., instead of $(A_n)_{n\in\mathbb{N}}$.

Definition 3. *Let ν be a probability measure and let A be a ν-measurable event. We say that A is true **almost surely**, abbreviated **a.s.**, if*

$$\mathbb{P}(A) = 1.$$

Let ν be a probability measure on D with continuous density $q : D \to \mathbb{R}_+$, satisfying

$$0 < q_{\min} \leq q_{\max} < \infty,$$

and let $X_n := \{x_1, \ldots, x_n\}$ be an i.i.d. sample with respect to ν.

Finally, let $\eta : \mathbb{R}^d \to \mathbb{R}_+$ be a kernel function and for $r > 0$, let $\eta_r(h) := r^{-d}\eta(h/r)$. We should think of η as a function rapidly decaying to 0. Typical examples include a compactly supported indicator function or a Gaussian, or more generally a function which has exponential decay. The exact requirements imposed on η are specified below in (2.3).

2.2 Dirichlet energy and the Laplacian: a variational approach to the consistency of spectral clustering

A central and very common task in machine learning is that of clustering, i.e., grouping data points according to some affinities. In the case where the points

are sampled from a distribution ν supported on a domain in $\mathbb{R}^d$, we may define the affinity between two points x and y as $\eta((x-y)/r)$, for some small bandwidth parameter $r > 0$. This yields a (generalised) random geometric graph, considering the points X_n as vertices and $\{\eta((x-y)/r)\}$ as weighted edges. In particular, if the kernel function is given by $\eta((x-y)/r) = \mathbb{1}\left(|x-y| < r\right)$, this is exactly the random geometric graph studied by Penrose in [55], also known as the r-neighbourhood graph.

We may then define the (unnormalised) graph Laplacian as

$$\Delta_{\eta,r} : u \mapsto \left(x \mapsto \sum_{y \in X_n} \eta_r (x-y) (u(x) - u(y)) \right), \ u \in L^2(\nu_n),$$

where $\nu_n = \dfrac{1}{n} \displaystyle\sum_{x \in X_n} \delta_x$ is the empirical measure associated to ν.

Normalised versions of the Laplacian can be defined (the normalised Laplacian and the symmetric Laplacian), after suitably normalising the affinity matrix (obtained by η) with the degree matrix of the graph. Most results we address later generally hold for the various normalisation forms of the Laplacian. For simplicity and brevity, we choose to only state those results for the unnormalised Laplacian.

The graph Laplacian, in particular its spectrum, is known to contain valuable information on the graph. For instance, the multiplicity of its first non-trivial eigenvalue gives the number of connected components of the graph. Many successful algorithms performing clustering on graphs rely on the spectrum of the graph Laplacian. These clustering methods are known as *spectral clustering*. We refer to [67] for an introductory survey on spectral clustering. From a theoretical point of view, one seeks for conditions such that both the eigenvalues (suitably normalised) and eigenvectors of the graph Laplacian converge to those of the continuous Laplacian of the sampling domain, i.e., such that the graph Laplacian converges spectrally to the underlying continuous Laplacian. Suitably normalised here will mean divided by nr^2, but we note that this explicit scaling is only relevant to the specific definition of the graph Laplacian which we have opted for. The problem of convergence of the graph to the continuous Laplacian has already been carefully studied in several papers in the case where the domain is bounded (e.g., [7, 24, 59, 68, 64]).

Spectral analysis of the Laplacian is also closely related to non-linear dimension reduction techniques, such as locally linear embedding (LLE) ([57]) or diffusion maps ([7, 24]). More generally, the topic of nonlinear clustering is a large field which we shall not explore in this thesis.

There has been a significant amount of work done to provide theoretical evidence for the consistency of spectral methods based on the graph Laplacian. To

that end, one is interested in establishing convergence (pointwise, or better spectral) of the graph Laplacian (suitably normalised) to the continuous Laplacian defined on the sampling domain (the Laplace-Beltrami operator if the domain is a Riemannian manifold). The continuous Laplacian, also known as the heat diffusion operator, occupies a key role in many branches of mathematics and physics. It is defined as

$$\Delta : u \longmapsto -\tfrac{1}{q}\mathrm{div}(q^2 \nabla u),$$

We consider the following problem, where $\lambda \in \mathbb{R}$ and $u \in H^1(\nu) \setminus \{0\}$:

$$\begin{cases} \Delta u &= \lambda u, \text{ on } D \\ \dfrac{\partial u}{\partial \mathbf{n}} &= 0, \text{ on } \partial D. \end{cases} \tag{2.1}$$

Here $H^1 := W^{1,2}$ is the Sobolev space of L^2, consisting of L^2 functions whose weak derivative is also contained in L^2. More generally, we define the Sobolev space $W^{m,p}$ as

$$W^{m,p} := \left\{ u \in L^p \mid \forall |\alpha| \leq m,\ D^\alpha u \in L^p \right\}.$$

Suppose that $(\lambda, u) \in \mathbb{R} \times (H^1(\nu) \setminus \{0\})$ is a solution to (2.1), and let $v \in H^1(\nu)$ be a test function. Define also the inner product on $L^2(\nu)$

$$\langle u, v \rangle := \int_D u(x)v(x)q(x)dx,$$

where recall that q is the sampling density of ν, supported on the domain D. Assuming (λ, u) is a solution to (2.1), then $\langle \Delta u, v \rangle = \lambda \langle u, v \rangle$, and we find by integration by parts and using the boundary condition

$$\int_D \nabla u \nabla v q^2(x)dx = \lambda \int_D uvq(x)dx,\ \forall\ v \in H^1(\nu). \tag{2.2}$$

The condition given in (2.2) is the weak formulation of (2.1). Since functions in $H^1(\nu)$ are not necessarily differentiable in the classic sense, we shall say that $(\lambda, u) \in \mathbb{R} \times (H^1(\nu) \setminus \{0\})$ is a weak solution to (2.1) if it satisfies (2.2). In this case we say that λ is an eigenvalue of Δ, with associated eigenfunction u.

Just like its discrete counterpart, the continuous Laplacian contains valuable information on properties of the domain on which it is defined. When the domain is a manifold, its spectrum provides a generalisation of Fourier bases and allows us to perform spectral analysis on the manifold.

In [7, 24, 59], Belkin and Niyogi, Coifman and Lafon, and Singer establish pointwise convergence of the graph to the continuous Laplacian, providing quantitative rates for the value of the bandwidth parameter r as a function of n (seeking optimal rates of pointwise convergence).

In [68] Luxburg et al. first establish spectral convergence of the discrete to the continuous Laplacian. This consists of showing convergence of the eigenvalues and eigenvectors of the graph Laplacian matrices suitably normalised, to those of the underlying continuous Laplacian.

In [64] Trillos and Slepcev successfully establish that spectral convergence occurs provided $nr^d = \omega(\log n)$, $d \geq 3$, and for $nr^2 = \omega((\log n)^{3/2})$ when $d = 2$. This is done via the so-called *variational approach*, first introduced by the same authors in [63], from which they derive the convergence both of the eigenvalues of the graph Laplacian (any version) suitably normalised (divided by nr_n^2), to those of the continuous Laplacian, and of the associated eigenfunctions. The mode of convergence of the eigenfunctions must be specified, since the metric in which they are defined $L^2(\nu_n)$ changes with n (see the definition of the TL^2 topology in Definition 8).

Definition 4.
- *A function $f : \mathbb{R}^d \to \mathbb{R}$ is called **radial** (or isotropic, or symmetric), if*

$$\forall (x, y) \in \left(\mathbb{R}^d\right)^2,\ |x| = |y| \Rightarrow f(x) = f(y).$$

- *Given a radial function $f : \mathbb{R}^d \to \mathbb{R}$, define its radial profile to be the function $\boldsymbol{f} : \mathbb{R}_+ \to \mathbb{R}$ satisfying*

$$\forall x \in \mathbb{R}^d,\ \boldsymbol{f}(|x|) = f(x).$$

Suppose that the kernel η is radial and let $\boldsymbol{\eta} : \mathbb{R}_+ \to \mathbb{R}_+$ be its radial profile. Furthermore, assume the following conditions for the kernel.

$$\begin{cases} \text{a) } \boldsymbol{\eta}(0) > 0 \text{ and } \boldsymbol{\eta} \text{ is continuous on } [0, r_0] \text{ for some } r_0 > 0, \\ \text{b) } \boldsymbol{\eta} \text{ is non-increasing,} \\ \text{c) } \sigma_\eta := \displaystyle\int_0^\infty \boldsymbol{\eta}(x)|x_1|^2 dx < \infty. \end{cases} \tag{2.3}$$

In [64, 61], the authors propose a variational approach to establish the consistency of spectral clustering. To do so, they investigate (discrete or continuous) functionals related to the discrete or continuous Laplacian. Define the continuous *Dirichlet energy* as

$$G : u \longmapsto \langle u, \Delta u \rangle,\ u \in H^1(D, \nu), \tag{2.4}$$

and

$$G(u) := \infty \text{ if } u \in L^2(\nu) \setminus H^1(\nu).$$

Using integration by parts and the boundary condition in (2.1), we can rewrite the Dirichlet energy of $u \in H^1(\nu)$ as

$$\begin{aligned} \langle u, \Delta u \rangle &= -\int_D u(x) \tfrac{1}{q(x)} \mathrm{div}(q^2(x) \nabla u(x)) q(x) dx \\ &= \int_D \left|\nabla u(x)\right|^2 q^2(x) dx. \end{aligned}$$

The Dirichlet energy is an important functional in physics. It is to be compared with the total variation functional, defined below. In some sense, the Dirichlet energy is an L^2 version of the total variation. While the total variation measures the smoothness of an L^1 function by summing over its variations (i.e., looking at an L^1 norm of the gradient when the function is differentiable), the Dirichlet energy measures the smoothness of an L^2 function by summing over the square of its variations (i.e., looking at an L^2 norm of the gradient when the function is differentiable).

The eigenvalues of Δ satisfy

$$0 = \lambda_0 \leq \lambda_1 \leq \lambda_2 \ldots,$$

and there exists an orthonormal basis $(u_k)_{k\in\mathbb{N}}$ of $L^2(\nu)$ such that for each $k \in \mathbb{N}$, u_k is an eigenfunction of Δ associated with λ_k. We refer for instance, to [67] for a nice introduction to graph Laplacian matrices, their spectrum and other properties, and their applications to spectral clustering.

The eigenvalues of the Laplacian can be seen as minimal values attained by the the Dirichlet energy restricted to some subspaces of $L^2(\nu)$.

In fact for each $k \in \mathbb{N}^*$, we have by the Courant minimax principle (e.g., see Proposition 2.15 in [64])

$$\lambda_k = \min_{||u||=1,\ u\in S_k^\perp} G(u) = \sup_{S\in\sum_{k-1}} \min_{||u||=1,\ u\in S^\perp} G(u),$$

where $S_k = span\{u_i \mid i \in [k-1]\}$ and $\sum_{k-1}$ denotes the set of $(k-1)$-dimensional subspaces of $L^2(\nu)$.

Definition 5. *Similary, define the discrete Dirichlet energy functional as*

$$G_{\eta,r} : u \mapsto \frac{1}{n^2 r^2} \sum_{y\in X_n} \sum_{x\in X_n} \eta_r(x-y)\big(u(x)-u(y)\big)^2.$$

The same relation between the spectrum of Δ and the Dirichlet energy exists in the discrete setting. Thus, a natural first step towards showing spectral convergence of the graph Laplacian to Δ, consists in showing some type of convergence of the discrete to the continuous Dirichlet energy, in order to imply convergence of the eigenvalues (suitably normalised). In [64] the authors establish Γ-convergence of those functionals, characteristic of the so-called variational approach (cf, [63, 62]).

Definition 6 (Γ-convergence)**.** *Let X be a topological space and consider a sequence of functionals $(F_n)_{n\in\mathbb{N}}$ on X, i.e., $\forall n \in \mathbb{N}$, $F_n : X \to \mathbb{R}_+$. We say that $(F_n)_{n\in\mathbb{N}}$ Γ-converges to $F : X \to \mathbb{R}_+$, and write $\Gamma\text{-}\lim_{n\in\mathbb{N}} F_n = F$, if the*

following liminf and limsup properties hold true.
For every sequence $(x_n)_{n\in\mathbb{N}}$ converging in X to $x \in X$,

$$F(x) \leq \liminf_{n\in\mathbb{N}} F_n(x_n).$$

For every $x \in X$, there exists a sequence $(x_n)_{n\in\mathbb{N}}$ converging in X to x and such that

$$\limsup_{n\in\mathbb{N}} F_n(x_n) \leq F(x).$$

To show Γ-convergence of the discrete to the continuous Dirichlet energy, the authors of [64] define a distance allowing them to compare functions in $L^2(\nu_n)$ with functions in $L^2(\nu)$, where ν_n is the empirical measure associated to ν. More generally, how can one compare functions in $L^2(\mu)$ and functions in $L^2(\theta)$, for two different probability measures μ and θ on D? To address this issue, the authors in [64] choose to work under the TL^2 topology, which we now define.

Definition 7. *Given a probability measure $\mu \in \mathcal{P}(X)$ and a probability measure $\theta \in \mathcal{P}(Y)$, define the set of transportation plans bewtween μ and θ to be*

$$\Gamma(\mu,\theta) := \left\{\gamma \in \mathcal{P}(X\times Y) \mid \gamma(.,Y) = \mu \text{ and } \gamma(X,.) = \theta\right\}.$$

Here $\mathcal{P}(X)$ denotes the set of Borel probability measures on a given metric space X.

Definition 8. *Let*

$$TL^2(D) := \{(\mu, f) \mid \mu \in \mathcal{P}(D),\ f \in L^2(\mu)\}.$$

The space $TL^2(D)$ defined above can be endowed with the following distance

$$d_{TL^2}((\mu,f),(\theta,g)) := \inf_{\pi\in\Gamma(\mu,\theta)} \left(\int_D\int_D |x-y|^2 + \left|f(x)-f(y)\right|^2 d\pi(x,y)\right)^{1/2},$$

where $\Gamma(\mu,\theta)$ denotes the set of transportation plans between the measures μ and θ.

With the above definition, the authors in [64] show (see Theorem 4.1 in [64]) Γ-convergence of the discrete Dirichlet energy functional to its continuous version under the TL^2 topology (with suitable normalising factors), provided the bandwidth parameter r_n satisfies

$$r_n = \begin{cases} \omega((\log n)^{1/d} n^{-1/d}), \text{ if } d \geq 3, \\ \omega((\log n)^{3/4} n^{-1/2}), \text{ if } d = 2. \end{cases} \tag{2.5}$$

The above dichotomy between the cases where $d = 2$ and $d \geq 3$ is recurring in various variational approaches as introduced by Trillos and Slepcev in [63]. It comes from the use of a previous work by the same authors on the concentration of empirical measures under the ∞-Wasserstein distance (see Theorem 1.1 in [62]). They also show some compactness property relative to the Discrete energy under the TL^2 topology. Namely, we have the following result from [64].

Theorem 2.2.1 (Theorem 1.4 in [64]). *Suppose that η sastisfies conditions (2.3) and that r_n satisfies conditions (2.5). Then the graph Dirichlet energies $(G_{\eta,r_n})_{n\in\mathbb{N}}$ Γ-converge to $\sigma_\eta G$ in the TL^2 sense, where σ_η is defined in conditions (2.3) and G is the continuous Dirichlet energy.*

Furthermore, $(G_{\eta,r_n})_{n\in\mathbb{N}}$ satisfies the following compactness property in TL^2. Every sequence $((\nu_n, u_n))_{n\in\mathbb{N}}$ in TL^2 satisfying

$$\sup_{n\in\mathbb{N}} ||u_n||_{L^2(\nu_n)} + G_{\eta,r_n}(u_n) < \infty,$$

is precompact in TL^2.

To show the above, they define an intermediate functional which can be seen as the expectation of the discrete Dirichlet energy, and for which Γ-convergence to $\sigma_\eta G$ and a similar compactness property can be shown to hold under the classic L^2 norm. We do not give too many details about this intermediate functional here. A similar set up arises in the case of the total variation functional, an L^1 version of the Dirichlet energy, where we give more details (see next section).

Essentially, the Γ-convergence of the discrete Dirichlet energy is what allows us to show convergence of the eigenvalues of the discrete Laplacian (suitably normalised) to the eigenvalues of Δ, seeing each eigenvalue as a minimizer of the Dirichlet energy (as discussed above via the Courant minimax principle). Given convergence of the eigenvalues, the compactness property then allows us to deduce convergence of the eigenvectors under the TL^2 metric (up to a subsequence). Once the convergence of the eigenvalues and of the eigenvectors have been established, the authors in [62] are able to deduce some result on the consistency of some spectral clustering algorithms. For a full statement of the main results obtained in [62], see Theorem 1.2 in [62] for the unnormalised graph Laplacian, which follows as a consequence of Theorem 2.2.1, and see Theorem 1.5 in [62] for an equivalent result for the normalised graph Laplacian.

2.3 Total variation: consistency of the Cheeger constant and the minimal bisection functional

We have just discussed a variational approach to address the problem of spectral clustering. This approach consists in studying Γ-convergence (defined above) and some compactness property of the graph (or discrete) Dirichlet energy. More generally, the same can be done for graph functionals of the form

$$\sum_{x\in X_n}\sum_{y\in X_n} \eta_r(x-y)(u(x)-u(y))^\alpha.$$

The success of the variational approach to tackle machine learning problems, as first proposed by Trillos and Slepcev in [63], comes from the fact that several

problems in machine learning can be formulated as an optimisation problem on such graph functionals. For instance, as discussed in the previous section, the eigenvalues of the graph Laplacian can be seen as minimizers of the graph Dirichlet energy via the Courant minimax principle. We now present other common graph quantities of interest arising in a similar way, and see how their consistency can similarly be established by a variational approach. Namely, we present some consistency results on the Cheeger constant and on the minimal bisection functional. The results we present were first established by Trillos and Slepcev in [63], in [66] for Cheeger consistency, and more recently improved by Müller and Penrose in [49]. They can be seen as L^1 versions of the results mentioned in the previous section. The functional of interest, the total variation, can be thought as an L^1 version of the Dirichlet energy.

The Cheeger constant and the minimal bisection functional are popular graph quantities arising in this way. The continuous Cheeger constant is defined as the minimum perimeter to volume ratio over all subsets with induced volume at most $1/2$. On a graph, one may define the Cheeger constant analogously, after defining analogous versions of the perimeter of a subgraph (the graph cut) and of its volume. The Cheeger constant is named after Cheeger's inequality, where it appears in the lower bound term for the smallest positive eigenvalue of the Laplacian of a compact Riemannian manifold, proved by Jeff Cheeger in [23].

In both the discrete and the continuous setting, the Cheeger constant can be used to provide eigengaps on the (graph or the continuous) Laplacian. As such, Cheeger constants are important quantities, relevant in spectral clustering methods on graphs, and the question of whether one can approximate well the Cheeger constant of an underlying manifold from the Cheeger constant of a graph (i.e., Cheeger consistency) is thus of interest from the point of view of machine learning, and has already been investigated in the case of a bounded domain $D \subset \mathbb{R}^d$ in [66, 49]. In [49], the authors show how their results can be applied to a broader class of graph functionals and problems. In particular they also provide consistency results for the minimal bisection functional of a graph, which can be seen as a special case of the Cheeger constant.

Assume that η satisfies the following properties (e.g., [63]).

$$\begin{cases} \text{a) } \boldsymbol{\eta}(0) > 0 \text{ and } \boldsymbol{\eta} \text{ is continuous on } [0, r_0] \text{ for some } r_0 > 0, \\ \text{b) } \boldsymbol{\eta} \text{ is non-increasing,} \\ \text{c) } \sigma_\eta := \displaystyle\int_0^\infty \boldsymbol{\eta}(x)|x_1|dx < \infty, \end{cases} \tag{2.6}$$

where x_1 above denotes the first coordinate of x, and $\boldsymbol{\eta}$ is, as before, the radial profile of η.

The following definitions can all be found in [49].

Definition 9. *Given $u \in L^1(\nu)$, consider the functional*

$$\eta_r(u, X_n) := \sum_{y \in X_n} \sum_{x \in X_n} \eta_r(x, y) \big| u(x) - u(y) \big| ,$$

and define its normalised version

$$G\eta_r(u, X_n) := \frac{2}{n(n-1)r} \eta_r(u, X_n).$$

A special case of interest to us is given when we take $u = \mathbb{1}_A$, $A \subset D$. In this case, define the graph cut of A by

$$(\mathrm{Cut}_{\eta,r}(A) =)\ \mathrm{Cut}_{\eta,r}(A, X_n) := \eta_r(\mathbb{1}_A, X_n) = \sum_{y \in Y} \sum_{x \in X_n \setminus Y} \eta_r(x, y);$$

where $Y := A \cap X_n$.

Definition 10. *The discrete minimal bisection functional is defined as follows.*

$$\mathrm{MBIS}(G(X_n, r_n)) := \min \left\{ \mathrm{Cut}_{\eta,r}(Y_n) | Y_n \subset X_n,\ |Y_n| = \lfloor n/2 \rfloor \right\}.$$

Definition 11. *The balance term of $Y \subset X_n$ is defined as*

$$\mathrm{Bal}(Y, X_n) := \frac{\min\{|Y|, |X_n \setminus Y|\}}{n}.$$

Definition 12. *Having defined the graph cut in Definition 9, and the balance term in Definition 11, we may now define the graph Cheeger constant as*

$$\mathrm{CHE}(G(X_n, r_n)) := \min \left\{ \frac{\mathrm{Cut}_{\eta,r}(Y)}{\mathrm{Bal}(Y, X_n)} \mid Y \subset X_n,\ Y \notin \{\emptyset, X_n\} \right\}.$$

Definition 13. *Recall also the definition of the total variation of $u \in L^1(D, q)$:*

$$TV(u; q) := \sup \left\{ \int_D u(x) \mathrm{div}(\phi)(x) dx \mid \phi \in C_c^1(D; \mathbb{R}^d), \forall x \in D, |\phi(x)| \leq q^2(x) \right\}.$$

Let ν be the (probability) measure associated to the sampling density q:

$$\nu(A) := \int_A q(x) dx,\ A \in \mathcal{B}(D).$$

Definition 14. *Define the minimal bisection functional on D with respect to q as*

$$\mathrm{MBIS}(D, q) := \inf\{TV(\mathbb{1}_A; q) | A \in \mathcal{B}(D), \nu(A) = 1/2\}.$$

Definition 15. *As in the discrete case, define for $A \in \mathcal{B}(D)$*

$$\mathrm{Bal}(A) := \min\{\nu(A), \nu(A^c)\}.$$

Definition 16. *Likewise, define the Cheeer constant of D with respect to q as*

$$\mathrm{CHE}(D,q) := \inf\left\{ \frac{TV(\mathbb{1}_A;q)}{\mathrm{Bal}(A)} \mid A \in \mathcal{B}(D),\ \nu(A) \in (0,1) \right\}.$$

In [49] the authors show in particular the following theorems on the consistency of the Cheeger constant and the minimal bisection functionals.

Theorem 2.3.1 (Theorem 2.1 in [49])**.** *Let $d \geq 2$ and let $D \subset \mathbb{R}^d$ be nonempty, open, bounded, connected with Lipschitz boundary. Let $q : D \to \mathbb{R}_+$ be a continuous probability density with $q_{\max} < \infty$ and $q_{\min} > 0$. For every $(r_n)_{n\in\mathbb{N}}$ such that $r_n = o(1)$ and $nr_n^d = \omega(\log n)$, we have a.s.*

$$\lim_{n\to\infty} \left(\frac{\mathrm{CHE}(G(X_n, r_n), q)}{n^2 r_n} \right) = (\sigma_\eta/2)\mathrm{CHE}(D,q).$$

Theorem 2.3.2 (Theorem 2.4 in [49])**.** *Under the same assumptions as above on D, q and r_n, we have a.s.*

$$\lim_{n\to\infty} \left(\frac{\mathrm{MBIS}(G(X_n, r_n), q)}{n^2 r_n} \right) = (\sigma_\eta/2)\mathrm{MBIS}(D,q).$$

One of the novelties of the above results is that that the argument does not use the ∞-Wasserstein distance as in [63, 64, 61], and as such attains consistency results for r_n barely faster than the connectivity threshold value of $(\log n)^{1/d}(n)^{-1/d}$ for every $d \geq 2$ (thus avoiding a dichotomy between the cases $d = 2$ and $d \geq 3$ inherent to the use of the ∞-Wasserstein distance).

2.4 Conclusion

In this chapter, we presented various consistency results of random geometric graph quantities and functionals. We saw that all of these problems arise as optimisation problems on graph functionals of the form

$$\sum_{x\in X_n} \sum_{y\in X_n} \eta_r(x-y)(u(x)-u(y))^\alpha.$$

For instance, in the case of spectral clustering, the Dirichlet energy is the funtional above with $\alpha = 2$ and $u \in L^2(\nu_n)$, where ν_n is the empirical measure associated to the underlying measure ν. Eigenvalues of the graph Laplacian are given by minimizers of this functional, hence the problem of spectral clustering, which is that of establishing spectral convergence of the graph to the continuous Laplacian, can indeed be seen as an optimisation problem on functionals as above. Likewise, taking $\alpha = 1$ we can study the problem of Cheeger consistency or that of the minimal bisection functional.

To study convergence properties of such functionals, we presented some tools (e.g., Γ-convergence) drawn from the calculus of variations. This approach to consistency problems on graphs was first investigated by Trillos and Slepcev in [63].

Chapter 3

Some consistency results on k-NN graphs

As we have seen in Chapter 2, there exists a plethora of consistency results on graphs, with various objectives and applications in machine learning. In practice however, one is generally interested in building sparse graphs (with few edges) for computational reasons. Two typical constructions yield sparse geometric graph representations.

One of them, called the r-neighbourhood graph, consists in imposing a compact support on the kernel function η inducing the weighted edges of the graph. In this case, the above kernel conditions (2.3) or (2.6) are still satisfied by the compactly supported kernel, hence the classic consistency results mentioned in the previous chapter still hold for the r-neighbourhood graph.

The second classic construction yielding a sparse geometric graph representation, widely used in practice, is known as the k nearest neighbours graph (k-NN). Two constructions exist. One consists in connecting each node only to its k nearest neighbours. An alternative construction, known as the mutually k-NN graph, consists in only connecting vertices which are each among the k nearest neighbours of each other. We address both constructions, as we will see that they are both dealt with in the same way in the arguments below.

Let $X_n := \{x_1, \dots, x_n\} \subset D$ be an i.i.d. sample with respect to some probability measure ν, supported on a bounded, open, connected and Lipschitz domain $D \subset \mathbb{R}^d$. Suppose that ν is absolutely continuuous with respect to the Lebesgue measure, denote by q the sampling density associated to ν, and suppose that

$$0 < q_{\min} \leq q_{\max} < \infty, \tag{3.1}$$

where

$$q_{\min} := \inf\{q(x) \mid x \in D\},$$

and

$$q_{\max} := \sup\{q(x) \mid x \in D\}.$$

In this chapter, we show how basic results comparing the regularity of the empirical measure with respect to the underlying measure can provide us with ways to infer some consistency results on a k-NN graph from the better known consistency results on r-neighbourhood graphs.

In Section 3.2 we show how one may easily use some concentration results on bounded domains in order to deduce connectivity results for k-NN graphs from better known connectivity results of r-neighbourhood graphs. Sharp connectivity threshold values are already known for k-NN graphs (see [6]).

In Section 3.3 we extend this idea and show how it can also be used successfully to convert various consistency results on r-neighbourhood graphs to k-NN graphs.

The main results of this Section are Theorems 3.3.4 and 3.3.5, providing conditions for the consistency results to hold on k-NN constructions (spectral clustering, Cheeger consistency).

3.1 Preliminaries

We first present a few results on the concentration of empirical measures, which are helpful for our later derivations. One such result, given in [62], is on the concentration of empirical measures under the ∞-Wasserstein distance.

Theorem 3.1.1 ([62])**.** *Suppose that ν has sampling density q satisfying (3.1). With high probability, the following holds. If the dimension $d = 2$, then*

$$W_\infty(\nu_n, \nu) \leq C\frac{(\log n)^{3/4}}{n^{1/2}};$$

if $d \geq 3$, then

$$W_\infty(\nu_n, \nu) \leq C\frac{(\log n)^{1/d}}{n^{1/d}},$$

where $C > 0$ depends on D and on the density q.

Another interesting concentration result, which in fact can be used to deduce the above theorem, is the discrepancy-type result in Lemma 3.2 of [49], which we mention below - see Lemma 4.2.1. Let us now derive similar discrepancy-type results to this lemma, which we will see to be useful to derive some consistency results for k-NN graphs.

Assume that

$$r_n = \omega((\log n)^{1/d} n^{-1/d})$$

and that $\gamma_n = o(1)$ is sufficiently slowly decaying such that

$$nr_n^d \gamma_n^{d+2} = \omega(\log n).$$

Assume furthermore that $\gamma_n = \omega((\log n)^{-1})$. For each $i \in [n+1]$, let

$$X_n^{(i)} := X_n \setminus \{x_i\}.$$

Adjusting the proof of the discrepancy-type result of Lemma 3.2 in [49], we have the following.

Lemma 3.1.2. *There exists an almost surely finite random variable $n_0 \in \mathbb{N}^*$, such that for all $n \geq n_0$ and all $i \in [n]$*

$$(n-1)(1-\gamma_n)\nu(B(x_i,r)) \leq \left|X_n^{(i)} \cap B(x_i,r)\right| \leq (n-1)(1+\gamma_n)\nu(B(x_i,r)).$$

From this, using the assumptions on q and r which imply that

$$n(\nu(B(x_i,r))) = \omega(\log n),$$

we easily deduce that for all $n \geq n_0$ and all $i \in [n]$

$$\begin{aligned}\left|X_n \cap B(x_i,r)\right| &= 1 + \left|X_n^{(i)} \cap B(x_i,r)\right| \\ &\leq n\nu(B(x_i,r))(1+\gamma_n + o((\log n)^{-1})) \\ &\leq n\nu(B(x_i,r))(1+\gamma_n(1+o(1))).\end{aligned}$$

The same can be done for the lower bound, so that we have the following estimations for the number of sampled points in each of the random balls.

Corollary 3.1.3. *For all $n \geq n_0$ and all $i \in [n]$*

$$\left|X_n \cap B(x_i,r)\right| \geq (1-\gamma_n(1+o(1)))n\nu(B(x_i,r))$$

and

$$\left|X_n \cap B(x_i,r)\right| \leq (1+\gamma_n(1+o(1)))n\nu(B(x_i,r)).$$

A similar result can be deduced if we take merely r such that

$$(n\omega_d r^d)(\log n)^{-1} \geq K + o(1),$$

for some constant $K > 0$ and keep $\gamma_n = \gamma > \sqrt{\frac{1}{K}}$. Then, by the same proof than the one of Lemma 3.2 in [49], one may obtain

Lemma 3.1.4. *For all $x \in X_n$,*

$$\left|X_n \cap B(x,r)\right| \geq (1-\gamma)n\nu(B(x,r))$$

and

$$\left|X_n \cap B(x,r)\right| \leq (1+\gamma)n\nu(B(x,r)).$$

The above results, i.e., Theorem 3.1.1, Corollary 3.1.3 and Lemma 3.1.4, are results on the concentration of the empirical measures. In particular, they provide us with a way to control the number of sampled points in any of the random balls very precisely in terms of the measure of the balls themselves. This elementary observation yields a natural way to infer consistency results on k-NN graphs from better known consistency results on r-neighbourhood graphs.

As a first example, let us see how the above results immediately give us some connectivity threshold values for k.

3.2 Connectivity of the k-NN graph

It is already known that the k-NN graph is connected with high probability, provided $k = \Omega(\log n)$ (e.g., [15] for the mutual k-NN graph). In fact it was shown in [69] that connectivity occurs w.h.p. for $k = \Theta(\log n)$. This was refined in [5] and later in [6], where the authors eventually establish a sharp connectivity threshold value for k as a function of $\log n$, such that beyond this value the graph is connected w.h.p., while below this same threshold value, the graph is not connected w.h.p..

This sharp transition is reminiscent of the well known sharp transition threshold probability value for the connectivity of Erdős-Rényi random graphs (although the setting and the techniques are different). Such a phenomenon also occurs for the r-neighbourhood graph. Indeed, Penrose proved in [54] that the smallest r such that the r-neighbourhood graph is connected satisfies

$$\lim_{n\to\infty} (n\omega_d r^d)(\log n)^{-1} = K \text{ a.s.}, \tag{3.2}$$

where $K := \max\{q_{\min}^{-1}, 2(1-1/d)((q|_{\partial D})_{\min})^{-1}\}$ and ω_d is the volume of the unit ball in $\mathbb{R}^d$. Here ∂D denotes the boundary of D and

$$(q|_{\partial D})_{\min} := \inf\{q(x) \mid x \in \partial D\} > 0.$$

Likewise, random geometric complexes (generalisation of random geometric graphs, cf, Chapter 5) exhibit similar sharp transition phenomena.

In this subsection we do not claim to obtain a sharp transition value for k as precise as the one obtained in [6]. The results presented below, on the connectivity of k-NN graphs, are to be taken as simple illustrations of how one can easily derive consistency results for k-NN constructions from better known consistency results on random geometric graphs.

This can be done in our case, with either Theorem 3.1.1 or Lemma 3.1.4.

Theorem 3.2.1 (k-NN connectivity from Theorem 3.1.1)**.** *Let $d \geq 3$. Suppose that*

$$k \geq \left(q_{max} \left(C + (K + o(1))\omega_d^{-1/d} \right)^d \right) \log n,$$

where the constants C and K are given as above (cf, [54] and Theorem 3.1.1), then w.h.p. the k-NN graph (both constructions mentioned above) is connected.

In fact this could be made into an a.s. statement by Borel-Cantelli. Indeed, referring Theorem 1.1 in [62], we see that the probabilities are summable over n.

Proof. First, let us note that the ∞-Wasserstein distance can be reformulated as (cf, [36])

$$W_\infty(\nu, \nu_n) := \inf\{r > 0 | \forall A \in \mathcal{B}(D),\ \nu(A^r) \geq \nu_n(A)\},$$

where

$$A^r := \{x \in D \mid dist(x, A) < r\}.$$

Let r satisfy the above connectivity condition of Penrose given in (3.2), and let $r_\infty := W_\infty(\nu_n, \nu)$. Using the above formulation for W_∞, we have the following bounds. For every $i \in [n]$

$$\begin{aligned} n\nu_n(B(x_i, r)) &\leq n\nu(B(x_i, r + r_\infty)) \\ &\leq nq_{\max}\omega_d(r + r_\infty)^d. \end{aligned}$$

We know by Theorem 3.1.1, that w.h.p.

$$r_\infty \leq C \left(\frac{\log n}{n} \right)^{1/d},$$

and

$$r \leq (K + o(1))\omega_d^{-1/d} \left(\frac{\log n}{n} \right)^{1/d},$$

so w.h.p.

$$r + r_\infty \leq \left(\frac{\log n}{n} \right)^{1/d} (C + (K + o(1))\omega_d^{-1/d}),$$

and for all $x \in X_n$

$$n\nu_n(B(x, r)) \leq \left(q_{max} \left(C + (K + o(1))\omega_d^{-1/d} \right)^d \right) \log n.$$

Choose k greater than (or equal to) the RHS above. Then, the above indicates that

$$X_n \cap B(x, r) \subset N_k(x),$$

where $N_k(x)$ denotes the set of k nearest neighbours of x, together with x itself; hence that for all $x, y \in X_n$

$$\begin{aligned}\mathbb{1}\left(|x-y| < r\right) &\leq \mathbb{1}\left(y \in N_k(x)\right)\mathbb{1}\left(x \in N_k(y)\right)\\ &\leq \mathbb{1}\left(y \in N_k(x) \text{ or } x \in N_k(y)\right).\end{aligned}$$

In other words, we know that w.h.p. the k-NN graph (either construction) thus constructed contains as a subgraph the r-neighbourhood graph, with r satisfying the connectivity condition in (3.2). In particular, this means that the k-NN graph is connected w.h.p. for such values of k.

□

By a very similar argument to above, we can use Lemma 3.1.4 instead and obtain the following, which is sharper than the above.

Theorem 3.2.2 (k-NN connectivity from Lemma 3.1.4)**.** *Let $d \geq 2$ and suppose that*

$$k \geq q_{\max}\left((K + o(1) + \sqrt{\frac{1}{K}}\right)\omega_d^{-1/d}\log n,$$

then w.h.p. the k-NN graphs are connected.

Note that in fact, using Lemma 3.1.4 we could similarly derive a lower threshold value for k.

The above is just to illustrate with a simple example, how one may use the concentration results mentioned in the previous section to derive consistency results for k-NN graphs from better known consistency results for random geometric graphs. While these results are not as sharp as those obtained in [6], their proofs are elementary.

Next, let us see how Corollary 3.1.3 can be used similarly to derive other consistency results on k-NN graphs. As far as we know, the consistency results presented below have not been studied before. In particular, we derive asymptotic conditions on k to guarantee spectral clustering consistency with a Laplacian sparsified by a k-NN construction. We note the work of Maier, Hein and Luxburg in [45] on choices of k for k-NN clustering, but emphasise that our results below are set in a different paradigm. First of all, we are concerned with spectral clustering, i.e., via the spectrum of a Laplacian (in particular, we do not perform the same construction nor do we follow the same clustering algorithm). Secondly, the authors in [45] have a different definition of clusters and different assumptions on the sampling density. Consequently, they are set to answer a different question to the one addressed below, and indeed find different answers.

3.3 Other consistency results on the k-NN graph

As seen in the previous chapter, various quantities of interest on geometric graphs (e.g. graph cut, Dirichlet energy) arise as functionals of the form

$$\sum_{x\in X_n}\sum_{y\in X_n}\eta_r(x-y)(u(x)-u(y))^\alpha.$$

Analogous quantities can be defined for k-NN graphs in a similar fashion. We present them in the following list of definitions.

Definition 17 (k-NN Dirichlet energies)**.**

$$G^{(1)}_{\eta,r}:u\mapsto\frac{1}{n^2r^2}\sum_{x\in X_n}\sum_{y\in X_n}\eta_r(x-y)\mathbb{1}\left(x\in N_k(y)\right)\mathbb{1}\left(y\in N_k(x)\right)(u(x)-u(y))^2$$

$$G^{(2)}_{\eta,r}:u\mapsto\frac{1}{n^2r^2}\sum_{x\in X_n}\sum_{y\in X_n}\eta_r(x-y)\mathbb{1}\left(x\in N_k(y)\ \textit{or}\ y\in N_k(x)\right)(u(x)-u(y))^2$$

Definition 18 (k-NN (unnormalised) graph Laplacians)**.**

$$\Delta^{(1)}_n:u\mapsto\left(x\mapsto\sum_{y\in X_n}\eta_r(x-y)\mathbb{1}\left(x\in N_k(y)\right)\mathbb{1}\left(y\in N_k(x)\right)(u(x)-u(y))\right)$$

$$\Delta^{(2)}_n:u\mapsto\left(x\mapsto\sum_{y\in X_n}\eta_r(x-y)\mathbb{1}\left(x\in N_k(y)\ \textit{or}\ y\in N_k(x)\right)(u(x)-u(y))\right)$$

Definition 19 (k-NN graph cuts)**.**

$$\mathrm{Cut}^{(1)}_{\eta,r}:A\mapsto\sum_{x\in X_n\setminus Y}\sum_{y\in Y}\eta_r(x-y)\mathbb{1}\left(x\in N_k(y)\right)\mathbb{1}\left(y\in N_k(x)\right)$$

$$\mathrm{Cut}^{(2)}_{\eta,r}:A\mapsto\sum_{x\in X_n\setminus Y}\sum_{y\in Y}\eta_r(x-y)\mathbb{1}\left(x\in N_k(y)\ \textit{or}\ y\in N_k(x)\right),$$

where $Y:=A\cap X_n$.

In the above definitions, the graph Laplacians and Dirichlet energies are defined in two different ways. These two definitions correspond to the two different ways one could construct the k-NN graph, where the edge set is either given by

$$[\mathbb{1}(x\in N_k(y))\mathbb{1}(y\in N_k(x))]_{x,y},$$

or by

$$[\mathbb{1}(x\in N_k(y)\text{ or }y\in N_k(x))]_{x,y}.$$

Definition 20 (k-NN graph Cheeger constants and minimal bisection functionals)**.**

$$\mathrm{CHE}(k\text{-}NN^{(i)}) := \min\left\{ \frac{\mathrm{Cut}^{(i)}_{\eta,r}(Y)}{\mathrm{Bal}(Y, X_n)} \mid Y \subset X_n,\ Y \notin \{\emptyset, X_n\} \right\},\ i \in [2]$$

$$\mathrm{MBIS}(k\text{-}NN^{(i)}) := \min\left\{ \mathrm{Cut}^{(i)}_{\eta,r}(Y_n) | Y_n \subset X_n,\ |Y_n| = \lfloor n/2 \rfloor \right\},\ i \in [2]$$

From now on, assume that ν is the uniform distribution.

Let r be such that $nr^d = \omega(\log n)$ and let $\gamma_n = o(1)$ be as above, so slow that $\gamma_n = \omega((\log n)^{-1})$ and

$$nr^d\gamma_n^{d+2} = \omega(\log n).$$

Finally let $C > 0$ and let $k \in [Cn\omega_d r^d, Cn\omega_d r^d + 1] \cap \mathbb{N}$.

Lemma 3.3.1. *Let $r_1 := r(1 + 2\gamma_n)^{-1/d}$ and $r_2 := r(1 - \frac{1}{2}\gamma_n)^{-1/d}$. For all $n \geq n_0$ and all $x \in X_n$*

$$n\nu_n(B(x, C^{1/d}r_1)) \leq k \leq n\nu_n(B(x, C^{1/d}r_2));$$

from which it follows that for all $x, y \in X_n$

$$\begin{aligned}
\mathbb{1}\left(|x - y| < C^{1/d}r_1\right) &\leq \mathbb{1}\left(x \in N_k(y)\right)\mathbb{1}\left(y \in N_k(x)\right) \\
&\leq \mathbb{1}\left(y \in N_k(x) \ \text{ or } x \in N_k(y)\right) \\
&\leq \mathbb{1}\left(|x - y| < C^{1/d}r_2\right).
\end{aligned}$$

Proof. Using Corollary 3.1.3, the fact that ν is the uniform measure, and the assumptions on r which imply that

$$n\nu(B(x, r)) = \omega(\log n),$$

we have for all $x \in X_n$

$$\begin{aligned}
n\nu_n(B(x, C^{1/d}r_1)) &\leq n\nu(B(x, C^{1/d}r)) \\
&= Cn\omega_d r^d \\
&\leq k \\
&\leq Cn\omega_d r^d + 1 \\
&= n\nu(B(x, C^{1/d}r))\left(1 + o\left((\log n)^{-1}\right)\right) \\
&\leq n\nu(B(x, C^{1/d}r_2))(1 - \gamma_n(1 + o(1)))(1 + o(\gamma_n)) \\
&\leq n\nu_n(B(x, C^{1/d}r_2)).
\end{aligned}$$

Hence for each $x \in X_n$, all the points of X_n contained in $B(x, C^{1/d}r_1)$ are among the k nearest neighbours of x in X_n, which are themselves contained in $B(x, C^{1/d}r_2)$. In other words, we have

$$X_n \cap B(x, C^{1/d}r_1) \subset N_k(x) \subset X_n \cap B(x, C^{1/d}r_2).$$

Now suppose that $x, y \in X_n$ satisfy

$$\mathbb{1}\left(|x-y| < C^{1/d}r_1\right) = 1.$$

By the above observation, it follows that $x \in N_k(y) \subset B(y, C^{1/d}r_2)$ and $y \in N_k(x) \subset B(x, C^{1/d}r_2)$, which gives the second claim of the lemma.

□

From this lemma, we have the following immediate corollary.

Corollary 3.3.2. *Let η be a kernel function satisfying the conditions (2.3) or (2.6) mentioned in Chapter 2. Let $\alpha \in \mathbb{N}$ and let $u \in L^\alpha(D)$. As above, let r be such that*

$$nr^d = \omega(\log n)$$

and suppose that

$$k \in [Cn\omega_d r^d, Cn\omega_d r^d + 1] \cap \mathbb{N},$$

then we have

$$\begin{aligned}
&\frac{r_1^d}{r^d} \sum_{x\in X_n} \sum_{y\in X_n} \eta_{r_1}(x-y)\mathbb{1}\left(|x-y| < C^{1/d}r_1\right)(u(x)-u(y))^\alpha \\
&\leq \sum_{x\in X_n} \sum_{y\in X_n} \eta_r(x-y)\mathbb{1}\left(x \in N_k(y)\right)\mathbb{1}\left(y \in N_k(x)\right)(u(x)-u(y))^\alpha \\
&\leq \sum_{x\in X_n} \sum_{y\in X_n} \eta_r(x-y)\mathbb{1}\left(x \in N_k(y) \text{ or } y \in N_k(x)\right)(u(x)-u(y))^\alpha \\
&\leq \frac{r_2^d}{r^d} \sum_{x\in X_n} \sum_{y\in X_n} \eta_{r_2}(x-y)\mathbb{1}\left(|x-y| < C^{1/d}r_2\right)(u(x)-u(y))^\alpha,
\end{aligned}$$

where k, r_1 and r_2 are chosen as above.

Proof. From Lemma3.3.1, we can already deduce that

$$\begin{aligned}
&r^{-d} \sum_{x\in X_n} \sum_{y\in X_n} \eta((x-y)/r)\mathbb{1}\left(|x-y| < C^{1/d}r_1\right)(u(x)-u(y))^\alpha \\
&\leq \sum_{x\in X_n} \sum_{y\in X_n} \eta_r(x-y)\mathbb{1}\left(x \in N_k(y)\right)\mathbb{1}\left(y \in N_k(x)\right)(u(x)-u(y))^\alpha \\
&\leq \sum_{x\in X_n} \sum_{y\in X_n} \eta_r(x-y)\mathbb{1}\left(x \in N_k(y) \text{ or } y \in N_k(x)\right)(u(x)-u(y))^\alpha \\
&\leq r^{-d} \sum_{x\in X_n} \sum_{y\in X_n} \eta((x-y)/r)\mathbb{1}\left(|x-y| < C^{1/d}r_2\right)(u(x)-u(y))^\alpha.
\end{aligned}$$

Since $r_1 \leq r \leq r_2$, we have

$$\eta((x-y)/r_1) \leq \eta((x-y)/r) \leq \eta((x-y)/r_2),$$

from which the claimed result now follows.

□

The above corollary provides us with a natural way to deduce consistency results for the k-NN graphs via analogous consistency results for r-neighbourhood graphs, with suitable choices for α and u. Below, we provide a few examples which follow immediately from the above corollary and the known consistency results for random geometric graphs (cf, Chapter 2).

3.3.1 The Dirichlet energy and spectral clustering consistency for k-NN graphs

As we have seen above, the Dirichlet energy can be defined on a k-NN graph in a similar way to before (cf, Definition 17). Combined with Corollary 3.3.2, this immediately gives us similar consistency results to the ones mentioned in the Chapter 2 for random geometric graphs.

Namely, we have the following result (to be compared with Theorem 2.2.1 above).

Theorem 3.3.3. *Suppose that η sastisfies conditions (2.3) (cf, previous chapter) and that r_n satisfies conditions (2.5), and suppose that*

$$k \in [Cn\omega_d r^d, Cn\omega_d r^d + 1] \cap \mathbb{N},$$

for some $C > 0$. Then the k-NN graph Dirichlet energies $(G^{(1)}_{\eta,r_n})_{n\in\mathbb{N}}$ and $(G^{(2)}_{\eta,r_n})_{n\in\mathbb{N}}$ both Γ-converge to $\sigma_{\tilde{\eta}}G$ in the TL^2 sense, where $\sigma_{\tilde{\eta}}$ is defined in conditions (2.3), G is the continuous Dirichlet energy, and

$$\tilde{\eta}(z) := \eta(z)\mathbb{1}(|z| < C^{1/d}),\ z \in \mathbb{R}^d.$$

Furthermore, $(G^{(1)}_{\eta,r_n})_{n\in\mathbb{N}}$ and $(G^{(2)}_{\eta,r_n})_{n\in\mathbb{N}}$ both satisfy the following compactness property in TL^2. Every sequence $((\nu_n, u_n))_{n\in\mathbb{N}}$ in TL^2 satisfying

$$\sup_{n\in\mathbb{N}} ||u_n||_{L^2(\nu_n)} + G^{(i)}_{\eta,r}(u_n) < \infty,$$

is precompact in TL^2, $i \in [2]$.

We refer the reader back to Chapter 2, Definition 8, for the definition of the TL^2 topology and to Chapter 2, Definition 6, for the definition of Γ-convergence.

Proof. Let $i \in [2]$. Let us first show the compactness property. Suppose that $((\nu_n, u_n))_{n\in\mathbb{N}}$ in TL^2 satisfies

$$\sup_{n\in\mathbb{N}} ||u_n||_{L^2(\nu_n)} + G^{(i)}_{\eta,r}(u_n) < \infty,$$

then, by Corollary 3.3.2 it follows that

$$\sup_{n\in\mathbb{N}} ||u_n||_{L^2(\nu_n)} + \frac{r_1^{d+2}}{r^{d+2}} G_{\tilde{\eta},r_1}(u_n) < \infty.$$

Since r_1 and r are asymptotically equivalent, we deduce by Theorem 2.2.1 that $((\nu_n, u_n))_{n\in\mathbb{N}}$ is precompact in TL^2.

Let us now show the Γ-convergence.

Liminf lower bound

Let $(u_n)_{n\in\mathbb{N}}$, with $u_n \in L^2(\nu_n)$, be converging in TL^2 to some $u \in L^2(\nu)$. By the Γ-convergence in Theorem 2.2.1, we know that

$$\liminf_{n\to\infty} G_{\tilde{\eta},r_1}(u_n) \geq \sigma_{\tilde{\eta}} G(u).$$

Again using Corollary 3.3.2 and the fact that r_1 and r are asymptotically equivalent, it then follows that

$$\liminf_{n\to\infty} G^{(i)}_{\eta,r}(u_n) \geq \sigma_{\tilde{\eta}} G(u),$$

which proves the required liminf property of the Γ-convergence.

Limsup upper bound

Similarly by Theorem 2.2.1, for every $u \in L^2(\nu)$, there exists a sequence $(u_n)_n$, with $u_n \in L^2(\nu_n)$, such that

$$\limsup_{n\to\infty} G_{\tilde{\eta},r_2}(u_n) \leq \sigma_{\tilde{\eta}} G(u).$$

Using Corollary 3.3.2 and the fact that r_2 and r are asymptotically equivalent, it follows that

$$\limsup_{n\to\infty} G^{(i)}_{\eta,r}(u_n) \leq \sigma_{\tilde{\eta}} G(u).$$

□

In [64], it is shown that Theorem 2.2.1 together with the Courant minimax principle, viewing the eigenvalues of the (discrete or continuous) Laplacian as minimizers of the (discrete or continuous) Dirichlet energy, imply Theorem 1.2 in [64]: spectral convergence (both eigenvalues suitably normalised and eigenvectors in the TL^2 metric) of the graph to the continuous Laplacian and consistency

of the spectral clustering Algorithm 1 proposed in that paper. By the same token, it is easy to verify that our Theorem 3.3.3 above similarly implies spectral convergence (both eigenvalues suitably normalised and eigenvectors in the TL^2 metric) of the unnormalised k-NN graph Laplacians to the continuous unnormalised Laplacian. As discussed before, the restriction to the unnormalised case is for the sake of clarity and brevity of the presentation, but a similar argument establishes likewise the spectral convergence of the normalised k-NN graph Laplacians to the continuous normalised Laplacian. Let us summarize the above discussion in a statement.
Let $\lambda_1 \leq \lambda_2 \leq \dots$ be the eigenvalues of the (unnormalised) Laplacian Δ, and let $u_1, u_2, \dots$ be the associated othonormal eigenvectors (or eigenfunctions).
For $i \in [2]$, let $\lambda_1^{(n,i)} \leq \lambda_2^{(n,i)} \leq \dots$ be the eigenvalues of the (unnormalised) k-NN graph Laplacians $\Delta_n^{(i)}$ (cf, Definition 18), and let $u_1^{(n,i)}, u_2^{(n,i)}, \dots$ be the associated eigenvectors. We have the following theorem (to be compared with Theorem 1.2 in [64]).

Theorem 3.3.4. *Let $i \in [2]$. Suppose that η satisfies conditions (2.3), that r_n satisfies conditions (2.5), and that $k \in [Cn\omega_d r^d, Cn\omega_d r^d + 1] \cap \mathbb{N}$, for some $C > 0$.*
Convergence of the eigenvalues: *For all $\ell \in \mathbb{N}^*$, we have*

$$\lim_{n\to\infty} \frac{2\lambda_\ell^{(n,i)}}{nr_n^2} = \sigma_{\tilde{\eta}} \lambda_\ell.$$

Convergence of the eigenvectors: *For every $\ell \in \mathbb{N}^*$, the sequence $(u_\ell^{(n,i)})_n$ is precompact in TL^2, and for every converging subsequence $(u_\ell^{(\varphi(n),i)})_n$ to some $u_\ell \in L^2(\nu)$, it holds that $||u_\ell||_{L^2(\nu)} = 1$ and u_ℓ is an eigenfunction of Δ associated to λ_ℓ.*
Convergence of the eigenprojections: *The same statement as in Theorem* 1.2 *of [64] holds.*
Consistency of spectral clustering: *We refer to Algorithm* 1 *in [64], where we change the unnormalised graph Laplacian Δ_n to $\Delta_n^{(i)}$. Then the same statement as in Theorem* 1.2 *[64] holds as well. This establishes, under the above conditions on r and k, consistency of spectral clustering done with a sparsified Laplacian via a k-NN construction.*

3.3.2 Consistency of the k-NN Cheeger constant and minimal bisection functional

Likewise, the same observations as those above yield consistency results for the Cheeger constants and the minimal bisection functionals on a k-NN graph.

We obtain, by the same token as in the previous section, the following a.s. convergence results (to be compared with the results in [49]).

Theorem 3.3.5. *Let $i \in [2]$. Suppose that η satisfies conditions (2.6), that $r_n = \omega((\log n)^{1/d} n^{-1/d})$, $d \geq 2$, and that $k \in [Cn\omega_d r^d, Cn\omega_d r^d + 1] \cap \mathbb{N}$. Then*

$$\lim_{n\to\infty} \left(\frac{1}{n^2 r} \mathrm{CHE}\left(k\text{-}NN^{(i)} \right) \right) = \mathrm{CHE}(D) \ a.s.,$$

and

$$\lim_{n\to\infty} \left(\frac{1}{n^2 r} \mathrm{MBIS}\left(k\text{-}NN^{(i)} \right) \right) = \mathrm{MBIS}(D) \ a.s..$$

3.4 Conclusion

For computational reasons, sparse graph constructions (few edges) are very important in practice. A classic sparsification method on graphs is the k-NN construction (two possible variations) which we have discussed above.

While many consistency results are known to hold for random geometric graphs, constructions are not assumed to be sparse. Generally, if one is given a smooth kernel function η, then every pair of vertices is assigned to a strictly positive weight (even if negligible when the points are far away). As far as we know, little is known from a theoretical point of view about consistency results (e.g., spectral clustering done with a sparsified Laplacian, k-NN Cheeger consistency, etc.) for sparse graph constructions such as k-NN graphs.

In this chapter, we showed how simple results on the regularity of empirical measures with respect to the underlying measure can be used to derive such consistency results of k-NN graph constructions from the better known consistency results for geometric graphs (presented in Chapter 2). In particular, we show if $k = \omega(\log n)$ then k-NN spectral clustering is consistent for appropriate r, and we also have Cheeger consistency and consistency of minimal bisection functionals.

Chapter 4

On the consistency of random geometric graphs quantities on $\mathbb{R}^d$

4.1 Introduction

The variational approaches described above, followed in [63, 64, 61, 49] to establish spectral clustering consistency and consistency of the Cheeger constants and the minimal bisection functionals all rely on a few key ideas. One of them deals with convergence and compactness properties of certain functionals. In particular the Γ-convergence of functionals, mentioned above, is characteristic of the variational approach. Another key idea is a discrepancy-type result, comparing the regularity of the empirical measure on a partition of small cubes of the domain, with respect to the underlying sampling measure, which is similar to Lemma 3.1.2 above. We state these key results below. Then, we provide some extensions of these results to the case where the domain is $\mathbb{R}^d$ and discuss some current limitations. We then illustrate an application of these generalisations to $\mathbb{R}^d$ by showing some type of consistent perimeter estimation on $\mathbb{R}^d$ using graph cut, a generalisation of the work in [65].

4.2 Some key results in the variational approach

It is worth noting a few key results which serve as cornerstones in all of the successful variational approaches followed to establish spectral consistency or Cheeger consistency.

One of them is a discrepancy-type result derived from Chernoff-type bounds, measuring the regularity of the empirical measure on a partition of small cubes of the domain, with respect to the underlying sampling measure. Some version

of this result is used in order to derive the concentration bounds obtained in [63] on empirical measures under the ∞-Wasserstein distance. This result is also repeatedly used throughout the argument in [49].

The second key result, characteristic of the variational approach, deals with properties of certain functionals. This is the so-called Γ-convergence, alluded to in the previous chapters (cf, Theorem 2.2.1). Together with a compactness property, this type of convergence proves itself very useful to establish various consistency results, as discussed above.

Before stating the discrepancy-type result, let us construct a partition of the domain D into small cubes, as it is done in [49].

Let $(\gamma_n)_{n\in\mathbb{N}}$ be non-increasing and such that

$$n\gamma_n^{d+2}r_n^d = \omega(\log n),$$

where r_n satisfies

$$nr_n^d = \omega(\log n).$$

In particular, $(\gamma_n)_{n\in\mathbb{N}}$ can be constant or tending to 0 arbitrarily slowly. In [49], the added conditions that $\gamma_n \to 0$ and

$$\gamma_n^{d+4} = \omega(r_n)$$

need to be satisfied for reasons related to later parts of the argument, but not directly impacting the general discrepancy-type result which we present below.

Let $n \in \mathbb{N}$ and divide $\mathbb{R}^d$ into a grid of cubes of side $\gamma_n r_n$. Call the cubes $\{Q'_{i,n} | i \in \mathbb{N}\}$ and for each $i \in \mathbb{N}$ denote the centre of $Q'_{i,n}$ by $z_{i,n}$. Without loss of generality, we may assume that the origin is one of the centres.

Given $D \subset \mathbb{R}^d$ connected, open, with Lipschitz boundary, let $S_n(D) := \{i \in \mathbb{N} \mid Q'_{i,n} \subset D\}$ and $D_n(D) := \{z_{i,n} \mid i \in S_n\}$. Suppose that $S_n(D) \neq \emptyset$ (which is necessarily true for n sufficiently large, given D).

For $i \in S_n(D)$, let

$$I(i,n) := \{j \in \mathbb{N} \mid \forall i' \in S_n(D) \setminus \{i\},\, |z_{i,n} - z_{j,n}| < |z_{i',n} - z_{j,n}|\},$$

and let

$$Q_{i,n} = \cup_{j\in I(i,n)}(Q'_{j,n} \cap D).$$

These boxes in particular partition the domain D. We have the following discrepancy-type result.

Lemma 4.2.1 (Lemma 3.2 in [49]). *Almost surely, there exists a finite random variable n_0 such that for all $n \geq n_0$ and all $i \in S_n$ we have*

$$|X_n \cap Q_{i,n}| \leq (1 + \gamma_n) n\nu(Q_{i,n})$$

and

$$|X_n \cap Q_{i,n}| \geq (1 - \gamma_n) n\nu(Q_{i,n}).$$

Note the similarity with our Lemma 3.1.2 and Corollary 3.3.2, in the previous chapter, which proofs are an easy adaptation of the proof of the above lemma.

The second key result, following results first established in [63], is about the Γ-convergence of some functionals. Here, we will focus on the setting of [49], which investigates the total variation functionals rather than the Dirichlet energies (cf, persentation in the previous chapters). In particular, we will focus on the Γ-convergence of the functionals $(TV_{r_n})_{n\in\mathbb{N}}$ and on some compactness properties relative to these functionals. These functionals are defined as follows.

Definition 21. *Given $u \in L^1(D, \nu)$, consider the functional*

$$TV_r(u; q) := \mathbb{E}\left[G\eta_r(u)\right] = \frac{2}{r}\int_D \int_D \eta_r(x, y)|u(x) - u(y)|q(x)q(y)dxdy.$$

These functionals can indeed be seen as the expectations of the (normalised) graph cut functionals mentioned in the previous chapters. In [49], they serve as an intermediate functional between graph cut and the total variation functional.

The convergence properties of the functionals $(TV_{r_n})_{n\in\mathbb{N}}$ (as first proved in [63] and also used in [49]) are gathered in the following lemma.

Lemma 4.2.2. *For every $r_n = o(1)$, the following holds.*
*(i)**[liminf lower bound]** For all sequences $(u_n)_{n\in\mathbb{N}}$ converging in $L^1(D, q)$ to some $u \in L^1(D, q)$,*

$$\liminf_{n\to\infty} TV_{r_n}(u_n; q) \geq \sigma_\eta TV(u; q). \tag{4.1}$$

*(ii)**[limsup upper bound]** For all $u \in L^1(D, q)$,*

$$\limsup_{n\to\infty} TV_{r_n}(u; q) \leq \sigma_\eta TV(u; q). \tag{4.2}$$

(ii)' We deduce from (i) and (ii) that for all $u \in L^1(D, q)$,

$$\lim_{n\to\infty} TV_{r_n}(u; q) = \sigma_\eta TV(u; q). \tag{4.3}$$

*(iii)**[compactness property]** If $(u_n)_{n\in\mathbb{N}}$ is a sequence bounded in $L^1(D, q)$ such that $(TV_{r_n}(u_n; q))_{n\in\mathbb{N}}$ is also bounded, then we can extract a subsequence $(u_{\varphi(n)})_{n\in\mathbb{N}}$ converging in $L^1(D, q)$ to some $u \in L^1(D, q)$.*

We refer the reader back to equation 2.6 for the general conditions assumed on the kernel function η, and more specifically to c) for the definition of σ_η. The discrepancy-type result and the above properties of functionals such as $(TV_{r_n})_{n\in\mathbb{N}}$, are essential steps towards establishing consistency results such as those presented in the previous chapters. Towards establishing consistency results for random geometric graphs quantities supported on $\mathbb{R}^d$, it is thus natural to first seek analogues of these results in the case of an unbounded domain. We choose to work with a homogeneous Poisson point process $\mathcal{P}_n$ of intensity n, sampled from a radial density q with support $\text{supp}(q) = \mathbb{R}^d$. The particular choice of q is of course an important matter. Let us now discuss some motivation behind the setting that we consider in $\mathbb{R}^d$.

4.3 A geometric interpretation for the ∞-Wasserstein distance

In [66] part of the approach relies on results previously obtained in [62], on the concentration of empirical measures under the ∞-Wasserstein distance. This forces a dichotomy between the cases where the dimension verifies $d = 2$ and where it verifies $d \geq 3$, due to the rates of convergence of empirical measures under the ∞-Wasserstein distance scaling differently with the dimension in those two cases (see Theorem 3.1.1). In the case where $d \geq 3$, the authors of [66] are able to obtain Cheeger consistency beyond the following threshold value of the bandwidth parameter of the graph: $r_n \sim (\log n)^{1/d} n^{-1/d}$. This threshold is sharp since this is the well known connectivity threshold value for random geometric graphs on bounded domains. On the other hand, they could only obtain Cheeger consistency for $r_n \sim (\log n)^{3/4} n^{-1/2}$ in the case where $d = 2$, due to the dichotomy inherent to the use of the ∞-Wasserstein distance. These bandwidth threshold values are similar to the ones obtained for the consistency of spectral clustering by the same authors in [64] (which we presented in more detail in Chapter 2). This is not surprising, since as we have already seen, both problems (Cheeger consistency and spectral clustering consistency) arise as optimisation problems on graph functionals of the same form, and both arguments use the concentration results on the ∞-Wasserstein distance (thus in both cases, there is a dichotomy between the cases $d = 2$ and $d \geq 3$).

Part of the point of the recent approach followed by Müller and Penrose in [49], is to avoid using the ∞-Wasserstein distance in the argument yielding to Cheeger consistency, so as to avoid such a dichotomy. They successfully obtain Cheeger consistency for $r_n = \omega((\log n)^{1/d} n^{-1/d})$ for every $d \geq 2$. Another advantage of their approach, is that it can be applied to a variety of other consistency problems. In particular, as discussed above, the consistency of minimal bisection functionals is also showed in [49] with the same approach.

It is interesting to note that in the case where $d \geq 3$, the concentration

results on the ∞-Wasserstein distance obtained in [62] essentially indicate that

$$W_\infty(\nu, \nu_n) \lesssim d_H(D, X_n),$$

where d_H is the Hausdorff distance and

$$\nu_n := \frac{1}{n} \sum_{x \in X_n} \delta_x$$

is the normalised empirical measure associated to X_n.
Indeed the Hausdorff distance between two sets A and B may be defined as

$$d_H(A, B) := \inf\{r > 0 \mid A \subset B^r \text{ and } B \subset A^r\},$$

where

$$A^r := \{x \mid dist(x, A) < r\}.$$

The Hausdorff distance between D and an i.i.d. sampled set of points $X_n \subset D$ is thus given by

$$d_H(D, X_n) = \inf\left\{r > 0 | D \subset \cup_{i=1}^n B(x_i, r)\right\},$$

which is the minimum radius needed for balls sampled at random (i.e., their centres) to cover a bounded domain (or a compact manifold).

This *minimal random covering radius* is known to be $\sim (\log n)^{1/d}(n)^{-1/d}$ w.h.p. (cf, [63]). This quantity will play an important role in our later investigations of random geometric complexes, in particular in Chapter 7 where we extend, following our joint work with Ulrike Tillmann and Oliver Vipond ([42]), some previous results of Bobrowski and Weinberger ([11]) and of Bobrowski and Oliveira ([10]).

It is well known that the Hausdorff distance serves as a lower bound for the ∞-Wasserstein distance. In our setting, this is immediately verified from the above formulation for $d_H(D, X_n)$ and using the following formulation for the ∞-Wasserstein distance (cf, [36])

$$W_\infty(\nu, \nu_n) := \inf\{r > 0 | \forall A \in \mathcal{B}(D),\ \nu(A^r) \geq \nu_n(A)\}.$$

The results in [62] which are sharp in the case where $d = 2$, illustrate that the reverse inequality is not expected to hold in general.

This observation yields a geometric interpretation of the ∞-Wasserstein distance when $d \geq 3$, in terms of this *minimal random covering radius*. We note that similar geometric interpretations have already been investigated in the case of the Wasserstein distances W_p, $p \in (1, \infty)$. In this case, it is shown that when $d \geq 3$, $W_p(\nu, \nu_n) \sim n^{-1/d}$, and is shown to be directly related to the minimal covering radius of a set of n points (equivalently, to the cover number of a manifold, given a radius) (e.g., [26]). We ask whether the geometric interpretation for the ∞-Wasserstein distance in terms of the above minimal random covering radius can be made more explicit, as in the other cases.

4.4 Consistency results on $\mathbb{R}^d$

These geometric interpretations for threshold values of the bandwidth parameter for consistency results in terms of the Hausdorff distance, raise natural questions in the case where we consider an unbounded domain such as $\mathbb{R}^d$. There, we obviously have $\infty = d_H(\mathbb{R}^d, X_n) \leq W_\infty(\nu, \nu_n)$. In other words, the same geometric pictures which helped to guide our guesses in the case of a bounded domain, do not carry over to $\mathbb{R}^d$. On the other hand, note that if $r_n \sim d_H(D, X_n)$, then in particular the union of the balls centered at the vertices with radius r_n is contractible (assuming D is connected), which in turn is also a sufficient condition for the graph to be connected. It turns out that if a sampling density supported on $\mathbb{R}^d$ has superexponential decay, conditions on r_n can be found such that the union of the balls centered at the vertices and of radius r_n are indeed contractible, hence such that the graph is connected. This is the object of Theorem 4.5 in [52], which we mention more in details below. This naturally raises the question of whether such conditions on the density and on r_n can yield consistency results on $\mathbb{R}^d$, similar to the ones obtained for bounded domains in [63, 64, 49] (spectral clustering consistency, Cheeger consistency, minimal bisection consistency, etc.).

There are several reasons to be interested in consistency results on $\mathbb{R}^d$. First of all, because definitions for the Cheeger constant, the minimal bisection functional or the Laplacian naturally extend to the case where the domain is $\mathbb{R}^d$ instead of, say, a compact manifold. Another reason is to consider the more realistic possibility of sampling from a bounded domain with noise. One may wonder if it is still possible to obtain consistency results on a bounded domain, in the case where we sample from the domain with unbounded noise. In this latter setting, we could assume the bounded domain to be just the origin, and model the unbounded noise by a sampling density supported on $\mathbb{R}^d$. This setting was already investigated with views on applications to Topological Data Analysis in [1, 52], where the authors investigated the effect on homology recovery when sampling a manifold with noise.

4.5 Topological crackle and setting on $\mathbb{R}^d$

In [1] the authors consider densities supported on $\mathbb{R}^d$ with various decays, and investigate the effect on capturing the homology of a manifold, when sampling with unbounded noise. The answer to whether noisy samples affect the recovery of the manifold homology, i.e., whether the homology of the noise eventually vanishes, depends on the density considered. This work itself builds on previous works by Niyogi, Smale and Weinberger ([50, 51]) which considered cleaning algorithms for random samples from low dimensional manifolds sampled with Gaussian noise. Here the setting is more general. Assume the submanifold to be reduced to be the origin and let q be a density supported on $\mathbb{R}^d$, not necessarily Gaussian.

The authors of [1] identify sequences of radii $(R_n^c)_{n\in\mathbb{N}}$ depending on the density, such that homology vanishes in $B(0, R_n^c)$ but not necessarily outside the ball, in which case they call this phenomenon *topological crackle*. Crackle generally happens except if the density has superexponential decay, in which case there exists a sequence of radii $(R_n^{(1)})_{n\in\mathbb{N}}$ and choices for $(r_n)_{n\in\mathbb{N}}$ depending on the density, such that w.h.p. $\mathcal{P}_n \subset B(0, R_n^{(1)})$ and $\cup_{x\in\mathcal{P}_n} B(x, r_n)$ is contractible; here $\mathcal{P}_n$ is a homogenous Poisson point process.

In [1], the authors focus on three particular densities, illustrating various decaying rates: *power law*, *exponential*, *Gaussian*. They identify crackling sequences of radii $(R_n^c)_{n\in\mathbb{N}}$ in each case. This is the content of Theorem 1 in [1]. Note that there is nothing special about these specific examples, other that they are representative of three general rates of decay (power law, exponential, superexponential). In fact, one can extract from the proof of Theorem 1 in [1] the following lemma.

Lemma 4.5.1 (from proof of Theorem 1 in [1])**.** *Let $q : \mathbb{R}^d \to \mathbb{R}_+$ be a sampling density supported on $\mathbb{R}^d$ and let $\mathcal{P}_n$ be a homogenous Poisson point process of intensity n, sampled with respect to q. Suppose that the sequences $(R_n^c)_{n\in\mathbb{N}}$ and $(r_n)_{n\in\mathbb{N}}$ satisfy*

$$(R_n^c/gr_n)^d \exp\left(-g^d n r_n^d q(R_n^c e_1)\right) = o(1),$$

where g is a small constant depending only on the dimension d, then

$$\lim_{n\to\infty} \mathbb{P}\left(B(0, R_n^c) \subset \cup_{x\in\mathcal{P}_n} B(x, r_n)\right) = 1.$$

(We note early works, [50, 51], in this direction, with the difference that they deal with compact Riemannian submanifolds of $\mathbb{R}^d$.) Building from this argument, the authors in [52] pursued the investigation of the topological structure of noise with more general settings assumed for the density, using tools from Extreme Value Theory. In particular, they exhibit sequences of annulii in $\mathbb{R}^d$, where each annulus generates homology of a certain degree. Furthermore if the density has superexponential decay (a generalisation of the case where the density was a Gaussian in [1]), they show (cf, Theorem 4.5 in [52]) as mentioned above that the union of the balls (centered at the vertices) is contractible, thus in particular that there is no topological crackling in this case.

As we have just seen, for every density one can find radii R_n^c such that the union of the balls covers $B(0, R_n^c)$ (in particular is contractible and the graph restricted to this ball is connected). For clarity and brevity of the presentation below, we restrict our attention to the above mentioned case of densities with superexponential decay. With this restriction slightly more can be said (cf, Theorem 4.5 in [52]), but we emphasise that similar results could be derived with any density, following the work in [52]. Let us recall the setting for light tail

densities used in [52], which we follow to derive our results in this chapter.

Definition 22. *Given a function $\psi \in C^2(\mathbb{R}_+;\mathbb{R})$, we say that it is of* ***von Mises type*** *if for all $z \to \infty$,*

$$\begin{cases} \psi'(z) > 0, \\ \psi(z) \to \infty, \\ a'(z) \to 0, \ \textit{where } a(z) := \dfrac{1}{\psi'(z)}. \end{cases}$$

Let the (light tail) density distribution function be given as in [52], by

$$\begin{aligned} q : \mathbb{R}^d &\longrightarrow \mathbb{R}_+ \\ x &\longmapsto L(|x|)\exp\left(-\psi(|x|)\right), \end{aligned}$$

where $\psi \in C^2(\mathbb{R}_+;\mathbb{R})$ is a function of von Mises type.

We suppose furthermore that

$$\frac{L(t+a(t)v)}{L(t)} \to 1 \text{ as } t \to \infty \text{ uniformly on intervals,} \tag{4.4}$$

and that $\exists(\gamma, z_0, C) \in \mathbb{R}_+ \times \mathbb{R}_+^* \times [1,\infty)$,

$$\forall t > 1, \forall z \geq z_0, \ \frac{L(zt)}{L(z)} \leq Ct^\gamma. \tag{4.5}$$

As discussed in [52], the above assumptions are made to ensure that the tail behaviour of q is governed by ψ, while the behaviour of L becomes asymptotically negligible. Hence, for clarity of the results presented below, we shall assume as in [52] that $L \equiv C$, where C is a suitable normalising constant.

Definition 23. *We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is regularly varying with exponent α, and we write $f \in RV_\alpha$, if for all $x \in S^{d-1}$ and all $t > 0$*

$$\lim_{R\to\infty} \frac{f(tRx)}{f(Rx)} = t^\alpha.$$

Assuming that $\psi \in RV_v$ for some $v > 0$, we say that the density q has a light tail with **subexponential decay** if $v < 1$, with **exponential decay** if $v = 1$, and with **superexponential decay** if $v > 1$. We shall consider later the special case where $\psi \in RV_v$ with $v > 1$, where some interesting concentration results can be extended from a bounded domain to all of $\mathbb{R}^d$.

Since ψ is eventually monotone (increasing), its inverse function $\psi^{\leftarrow}$ is well-defined asymptotically.

Let $(r_n)_{n\in\mathbb{N}}$ be a regularly varying sequence of positive real numbers, decreasing to 0 and such that

$$\lim_{n\to\infty} \frac{a \circ \psi^{\leftarrow}(\log n)}{r_n} \log\log n = 0. \tag{4.6}$$

Furthermore, let $(\gamma_n)_{n\in\mathbb{N}}$ be decreasing to 0 (arbitrarily slowly) such that

$$n r_n^d \gamma_n^{d+2} = \omega(\log n).$$

For $n \in \mathbb{N}$, let $\mathcal{P}_n$ be a homogenous Poisson point process with intensity n with respect to the density q, i.e.,

$$\mathcal{P}_n := \{x_1, \ldots, x_N\},$$

where the x_i's are i.i.d. samples with respect to q and independent from $N \sim Po(n)$.

Assuming that $\psi \in RV_v$ with $v > 1$, i.e., that q has a light tail with superexponential decay, and under conditions (4.6) assumed for the bandwidth parameter r_n, we have the following theorem from [52].

Theorem 4.5.2 (Theorem 4.5 in [52])**.** *There exists $g > 0$ sufficiently small (independent of n), such that choosing δ satisfying*

$$d - e^{\delta} g^d C < 0,$$

letting

$$R_n^{(0)} := \psi^{\leftarrow}\left(\log n + d\log r_n - \log\log\left(r_n^{-1}\psi^{\leftarrow}(\log n)\right) - \delta\right)$$

and

$$R_n^{(1)} := \psi^{\leftarrow}\left(\log n + (d-1)\log\psi^{\leftarrow}(\log n) + \log\left(a \circ \psi^{\leftarrow}(\log n)\right) + \log\log n\right),$$

we have

$$\lim_{n\to\infty} \mathbb{P}\left(B(0, R_n^{(0)}) \subset \bigcup_{x\in\mathcal{P}_n \cap B(0,R_n^{(0)})} B(x, r_n),\ \mathcal{P}_n \cap B(0, R_n^{(1)})^c = \emptyset\right) = 1$$

and

$$R_n^{(1)} - R_n^{(0)} = o(r_n).$$

As noted above, it is straightforward to deduce from this theorem that w.h.p. $\cup_{x\in\mathcal{P}_n} B(x, r_n)$ is contractible, hence that w.h.p. the graph $G(\mathcal{P}_n, r_n)$ is connected.

4.6 Some discrepancy-type results on $\mathbb{R}^d$

Here, we present some extensions of the discrepancy-type result used in the variational approaches for bounded domains (cf, Lemma 3.2 in [49]) to the case where the domain is $\mathbb{R}^d$ and the density q has superexponential decay, as specified above. To do this we identify as in [1, 52], sequences of radii beyond which the discrepancy-type inequalities cease to hold (while they hold before these radii).

As noted above, we could provide similar discrepancy-type results for more general densities $q : \mathbb{R}^d \to \mathbb{R}_+$, using the criterion given in Lemma 4.5.1. Depending on the exact results one wishes to deduce, one may relax the conditions on the bandwidth r_n.

The following Chernoff-type bounds for Poisson random variables, which are the analogue of the bounds used in [49] for binomial random variables (see also Lemmas 1.1 and 1.2 in [33]), hold.
Recall that $N \sim Po(n)$ is a Poisson random variable with intensity n. For $x > 0$ let

$$H(x) := 1 - x + x \log x,$$

and set $H(0) := 1$.
We have

$$\mathbb{P}(N \geq k) \leq \exp\left(-nH\left(\frac{k}{n}\right)\right), \, k \geq n; \tag{4.7}$$

$$\mathbb{P}(N \leq k) \leq \exp\left(-nH\left(\frac{k}{n}\right)\right), \, k \leq n. \tag{4.8}$$

We start with the following discrepancy-type result on $\mathbb{R}^d$ derived from the above Chernoff-type bounds, which is obtained by combining Lemma 3.2 in [49] and Theorem 4.5 in [52] (cf, Theorem 4.5.2 above).

Theorem 4.6.1. *Let q and r_n be as specified in the previous section. There exists sequences $(R_n^{(0)})_{n\in\mathbb{N}}$ and $(R_n^{(1)})_{n\in\mathbb{N}}$ such that*

$$R_n^{(1)} - R_n^{(0)} = o(r_n),$$

and such that with probability going to 1 as $n \to \infty$, the following holds.
For all $i \in S_n(B(0, R_n^{(0)}))$

$$\mathcal{P}_n(Q(i,n)) \leq (1 + \gamma_n) n\nu(Q(i,n)) \tag{4.9}$$

and

$$\mathcal{P}_n(Q(i,n)) \geq (1 - \gamma_n) n\nu(Q(i,n)). \tag{4.10}$$

Furthermore,

$$\mathcal{P}_n \cap B(0, R_n^{(1)})^c = \emptyset.$$

Proof. Let $R_n^{(1)}$ be chosen as in Theorem 4.5.2 above, and let

$$R_n^{(0)} := \psi^{\leftarrow}\left(\log n + d\log(r_n) + (d+2)\log(\gamma_n) - \log\log((\gamma_n r_n)^{-1}\psi^{\leftarrow}(\log n)) - \delta\right),$$

where δ is chosen such that

$$(d - e^{\delta}C'/3) < 0.$$

One can verify as in the proof of Theorem 4.5.2, that

$$R_n^{(1)} - R_n^{(0)} = o(r_n)$$

and that with probability going to 1

$$\mathcal{P}_n \cap B(0, R_n^{(1)}) = \emptyset.$$

Furthermore, note that for $i \in S_n(B(0, R_n^{(0)}))$

$$\int_{Q_{i,n}} q(x)dx \geq \int_{Q_{i,n}\cap B(0,R_n^{(0)})} q(x)dx \geq C'q(R_n^{(0)}e_1)(\gamma_n r_n)^d,$$

where $C' > 0$ is an absolute constant.
Thus, picking n sufficiently large so that $H(1+\gamma_n) > (1/3)\gamma_n^2$, we have for all $i \in S_n(B(0, R_n^{(0)})$

$$\begin{aligned}\mathbb{P}\left(\mathcal{P}_n(Q_{i,n}) > (1+\gamma_n)n\nu(Q_{i,n})\right) &\leq \exp\left(-n\nu(Q_{i,n})H(1+\gamma_n)\right)\\ &\leq \exp\left(-C'nq(R_n^{(0)}e_1)r_n^d\gamma_n^{d+2}/3\right).\end{aligned}$$

Summing over all $i \in S_n(B(0, R_n^0))$, we find by a union bound that the probability that (8) does not hold for some $i \in S_n(B(0, R_n^{(0)})$ is bounded above by

$$\left(\frac{R_n^{(0)}}{\gamma_n r_n}\right)^d \exp\left(-C'nq(R_n^{(0)}e_1)r_n^d\gamma_n^{d+2}/3\right) = \exp\left(d\log(r_n^{-1}R_n^{(0)}) - C'nq(R_n^{(0)}e_1)r_n^d\gamma_n^{d+2}/3\right).$$

To prove that the above goes to 0 as $n \to \infty$, it suffices to show that the exponent on the RHS goes to $-\infty$. We have indeed

$$d\log(r_n^{-1}R_n^{(0)}) - C'nq(R_n^{(0)}e_1)r_n^d\gamma_n^{d+2}/3 \leq (d - e^{\delta}C'/3)\log\left(r_n^{-1}\psi^{\leftarrow}(\log n)\right) \to -\infty.$$

This shows that (8) holds w.h.p. and (9) is shown similarly.

□

While interesting in its own right, this theorem is not directly interesting for the purpose of establishing a.s. consistency results. Indeed, the above probabilities converge to 1 too slowly in n for us to apply the Borel-Cantelli lemma.

This kind of result is required to obtain an a.s. convergence result similar to Theorem 2.1 above, which is only possible if with probability 1, the above events are true for all but finitely many n. To improve the rate of convergence of these probabilities (to make the probabilities summable over n, so that we can invoke the Borel-Cantelli lemma), we may restrict the setting to $B(0, \tilde{R}_n^{(0)})$, where $\tilde{R}_n^{(0)}$ is barely smaller than $R_n^{(0)}$. We then find the following theorem.

Theorem 4.6.2. *Suppose that*

$$\lim_{n\to\infty} \frac{a \circ \psi^{\leftarrow}(\log n)}{r_n} \log\log n = 0,$$

and let

$$\tilde{R}_n^{(0)} := \psi^{\leftarrow}\left(\log n + d\log(r_n) + (d+2)\log(\gamma_n) - \log\log n - \delta\right),$$

where δ is chosen such that

$$d - (C'/3)e^{\delta} \leq -2.$$

Note that it still holds that

$$R_n^{(1)} - \tilde{R}_n^{(0)} = o(r_n),$$

with $R_n^{(1)}$ chosen as above.
There exists an almost surely finite random variable $N_0 > 0$ such that for all $n \geq N_0$ and all $i \in S_n(B(0, \tilde{R}_n^{(0)}))$

$$\mathcal{P}_n(Q(i,n)) \leq (1+\gamma_n) n\nu(Q(i,n)) \tag{4.11}$$

and

$$\mathcal{P}_n(Q(i,n)) \geq (1-\gamma_n) n\nu(Q(i,n)). \tag{4.12}$$

Proof. We repeat the proof of the previous theorem with $\tilde{R}_n^{(0)}$ instead of $R_n^{(0)}$. We find similarly that the probability that (10) does not hold for some $i \in S_n(B(0, \tilde{R}_n^{(0)}))$ is bounded above by

$$\exp\left(d\log((\gamma_n r_n)^{-1}\tilde{R}_n^{(0)}) - C'nq(\tilde{R}_n^{(0)}e_1) r_n^d \gamma_n^{d+2}/3\right).$$

From the choice of $\tilde{R}_n^{(0)}$, we have

$$C'nq(\tilde{R}_n^{(0)}e_1) r_n^d \gamma_n^{d+2}/3 = (C'/3)\log(n) e^{\delta}.$$

Furthermore, since

$$\begin{aligned}\frac{\tilde{R}_n^{(0)}}{n} &\leq \frac{\psi^{\leftarrow}(\log n)}{n} \\ &\leq \frac{\log n}{n} \\ &\leq \gamma_n r_n,\end{aligned}$$

the exponent above is bounded above by

$$\left(d - (C'/3)e^{\delta}\right) \log n \leq -2 \log n,$$

hence this probability is bounded above by

$$n^{-2}.$$

The probabilities being summable over n, we can thus apply the Borel-Cantelli lemma and conclude that there exists an almost surely finite random variable N_0 such that for all $n \geq N_0$ and all $i \in S_n(B(0, \tilde{R}_n^{(0)}))$

$$\mathcal{P}_n(Q(i,n)) \leq (1 + \gamma_n) n \nu(Q(i,n)).$$

The second inequality is obtained similarly.

□

4.7 Properties of the functionals $(TV_{r_n})_{n\in\mathbb{N}}$ on $\mathbb{R}^d$

We have presented above some extensions of the discrepancy-type result, in Lemma 3.2 in [49]. To do this we followed the setting introduced [52] and focused our attention on densities with superexponential decay. We now present some extensions of the properties of the functionals $(TV_{r_n})_{n\in\mathbb{N}}$ (defined above), which are key properties in the variational approach (Γ-convergence and a compactness property). We start by extending some definitions to this new setting.

Throughout, suppose as before that ν has sampling density q satisfying (3.1). We assume that q is Lipschitz continuous and that is has superexponential decay, as in the two previous sections.

Let as before, the kernel function η satify (2.6), and let

$$\eta_r(z) := r^{-d}\eta(z),\ r > 0.$$

Definition 24. *Let $\Omega \subset \mathbb{R}^d$. For $u \in L^1(\mathbb{R}^d, q)$, define the functional*

$$\eta_r(u,\Omega) := \sum_{y\in\mathcal{P}_n}\sum_{x\in\mathcal{P}_n} \eta_r(x,y)|u(x)-u(y)|\mathbb{1}_\Omega(x)\mathbb{1}_\Omega(y).$$

Define its normalised versions

$$G\eta_r(u) := \frac{2}{n(n-1)r}\eta_r(u,\mathcal{P}_n) \tag{4.13}$$

and

$$\overline{G}\eta_{r_n}(u) := \frac{2}{N(N-1)r}\eta_r(u,\mathcal{P}_n), \tag{4.14}$$

where recall that $|\mathcal{P}_n| = N \sim Po(n)$.
Given $u = \mathbb{1}_A$, $A \subset \mathbb{R}^d$, define similarly the graph cut of A by

$$\mathrm{Cut}_{\eta,r}(A,\Omega) := \eta_r(\mathbb{1}_A,\Omega) = \sum_{y\in Y}\sum_{x\in\mathcal{P}_n\setminus Y} \eta_r(x,y)\mathbb{1}_\Omega(x)\mathbb{1}_\Omega(y), \tag{4.15}$$

where $Y := A\cap\mathcal{P}_n$.

Definition 25. *Given $u \in L^1(\mathbb{R}^d, q)$ and $\Omega \subset \mathbb{R}^d$, consider as before the functional*

$$TV_r(u,q,\Omega) := \frac{2}{r}\int_\Omega\int_\Omega \eta_r(x,y)|u(x)-u(y)|q(x)q(y)dxdy.$$

The total variation of u in Ω is given by

$$TV(u,q,\Omega) := \sup\left\{\int_\Omega u(x)div(\phi)(x)dx \mid \phi\in C_c^1(\Omega;\mathbb{R}^d), \forall x\in\Omega, |\phi(x)|\le q^2(x)\right\}.$$

The results of this section are gathered in the following lemma and are to be compared to those presented in Lemma 3.3 in [49] and Theorem 4.1 in [63]. We note in particular the difference in the compactness property, where we could only attain convergence in $L^1(\mathbb{R}^d, q^2)$ instead of $L^1(\mathbb{R}^d, q)$, and under restricted assumptions.

Lemma 4.7.1. *Let $(r_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{R}_+$ with $r_n = o(1)$.*
*(i)**[liminf lower bound]** For all sequences $(u_n)_{n\in\mathbb{N}}$ converging in $L^1(\mathbb{R}^d, q^2)$ to some $u \in L^1(\mathbb{R}^d, q)$,*

$$\liminf_{n\to\infty} TV_{r_n}(u_n, q) \geq \sigma_\eta TV(u, q). \tag{4.16}$$

*(ii)**[limsup upper bound]** For all $u \in L^1(\mathbb{R}^d, q) \cap L^\infty(\mathbb{R}^d)$,*

$$\limsup_{n\to\infty} TV_{r_n}(u, q, \Omega_n) \leq \sigma_\eta TV(u, q, \Omega),$$

where $\Omega_n := \Omega \cap B(0, \tilde{R}_n)$ and $\tilde{R}_n$ is such that

$$r_n q_{\min}^{-1}(\tilde{R}_n) = o(1).$$

(ii)' We deduce from (i) and (ii) that for all $u \in L^1(\mathbb{R}^d, q)$,

$$\lim_{n\to\infty} TV_{r_n}(u, q) = \sigma_\eta TV(uq). \tag{4.17}$$

*(iii)**[compactness property]** Suppose that $(u_n)_{n\in\mathbb{N}}$ is a sequence bounded in $L^1(\mathbb{R}^d, q)$ and in $L^\infty(\mathbb{R}^d)$, with $\operatorname{supp}(u_n) \subset B(0, \tilde{R}_n)$ where $\tilde{R}_n$ is as above, such that*

$$r_n q_{\min}(\tilde{R}_n)^{-1} = o(1);$$

suppose that $(TV_{r_n}(u_n, q))_{n\in\mathbb{N}}$ is also bounded and that $\forall \delta > 0$, $\exists R_\delta > 0$, $\forall n \in \mathbb{N}$

$$TV_{r_n}(u_n, q, \overline{B(0, R_\delta)}^c) < \delta,$$

then we can extract a subsequence $(u_{\varphi(n)})_{n\in\mathbb{N}}$ converging in $L^1(\mathbb{R}^d, q^2)$ to some $u \in L^1(\mathbb{R}^d, q)$.

Note that (i) deals with convergence in $L^1(q^2)$ instead of convergence in $L^1(q)$ (it is thus a stronger result). This change of weight simply does not affect the proof of the liminf lower bound; we chose to present it with $L^1(q^2)$-convergence in order to relate it better to the compactness property.

The fact that we can currently only establish the compactness property in $L^1(q^2)$ is a major limitation to the potential applications towards establishing consistency results for spectral clustering or Cheeger constants, as presented above in the case of a bounded domain.

4.7.1 The liminf lower bound

Proposition 4.7.2. *Let $\Omega \subset \mathbb{R}^d$ be (not necessarily bounded,) open, connected, with empty or Lipschitz boundary, and let $u \in L^1(\mathbb{R}^d, q)$. Let $(u_n)_{n\in\mathbb{N}}$ be a sequence converging in $L^1(\Omega, q^2)$ to u. We have*

$$\liminf_{n\to\infty} TV_{r_n}(u_n; q; \Omega) \geq \sigma_\eta TV(u; q; \Omega).$$

Proof. Let $\epsilon > 0$ and let $\phi_0 \in C_c^1(\Omega; \mathbb{R}^d)$ be such that $\forall x \in \Omega$, $|\phi_0(x)| \leq q^2(x)$, and

$$\int_\Omega u(x)\mathrm{div}\phi_0(x)dx \geq TV(u; q; \Omega) - \epsilon.$$

Let $R_0 > 0$ be such that $supp(\phi_0) \subset B(0, R_0) \subset \Omega$. Since $B(0, R_0) \subset \mathbb{R}^d$ is an open, bounded, connected Lipschitz set, we can use Theorem 4.1 from [63], where we note that replacing the $L^1(q)$-convergence by $L^1(q^2)$ is just a change of weight and the proof is not affected. We then find

$$\begin{aligned}
\liminf_{n\to\infty} TV_{r_n}(u_n; q; \Omega) &\geq \liminf_{n\to\infty} r^{-1} \int_{B(0,R_0)} \int_{B(0,R_0)} \eta_r(x-y)|u_n(x) - u_n(y)|q(x)q(y)dxdy \\
&\geq TV(u; q; B(0, R_0)) \\
&\geq \int_{B(0,R_0)} u(x)\mathrm{div}(\phi_0)(x)dx \\
&\geq TV(u; q; \Omega) - \epsilon,
\end{aligned}$$

and this holds for every $\epsilon > 0$.

□

4.7.2 The limsup upper bound

In this section, we prove the following limsup upper bound. From now on, let $\tilde{R}_n$ be such that

$$r_n q_{min}(\tilde{R}_n)^{-1} = o(1).$$

Let $\Omega \subset \mathbb{R}^d$ be (not necessarily bounded,) open, connected, with empty or Lipschitz boundary, and let $\Omega_n := \Omega \cap B(0, \tilde{R}_n)$.

Proposition 4.7.3. *Let $u \in BV(\mathbb{R}^d, q) \cap L^\infty(\mathbb{R}^d)$. If $u \in W^{1,1}(\mathbb{R}^d, q)$, then*

$$\limsup_{n\to\infty} TV_{r_n}(u, q, \Omega) \leq \sigma_\eta TV(u, q, \Omega).$$

More generally, we have

$$\limsup_{n\to\infty} TV_{r_n}(u, q, \Omega_n) \leq \sigma_\eta TV(u, q, \Omega).$$

Definition 26. *Let $u \in BV(\mathbb{R}^d, q)$. Define for $x, y \in \mathbb{R}^d$*

$$\phi_{r_n}(x,y) := r_n^{-1}\eta_{r_n}(x,y)|u(x)-u(y)|\mathbb{1}_\Omega(x)\mathbb{1}_{\Omega\cap\Omega_n}(y),$$

and define for $x \in \mathbb{R}^d$

$$\overline{\phi}_{r_n}(x) := \int_{\mathbb{R}^d} \phi_{r_n}(x,z)q(z)dz.$$

The limsup upper bound follows from the following lemma, which can be seen as an extension of Lemma 2.1 in [65] to the case where the sampling density is (not uniform and) supported on $\mathbb{R}^d$.

Lemma 4.7.4. *Let $p \geq 1$, let $u \in W^{1,1}(\mathbb{R}^d, q) \cap L^\infty(\mathbb{R}^d)$, and define*

$$u_\infty := \sup\{|u(x)-u(y)| \mid x, y \in \mathbb{R}^d\}.$$

For all $r > 0$, we have

$$\int_{\mathbb{R}^d} \overline{\phi}_r^p(x)q(x)dx \leq \frac{(u_\infty q_{\max}\omega_d)^{p-1}\sigma_\eta}{r^{p-1}}\left(\int_\Omega \left|\nabla u(x)\right| q^2(x)dx + C'r\int_\Omega \left|\nabla u(x)\right| q(x)dx\right),$$

for some constant $C' > 0$. In particular, for $p = 1$

$$TV_{r_n}(u,q,\Omega) \leq \sigma_\eta TV(u,q,\Omega) + o(1).$$

Suppose that the lemma holds for $u \in W^{1,1}(\mathbb{R}^d, q) \cap L^\infty(\mathbb{R}^d)$, so that the first claim of the above proposition holds. Then this implies the second statement in the proposition above for general $u \in BV(\mathbb{R}^d, q) \cap L^\infty(\mathbb{R}^d)$. To see this, let $n \in \mathbb{N}$, let $u \in BV(\mathbb{R}^d, q) \cap L^\infty(\mathbb{R}^d)$ and let $(u_k)_k$ be a sequence in $BV(\mathbb{R}^d, q) \cap L^\infty \cap C^\infty$ approximating u as in Theorem 5.3 of the book [29], i.e.,

$$\lim_{k\to\infty} ||u - u_k||_{L^1(\Omega_n, q)} = 0$$

and

$$\lim_{k\to\infty} \int_{\Omega_n} |\nabla u_k(x)| q^2(x)dx = TV(u,q,\Omega_n);$$

recall that n is fixed here.
By construction of $\tilde{R}_n$, we have independently of k

$$r\int_{\Omega_n} |\nabla u_k(x)| q(x)dx = o\left(\int_{\Omega_n} |\nabla u_k(x)| q^2(x)dx\right),$$

hence for all $k \in \mathbb{N}$, we deduce from the lemma

$$\int_{\mathbb{R}^d} \overline{\phi}_r^p(x)q(x)dx \leq \frac{(u_\infty q_{\max}\omega_d)^{p-1}\sigma_\eta}{r^{p-1}}\left(\int_{\Omega_n} \left|\nabla u_k(x)\right| q^2(x)dx\right)(1+o(1)).$$

Passing to the limit in k, we attain the desired result for general $u \in BV \cap L^\infty$.

Proof of Lemma 4.7.4. Suppose that $u \in W^{1,1}(\mathbb{R}^d, q) \cap L^\infty(\mathbb{R}^d)$ and let us show that

$$\int_{\mathbb{R}^d} \overline{\phi}^p_{r_n}(x) q(x) dx \leq \frac{(u_\infty q_{\max} \omega_d)^{p-1} \sigma_\eta}{r^{p-1}} \left(\int_\Omega \left| \nabla u(x) \right| q^2(x) dx + C' r \right),$$

for some constant $C' > 0$. This shows the proposition when $u \in W^{1,1} \cap L^\infty$. The result for a general $u \in BV \cap L^\infty$ follows after picking an approximating sequence $u_k \to u$ (converging in $L^1(\mathbb{R}^d, q)$), where $u_k \in C^\infty$, as it is done in Theorem 5.3 of the book [29], and as explained above. We have

$$\begin{aligned}
\int_{\mathbb{R}^d} \overline{\phi}^p_{r_n}(x) q(x) dx &\leq r^{-p} \int_\Omega \left(\int_{B(0,1)} |u(x+rh) - u(x)| q(x+rh) dh \right)^p q(x) dx \\
&\leq q_{\max}^{p-1} r^{-p} \omega_d^{p-1} \int_\Omega \int_{B(0,1)} |u(x+rh) - u(x)|^p q(x+h) dh q(x) dx \\
&\leq (u_\infty q_{\max})^{p-1} \omega_d^{p-1} r^{-p} \\
&\times \int_\Omega \int_{B(0,1)} |u(x+rh) - u(x)| q(x+h) dh q(x) dx \\
&\leq (u_\infty q_{\max})^{p-1} \omega_d^{p-1} r^{-p} \\
&\times \int_\Omega \int_{B(0,1)} \int_0^1 \left| \nabla u(x+trh).hq(x+rh) \right| dt dh q(x) dx \\
&\leq (u_\infty q_{\max})^{p-1} \omega_d^{p-1} r^{-p} \\
&\times \Big(\int_\Omega \int_{B(0,1)} \int_0^1 \left| \nabla u(x+trh).h \right| dt dh q^2(x) dx \\
&+ r \mathrm{Lip}(q) \int_\Omega \int_{B(0,1)} \int_0^1 \left| \nabla u(x+trh).h \right| dt dh q(x) dx \Big) \\
&\leq \frac{(u_\infty q_{\max} \omega_d)^{p-1} \sigma_\eta}{r^{p-1}} \left(\int_\Omega \left| \nabla u(x) \right| q^2(x) dx + C' r \int_\Omega \left| \nabla u(x) \right| q(x) dx \right).
\end{aligned}$$

□

4.7.3 The compactness property

Proposition 4.7.5. *Suppose that $(u_n)_{n\in\mathbb{N}}$ is a sequence bounded in $L^1(\mathbb{R}^d, q)$ and in $L^\infty(\mathbb{R}^d)$, let $\Omega \subset \mathbb{R}^d$ be as above and let $\Omega_n := \Omega \cap B(0, \tilde{R}_n)$ where $\tilde{R}_n$, as above, is such that*

$$r_n q_{\min}(\tilde{R}_n)^{-1} = o(1).$$

Suppose furthermore that $(TV_{r_n}(u_n; q))_{n\in\mathbb{N}}$ is bounded and that $\forall \delta > 0$, $\exists R_\delta > 0$, $\forall n \in \mathbb{N}$

$$TV_{r_n}(u_n, q, \overline{B(0, R_\delta)}^c) < \delta,$$

then we can extract a subsequence $(u_{\varphi(n)})_{n\in\mathbb{N}}$ converging in $L^1(\mathbb{R}^d, q^2)$ to some $u \in L^1(\mathbb{R}^d, q)$.

The proof of Proposition 4.7.5 draws from the argument of Theorem 3.1 in [2]. Nonetheless, several steps are affected by the addition of a weight (i.e., the density function q) and by the fact that the domain is now unbounded. These new considerations demand a more careful analysis at parts. We start with a few lemmas which we require in the proof of Proposition 4.7.5.

Lemma 4.7.6. *Let $\Omega \subset \mathbb{R}^d$. For every $u \in L^1(\mathbb{R}^d, q) \cap L^\infty(\mathbb{R}^d)$ and every $r > 0$, we have*

$$\begin{aligned}
&\int_\Omega \int_{\mathbb{R}^d} (\eta_r * \eta_r)(y)|u(x+y) - u(x)|q(x+y)q(x)dydx \\
&\leq 2\left(\int_\Omega \int_{\mathbb{R}^d} \eta_r(w)|u(x+w) - u(x)|q(x+w)q(x)dwdx + ru_\infty \mathrm{Lip}(q)\right) \\
&\leq 2rTV_r(u;\Omega^r) + 2ru_\infty \mathrm{Lip}(q).
\end{aligned}$$

Recall that given two functions $f, g : \mathbb{R}^d \to \mathbb{R}$, their convolution is defined as

$$\begin{aligned}
f * g : \mathbb{R}^d &\longrightarrow \mathbb{R} \\
x &\longmapsto \int_{\mathbb{R}^d} f(x-y)g(y)dy.
\end{aligned}$$

Proof. By the triangle inequality

$$\begin{aligned}
&\int_\Omega \int_{\mathbb{R}^d} (\eta_r * \eta_r)(y)|u(x+y) - u(x)|q(x+y)q(x)dydx \\
&\qquad \leq \int_\Omega \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \eta_r(z)\eta_r(y-z)|u(x+(y-z)) - u(x)|q(x+y)q(x)dzdydx \\
&\qquad + \int_\Omega \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \eta_r(z)\eta_r(y-z)|u((x+y)-z) - u(x+y)|q(x+y)q(x)dzdydx.
\end{aligned}$$

Denote the first integral on the RHS above by I_1 and the second integral by I_2. Using Fubini, we have

$$I_1 = \int_\Omega \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \eta_r(z)\eta_r(y-z)|u(x+(y-z)) - u(x)|q(x+y)q(x)dy\right) dzdx;$$

using the change of variable $w := y - z$, the compact support of η_r and the following inequality due to the Lipschitz continuity of q

$$q(x+y) \leq q(x+w) + r\mathrm{Lip}(q),\ \forall |w-y| < r,$$

we find

$$\begin{aligned}
I_1 \leq &\int_\Omega \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \eta_r(z)\eta_r(w)|u(x+w) - u(x)|q(x+w)q(x)dwdzdx \\
&+ ru_\infty \mathrm{Lip}(q)||\eta||^2_{L^1(\mathbb{R}^d,q)}.
\end{aligned}$$

Without loss of generality we are assuming that $||\eta||_{L^1(\mathbb{R}^d,q)} = 1$, hence (using Fubini again)

$$I_1 \leq \int_\Omega \int_{\mathbb{R}^d} \eta_r(w)|u(x+w) - u(x)|q(x+w)q(x)dwdx + ru_\infty \mathrm{Lip}(q).$$

The same holds for I_2, which proves the lemma.

□

Lemma 4.7.7. *Let $\Omega \subset \mathbb{R}^d$. We have*

$$r_n \int_\Omega \int_\Omega \eta_{r_n}(x-y)|u(x) - u(y)|q(x)dydx = o\left(TV_{r_n}(u,q,\Omega)\right).$$

Proof. We have

$$\begin{aligned} &r_n \int_\Omega \int_\Omega \eta_{r_n}(x-y)|u(x) - u(y)|q(x)dydx \\ &\leq (r_n q_{\min}(\tilde{R}_n + r_n)^{-1}) \int_\Omega \int_\Omega \eta_{r_n}(x-y)|u(x) - u(y)|q(y)q(x)dydx; \end{aligned}$$

by construction of $\tilde{R}_n$ and using the Lipschitz continuity of q, we find

$$r_n q_{\min}(\tilde{R}_n + r_n)^{-1} = o(1).$$

□

We also require the following result from [38].

Lemma 4.7.8 (cf, Theorem 10 in [38])**.** *Suppose that $(w_n)_{n\in\mathbb{N}}$ is a sequence bounded in $W^{1,1}(\mathbb{R}^d,\rho)$ and that $\forall \delta > 0,\ \exists R_\delta > 0,\ \forall n \in \mathbb{N}$,*

$$\int_{|x|>R_\delta} (|w_n(x)| + |\nabla w_n(x)|)\rho(x)dx < \delta,$$

then $(w_n)_{n\in\mathbb{N}}$ is relatively compact in $L^1(\mathbb{R}^d,\rho)$.

Proof of Proposition 4.7.5. We may find a non-negative smooth function φ with compact support, such that

$$\varphi \leq \eta * \eta \text{ and } |\nabla\varphi| \leq \eta * \eta.$$

Without loss of generality, we may assume that $\mathrm{supp}(\varphi) \subset \overline{B(0,1)}$ and that $\int_{\mathbb{R}^d} \varphi(x)dx = 1$. Define then, for $y \in \mathbb{R}^d$ and $r > 0$

$$\varphi_r(y) := r^{-d}\varphi(y/r) \text{ and } w_n(y) := \varphi_{r_n} * u_n(y).$$

Note that $(\varphi_{r_n})_{n\in\mathbb{N}}$ is a standard sequence of mollifiers. We claim that

$$||w_n - u_n||_{L^1(\mathbb{R}^d,q^2)} \to 0, \text{ as } n \to \infty,$$

that $(w_n)_{n\in\mathbb{N}}$ is bounded in $W^{1,1}(\mathbb{R}^d, q^2)$, and that for every $\delta > 0$, there exists $R_\delta > 0$ such that for every $r > 0$

$$\int_{|x|>R_\delta} |\nabla w_r(x)| q^2(x) < \delta.$$

In other words, we claim that the two sequences $(u_n)_{n\in\mathbb{N}}$ and $(w_n)_{n\in\mathbb{N}}$ are asymptotically equivalent in $L^1(\mathbb{R}^d, q^2)$, and that $(w_n)_{n\in\mathbb{N}}$ verifies the required conditions stated in the above Lemma 4.7.8, taking $\rho = q^2$, such that it is relatively compact in $L^1(\mathbb{R}^d, q^2)$. From these claims it follows that $(u_n)_{n\in\mathbb{N}}$ is relatively compact in $L^1(\mathbb{R}^d, q^2)$, as required.

It remains to prove the above claims. Denoting $w_r = w_{r_n} = w_n$ and $u_r = u_{r_n} = u_n$, we have

$$\begin{aligned}
\int_{\mathbb{R}^d} |w_r - u_r| q^2(x)dx &= \int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} \varphi_r(y)(u_r(x+y) - u_r(x)dy \right| q^2(x)dx \\
&\leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left|\varphi_r(y)\right| \left|u_r(x+y) - u_r(x)\right| q^2(x)dydx \\
&\leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left|\varphi_r(y)\right| \left|u_r(x+y) - u_r(x)\right| q(x+y)q(x)dydx \\
&+ r\mathrm{Lip}(q) \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left|\varphi_r(y)\right| \left|u_r(x+y) - u_r(x)\right| q(x)dydx.
\end{aligned}$$

Since $\varphi_r \leq \eta_r * \eta_r$, using Lemmas 4.7.6 and 4.7.7 and the boundedness of the functionals $(TV_n(u_n, q))_{n\in\mathbb{N}}$, the RHS above is $O(r)$. Hence, we have as $n \to \infty$

$$||w_n - u_n||_{L^1(\mathbb{R}^d, q^2)} \to 0.$$

This also shows that $(w_n)_{n\in\mathbb{N}}$ is bounded in $L^1(\mathbb{R}^d, q^2)$. Next, we have

$$\begin{aligned}
\int_{\mathbb{R}^d} |\nabla w_r(x)| q^2(x)dx &= \int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} \nabla\varphi_r(y) u_r(x+y)dy \right| q^2(x)dx \\
&= \int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} \nabla\varphi_r(y)(u_r(x+y) - u_r(x)) \right| q^2(x)dx \\
&\leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |\nabla\varphi_r(y)| \left|u_r(x+y) - u_r(x)\right| q^2(x)dx \\
&\leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left|\nabla\varphi_r(y)\right| \left|u_r(x+y) - u_r(x)\right| q(x+y)q(x)dydx \\
&+ r\mathrm{Lip}(q) \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left|\nabla\varphi_r(y)\right| \left|u_r(x+y) - u_r(x)\right| q(x)dydx.
\end{aligned}$$

Since this time

$$|\nabla\varphi_r| \leq r^{-1}\eta_r * \eta_r,$$

we deduce as above that the RHS above is $O(1)$.

Repeating the above argument with $\{|x| > R_\delta + r\}$ instead of $\mathbb{R}^d$, we find

$$\int_{|x|>R_\delta+r} \left|\nabla w_{r_n}(x)\right| q^2(x)dx \lesssim TV_{r_n}(u_n, q, \overline{B(0,R_\delta)}^c) < \delta,$$

where the implied constant above is independent of n, and the rest follows by Lemma 4.7.8.

□

4.8 Perimeter estimation on $\mathbb{R}^d$ using graph cut

As an application of the above generalisations to $\mathbb{R}^d$, we present a result on the estimation of perimeter using graph cut ($\mathrm{Per}(A,\Omega) := TV(\mathbb{1}_A,\Omega)$) for a set $A \subset \mathbb{R}^d$ with respect to some domain $\Omega \subset \mathbb{R}^d$ (possibly unbounded). This can be seen as a generalisation of the work in [65].

Theorem 4.8.1. *Let $p \geq 1$ and let $A \subset \mathbb{R}^d$. There exists $C_{p,d} > 0$ depending only on p and the dimension d, such that*

$$\mathbb{E}\left[\left|\overline{G}\mathrm{Cut}_n(Y_n,\Omega_n) - TV_{r_n}(\mathbb{1}_A,q,\Omega_n)\right|^p\right] \leq C_{p,d}(\max\left\{1,\mathrm{Per}(A,q,\Omega)\right\} f(n,r_n))^p,$$

where $Y_n = A \cap \mathcal{P}_n$ and $f(n,r_n) := \sqrt{nr_n}$, and where $\Omega_n := \Omega \cap B(0,\tilde{R}_n)$ is as above.

In particular if $\epsilon_n = \omega(\sqrt{nr_n})$, then a.s.

$$\left|G\mathrm{Cut}_n(Y_n,\Omega_n) - TV_{r_n}(\mathbb{1}_A,q,\Omega_n)\right| \leq \epsilon_n + o(n^{-1}).$$

We refer the reader back to Definition 24 for the definition of the graph cut of a set (see (4.15), and to (4.13) and (4.14) for its normalised versions $G\mathrm{Cut}_n(Y_n)$, respectively $\overline{G}\mathrm{Cut}_n(Y_n)$.

Proof. Note that this theorem is similar to Theorem 1.1 in [65]. Most of the steps of the proof given there follow identically here and we do not repeat them. We note the few changes that must be taken into account in our setting.

First of all, part of the proof of Theorem 1.1 in [65] relies on Lemma 2.1 proved later in the same paper. This lemma establishes that $TV_{r_n}(\mathbb{1}_A,D) \leq \sigma_(\mathbb{1}_A,D)$, where $A \subset D$, D is a bounded domain and the underlying measure is the uniform distribution. We already provided the required extension of this lemma to the case where the sampling density q is (not uniform and) supported on $\mathbb{R}^d$ in Lemma 2.11, in order to establish the limsup upper bound claimed in Section 2.7.2.

The second difference to note between the setting of [65] and the current setting, is that we are considering a homogenous Poisson point process $\mathcal{P}_n$ of intensity n instead of an i.i.d. sample X_n of size n. The main difference, as far as we are concerned with the impact on the proof of Theroem 1.1, is that the sampled set of points now has size $N \sim Po(n)$ instead of n. This explains why the control on the pth moments in the statement of the above theorem involves $\overline{G}\mathrm{Cut}_n$ instead of $G\mathrm{Cut}_n$.

With these two changes in mind, the estimates on the pth moments stated above follow as in the proof of Theorem 1.1 in [65].

Since we are rather interested to compare $G\mathrm{Cut}_n$ with TV_{r_n}, it suffices to have a good control on

$$\left|\overline{G}\mathrm{Cut}_n(Y_n) - G\mathrm{Cut}_n(Y_n)\right|,$$

and to use the estimates on the pth moments. This is what we shall do now. We show that if $\epsilon_n = \omega(\sqrt{nr_n})$, then a.s.

$$\left|\overline{G}\mathrm{Cut}_n(Y_n) - G\mathrm{Cut}_n(Y_n)\right| = o(n^{-1}).$$

Since $N \sim Po(n)$, a.s. there exists $w(n) = o(1)$ and a finite random variable n' such that for all $n \geq n'$

$$N = n + w(n).$$

This holds by a similar argument to the one used in the proof of the discrepancy-type results as discussed above, using the concentration of $N \sim \mathrm{Po}(n)$ given by the Chernoff-type bounds mentioned above, and using the Borel-Cantelli lemma.

We then have

$$N(N-1) = n(n-1) + w(n)(2n-1) + w(n)^2,$$

hence

$$\begin{aligned}\left|G\eta_r(u) - \overline{G}\eta_r(u)\right| &= \frac{(2n-1)w(n) + w(n)^2}{n^4 + o(n^4)} 2r^{-1}\eta_r(u) \\ &= O\left(w(n)n^{-1}G\eta_r(u)\right) \\ &= o(n^{-1}).\end{aligned}$$

□

Corollary 4.8.2. *For every $\Omega \subset \mathbb{R}^d$ and every $A \subset \mathbb{R}^d$, we have the following graph cut estimation of the perimeter of A with respect to Ω:*

$$\lim_{n\to\infty} G\mathrm{Cut}_{r_n}(A, q, \Omega_n) = \sigma_\eta TV(\mathbb{1}_A, q, \Omega) =: \sigma_\eta \mathrm{Per}(A, q, \Omega),$$

for instance with r_n satisfying the conditions mentioned in Theorem 4.6.2 (in fact, better conditions can be found for r_n).

4.9 Conclusion

While various consistency results are well understood on bounded domains (cf, Chapter 2), little is known about possible extensions on $\mathbb{R}^d$, or more generally on unbounded domains.

In this chapter, we investigated possible extensions of some of the key results used in the successful consistency results seen in Chapter 2 to the case where the sampling domain is $\mathbb{R}^d$. In this case, extra care must be put into choosing the sampling density. Depending on its properties, such as how fast it decays, one may obtain different results. We have proposed a similar setting to the one studied by Owada and Adler in [52]. In the case where the density has a light tail, we showed some extension to $\mathbb{R}^d$ of the discrepancy-type results (cf, Lemma 3.2 in [49]) seen in the previous chapters. Then, we studied some extensions of the Γ-convergence properties of some functionals. While the liminf porperty is shown to easily extend, one must add some extra restrictions in order to derive a limsup property, as well as a compactness property similar to the one used in the case of a bounded domain. In fact, the results obtained on the compactness property are not completely satisfactory, since we can only find a converging subsequence in $L^1(\nu^2)$ instead of $L^1(\nu)$ as needed to establish a.s. convergence results such as those of [49].

Chapter 5

Random geometric complexes

5.1 Introduction

A random geometric complex is a simplicial complex built on a random geometric graph. Topological properties of random geometric complexes have been investigated (e.g., [41, 11, 10, 9, 42]) as generalisations of properties of random geometric graphs, as studied by Penrose in [55]. See also [50, 51, 31, 32, 30, 22, 17, 4] for early works in this direction.

As observed in [41], the investigation of topological properties of random geometric complexes is motivated by applications in Topological Data Analysis and persistent homology (e.g., [71, 28, 20, 16, 13]). In particular, it is desirable to find asymptotic bounds on the expected Betti numbers (the expected ranks of the homology groups) under various settings, and to exhibit w.h.p. connectivity thresholds (asymptotic values of the bandwidth beyond which homology vanishes).

The two most commonly studied complexes are the Čech complex and the Vietoris-Rips (VR) complex. The Čech complex has the advantage of yielding a geometric interpretation via the Nerve Lemma (see [14] and 5.2.1), making it the more natural complex to study when points are sampled from a Riemanian submanifold $M \subset \mathbb{R}^d$ (e.g., [11, 10, 42]). The Vietoris-Rips complex, on the other hand, can be thought as a completely combinatorial complex - it is the clique complex of a random graph (also known as the flag complex) - and is the preferred choice from a computational point of view (since all the information of the complex is contained within the underlying graph). In other words, the Čech complex has a direct relationship to the geometry of the manifold we wish to recover via the Nerve Lemma, while the VR complex is a better choice for computational reasons.

Expected topological features of the Čech and the VR complex can be investigated for various phases of the bandwidth parameter r. As long as the bandwidth satisfies $r = O(n^{-1/d})$, i.e., that we are in the subcritical or critical phase, topological investigations of complexes can be done with fairly general assumptions on the sampling density, which choice does not affect greatly the asymptotic expected values of the Betti numbers (see [41, 9]). On the other hand, as observed in [55, 41, 9], a change in the sampling density may significantly alter the resulting connectivity threshold of a graph. In particular in the supercritical phase, i.e., $r = \omega(n^{-1/d})$, one generally restricts the density for instance to being compactly supported and bounded away from 0, in order to identify a connectivity threshold value for r (beyond which the homology groups vanish w.h.p.).

It is shown in [41] that for a uniform sampling density on a smooth convex body of $\mathbb{R}^d$, the expected Betti numbers of the Čech and the VR complex grow sublinearly in the supercritical phase. It is also shown that the homology of the Čech complex vanishes beyond the threshold value of $r \sim (\log n)^{1/d} n^{-1/d}$. In order to identify such a threshold value for the Čech complex, it suffices by the Nerve Lemma to identify a value beyond which the union of the balls $\cup_{x \in X_n} B(x, r)$ is contractible. For the Vietoris-Rips complex we cannot use the Nerve Lemma and can only show k-connectivity of the complex (see Definition 6.4 in and Theorem 6.5 in [41]), which is weaker.

Such results are classic and well known by now; they have been expanded in various ways. The following list of works is by no means exhaustive.
Bobrowski and Weinberger [11] and later Bobrowski and Oliveira [10] continued the investigation of homology of Čech complexes with more generality assumed on the sampling domain. They consider respectively the case where the domain is a torus and more generally when it is a compact and closed (without boundary) Riemannian manifold. The task in those settings is to identify upper and lower thresholds, such that if the bandwidth parameter r is below the lower threshold, then w.h.p. the homology of the Čech complex does not recover that of the underlying manifold, while if r is beyond the upper threshold it does so w.h.p.. The upper and lower threshold values have tight gaps and these sharp transitions of states are reminiscent of the well known sharp transition threshold value for the connectivity of a random graph, given by $\Lambda = \log n$ in the case of a random geometric graph and by $np = \log n$ in the case of a combinatorial graph. In fact the recent work of Bobrowski in [12], refined the sharpness of those threshold values to exhibit sharp transitions for the recovery of the kth homology group, occuring exactly at $\Lambda := n\omega_d r^d = \log n + (k-1)\log\log n$, where ω_d denotes the Euclidean volume of the unit sphere (d being the dimension of the manifold). The case of a compact Riemannian manifold with non-empty boundary was recently investigated in [42]. Upper and lower threshold values were provided as in the case of a closed compact manifold, but with a different scaling due to the presence of a non-empty boundary, complicating several

steps. Although the difference between the two thresholds remains small, no sharp transition threshold has yet been identified in this setting. The work in [42] is the content of the next section. It will be discussed in more detail there. In all of the above mentioned works, only the case of a uniform density was considered, with the extra requirements that it remains bounded and strictly positive on the domain. We note (as observed in [9]) that by compactness these results easily generalise to an arbitrary sampling density q satisfying the classic assumptions

$$0 < q_{\min} \leq q_{\max} < \infty,$$

where $q_{\min}$ and $q_{\max}$ denote respectively, as before, the inf and sup of the range of q in $\mathbb{R}_+$.

5.2 Geometric complexes

To motivate the definition of Čech and Vietoris-Rips complexes, we define a geometric graph as follows.

Definition 27 (Geometric graph)**.** *Given vertices $\mathcal{P} \subset \mathbb{R}^d$ and a radius $r > 0$ (also called bandwidth parameter), define the geometric graph $G(\mathcal{P}, r)$ where $\{x, y\} \subset \mathcal{P}$ is an edge if*

$$B(x, r/2) \cap B(y, r/2) \neq \emptyset.$$

Geometric complexes can be thought as generalisations of geometric graphs, where we not only take into considerations vertices and edges, but also triangles and higher dimensional simplices. As mentioned above, there are two common geometric complexes, which we now define.

Definition 28 (Čech complex)**.** *Given vertices $\mathcal{P} \subset \mathbb{R}^d$ and bandwidth parameter $r > 0$, define the Čech complex $\mathcal{C}(\mathcal{P}, r)$ to be the simplicial complex where for all $k \in [d]$, $[x_0, \ldots, x_k]$ is a k-face of the complex if*

$$\cap_{i=0}^{k} B(x_i, r) \neq \emptyset.$$

Definition 29 (Vietoris-Rips complex)**.** *Given vertices $X \subset \mathbb{R}^d$ and $r > 0$, define the Vietoris-Rips complex $R(\mathcal{P}, r)$ to be the clique complex of the graph $G(\mathcal{P}, r)$, i.e., for all $k \in \mathbb{N}$, $[x_0, \ldots, x_k]$ is a k-face of the complex if it is a clique in $G(\mathcal{P}, r)$, i.e., if*

$$\forall i_1 \neq i_2,\ B(x_{i_1}, r/2) \cap B(x_{i_2}, r/2) \neq \emptyset.$$

Those two complexes are closely related. In fact, it is immediate to verify the following inclusions:

$$\mathcal{C}(\mathcal{P}, r) \subset R(\mathcal{P}, 2r) \subset \mathcal{C}(\mathcal{P}, 2r).$$

For a tighter nested inclusion than above, see Jung's lemma in [40] and also [25].

While the VR complex is completely combinatorial, in that all the information is contained in the underlying graph (i.e., it is the clique complex of the graph), the Čech complex has the advantage of yielding a nice geometric interpretation via the following celebrated Nerve Lemma. Below, we phrase it in a way directly related to our purposes. A more general and well known topological result holds, originating in [14].

Lemma 5.2.1 (Nerve Lemma, [14]). *Let $\mathcal{P} \subset \mathbb{R}^d$, and let $\mathcal{B}(\mathcal{P}, r) := \{B(x, r) \mid x \in \mathcal{P}\}$. The Čech complex $\mathcal{C}(\mathcal{P}, r)$ is homotopy equivalent to $\mathcal{B}(\mathcal{P}, r)$.*

Recall the definition of a nerve.

Definition 30 (Nerve, [3]). *Given open sets $\{U_i \mid i \in I\}$ from a topological space $\mathcal{T}$, define the **nerve** as*

$$N := \{J \subset I \mid \cap_{j \in J} U_j \neq \emptyset\}.$$

The Nerve Lemma ([14]) gives general conditions for which the nerve, which is an abstract simplicial complex (this can be checked from the definition), captures the topology of the associated topological space. In our setting, the Čech complex $\mathcal{C}(\mathcal{P}, r)$ is the nerve of $\mathcal{B}(\mathcal{P}, r)$ (by definition) and the general conditions of the Nerve Lemma are satisfied (non-empty intersections of balls in $\mathcal{B}(\mathcal{P}, r)$) are contractible), the Nerve Lemma reads that

$$C(\mathcal{P}, r) \cong \mathcal{B}(\mathcal{P}, r),$$

as stated in Lemma 5.2.1, i.e., that the Čech complex is homotopy equivalent to its geometric realization. Hence the Nerve Lemma gives us a nice geometric interpretation for the Čech complex, and a simple criterion for the vanishing of homology of Čech complexes: the contractibility of the union of the balls $\cup_{x \in \mathcal{P}} B(x, r)$.

Chapter 6

Random geometric complexes on $\mathbb{R}^d$: decrackling the noise

6.1 Introduction

Here we focus our topological investigation to the supercritical phase, where the sampling domain is $\mathbb{R}^d$. A similar setting has already been carefully investigated in [1] and [52]. (Early works in this direction, on compact Riemannian submanifolds of $\mathbb{R}^d$, can be found in [50, 51].) In [1] the authors are interested in the effect of unbounded noise on homology recovery of a bounded domain, building on [50, 51], which considered the case of homology recovery for a low dimensional submanifold sampled with Gaussian noise. In [1] the authors extend these previous works to consider a larger class of probability distributions than just a Gaussian. They take the low dimensional submanifold to be just the origin in the higher dimensional ambient space $\mathbb{R}^d$, and model the unbounded noise by sampling densities with various decays on $\mathbb{R}^d$ (*power law*, *exponential*, *Gaussian*). Thus, they identify sequences of radii $(R_n^c)_{n\in\mathbb{N}}$ depending on the density, such that homology vanishes in $B(0, R_n^c)$ but not necessarily outside the ball, in which case they call this phenomenon *topological crackle*. In practice if topological crackle occurs for a given density, then cycles still form far away from the origin, even for $r_n >> (\log n)^{1/d} n^{-1/d}$, i.e., the homology of the noise does not vanish and we cannot hope to recover the true homology from a noisy sample. In [1] the authors show that crackling occurs for the power law and the exponential decay, but not for the Gaussian. In [52] the authors pursue the work initiated in [1] with more generality assumed on the density, using tools from Extreme Value Theory. They exhibit finite sequences of annuli splitting $\mathbb{R}^d$ (as in [1]), where each annulus generates homology of a certain degree. Their analysis confirms that topological crackle generally occurs (in which case, as

mentioned above, we cannot recover the true homology from a noisy sample) unless the density has superexponential decay, a generalisation of the Gaussian case analyzed in [1] (see Theorem 4.5 in [52]).

The goal of this chapter is to show that well-chosen variable bandwidth constructions can be used to remove the crackling phenomenon observed in [1, 52] in a non-trivial way. While suitable conditions on the bandwdith parameter in order to avoid the crackling phenomenon can only be found in the case where the sampling density has Gaussian or more generally superexponential decay following the results in [1, 52], our constructions will allow us to identify such conditions on the bandwidth parameter for any light tail distribution. Furthermore if the density has superexponential decay, we present a construction which can weaken the asymtpotic conditions on the bandwith parameter, as asked by the authors of [52]. More generally, with a variable bandwidth construction one may find suitable conditions to remove the crackling phenomenon observed in [1, 52] for any radial density, however as we will see the conditions obtained for the bandwidth in the case of a heavy tail distribution are trivial and of limited interest.

6.2 Variable bandwidth constructions

Variable bandwidth constructions are well known and have proven to be useful in the past. They have been successfully used in various topics, for instance in non-parametric kernel density estimations (e.g., [39, 60, 58]) and spectral clustering (e.g., [70]).

There are various ways one can define a so-called self-tuning or variable bandwidth geometric graph. One way is the following.

Definition 31 (Variable bandwidth geometric graph)**.** *Given vertices $X \subset \mathbb{R}^d$, a radius $r > 0$, and a scaling function $\varphi : \mathbb{R}^d \to \mathbb{R}^+$, define the the variable bandwidth geometric graph $G(X; r\varphi)$ to be the graph with vertex set X and where $\{x, y\} \subset X$ is an edge if*

$$B(x, r\varphi(x)/2) \cap B(y, r\varphi(y)/2) \neq \emptyset.$$

Note that the associated affinity matrix in this case is not given as in [70], by

$$\mathbb{1}\left(|x-y|^2 \leq r^2\varphi(x)\varphi(y)\right).$$

However, it is easy to verify that the above variable bandwidth geometric graph is contained in the graph induced by the affinity matrix

$$\mathbb{1}\left(|x-y| \leq r\max\{\varphi(x), \varphi(y)\}\right).$$

Likewise, the definitions of the Čech and the VR complex naturally extend to a variable bandwidth setting.

Definition 32 (Variable bandwidth Čech complex)**.** *Given vertices $X \subset \mathbb{R}^d$, a radius $r > 0$ and a scaling $\varphi : \mathbb{R}^d \to \mathbb{R}^+$, define the Čech complex $\mathcal{C}(X, r\varphi)$ to be the simplicial complex where for all $k \in [d]$, $[x_0, \dots, x_k]$ is a k-face of the complex if*

$$\cap_{i=0}^{k} B(x_i, r\varphi(x_i)) \neq \emptyset.$$

Definition 33 (Variable bandwidth VR complex)**.** *Given vertices $X \subset \mathbb{R}^d$, a radius $r > 0$ and a scaling $\varphi : \mathbb{R}^d \to \mathbb{R}^+$, define the Vietoris-Rips complex $R(X, r\varphi)$ to be the clique complex of the graph $G(X, r\varphi)$, i.e., for all $k \in \mathbb{N}$, $[x_0, \dots, x_k]$ is a k-face of the complex if it is a clique in $G(X, r\varphi)$, i.e., if*

$$\forall i_1 \neq i_2,\ B(x_{i_1}, r\varphi(x_{i_1})/2) \cap B(x_{i_2}, r\varphi(x_{i_2})/2) \neq \emptyset.$$

Picking a suitable scaling for φ is not straightforward and different choices may yield very different results. A classic scaling (e.g., [58, 60, 39, 70, 8]) generally consists in choosing $\varphi \sim \rho := q^{-1/d}$ on the sampled points, where q is the sampling density and d the dimension of the ambient space. Indeed it is well known that $\rho(x_i)$ can be asymptotically approximated in practice by $\left|x_i - x_i^{(k)}\right|$, where $x_i^{(k)}$ is the k-nearest neighbour of x_i from the sampled points X. Furthermore, one easily verifies that this local scaling makes the measure of the balls (i.e., the probability for a random point to lie in that ball) roughly constant everywhere in the domain, for fixed bandwidth r, a property which is characteristic of the uniform distribution and which may facilitate calculations in various settings.

A first naive approach to our current problem of investigating random geometric complexes on $\mathbb{R}^d$ thus consists in following in the steps of [58, 60, 39, 70, 8], picking similarly $\varphi \sim \rho$. Note however that in the case where the domain is unbounded, this scaling choice presents major drawbacks. Indeed, it is not hard to see that such a scaling will have the effect of making the radii of balls far from the origin grow too fast, such that beyond a certain radius value, the balls will all contain the origin, at which point our construction has limited interest. In particular, this trivially implies for instance that the union of the balls thus created is contractible and that the underlying graph is connected.

Motivated by numerous previous works ([58, 60, 39, 70, 8]) using variable bandwidth constructions with such a scaling (this scaling is classic and perfectly reasonable and judicious in the case where the sampling domain is bounded, as discussed above), we have first approached the problem of random geometric complexes on $\mathbb{R}^d$ with this naive scaling choice. Thus, we went on to revisit the classic argument of Kahle in [41] in the case of a variable bandwidth construction and an arbitrary radial density supported on $\mathbb{R}^d$. In particular we could establish a result analogous to Theorem 6.5 in [41], giving explicit conditions on the bandwith parameter r for the k-connectivity of a variable bandwidth

complex on $\mathbb{R}^d$. As mentioned above however, such a scaling choice has limited interest, and one may indeed check that the conditions we obtain on r imply that beyond a certain radius, the radius $r\rho(x_i)$ at which we grow our balls eventually becomes larger than $|x_i|$. We present this naive approach in the appendix below, following the work of Kahle in [41].

Below, we discuss the cases where a more advisable choice of scaling can be made, in order to "decrackle" the noise in a non-trivial way.

We will usually denote the vertices by X_n, when $|X_n| = n$. Furthermore, if the vertices are randomly sampled, we will refer for instance to $G(X_n; r)$ and $G(X_n; r\varphi)$ as random geometric graphs and to $R(X_n; r)$ and $R(X_n; r\varphi)$ as random complexes.

Our setting relates to and follows closely the work in [52]. The same setting was already introduced in Chapter 4.

6.3 Random Čech complexes on $\mathbb{R}^d$: decrackling the noise

6.3.1 Introduction

As mentioned in the introduction of this section, the above result only has limited interest in the sense that the chosen scaling $\varphi = \rho = q^{-1/d}$ is naive, yielding balls far from the origin with radii so large that they necessarily contain the origin, in which case the graph thus constructed is not sparse enough and most of the properties that we seek to establish will trivially follow for this reason. We now show how non-trivial results can be obtained from a variable bandwidth construction with a well-chosen scaling. To do this, we relate to and follow closely the work in [52]. There, sequences of radii $(R_n^c)_{n\in\mathbb{N}}$ and $(\overline{R_n})_{n\in\mathbb{N}}$ are more or less explicitly given, depending on the density, such that the union of the balls which centre lie in $B(0, R_n^c)$ is contractible (in fact covers the ball) w.h.p., while the probability to find points outside the ball $B(0, \overline{R_n})$ tends to 0. Furthermore, densities with superexponential decay are showed to satisfy the extra property that $\overline{R_n} - R_n^c = o(r_n)$, under specified asymptotic conditions on the bandwidth parameter r_n of the graph, such that in fact the union of the balls $\cup_{x\in\mathcal{P}_n} B(x, r_n)$ (not just those with centre contained in $B(0, R_n^c)$) is contractible, in which case by the Nerve Lemma, the associated Čech complex (that of the noise) vanishes, i.e., such that eventually there is no topological crackle. The elementary, yet key observation to keep, is that whenever the difference $\overline{R_n} - R_n^c$ decays faster than the bandwidth parameter, then one can deduce that the union of the balls is contractible hence that the homology of Čech complexes of the noise vanishes and that there is no topological crackle (as $n \to \infty$). This property, as explained above, is unfortunately showed in [52] to only happen if the sampling density has superexponential decay. In this section, we show how well-chosen variable bandwidth constructions can allow us to extend the class of densities for which topological crackle does not happen. We call this *decrackling the noise*. If the density has a heavy tail (cf, definition in Section 3 in [52]), then there is no hope to find any result much more interesting than the one obtained in the above "naive approach". This is because the radii $\overline{R_n}$ and R_n^c do not have the same asymptotic order, such that in fact $\overline{R_n} - R_n^c \sim \overline{R_n}$. This can easily be verified, for instance in the case of the power law density

$$q(x) = \frac{1}{1 + |x|^\alpha},$$

taking $r_n \equiv 1$ and referring to the obtained asymptotic values for $\overline{R_n}$ and R_n^c in [1]. Thus, any variable bandwidth construction will have to be such that $r_n\varphi(R_n^c) \sim \overline{R_n}$ if we wish for the homology of the noise to vanish, such that the balls with centre near the radius R_n^c contain the origin, which yields a trivial construction of limited interest, as discussed above.

In the case of light tail densities on the other hand, the key fact is that there

$\overline{R_n} \sim R_n^c$, hence in particular their difference grows much slower than the radii themselves, i.e.,

$$\overline{R_n} - R_n^c = o(R_n^c).$$

This creates some room for a well-chosen scaling φ such that

$$\overline{R_n} - R_n^c = o(r_n \varphi(R_n^c)),$$

while also

$$r_n \varphi(R_n^c) = o(R_n^c).$$

This last condition is to make sure that the graph remains non-trivial (that the radii of the balls far from the origin don't grow sufficiently fast to contain the origin).

6.3.2 Outline of the results

Below, we present first the case where the light tail density has subexponential or exponential decay and show that a variable construction can be found to effectively decrackle the noise in a non-trivial way (i.e., satisfying the conditions described above).

Next, we investigate the case of a density with superexponential decay. Even though this case is already known to be exempt from crackle, the authors in [52] asked whether the asymptotic conditions given on r_n could be improved from

$$\frac{a \circ \psi^{\leftarrow}(\log n) \log\log n}{r_n} = o(1)$$

to merely the condition already given in the other light tail cases:

$$\frac{a \circ \psi^{\leftarrow}(\log n)}{r_n} = o(1),$$

where recall that $q(x) \sim e^{-\psi(|x|)}$ and $a(z) = \frac{1}{\psi'(z)}$, $z \in \mathbb{R}_+$. While this question remains open in the case of the classic "constant" bandwidth construction investigated in [52], we will see that a well-chosen variable bandwidth construction can indeed allow us to weaken the asymptotic condition imposed on r_n as asked above.

6.3.3 Preliminaries

Recall the setting in [52], described in details above in this thesis, which we follow here as well. We suppose that our light tail sampling density satisfies

$$q(z) \sim e^{-\psi(|z|)},$$

where $\psi \in RV_v$ for some v. If $v < 1$ we say that the density has subexponential decay, if $v = 1$ that it has exponential decay, and if $v > 1$, we say that it has

superexponential decay. Recall also, that we define $a := 1/(\psi)'$ on $\mathbb{R}_+$.

Let us start with some preliminary observations, useful for the later parts of the argument.

Lemma 6.3.1. *It always holds that*

$$a \circ \psi^{\leftarrow}(z) = (\psi^{\leftarrow})'(z),$$

and that

$$\lim_{z\to\infty} \frac{a \circ \psi^{\leftarrow}(z) \log(z)}{\psi^{\leftarrow}(z)} = 0.$$

If $\psi \in RV_v$ with $v \leq 1$, then

$$\lim_{z\to\infty} a \circ \psi^{\leftarrow}(z) \log z = \infty.$$

Proof. Note that

$$1 = (\psi \circ \psi^{\leftarrow})'(z) = (\psi^{\leftarrow})'(z)\psi'(\psi^{\leftarrow}(z)),$$

hence

$$a \circ \psi^{\leftarrow}(z) = (\psi'(\psi^{\leftarrow}(z)))^{-1} = (\psi^{\leftarrow})'(z).$$

We have $\psi \in RV_v$ for some $v \in \mathbb{R}$, so $\psi^{\leftarrow} \in RV_{1/v}$ and

$$a \circ \psi^{\leftarrow} = (\psi^{\leftarrow})' \in RV_{1/v-1}.$$

The observations made thus far were already alluded to (without proof) in [52]. It then follows from the above, that

$$\frac{a \circ \psi^{\leftarrow}}{\psi^{\leftarrow}} \in RV_{-1},$$

hence in particular

$$\lim_{z\to\infty} \frac{a \circ \psi^{\leftarrow}(z) \log(z)}{\psi^{\leftarrow}(z)} = 0.$$

Finally, if $v \leq 1$ then $1/v - 1 \geq 0$ and as we have just seen, $a \circ \psi^{\leftarrow} \in RV_{1/v-1}$, hence

$$\lim_{z\to\infty} a \circ \psi^{\leftarrow}(z) \log z = \infty.$$

□

Let us define the sequences of radii of interest $(R_n^c)_n$ and $(\overline{R_n})_n$, alluded to in the introduction. We follow the setup in [52].

Definition 34. *Let $R_n^c := \psi^{\leftarrow}(A_n)$, where*

$$A_n := \log n + d \log r_n - \log\log r_n^{-1} \psi^{\leftarrow}(\log n) - \delta$$

and δ satisfies

$$d - e^{\delta} g^d C < 0;$$

and let $\overline{R_n} := \psi^{\leftarrow}(B_n)$, where

$$B_n := \log n + (d-1) \log \psi^{\leftarrow}(\log n) + \log a \circ \psi^{\leftarrow}(\log n) + \log\log n.$$

Lemma 6.3.2. *If* $\log r_n = o(\log n)$*, then*

$$\lim_{n\to\infty} \mathbb{P}\left(B(0, R_n^c) \subset \bigcup_{\mathcal{P}_n \cap B(0,R_n^c)} B(x, r_n) \right) = 1.$$

This follows as in the proof of Theorem 4.5 in [52].

Lemma 6.3.3. *We have*

$$\lim_{n\to\infty} \mathbb{P}\left(\mathcal{P}_n \cap B(0, \overline{R_n}) = \emptyset \right) = 1.$$

Note that this lemma is similar to part of the content of Theorem 4.5 in [52]. We provide the proof as we have assumed slightly weaker assumptions on a. The proof remains mostly similar to the argument provided in [52].

Proof. To ease notation, let us denote $R := \overline{R_n}$ for this proof.
Note that

$$\mathbb{P}\left(\mathcal{P}_n \cap B(0, R) = \emptyset \right) = \exp\left(-n \int_{|x|\geq R} q(x)dx \right);$$

we show $n \displaystyle\int_{|x|\geq R} q(x)dx \to 0$ as $n \to \infty$. Recalling that q is radial, this integral can be written as

$$\begin{aligned} n \int_{|x|\geq R} q(x)dx =& s_{d-1} n a(R) (R)^{d-1} q(R) \\ & \times \int_0^\infty \left(1 + \frac{a(R)}{R} z\right)^{d-1} \frac{q(R + a(R)z)}{q(R)} dz, \end{aligned}$$

where s_{d-1} denotes the surface area of the $(d-1)$ dimensional unit sphere in $\mathbb{R}^d$.

Let us estimate the integral on RHS, as $n \to \infty$. By the mean value theorem, there exists t between R and $R + a(R)z$ such that

$$\psi(R) - \psi(R + a(R)z) = -(\psi)'(t)(a(R)z).$$

By the preliminary observations made above on a we have $a(R) = o(R)$, so $t \sim R$ and by the regular variation of a

$$(\psi)'(t) = a(t)^{-1} \sim a(R)^{-1};$$

hence, as $n \to \infty$,

$$\psi(R) - \psi(R + a(R)z) \to -C'z,$$

for some constant $C' > 0$, from which we find by the dominated convergence theorem that the integral on the RHS above converges as $n \to \infty$ to

$$\int_0^\infty e^{-C'z} dz < \infty.$$

It remains to check that the remaining factor on the RHS above tends to 0, as $n \to \infty$. And indeed we have, as in the proof of Theorem 4.5 in [52], that

$$na(R)(R)^{d-1}e^{-\psi(R)} \sim (\log n)^{-1}.$$

□

6.3.4 Subexponential or exponential decay: decrackling the noise

From now on, let $(r_n)_{n\in\mathbb{N}}$ be a regularly varying sequence decaying to 0 and such that

$$\frac{a \circ \psi^{\leftarrow}(\log n) \log\log n}{r_n \psi^{\leftarrow}(\log n)} = o(1).$$

Such a choice is feasible by the preliminary observations in Lemma 6.3.1.

Lemma 6.3.4. *It follows, from the above assumptions on r_n, that for n sufficiently large*

$$-\log(r_n) < \log\log n.$$

Proof. By the above preliminary observations, since here $\psi \in RV_v$ with $v \leq 1$, we must have

$$a \circ \psi^{\leftarrow}(\log n) \log\log n = \omega(1).$$

Using the asymptotic conditions assumed for r_n, we then have

$$\begin{aligned}(\psi^{\leftarrow})^{-1}(\log n) &= o\left(\frac{a \circ \psi^{\leftarrow}(\log n) \log\log n}{\psi^{\leftarrow}(\log n)}\right) \\ &= o(r_n),\end{aligned}$$

from which it follows, since $r_n = o(1)$, that for n sufficiently large

$$\log\log n \sim \log \psi^{\leftarrow}(\log n) > -\log r_n.$$

□

In particular $\log r_n = o(\log n)$ is still true, so the assumptions of Lemma 6.3.2 hold.

Theorem 6.3.5 (Subexponential or exponential decay: decrackling the noise)**.** *Suppose that $(r_n)_{n\in\mathbb{N}}$ satisfies the above asymptotic conditions, then*

$$\lim_{n\to\infty} \mathbb{P}\left(\bigcup_{x\in\mathcal{P}_n} B(x, r_n\psi^{\leftarrow}(\log(e+\rho(x)))) \text{ is contractible}\right) = 1.$$

Note that by regular variation $\psi^{\leftarrow}(\log(\rho(R_n^c))) \sim \psi^{\leftarrow}(\log n) \sim \overline{R_n}$, and since $r_n = o(1)$, we have

$$r_n \psi^{\leftarrow}(\log(\rho(R_n^c))) = o(\overline{R_n})$$

as desired (cf, above discussion in the introduction section).

The theorem follows from the lemma below.

Lemma 6.3.6. *With the same assumptions on $(r_n)_{n\in\mathbb{N}}$ as above, we have*

$$\overline{R_n} - R_n^c = o(r_n \psi^{\leftarrow} \log(\rho(R_n^c))).$$

Proof. Using the assumptions on r_n, we find by direct computations

$$\begin{aligned} \log(\rho(R_n^c)) &\sim \log n + \log r_n - \log\log(r_n^{-1}\psi^{\leftarrow}(\log n)) \\ &\sim \log n, \end{aligned}$$

hence by the regular variation properties of $\psi^{\leftarrow}$,

$$\psi^{\leftarrow}(\log(\rho(R_n^c)))) \sim \psi^{\leftarrow}(\log n).$$

By the mean value theorem, there exists $t_n \in [A_n, B_n]$ such that

$$\overline{R_n} - R_n^c = (\psi^{\leftarrow})'(t_n)(B_n - A_n),$$

and ($A_n \sim \log n$ and $B_n \sim \log n$) $\Rightarrow t_n \sim \log n$. Using the above preliminary observations, we then find

$$(\psi^{\leftarrow})'(t_n) \sim (\psi^{\leftarrow})'(\log n) = a \circ \psi^{\leftarrow}(\log n).$$

Thus

$$(r_n \psi^{\leftarrow} \log(\rho(R_n^c)))^{-1}(\overline{R_n} - R_n^c) \sim \frac{a \circ \psi^{\leftarrow}(\log n)}{r_n} \frac{B_n - A_n}{\psi^{\leftarrow}(\log n)}.$$

By Lemma 6.3.4, we know that for n sufficiently large

$$-\log r_n < \log\log n,$$

hence

$$B_n - A_n \sim \log\log n - \log r_n \sim \log\log n.$$

Therefore

$$(r_n \psi^{\leftarrow} \log(\rho(R_n^c)))^{-1}(\overline{R_n} - R_n^c) \sim \frac{a \circ \psi^{\leftarrow}(\log n)\log\log n}{r_n \psi^{\leftarrow}(\log n)} = o(1).$$

□

Combining Lemmas 6.3.2, 6.3.3 and 6.3.6, Theorem 6.3.5 holds as follows.

Proof of Theorem 6.3.5. For all $x \in \mathbb{R}^d$ note that

$$r_n < r_n \psi^{\leftarrow}(\log(e + \rho(x)));$$

hence for $z \in B(0, \overline{R_n})$, using Lemmas 6.3.6 and 6.3.2, and

$$dist(z, \mathcal{P}_n) \leq dist(z, B(0, R_n^c)) + dist(B(0, R_n^c), \mathcal{P}_n),$$

we deduce that with probability going to 1 as $n \to \infty$, there exists $x \in \mathcal{P}_n$ such that

$$|x - z| < r_n \psi^{\leftarrow}(\log(e + \rho(x))).$$

Combined with Lemma 6.3.3, we have with probability going to 1 as $n \to \infty$

$$\mathcal{P}_n \subset B(0, \overline{R_n}) \subset \bigcup_{x \in \mathcal{P}_n} B(x, r_n \psi^{\leftarrow}(\log(e + \rho(x)))),$$

which gives the theorem.

□

6.3.5 Superexponential decay

In the case where $\psi \in RV_v$ and $v > 1$, so that the sampling density has super-exponential decay, the authors in [52] showed that the homology of the noise vanishes provided r_n satisfies the following asymptotic condition

$$\frac{a \circ \psi^{\leftarrow}(\log n) \log\log n}{r_n} = o(1).$$

They asked whether this condition could be weakened to merely

$$\frac{a \circ \psi^{\leftarrow}(\log n)}{r_n} = o(1).$$

Here we show that under a suitable variable bandwidth construction, one may weaken the asymptotic conditions for r_n as asked above. Namely, we show the following theorem.

Theorem 6.3.7. *Suppose that*

$$\frac{a \circ \psi^{\leftarrow}(\log n)}{r_n} = o(1),$$

then

$$\lim_{n \to \infty} \mathbb{P}\left(\bigcup_{x \in \mathcal{P}_n} B(x, r_n \mathcal{L}(\rho(x))) \text{ is contractible} \right) = 1,$$

where $\mathcal{L}(z) := \log(e + \log(1 + z))$.

Proof. We find as before by direct computations,

$$\log\log\rho(R_n^c) \sim \log\log n.$$

As in the proof of Theorem 6.3.5, by the mean value theorem, there exists $t_n \sim \log n$ such that

$$\overline{R_n} - R_n^c = (\psi^{\leftarrow})'(t_n)(B_n - A_n).$$

We have as before

$$(\psi^{\leftarrow})'(t_n) \sim (\psi^{\leftarrow})'(\log n) = a \circ \psi^{\leftarrow}(\log n),$$

hence

$$(r_n \log\log\rho(R_n^c))^{-1}(\overline{R_n} - R_n^c) \sim \frac{a \circ \psi^{\leftarrow}(\log n)}{r_n} \frac{B_n - A_n}{\log\log n}.$$

The first factor on the RHS above tends to 0 as $n \to \infty$, by assumptions. This also implies that for n sufficiently large

$$\log a \circ \psi^{\leftarrow}(\log n) < \log r_n < 0,$$

hence

$$-\log r_n \lesssim \log\log n.$$

Thus, we have

$$\frac{B_n - A_n}{\log\log n} \lesssim \frac{\log\log n - \log r_n}{\log\log n} = O(1).$$

Wrapping up, we have just showed that

$$\overline{R_n} - R_n^c = o(r_n \log\log(\rho(R_n^c))),$$

from which it follows as in Theorem 6.3.5, that

$$\lim_{n\to\infty} \mathbb{P}\left(\bigcup_{x\in\mathcal{P}_n} B(x, r_n\mathcal{L}(\rho(x))) \text{ is contractible} \right) = 1.$$

□

6.4 Conclusion

In this chapter, we extended some of the work done in [52], investigating conditions under which homology of noise on $\mathbb{R}^d$ may vanish (which is the case when the union of the balls is contractible). In [52] it was shown that noise introduces non-vanishing homology in general, what the authors call topological crackle, unless the sampling density has superexponential decay.

We showed how some well-chosen variable bandwidth constructions allow us to *decrackle* the noise for light tail densities, in a non-trivial way. In the case of a density with superexponential decay, which is already known to be exempt from crackle, a well-chosen construction also allows one to weaken the asymptotic conditions on the bandwidth parameter, to satisfy some conditions asked by the authors in [52].

Chapter 7

Random Čech complexes on compact manifolds with boundary

The goal of this chapter is to investigate the impact on homology recovery when sampling from a compact manifold with non-empty boundary.

The content of this chapter follows the argument developed in [42], a joint work with Ulrike Tilmann and Oliver Vipond.

This argument itself builds from previous works already mentionned above: [11] which studied homology recovery when sampling from a torus, and [10] which generalised [11] to compact and closed (empty boundary) Riemannian manifolds.

7.1 Outline of the main results

The main result of this chapter is the following theorem.

Theorem 7.1.1. *Let $d \geq 2$ and let $M \subset \mathbb{R}^d$ be a compact Riemannian manifold with smooth non-empty boundary. Let $\Lambda := n\omega_d r^d$, let $w(n) \to \infty$ arbitrarily slowly, and let $\mathcal{P}_n$ be a homogenous Poisson point process with intensity of n, sampled with respect to the uniform distribution on M. For every $k \in [d-1]$ we have*

$$\lim_{n\to\infty} \mathbb{P}\left(H_k(\mathcal{C}(n,r)) \cong H_k(M)\right) = \begin{cases} 1 \text{ if } \Lambda = (2-2/d)\log n + 2k\log\log n + w(n), \\ 0 \text{ if } \Lambda = (2-2/d)\log n + 2(1/d-3)\log\log n - w(n). \end{cases}$$

Similarly to the results obtained in [11, 10], this theorem yields upper and lower threshold values for the bandwidth parameter r_n and for each $k \in [d-1]$,

such that if r is beyond the upper threshold value, the kth homology group of the Čech complex is isomorphic to that of the underlying manifold with probability tending to 1, while if r is below the lower threshold value, this probability tends to 0. Similarly to when studying the connectivity of random geometric graphs or the homology of compact manifolds without boundary, we obtain a tight gap between the upper and the lower threshold, which both behave asymptotically like $(2-2/d)\log n$. We note however, that there is still some room in the $\log\log n$ coefficients, and that it is not known yet whether a sharp transition occurs as it does when the manifold has empty boundary (see [12]).

The upper threshold for Λ, beyond which the above probability tends to 1, is established using an upper bound estimate on the expected kth Betti numbers of the Čech complex in terms of $\beta_k(M)$. Namely, we will show the following upper bound estimate.

Proposition 7.1.2. *Let $M \subset \mathbb{R}^d$ ($d \geq 2$) be a compact manifold with boundary. Let $n \to \infty$ and $r, r_0 \to 0$ such that $\Lambda \to \infty$, $\Lambda_{r_0} := n\omega_d r_0^d \to 0$, $\Lambda_{r_0} r_0^2 \to 0$ and $r_0 \geq r(\omega_d/\kappa(1+|\log r|))^{1/d}$, for some constant κ. For every $k \in [d-1]$ we have*

$$\mathbb{E}[\beta_k] \leq \beta_k(M) + O\left(n\Lambda^k e^{-\Lambda}, n^{1-1/d}\Lambda^k e^{-1/2\Lambda}\right).$$

To that end, we will establish the following upper bound on the expected number of critical points of the distance function.

Lemma 7.1.3. *Let $r_0 = o(1)$ and $r = o(r_0)$, suppose that as $n \to \infty$, $\Lambda \to \infty$, $\Lambda_{r_0} r \to 0$ and $\Lambda_{r_0} r_0^2 \to 0$ where $\Lambda_{r_0} := \omega_d n r_0^d$. Then for all $k \in [d-1]$*

$$\mathbb{E}[C_k(r, r_0)] = O\left(n^{1-1/d}\Lambda^{k-1}e^{-\Lambda/2}, n\Lambda^{k-1}e^{-\Lambda}\right).$$

Likewise the lower threshold for Λ, below which the above probability tends to 0, is established using a lower bound estimate on the kth Betti numbers of the complex. We will show

Proposition 7.1.4. *Let $M \subset \mathbb{R}^d$ ($d \geq 2$) be a compact manifold with boundary. Let $n \to \infty$ and $r \to 0$ such that $\Lambda \to \infty$ and $\Lambda r^2 \to 0$.*
There exists $\alpha := 1/2 + O((\log n)^{-1})$ such that for all $k \in [d-1]$, we have w.h.p.

$$\beta_k = \Omega\left(n\Lambda^{k-2}e^{-\alpha\Lambda} r (\log n)^{-(k+1)}\right).$$

Note that this result yields a lower bound on β_k instead of $\mathbb{E}[\beta_k]$ as in the upper bound. This will be made possible, analogously to [10], by showing some second moment result for the lower threshold (cf, Section 8 in [10] and Section 8 in [42], and the discussion in the section on the second moment calculations below). In the upper bound case, we can instead use directly Markov's inequality (a first moment inequality).

Throughout, a common strategy to address the issue of the presence of a non-empty boundary, consists in viewing the manifold M embedded in its double manifold

$$DM := M \cup_{\partial M} M',$$

where two copies of M are glued together at the boundary. Equivalently,

$$DM := M \times \{0, 1\}/\sim,$$

where $(x, 0) \sim (x, 1)$ if $x \in \partial M$. The double manifold is, by construction, closed. This provides us with a natural bridge, facilitating the extension of various results holding on closed manifolds to the case of manifolds with non-empty boundaries. In fact, we could just as well view M as embedded in a larger closed manifold, not necessarily DM.

Let us start with a a few preliminary results and necessary set ups, which will be needed in the later arguments.

7.2 Riemannian approximations

Here, we briefly review some basic Riemannian geometry tools and discuss some elementary Riemannian approximations results which are used later. We refer to [44] for a more thorough introduction to Riemannian geometry. We also point at [43, 56] as other references where the results asserted in this section can be found.

Throughout, we consider a smooth Riemannian manifold (M, g) of dimension d, where M is smooth and compact and the metric g is smooth. This means that at every $p \in M$, there is a smoothly varying inner product $g_p : T_pM \times T_pM \to \mathbb{R}_+$, where T_pM denotes the tangent space of M at p (thus inducing a norm on T_pM). This metric can be used to define the length of a path on M and we denote, for $p_1, p_2 \in M$, the shortest path length between them to be $\rho_M(p_1, p_2)$. This defines a natural distance on M, which we implicitly refer to when mentioning distances or balls on M.

For $p \in M$, let $\exp_p : T_pM \to M$ be the Riemannian exponential map. The exponential map $\exp_p$ is a local diffeomorphism, in the sense that for every $p \in M$, we can find a sufficiently small radius $R_p > 0$ (cf, injectivity radius) such that $\exp_p$ restricted to $B_{E_d}(0, R_p) \subset T_pM$ is a diffeomorphism. Consequently, the exponential map can be used to define normal coordinates $(x^1, \ldots, x^d)$ locally around p: the geodesic normal coordinates at p. Under these coordinates, the Riemannian metric can be written (using the Einstein notation for sums) as

$$g = g_{ij} dx^i \otimes dx^j,$$

with

$$g_{ij} := \delta_{ij} + \frac{1}{3} R_{iklj} x^k x^l + O(|x|^3),$$

where δ_{ij} is the Kronecker delta function, and R_{iklj} form the Riemann curvature tensor at p. The canonical measure on M, induced by the Riemannian density, is then given by

$$dvol_g(\bar{x}) = J_p(x) dvol_{E_d}(x),$$

where

$$J_p(x) := \sqrt{|\det(g_{ij})|} = 1 - \frac{1}{3} Ric_{ij} x^i x^j + O(|x|^3),$$

and where

$$Ric_{ij} := \sum_k R_{ikkj}$$

is called the Ricci curvature tensor at p.

Using this expression for the Riemannian density, one deduces the following results when M is closed (has empty boundary), which compare the Riemannian volume and surface area of small Riemannian balls with their Euclidean counterparts (see for instance [44, 43, 56]):

$$V(B(p,r)) = \omega_d r^d \left(1 - \frac{s(p)}{6(d+2)} r^2 + O(r^3)\right)$$

and

$$\sigma(S(p,r)) = d\omega_d r^{d-1} \left(1 - \frac{s(p)}{6d} r^2 + O(r^3)\right),$$

where σ is the induced $(d-1)$-dimensional surface area, ω_d denotes the volume of a Euclidean unit ball and $s(p) := \sum_i Ric_{ii}$ is the scalar curvature at p.

In fact, we have the following volume comparisons.

Lemma 7.2.1 ([10]). *Let M be closed, compact. Let $|Ric_p| := \sup_{v \in T_c M \setminus \{0\}} \frac{Ric(v,v)}{|v|}$. For every $\nu > 0$, there exists $r_\nu > 0$ depending continuously on ν, such that for all $r \leq r_\nu$ and all $p \in M$*

$$r^{d-1}\left(1 - \frac{|Ric_p| + \nu}{3} r^2\right) dvol_{\mathbb{S}^{d-1}} \leq dvol_{S_r(p)} \leq r^{d-1}\left(1 + \frac{|Ric_p| + \nu}{3} r^2\right) dvol_{\mathbb{S}^{d-1}}$$

holds on $B(p,r)$.

Corollary 7.2.2 ([10]). *For $\nu > 0$, let $s_{\min}(\nu) := \inf_{p \in M} \frac{s(c)}{6(d+2)} - \nu$ and $s_{\max}(\nu) := \sup_{p \in M} \frac{s(c)}{6(d+2)} + \nu$. For all $\nu > 0$, there exists $r_\nu > 0$ such that for all $r \leq r_\nu$ and all $p \in M$*

$$\omega_d r^d (1 - s_{\max}(\nu) r^2) \leq V(B(p,r)) \leq \omega_d r^d (1 + s_{\min}(\nu) r^2).$$

Lemma 7.2.3 ([10]). *For all $\nu > 0$, there exists a continuous choice of $r\nu > 0$ such that for all $r \leq r_\nu$ and all $p \in M$, we have on $B(p,r)$*

$$(1 - \nu r^2)\left|dvol_{E_d}\right| \leq \left|dvol_g\right| \leq (1 + \nu r^2)\left|dvol_{E_d}\right|.$$

Let us also mention Lemma 2.10 in [10], comparing the union of two small Riemannian balls with the union of their Euclidean counterparts. This result is required in the proof of the second moment section which we do not write in full detail, as it is fairly heavy in integral calculations while these calculations do not bring any conceptual insight to the argument, and mostly similar to the second moment section in [10]. Consequently, it is not necessary to restate the content of Lemma 2.10 from [10].

Viewing our manifold M with non-empty boundary embedded in its double manifold DM which is closed, the above results naturally extend to our setting for balls having empty intersection with the boundary. Let us see how these volume comparisons are affected for small Riemannian balls intersecting the boundary.

Let $c \in M \subset DM$ and let $r > 0$ be sufficiently small.

Lemma 7.2.4. *Let $\delta := \rho(c, \partial M)$ and let $q := \delta/r$. There exists $\epsilon \in [q/2, q]$ such that*

$$\left| V(B(c,r) \cap M) - \frac{1}{2}(1+\epsilon)\omega_d r^d \right| = O(r^{d+1}).$$

Proof. By the above volume approximation results between $dvol_g$ and $dvol_{E_d}$, we have

$$\left| V(B(c,r) \cap M) - V_{E_d}(\exp_c^{-1}(V(B(c,r) \cap M))) \right| = O(r^{d+2}),$$

hence it suffices to show that

$$\left| V_{E_d}(\exp_c^{-1}(B(c,r) \cap M)) - \frac{1}{2}(1+\epsilon)\omega_d r^d \right| = O(r^{d+1}).$$

Assume first that the image of the boundary $\exp_c^{-1}(\partial M)$ on T_cM is flat, such that $\exp_c^{-1}(B(c,r) \cap M)$ is a Euclidean ball from which we have removed a spherical cap of height $r - \delta$. In this case, the volume of is known to be given by

$$V_{E_d}(\exp_c^{-1}(B(c,r) \cap M)) = \omega_d r^d \frac{1}{2}\left(1 + \frac{G_d(q)}{G_d(1)}\right),$$

where $G_d(u) = \displaystyle\int_0^u (1-t^2)^{(n-1)/2} dt \in [u/2, u]$, for u sufficiently small.

Indeed the integrand above is bounded above by 1, which gives $G_d(u) \leq u$. Conversely we have

$$G_d(u) \geq (1-u^2)^{(n-1)/2}.u \geq u/2,$$

for u sufficiently small.

Thus, it remains to estimate the error induced by the non-flatness of the boundary. We may model this boundary as

$$f : B^{(d-1)}(0,r) \to \mathbb{R},$$

with $f(0) = 0$ (if the boundary were flat as above, we would have $f \equiv 0$), where $B^{(d-1)}(0,r)$ denotes a $(d-1)$-dimensional Euclidean ball. Thus, the error induced by the non-flatness of the boundary is given by

$$\int_{B^{(d-1)}(0,r)} f(x)dx.$$

The boundary being smooth, we can Taylor expand f locally around 0. Using the symmetry of $B^{(d-1)}(0,r)$, note that all odd degree terms of the Taylor expansion vanish when integrated (these terms being odd functions integrated over a symmetric set).

Hence we have

$$\int_{B^{(d-1)}(0,r)} f(x)dx = \omega_{d-1} r^{d-1} f(0) + O(r^{d+1}) = O(r^{d+1}),$$

since $f(0) = 0$.

□

We remark from the above, that the boundary f is such that

$$r^{-(d-1)} \int_{B^{(d-1)}(0,r)} f(x)dx = O(r^2).$$

Furthermore, for r sufficiently small f is an increasing function on $B^{(d-1)}(0,r)$. From this, it is immediate to see that we must have $f(x) = O(r^2)$ everywhere on $B^{(d-1)}(0,r)$ for r sufficiently small (arguing by contradiction and using the continuity of f). We will use this remark later in the argument (cf, Lower threshold section).

7.3 Random coverings

Throughout the thesis, we have referred to an interesting random covering result, which also plays a key role in this argument. Given a fixed radius $r > 0$ and a compact manifold M, what is the minimum number of random balls of radius r that we should sample in order for M to be covered? Several papers have investigated this question (see [33, 37, 48, 11, 21]). In [33] for instance, the authors find an asymptotic expression for the expected number of balls in terms of the radius, as well as some concentration results around this expected value. For the sake of a better intuition, it is helpful to reformulate the result obtained in [33] by the dual problem of estimating the minimal radius needed in order

for n random balls to cover M. In other words, given an i.i.d. sample $X_n \subset M$ of size n, what is the Hausdorff distance $d_H(M, X_n)$? The main result obtained in [33] then reads that this *random covering radius* will be given w.h.p. by

$$r \sim (\log n)^{1/d} n^{-1/d}.$$

Note that in our case, we are not dealing with X_n, but rather with an homogeneous Poisson point process $\mathcal{P}_n$ whose size is a Poisson random variable $N \sim Po(n)$. However, due to the concentration results of Poisson random variables around their expected value (cf, the Chernoff bounds given in Section 2), this does not affect the asymptotic formulas. In fact, we have the following reformulation of the results of [33] from [11], which is the most suitable formulation for our purposes.

Theorem 7.3.1 ([11]). *Let M be compact, closed, let $w(n) \to \infty$ arbitrarily slowly, let $\Lambda : \omega_d n r^d$, and let $\mathcal{P}_n$ a uniform homogeneous Poisson process on M of intensity n. We have*

$$\lim_{n\to\infty} \mathbb{P}\left(M \subset \bigcup_{x\in\mathcal{P}_n} B(x,r)\right) = \begin{cases} 1 \text{ if } \Lambda = \log n + (d-1)\log n \log n + w(n) \\ 0 \text{ if } \Lambda = \log n + (d-1)\log\log n - w(n). \end{cases}$$

It is interesting to note that, up to multiplicative constants, this is the same asymptotic as the well known connectivity threshold for random geometric graphs. This provides us with an intuitive explanation of why there is a sharp threshold for the various homology groups, all occurring around $\Lambda := \omega_d n r^d = \Theta(\log n)$ (even though each homology group has a different threshold value). Indeed, as long as $\Lambda << \log n$, the underlying random geometric graph is not connected, hence we expect that none of the homology groups of the Čech complex will match those of M. This is easy to tell for instance, with the 0th homology group, whose dimension indicates the number of connected components. On the other hand, when $\Lambda >> \log n$ so that $r > d_H(\mathcal{P}_n, M)$, then the union of the balls cover M and by the Nerve Lemma this implies that all of their homology groups match.

The authors of [33] implicitly assumed the manifold to have empty boundary. In the case where $\partial M \neq \emptyset$, as already observed above, balls near the boundary will have volume about $1/2$ that of another ball of same radius far from the boundary. This suggests that a similar result to [33] holds with a suitable normalisation in the asymptotic formulas. Such a result can be obtained as an immediate corollary of the above random covering theorem. The following sharp covering threshold was obtained in [21].

Theorem 7.3.2 ([21]). *Let M be compact with non-empty boundary. Let $w(n) \to \infty$, and let Λ and $\mathcal{P}_n$ be as above. We have*

$$\lim_{n\to\infty} \mathbb{P}\left(M \subset \bigcup_{x\in\mathcal{P}_n} B(x,r)\right) = \begin{cases} 1 \text{ if } \Lambda = (2-2/d)\log n + 2(d-2)\log\log n + w(n) \\ 0 \text{ if } \Lambda = (2-2/d)\log n + 2(d-2)\log\log n - w(n). \end{cases}$$

7.4 Palm theory

The spatial independence property of the homogeneous Poisson point process $\mathcal{P}_n$ is a key feature facilitating its use over that of a deterministic i.i.d. sample X_n. In our argument we shall use the following results from Palm theory, which follow from the spatial independence property of $\mathcal{P}_n$. Palm theory was named after Conrad Conny Palm, whose work can be seen as a generalisation of Norman Robert Campbell's work. In particular Campbell's theorem, or the Campbell-Hardy theorem (see [18, 19]), gives a way to compute the expectation of the sum of a real valued function over a Euclidean point process.

Lemma 7.4.1 ([55]). *Suppose that $h(\mathcal{Y}, \mathcal{P}_n)$ is measurable for all $\mathcal{Y} \subset \mathcal{P}_n$ with $|\mathcal{Y}| = k+1$, then*

$$\mathbb{E}\left[\sum_{|\mathcal{Y}|=k+1} h(\mathcal{Y}, \mathcal{P}_n)\right] = \frac{n^{k+1}}{(k+1)!}\mathbb{E}\left[h(\mathcal{Y}', \mathcal{Y}' \cup \mathcal{P}_n)\right],$$

where $\mathcal{Y}' \subset M$ is an i.i.d. set with $|\mathcal{Y}'| = k+1$, independent from $\mathcal{P}_n$.

Lemma 7.4.2 ([55]). *Suppose that h is as above, then we also have*

$$\mathbb{E}\left[\sum_{\substack{|\mathcal{Y}_1|=|\mathcal{Y}_2|=k+1,\\ |\mathcal{Y}_1\cap\mathcal{Y}_2|=j}} h(\mathcal{Y}_1, \mathcal{P}_n)h(\mathcal{Y}_2, \mathcal{P}_n)\right] = \frac{n^{2k-j}}{j!(k-j)!}\mathbb{E}\left[h(\mathcal{Y}_1', \mathcal{Y}' \cup \mathcal{P}_n)h(\mathcal{Y}_2', \mathcal{Y}' cup \mathcal{P}_n)\right],$$

where $\mathcal{Y}' = \mathcal{Y}_1' \cup \mathcal{Y}_2'$ is a set of $2k-j$ i.i.d. points in M, independent from $\mathcal{P}_n$ and such that $\left|\mathcal{Y}_1 \cap \mathcal{Y}_2'\right| = j$.

7.5 Morse theory

A key idea in the argument of [10] is the use of Morse theory to study critical points of the distance function

$$\rho_{\mathcal{P}_n} : x \mapsto \min\{\rho_p(x) := \rho(p, x) \mid p \in \mathcal{P}_n\},\ x \in M,$$

in order to draw conclusions on the homology of the Čech complex.

Let us start with some definitions on critical points and smooth Morse functions.

Definition 35. *Let M be a smooth d-dimensional manifold (possibly with boundary) and let $f : M \to \mathbb{R}$ be smooth.*

We say that $c \in M$ is a critical point of f, if $\nabla f(c) = 0$.

A critical point c of f is called non-degenerate, if the Hessian $\nabla^2 f(c)$ is non-singular (i.e., its determinant is not equal to 0, i.e., is invertible); otherwise, c is called degenerate.

A smooth function $f : M \to \mathbb{R}$ is called a Morse function, if all its critical points are non-degenerate.

Given a critical point c of f, the index of c is the dimension of the space spanned by the eigenvectors associated to the negative eigenvalues of the Hessian $\nabla^2 f(c)$.

While the distance function is not smooth, the square distance function $\rho^2_{\mathcal{P}_n}$ is continuous and can be shown to be a Morse min-type function, i.e., at every point, $\rho^2_{\mathcal{P}}$ can be written as the minimum of finitely many smooth Morse functions.

In [10] the authors easily check that the square distance function is indeed a Morse min-type function.

Lemma 7.5.1 (Lemma 4.1 in [10])**.** *Let (M, g) be a compact manifold (possibly with non-empty boundary), there exists $r_{mt} > 0$ such that for every finite $\mathcal{P} \subset M \setminus \partial M$, the function $\rho^2_{\mathcal{P}}$ is a Morse min-type function on $\cup_{p \in \mathcal{P}} B(p, r_{mt})$, i.e., at every point, $\rho^2_{\mathcal{P}}$ can be written as the minimum of finitely many smooth Morse functions.*

Morse theory generally applies to smooth functions. However, the work of [35] essentially allows us to extend this study to Morse min-type functions. There, the authors of [35] generalise the notions of critical points and their index for Morse min-type functions, and show that a Morse min-type function on a closed, compact manifold can be approximated arbitrarily well by a smooth Morse function, in such a way that the critical points of the Morse min-type function are in one-to one index preserving correspondence with the critical points of the smooth Morse function.

After extending the notion of a smooth Morse function on a compact manifold with boundary as in [47], we may derive an analogous approximation result of min-type Morse functions by smooth Morse functions for a compact manifold with boundary M, by embedding it in its double manifold DM and invoking the approximation result of [35] (see Proposition 3.12 and Corollary 3.13 in [42]).

We thus have a Morse theoretic framework for the critical points of $\rho^2_{\mathcal{P}_n}$ in our setting.

7.5.1 Critical points for the distance function

Let $\mathcal{Y} = \{y_1, \dots, y_k\} \subset M$ be finite and define the following sets

$$E(\mathcal{Y}) := \{x \in M \mid \rho_{y_1}(x) = \cdots = \rho_{y_k}(x)\}$$
$$E_r(\mathcal{Y}) := E(\mathcal{Y}) \cap \bigcup_{y \in \mathcal{Y}} B(y, r).$$

Definition 36. *Given two sets A and B, define their Minkowski sum as*

$$A + B := \{a + b \mid a \in A,\ b \in B\}.$$

Definition 37. *Two submanifolds $N_1, N_2 \subset M$ are said to intersect transversally at a point $p \in N_1 \cap N_2$, if*

$$T_pN_1 + T_pN_2 = T_pM.$$

We say that N_1 and N_2 intersect transversally, if they do so at every point of $N_1 \cap N_2$.

Definition 38. *The set $\mathcal{Y} = \{y_1, \dots, y_k\} \subset M$ is called generic, if $E(\mathcal{Y}) \neq \emptyset$ and the $k-1$ submanifolds $ker(\rho_{y_i} - \rho_{y_1})$ intersect transversally at every point in $\cap_{i=1}^{k} B(y_i, r_{mt})$, where r_{mt} is chosen as in Lemma 7.5.1.*

Lemma 7.5.2 ([10])**.** *There exists $0 < r_{\max} < r_{mt}$ such that if $\mathcal{Y}$ is generic and $E_{r_{\max}}(\mathcal{Y}) \neq \emptyset$, then there exists a unique $c(\mathcal{Y}) \in M$ such that for all $y \in \mathcal{Y}$*

$$\rho_y(c(\mathcal{Y})) = \inf_{x \in E(\mathcal{Y})} \rho_{\mathcal{Y}}(x).$$

Definition 39. *Let $\mathcal{Y} = \{y_1, \dots, y_k\} \subset M$ and let $p \in M$. Define*

$$\Delta(\mathcal{Y}) := conv(\{\nabla \rho^2_{y_i}(p)\}_{i \in [k]}),$$

where $conv(.)$ denotes the convex hull.

We have the following proposition from [10], characterising critical points for the distance function.

Proposition 7.5.3 (Proposition 4.6 in [10])**.** *A critical point $c \in M$ of $\rho^2_{\mathcal{P}_n}$ has index k if and only if there exists $\mathcal{Y} \subset M$ of size $k+1$ such that*

$$\begin{cases} (1)\ c(\mathcal{Y}) = c, \\ (2)\ 0 \in \Delta(\mathcal{Y}) \subset T_pM, \\ (3)\ \mathcal{P}_n \cap B(c, \rho_{\mathcal{Y}}(c)) = \mathcal{Y}, \end{cases}$$

where $c(\mathcal{Y})$ is the centre of $\mathcal{Y}$ defined above.

7.5.2 Morse inequalities

Given a Morse min-type function f, define $C_k(a,b)$ to be the number of critical points of f of index k, such that $f(c) \in (a,b]$.

The following Morse inequalities, also presented in [10], motivate our considerations for critical points of the distance function, showing how they relate to the study of homology of Čech complexes.

Lemma 7.5.4 (Lemma 4.9 in [10])**.** *Let $f : M \to \mathbb{R}$ be a Morse min-type function. For $a \in \mathbb{R}$, let $M_a := f^{-1}((-\infty, a])$. For all $k \in \mathbb{N}$, we have*

$$\beta_k(M_a) \leq \beta_k(M) + C_{k+1}(a, +\infty).$$

Applying this result to the distance function $\rho^2_{\mathcal{P}_n}$ and using the Nerve Lemma, we find the following inequalities, relating critical points for the distance function with the study of homology of Čech complexes:

$$\beta_k(\mathcal{C}_r(\mathcal{P}_n)) = \beta_k(\cup_{x\in\mathcal{P}_n} B(x,r)) \leq \beta_k(M) + C_{k+1}(r, +\infty),$$

where $C_{k+1}(r, +\infty)$ denotes the number of critical points c of $\rho^2_{\mathcal{P}_n}$ of index $k+1$ such that $\rho^2_{\mathcal{P}_n}(c) > r^2 \Leftrightarrow \rho_{\mathcal{P}_n}(c) > r$.

7.6 Blascke-Petkantschin formulae

A key element of the success of the bounds obtained in our argument, as in the argument of [10] compared to [11], is the use of a change of variable integral formula known as the Blaschke-Petkantschin (B-P) formula. It is used repeatedly in various forms both in the calculations of the upper and the lower thresholds, as well as in the second moment calculations, discussed in the later sections.

In [10] the authors extend a Euclidean B-P formula (cf, [46]) to the case of a compact Riemannian manifold without boundary. We first present a heuristic argument recalling the Euclidean formula derived in [46]. Then, we recall the result obtained in [10], with a slight modification such that it can be applied to extending further the formula to the case of compact manifolds with non-empty boundary. We then present the formula in the case where the boundary is non-empty, and a multivariable version of the B-P formula, already used in [10] in the case where $\partial M = \emptyset$, without proof.

7.6.1 The Blaschke-Petkantschin formula in the Euclidean case

Let us first recall a derivation of the classical Blaschke-Petkantschin formula in the Euclidean case. Our derivation and Proposition 7.6.1, roughly follow Sections 2 and 3 of Miles in [46].

Let E_d be a d-dimensional Euclidean space, and let $(e_i)_{i=1}^d$ be an orthonormal moving frame in E_d, where for an infinitesimal rotation of the frame,

$$e_i \cdot de_i = 0, \, \forall i \in [d];$$

and set

$$\omega_{ij} := e_i \cdot de_j = -\omega_{ji}, \, \forall i, j \in [d].$$

Given r points $\{x_i : i \in [r]\} \subset E_d$, Miles derives heuristically the associated volume form

$$\bigwedge_{j=1}^{r} dV(x_j) = \bigwedge_{i=1}^{d} \bigwedge_{j=1}^{r} e_i . dx_j.$$

Furthermore, given the Grassmannian manifold $\mathrm{Gr}(r, d)$ with invariant measure $d\mu_{r,d}(V)$, Miles also derives

$$d\mu_{r,d} = \bigwedge_{i=1}^{r} \bigwedge_{j=r+1}^{d} \omega_{ij}.$$

Using the above, the Blaschke-Petkantschin formula expresses the Euclidean volume form $dV(x_i^d)$ on $\{x_i : i \in [r]\}$ in terms of the volume element associated to the r-plane containing $\{x_i : i \in [r]\}$, denoted by $dV(x_i^r)$.

Proposition 7.6.1 (Blaschke-Petkantschin formula Euclidean case ([46])). *Let* $\{x_i \mid i \in [r]\}$ *be a linearly independent set of vectors spanning* $V = Span(\{e_i \mid i \in [r]\}) \in Gr(r,d)$. *For each* $j \in [r]$, *let* $(\lambda_{jk})_{k\in[r]} \in E^r$ *be such that*

$$x_j = \sum_{k=1}^{r} \lambda_{jk} e_k, \text{ and let } \Upsilon := \big|det(\lambda_{jk})\big| > 0.$$

Then

$$\bigwedge_{i=1}^{r} dV(x_i^d) = \Upsilon^{d-r} d\mu_{r,d} \bigwedge_{i=1}^{r} dV(x_i^r).$$

Proof. Given $j \in [r]$ and $i \in \{r+1, \ldots, n\}$, we have

$$\begin{aligned}
dx_j &= \sum_{k=1}^{r} (d\lambda_{jk} e_k + \lambda_{jk} de_k) \\
e_i.dx_j &= \sum_{k=1}^{r} \lambda_{jk} \omega_{ik} \\
\bigwedge_{j=1}^{r} e_i.dx_j &= \big|det(\lambda_{jk})\big| \bigwedge_{k=1}^{r} \omega_{ik} \\
\bigwedge_{i=r+1}^{d} \bigwedge_{j=1}^{r} e_i.dx_j &= \Upsilon^{d-r} \bigwedge_{i=r+1}^{d} \bigwedge_{k=1}^{r} \omega_{ik} \\
\bigwedge_{i=r+1}^{d} \bigwedge_{j=1}^{r} e_i.dx_j &= \Upsilon^{d-r} d\mu_{r,d},
\end{aligned}$$

and multiplying on both sides above by $\bigwedge_{i=1}^{r} \bigwedge_{j=1}^{r} e_i.dx_j = \bigwedge_{j=1}^{r} dV(x_j^r)$, we obtain the desired result.

□

7.6.2 Blaschke-Petkantschin formula for Riemannian manifolds

Following [46], we obtain a Riemannian generalisation of the Blaschke-Petkantschin formula.

It enables us to reparametrise a $(k+1)$-tuple of points near the diagonal of M^{k+1} into local coordinates about their centre, according to the following decomposition:

$$M^{k+1} \longleftrightarrow M \times \mathbb{R} \times Gr(k,d) \times (S^{(k-1)})^{k+1}$$

$$\mathbf{y} \longleftrightarrow (c(\mathbf{y}), u, V, \mathbf{w}),$$

where $c(\mathbf{y})$ is the centre of the point $\mathbf{y}$, $u = \rho(\mathbf{y})$ is the distance of the points from their centre, $V \in Gr(k,d)$ is the k-dimensional subspace in which the pre-image of the points lie in the tangent space at the centre, and $\mathbf{w}$ are the $k+1$ points of the $(k-1)$-sphere upon which they lie.

Suppose that $M \subset \mathbb{R}^d$ is a closed Riemannian manifold and let $\mathbf{y} = (y_i)_{i=1}^{k+1} \in M^{k+1}$, with centre $c = c(\mathbf{y})$, with radius $\rho(\mathbf{y}) \leq r$, and local normal coordinates $(x^1, \ldots, x^d)$. For sufficiently small r, we can write for all $i \in [k+1]$

$$y_i = \exp_c(v_i),$$

with

$$v_i = \sum_{j=1}^{d} x^j(y_i) \left(\frac{\partial}{\partial x^i}\right)_c.$$

Let $\mathbb{1}_r(\mathbf{y}) = \mathbb{1}\{E_{r_{\max}}(y_0, ..., y_k) \neq \emptyset \text{ and } \rho(\mathbf{y}) \leq r\}$. It is shown in [10] that $\{v_i : i \in [k+1]\}$ have linear dependency and span a k-dimensional subspace $V \subset T_{c(\mathbf{y})}M$ when $c(\mathbf{y})$ is a critical point. We have the following change of variable formula.

Lemma 7.6.2 ([10])**.** *Let M be a compact closed Riemannian manifold with $M' \subset M$ a submanifold with or without boundary. Let $r_{\max}$ be as in Lemma 7.5.2, and $r < r_{\max}$. There exists an invariant measure $d\mu_{k,d}(V)$ on $Gr(k, T_cM) = Gr(k,d)$, such that for every $f : M^{k+1} \to \mathbb{R}$*

$$\begin{aligned}
&\int_{M^{k+1}} f(\mathbf{y})\mathbb{1}_r(\mathbf{y})\mathbb{1}\{c(\mathbf{y}) \in M'\}\left|dvol_g(\mathbf{y})\right| \\
&\qquad = \int_{M'} \left|dvol_g(c)\right| \int_0^r du u^{dk-1} \int_{Gr(k,T_cM)} d\mu_{k,d}(V) \\
&\qquad \times \left(\int_{S_1^{k+1}} \Upsilon_1^{d-k}(w) f(\exp_c(uw)) \prod_{i=1}^{k+1} \sqrt{\det(g_{\exp_c(uw_i)})} \left| dvol_{S_1(V)}(w_i) \right| \right).
\end{aligned}$$

Proof. First note that if M' has positive codimension then both expressions are zero, so assume M' has zero codimension. If $\mathbf{y} \in M^{k+1}$ is such that $E_{r_{\max}}(\mathbf{y}) \neq \emptyset$, then the induced centre $c(\mathbf{y})$ is uniquely defined; hence

$$\{\mathbf{y} \in M^{k+1} \mid c(\mathbf{y}) \in M' \text{ and } E_{r_{\max}}(\mathbf{y}) \neq \emptyset\} = \bigcup_{c \in M'} \mathcal{Y}(c),$$

where $\mathcal{Y}(c) := \{\mathbf{y} \in M^{k+1} \mid E_{r_{\max}}(\mathbf{y}) \neq \emptyset \text{ and } c(\mathbf{y}) = c\}$, and this union is disjoint (by uniqueness of the centre).
Thus

$$\int_{M^{k+1}} f(\mathbf{y})\mathbb{1}_r(\mathbf{y})\mathbb{1}\{c(\mathbf{y}) \in M'\}\left|dvol_g(\mathbf{y})\right| = \int_{M'} \left|dvol_g(c)\right| \int_{\mathcal{Y}(c)} f(\mathbf{y})\mathbb{1}\{\rho(\mathcal{Y}) \leq r\}\left|dvol_g(\mathbf{y})\right|.$$

Now fix $c \in M'$ with local normal coordinates $(x^1, \ldots, x^d)$; for $\mathbf{y} \in \mathcal{Y}(c)$ with $\rho(\mathbf{y}) \leq r < r_{\max}$ (this last condition ensures that y_i can be written as $\exp_c(v_i)$,

$v_i \in T_cM$ and $|v_i| = u \leq r$, for all $i \in [k+1]$), we find:

$$\begin{aligned}
\left|dvol_g(\mathbf{y})\right| &= \left|\wedge_{i=1}^{k+1} dvol_g(y_i)\right| \\
&= \left|\wedge_{i=1}^{k+1} \sqrt{\left|det(g_{y_i})\right|} dx^1(y_i) \wedge \cdots \wedge dx^d(y_i)\right| \\
&= \prod_{i=1}^{k+1} \sqrt{\left|det(g_{y_i})\right|} \left|\wedge_{i=1}^{k+1} dx^1(y_i) \wedge \cdots \wedge dx^d(y_i)\right| \\
&= \prod_{i=1}^{k+1} \sqrt{\left|det(g_{y_i})\right|} \left|\wedge_{i=1}^{k+1} dvol_{g_{E_d}}(v_i)\right| \\
&= \prod_{i=1}^{k+1} \sqrt{\left|det(g_{y_i})\right|} \left|dvol_{g_{E_d}}(\mathbf{v})\right|.
\end{aligned}$$

Note that we have the polar decomposition:

$$\left|dvol_{g_{E_d}}(\mathbf{v})\right| = du \left|dvol_{\mathcal{S}_u(E_d)}(\mathbf{v})\right|,$$

and so by the Blaschke-Petkantschin formula, since $\{v_i : i \in [k+1]\}$ lies in a k-dimensional subspace $V \subset T_cM$, we have

$$\left|dvol_{\mathcal{S}_u(E_d)}(\mathbf{v})\right| = \Upsilon_u(\mathbf{v})^{d-k} d\mu_{k,d}(V) \left|dvol_{\mathcal{S}_u(V)}(\mathbf{v})\right|,$$

hence, we deduce that

$$\left|dvol_g(\mathbf{y})\right| = \prod_{i=1}^{k+1} \sqrt{\left|det(g_{y_i})\right|} du \Upsilon_u(\mathbf{v})^{d-k} d\mu_{k,d}(V) \left|dvol_{\mathcal{S}_u(V)}(\mathbf{v})\right|.$$

This shows that for $c \in M'$

$$\begin{aligned}
&\int_{\mathcal{Y}(c)} f(\mathbf{y}) \mathbb{1}\{\rho(\mathcal{Y}) \leq r\} \left|dvol_g(\mathbf{y})\right| = \\
&\qquad \int_0^r du u^{dk-1} \int_{Gr(k,T_cM)} d\mu_{k,d}(V) \\
&\qquad \times \left(\int_{S_1^{k+1}} \Upsilon_1^{d-k}(\mathbf{w}) f(\exp_c(u\mathbf{w})) \prod_{i=1}^{k+1} \sqrt{\det(g_{\exp_c(uw_i)})} \left|dvol_{S_1(V)}(w_i)\right| \right),
\end{aligned}$$

and thus the result follows.

□

7.6.3 The Blaschke-Petkantschin formula for compact Riemannian manifold with non-empty boundary

Using the change of variable formula established in Lemma 7.6.2 for compact closed Riemannian manifolds, we derive a B-P formula for Riemannian manifolds with non-empty boundary.

Lemma 7.6.3. *Suppose that M is a compact Riemannian manifold with non-empty boundary, let DM be its double manifold and let $M' \subset M$ be a submanifold. Then we have:*

$$\int_{M^{k+1}} f(\mathbf{y})\mathbb{1}_r(\mathbf{y})\mathbb{1}\{c(\mathbf{y}) \in M'\}|dvol_g(\mathbf{y})|$$
$$= \int_{c\in M'} |dvol_g(c)| \int_0^r du \int_{Gr(k,T_cDM)} d\mu_{k,d}(V)$$
$$\times \int_{(S_u(V)\cap \exp_c^{-1}(M))^{k+1}} \Upsilon_u^{d-k}(v) f(\exp_c(v)) \prod_{i=1}^{k+1} \sqrt{|det(g_{\exp_c(v_i)})|}|dvol_{\mathcal{S}_u^{k+1}(V)}(v)|.$$

Proof. We have

$$\int_{M^{k+1}} f(\mathbf{y})\mathbb{1}_r(\mathbf{y})\mathbb{1}\{c(\mathbf{y}) \in M'\}|dvol_g(\mathbf{y})|$$
$$= \int_{DM^{k+1}} f(\mathbf{y})\mathbb{1}\{\mathbf{y} \in M^{k+1}\}\mathbb{1}_r(\mathbf{y})\mathbb{1}\{c(\mathbf{y}) \in M'\}|dvol_g(\mathbf{y})|;$$

the double manifold DM is closed, hence applying the change of variables formula in Lemma 7.6.2 to the function

$$\mathbf{y} \longmapsto f(\mathbf{y})\mathbb{1}\{\mathbf{y} \in M^{k+1}\},$$

we find

$$\int_{DM^{k+1}} f(\mathbf{y})\mathbb{1}\{y \in M^{k+1}\}\mathbb{1}_r(\mathbf{y})\mathbb{1}\{c(\mathbf{y}) \in M'\}|dvol_g(\mathbf{y})| =$$
$$\int_{c\in M'} |dvol_g(c)| \int_0^r du \int_{Gr(k,T_cDM)} d\mu_{k,d}(V)$$
$$\times \int_{\mathcal{S}_u(V)^{k+1}} \Upsilon_u^{d-k}(v) f(\exp_c(v))\mathbb{1}\{\exp_c(v) \in M^{k+1}\} \prod_{i=1}^{k+1} \sqrt{|det(g_{\exp_c(v_i)})|}|dvol_{\mathcal{S}_u^{k+1}(V)}(v)|,$$

which gives the lemma.

□

7.6.4 The multivariable Blaschke-Petkantschin formula

We require another change of variable formula, used in the second moment calculations in order to bound the variance of the number of critical points induced

by a point process. Although we skip some calculations from the second moment section below, we still present this multivariable B-P formula.

We show how to bound a change of variable formula when integrating over two variables in M^{k+1} where M has non-empty boundary. This formula is already used (without proof) in [10] in the case where $\partial M = \emptyset$.

Lemma 7.6.4. *Let M be a compact Riemannian manifold with non-empty boundary and let $\mathbf{y}_1, \mathbf{y}_2 \in M^{k+1}$. Denote the respective centres by c_1, c_2, and define*

$$\Omega := \left\{ (\mathbf{y}_1, \mathbf{y}_2) \in \left(M^{k+1}\right)^2 \mid a \leq \rho_M(c_1, c_2) \leq b \right\}.$$

The following bound holds:

$$\int_\Omega f_1(\mathbf{y}_1) f_2(\mathbf{y}_2) \mathbb{1}_r(\mathbf{y}_1, \mathbf{y}_2) \left| dvol_g(\mathbf{y}_1, \mathbf{y}_2) \right| \lesssim$$
$$\int_M \left| dvol_g(c_1) \right| \int_a^b ds \int_{S_1(T_{c_1}M)} s^{d-1} \left| dvol_{S_1(T_{c_1}M)}(w) \right|$$
$$\times \prod_{i=1}^{2} \int_0^r du_i u_i^{kd-1} \int_{Gr(k,d)} d\mu_{k,d}(V) \int_{S_1(V)^{k+1}} \left| dvol_{S_1(V)^{k+1}}(\mathbf{w}_i) \right| f_i(\exp_{c_i}(u_i \mathbf{w}_i)),$$

where the implied constant above only depends on M.

Proof. We use the Blaschke-Petkantschin formula for integrals over one variable in M^{k+1} to attain the result. Given $\mathbf{y}_1 \in M^{k+1}$ and c_1 the induced centre let

$$\Omega(\mathbf{y}_1) := \left\{ \mathbf{y}_2 \in M^{k+1} \mid a \leq \rho_M(c_1, c_2) \leq b \right\}.$$

Denote the above integral on the LHS by I, we have

$$I = \int_{M^{k+1}} f_1(\mathbf{y}_1) \mathbb{1}_r(\mathbf{y}_1) \left(\int_{\Omega(\mathbf{y}_1)} f_2(\mathbf{y}_2) \mathbb{1}_r(\mathbf{y}_2) \left| dvol_g(\mathbf{y}_2) \right| \right) \left| dvol_g(\mathbf{y}_1) \right|.$$

We first compute the inner integral, for fixed $c_1 \in M$.
Note that

$$\int_{\Omega(\mathbf{y}_1)} f_2(\mathbf{y}_2) \mathbb{1}_r(\mathbf{y}_2) \left| dvol_g(\mathbf{y}_2) \right| = \int_M f_2(\mathbf{y}_2) \mathbb{1}_r(\mathbf{y}_2) \mathbb{1}\{c_2 \in A_a^b(c_1)\} \left| dvol_g(\mathbf{y}_2) \right|,$$

where $A_a^b(c_1) := \overline{B_b(c_1)} \setminus B_a(c_1)$.
By the Blaschke-Petkantschin formula for manifolds with non-empty boundary,

we then find

$$\int_{\Omega(\mathbf{y}_1)} f_2(\mathbf{y}_2)\mathbb{1}_r(\mathbf{y}_2)\big|dvol_g(\mathbf{y}_2)\big|$$
$$= \int_{A_a^b(c_1)} \big|dvol_g(c_2)\big| \int_0^r du_2 \int_{Gr(k,T_cDM)} d\mu_{k,d}(V)$$
$$\times \left(\int_{(S_u(V)\cap \exp_{c_2}^{-1}(M))^{k+1}} \Big|dvol_{S_u(V)^{k+1}}(\mathbf{v}_2)\Big|\, f_2(\exp_{c_2}(\mathbf{v}_2))\Upsilon_u^{d-k}(\mathbf{v}_2) \prod_{j=1}^{k+1} \sqrt{\det(g_{\exp_{c_2}(v_j)})} \right)$$
$$\lesssim \int_{A_a^b(c_1)} \big|dvol_g(c_2)\big| \int_0^r du_2 u_2^{dk-1} \int_{Gr(k,d)} d\mu_{k,d}(V)$$
$$\times \left(\int_{S_1(V)^{k+1}} \Big|dvol_{S_1(V)^{k+1}}(\mathbf{w}_2)\Big|\, f_2(\exp_{c_2}(u_2\mathbf{w}_2))\Upsilon_1^{d-k}(\mathbf{w}_2) \prod_{j=1}^{k+1} \sqrt{\det(g_{\exp_{c_2}(u_2 w_j)})} \right).$$

Using the compactness of M, the above is

$$\leq C \int_{A_a^b(c_1)} \big|dvol_g(c_2)\big| \int_0^r du_2 u_2^{dk-1} \int_{Gr(k,d)} d\mu_{k,d}(V) \int_{S_1(V)^{k+1}} \Big|dvol_{S_1(V)^{k+1}}(\mathbf{w}_2)\Big|\, f_2(\exp_{c_2}(u_2\mathbf{w}_2)).$$

Furthermore, using the Riemannian approximation results and polar decomposition, we have

$$\begin{aligned}\int_{A_a^b(c_1)} \big|dvol_g(c_2)\big| &\leq C \int_a^b ds \int_{S_s(T_{c_1}DM)} \Big|dvol_{S_s(T_{c_1}DM)}(w)\Big| \\ &= \int_a^b ds \int_{S_1(T_{c_1}DM)} s^{d-1} \Big|dvol_{S_1(T_{c_1}DM)}(w)\Big|.\end{aligned}$$

The outer integral in the expression of I is estimated again with the Blaschke-Petkantschin formula for manifolds with non-empty boundary. Doing so and combining the result with the above expression for the inner integral of I yields the claimed formula.

□

We are now ready to show the claimed upper and lower thresholds.

7.7 Upper threshold

In this section we prove the upper bound on the expected Betti numbers $\mathbb{E}[\beta_k(r)]$, $k \in [d-1]$, claimed in Proposition 7.1.2. This follows as in [10] from the lemma below.

Lemma 7.7.1. *If $n \to \infty$ and $r, r_0 \to 0$ such that $r = o(r_0)$, $\Lambda \to \infty$, and $\Lambda_{r_0} r_0^2 \to 0$, then for every $k \geq 1$ we have*

$$\mathbb{E}\left[C_k(r, r_0)\right] = O\left(n\Lambda^{k-1}e^{-\Lambda}, n^{1-1/d}\Lambda^{k-1}e^{-\Lambda/2}\right).$$

Recall that $C_k(r, r_0)$ denotes the number of critical points for the distance function of index k and with critical values in $(r, r_0]$. We then deduce the bound claimed in Proposition 7.1.2 as in [10] by the above lemma and the Morse inequalities (see proof of Proposition 6.1 in [10]; this part of the argument remains the same in our setting and we do not repeat it).

Proof. The proof of the lemma follows a similar structure to [10], adapting the estimation of the volume of Riemannian balls near the boundary.

We can write the number of critical points as a sum of indicator functions:

$$C_k(r, r_0) := \sum_{|\mathcal{Y}_n|=k+1} g_{r,r_0}(\mathcal{Y}, \mathcal{P}_n),$$

where $g_{r,r_0}(\mathcal{Y}, \mathcal{P}_n)$ is defined as in the upper threshold section of [10]. This can be rewritten by Palm theory as

$$\begin{aligned}\mathbb{E}\left[C_k(r, r_0)\right] &= \frac{n^{k+1}}{(k+1)!}\mathbb{E}\left[g_{r,r_0}(\mathcal{Y}', \mathcal{Y}' \cup \mathcal{P}_n)\right] \\ &= \frac{n^{k+1}}{(k+1)!}\int_{M^{k+1}} h_{r,r_0}(\mathbf{y}) e^{-nV(B(\mathbf{y})\cap M)} \left|dvol_g(\mathbf{y})\right|.\end{aligned}$$

Using the Blaschke-Petkantschin formula for manifolds with boundary, we then have

$$\begin{aligned}\mathbb{E}\left[C_k(r, r_0)\right] \lesssim n^{k+1} &\int_M \left|dvol_g(c)\right| \int_r^{r_0} du u^{dk-1} e^{-nVol(B(c,u)\cap M)} \\ &\times \int_{S_1^{k+1}} \left|dvol_{S_1}(v)\right| h\left(\exp_c(v)\right).\end{aligned}$$

Let us split the above integral over M into two integrals, depending on whether c is near the boundary: the first integration being over $M_{r_0} := \{x \in M \mid \rho(x, \partial M) \geq r_0\}$ and the second over $\partial M_{r_0} := M \setminus M_{r_0}$.

If $c \in M_{r_0}$, then $V(B(c, u) \cap M) = V(B(c, u))$; we can thus proceed as in [10] and similarly obtain the bound

$$O\left(n\Lambda^{k-1}e^{-\Lambda}\right).$$

Let us now focus on the second integral

$$\begin{aligned}I := n^{k+1} &\int_{\partial M_{r_0}} \left|dvol_g(c)\right| \int_r^{r_0} du u^{dk-1} e^{-nV(B(c,u)\cap M)} \\ &\times \int_{S_1^{k+1}} \left|dvol_{S_1}(v)\right| h\left(\exp_c(v)\right).\end{aligned}$$

For n sufficiently large, it is clear that $B(c,u)\cap M$ must contain a half-ball from $B(c,u)$, so that

$$V\left(B(c,u)\cap M\right) \geq \frac{V\left(B(c,u)\right)}{2}.$$

Furthermore if $\rho(c,\partial M) =: \delta \leq r_0$, then $B(c,\delta) \subset B(c,u)\cap M$, and there is necessarily a half ball from $B(c,\delta)$ which does not intersect the half ball from $B(c,u)$ contained in $B(c,u)\cap M$ from above; hence we actually have

$$V(B(c,u)\cap M) \geq 1/2\left(V(B(c,u)) + V(B(c,\delta)\right).$$

Note that this lower bound is less sharp than what we could obtain by Lemma 7.2.4. However, the bound as provided above is more suitable for the sake of separating the variables u and δ, hence yields easier integrals to compute. We remark nonetheless that with extra considerations, using the estimates provided by Lemma 7.2.4 could yield sharper bounds for Proposition 7.1.2, hence ultimately a sharper upper threshold value for Λ in Theorem 7.1.1.

Using the Riemannian approximation results and that $nr_0^{d+2} \to 0$, we then find

$$\exp\left(-nV(B(c,u)\cap M)\right) \lesssim \exp\left(-\frac{n}{2}\omega_d u^d - \frac{n}{2}\omega_d\delta^d\right);$$

therefore

$$I \lesssim n^{k+1}\int_0^{r_0} d\delta\gamma(\delta)\exp\left(-\frac{n}{2}\omega_d\delta^d\right)\int_r^{r_0} duu^{dk-1}\exp\left(-\frac{n}{2}\omega_d u^d\right),$$

where $\gamma(\delta) = O(1)$ denotes the induced arclength of the set of points at distance δ from ∂M.
We compute the first integral on the RHS via an identification to an incomplete gamma function, after a suitable change of variable. With $t := \frac{n}{2}\omega_d\delta^d$, we find

$$d\delta \sim (n^{-1/d}t^{1/d-1})dt,$$

and thus

$$\int_0^{r_0} d\delta\exp\left(-\frac{n}{2}\omega_d\delta^d\right) \sim n^{-1/d}\int_0^{\Lambda_{r_0}/2} dt(t^{1/d-1}e^{-t});$$

now the integral on the RHS tends to $\Gamma(1/d) = O(1)$, as $n\to\infty$, hence

$$\int_0^{r_0} d\delta\exp\left(-\frac{n}{2}\omega_d\delta^d\right) \lesssim n^{-1/d}.$$

Therefore

$$I \lesssim n^{1-1/d}\left(n^k\int_r^{r_0} duu^{dk-1}\exp\left(-\frac{n}{2}\omega_d u^d\right)\right),$$

and we recognize a similar lower incomplete gamma function, in the parenthese on the RHS above, to the one investigated in [10] and which evaluates to

$$O\left(\Lambda^{k-1}e^{-\Lambda/2}\right).$$

□

7.8 Lower threshold

In this section we adapt to our setting the lower threshold argument developed in [10]; we derive a lower bound estimate for $\mathbb{E}[\beta_k(r)] = \mathbb{E}[\dim(H_k(\mathcal{C}(\mathcal{P}_n, r)))]$. In the upper threshold case, using the Morse inequalities it suffices to find an upper bound on $C_{k+1}(r, r_0)$: the number of critical points for the distance function of index $k+1$ with critical value in $(r, r_0]$. For the lower threshold on the other hand, we must make sure to only count critical points which induce new non-trivial cycles in $H_k(\mathcal{C}(\mathcal{P}_n, r))$. This motivates the restriction in [10] to what the authors call Θ-cycles.

Assume that $\partial M = \emptyset$, and let us recall some key results from [11, 10] used in their lower threshold arguments.

Let $\mathcal{Y} \subset \mathcal{P}_n \subset M$ be generic with centre $c = c(\mathcal{Y})$ and critical value $\rho = \rho(\mathcal{Y})$, and let $\epsilon \in (0, 1)$. Define the annulus

$$A_\epsilon(c) := \overline{B(c, \rho)} \setminus B(c, \epsilon\rho),$$

and define

$$\phi = \phi(\mathcal{Y}) := \frac{1}{2\rho} \min_{v \in \partial\Delta(\mathcal{Y})} |v| = \frac{1}{2} \sup\{\epsilon \geq 0 \mid \partial\Delta \subset A_\epsilon(c)\}.$$

Suppose furthermore that c has index k, $\rho < r_{\max}$, and that $A_\phi(c) \subset \bigcup_{x \in \mathcal{P}_n} B(x, \rho)$, then the authors in [11] (see Lemma 7.1) show that c induces a new non-trivial cycle in $H_k(\mathcal{C}(\mathcal{P}_n, r))$. Such cycles are the so-called Θ-cycles.

In Lemma 7.3 in [10], the following result shows how to use Θ-cycles to yield a lower bound on $\beta_k(r)$.

Lemma 7.8.1 (Lemma 7.3 in [10]). *Let $r_2 > r > 0$ and $r_1 > r\sqrt{1 - \frac{1}{c_g^2}(\frac{r_2}{r} - 1)^2}$. Suppose the following conditions hold:*

$$\begin{aligned} &\rho \in (r_1, r], \\ &B_{r_2}(c) \cap \mathcal{P}_n = \mathcal{Y}, \\ &\phi \geq \epsilon, \end{aligned}$$

and denote by $\beta_k^\epsilon(r)$ *the number of* Θ*-cycles satisfying the above. Then,* $\beta_k(r) \geq \beta_k^\epsilon(r)$.

Assume now that $\partial M \neq \emptyset$. The same results than above hold for critical points far from the boundary, i.e., $\rho(c, \partial M) > r_2$, and it is straightforward to see that counting Θ-cycles far from the boundary reproduces the same bounds than the ones obtained in [10].

On the other hand, as in the upper threshold argument, counting critical points near the boundary affects the estimates. Below we focus on obtaining a lower bound estimate for the number of Θ-cycles near the boundary (this bound will be sharper than the one we obtain by counting Θ-cycles far from the boundary).

Some of the above mentioned conditions to be met by the Θ-cycles are due to technicalities. However, as we have seen, the annulus A_ϵ being covered by the above union of balls is a crucial requirement, making sure that each such critical point induces a new non-trivial cycle in $H_k(\mathcal{C}(\mathcal{P}_n, \rho))$, where k is the index of c (see Lemma 7.2 in [10] and Lemma 7.1 in [11]).

In our setting, since we are interested in counting critical points near the boundary ∂M, the annulus $A_\epsilon(c) \subset DM$ around a critical point c will not entirely be contained in M. To resolve this issue, we must restrict our attention to the partial annulus formed by taking points not further away than a given angle φ from a hyperplane approximately parallel to the boundary, where φ is chosen in such a way that we do have indeed $A_\epsilon^{(\varphi)} \subset M$. To that end, given $c \in M$ near the boundary, let $n := \exp_c^{-1}(p)$, where p is the nearest point on ∂M to c, and let W be the hyperplane with normal n. Define the partial annulus mentioned above as

$$A_\epsilon^{(\varphi)}(c) := \{x \in A_\epsilon(c) \mid \angle(\exp_c^{-1}(x), W) \leq \varphi\},$$

and define the analogue of ϕ above

$$\psi = \psi(\mathcal{Y}, \varphi) := \frac{1}{2} \sup\{\epsilon \geq 0 \mid \partial\Delta \subset A_\epsilon^{(\varphi)}(c)\}.$$

How large we can afford to choose the angle φ of the partial annulus naturally depends on how far away the critical point c is from ∂M. Thus, we see that a suitable choice for φ and $\delta := \rho(c, \partial M)$ must be made, towards finding the best lower bound estimate possible. On the one hand, the larger φ the more critical points we allow to count. On the other hand, the larger δ the smaller $\exp(-V(B(c,r) \cap M))$, which we will see to also be a part of our lower bound estimate. Thus, one needs to find a suitable choice for δ as a function n, balancing out things in such a way that the lower bound estimate is as large as possible. Due to its exponential expression, keeping the term $\exp(-V(B(c,r) \cap M))$ as large as possible turns out to be of more importance than the gain we have by making φ large, hence we mostly want to choose δ relatively small. We choose

$\delta \sim (\log n)^{-1} r$. This choice will become clear during the proof.

Given our choice of δ, let us see how large φ can be taken such that the partial annulus is contained in M.

Let $x_0 \in A_\epsilon(c) \cap \partial M$ and let $\varphi_0 := \langle \exp_c^{-1}(x_0), W \rangle$. Assume first that the image of the boundary $\exp_c^{-1}(\partial M)$ is flat. Since $\delta, \varphi_0 \to 0$ as $r \to 0$, i.e., as $n \to \infty$, we have by trigonometry the following asymptotic formula

$$\varphi_0 \sim \tan(\varphi_0) \geq \delta / r.$$

If the boundary is not flat, this may affect the numerator δ in the above tangent formula. After modeling the boundary as before by $f : B^{(d-1)}(0, r) \to \mathbb{R}$, we know from the remark in Lemma 7.2.4 that $f(x) = O(r^2)$, which is $o(\delta)$ with our choice of $\delta \sim (\log n)^{-1} r$ (keeping in mind that $r \sim (\log n)^{1/d} n^{-1/d} = o((\log n)^{-1})$). Hence the above asymptotic formula remains unchanged and we may choose in all cases

$$\varphi \sim \delta / r \sim (\log n)^{-1}.$$

Following the above discussion, we shall adapt the conditions of Lemma 7.1 in [11] and Lemma 7.3 in [10] to the boundary case. To that end, define

$$\begin{aligned}
h(\mathcal{Y}) &:= \mathbb{1}\left(0 \in \Delta(\mathcal{Y})\right), \\
h_{r_1,r}(\mathcal{Y}) &:= h(\mathcal{Y}) \mathbb{1}\left(r_1 < \rho(\mathcal{Y}) \leq r\right), \\
h_r^{\epsilon,\delta}(\mathcal{Y}) &:= h_{r_1,r}(\mathcal{Y}) \mathbb{1}\left(\psi(\mathcal{Y}, \varphi) \geq \epsilon\right) \mathbb{1}\left(\delta \leq \rho(c(\mathcal{Y}), \partial M) \leq 2\delta\right), \\
g_r^{\epsilon,\delta}(\mathcal{Y}, \mathcal{P}_n) &:= h_r^{\epsilon,\delta}(\mathcal{Y}) \mathbb{1}\left(B_{r_2}(c(\mathcal{Y})) \cap (\mathcal{P}_n \setminus \mathcal{Y}) = \emptyset\right) \mathbb{1}\left(A_\epsilon^{(\varphi)} \subset \bigcup_{x \in \mathcal{P}_n} B(x, \rho(\mathcal{Y}))\right).
\end{aligned}$$

Let us denote by $\beta_k^{\epsilon,\delta}(r)$ the number of critical points counted by the above indicator functions, i.e.,

$$\beta_k^{\epsilon,\delta}(r) := \sum_{|\mathcal{Y}|=k+1} g_r^{\epsilon,\delta}(\mathcal{Y}, \mathcal{P}_n).$$

The two differences from [10] are the restriction to the partial annulus $A_\epsilon^{(\varphi)}$ (to make sure the detected cycles are contained in M) and the restriction to a given distance $\sim \delta$ away from the boundary ∂M.

Similarly to Lemma 7.1 in [11] and Lemma 7.2 in [10], the following lemma guarantees that each critical point counted above induces a non-trivial cycle in $H_k(\mathcal{C}(\mathcal{P}_n, \rho))$.

Lemma 7.8.2. *Let $\mathcal{Y} \subset \mathcal{P}_n \subset M$ be generic, inducing a critical point c of index $k \in [d-1]$. Let $\psi := \psi(\mathcal{Y})$ and suppose that $A_\psi^{(\varphi)}(c(\mathcal{Y})) \subset \bigcup_{x \in \mathcal{P}_n} B(x, \rho)$, where $\rho := \rho(\mathcal{Y}) = \rho_{\mathcal{P}_n}(c)$ is the critical value of c.*
Then c induces a non-trivial cycle in $H_k(\mathcal{C}(\mathcal{P}_n, \rho))$.

The proof of this lemma is similar to that of Lemma 7.1 in [11] and we do not repeat it here.

Me may now seek a lower bound for $\mathbb{E}[\beta_k^{\epsilon,\delta}(r)]$ and prove the following result.

Lemma 7.8.3. *Let $r = o(1)$. With our choice of $\delta \sim (\log n)^{-1} r$, there exists*

$$\alpha = \tfrac{1}{2} + O((\log n)^{-1}),$$

such that for all $k \in [d-1]$

$$\mathbb{E}[\beta_k^{\epsilon,\delta}(r)] = \Omega\left(n\Lambda^{k-2} e^{-\alpha\Lambda} r (\log n)^{-(k+1)}\right).$$

Proof. We have as in [10],

$$\begin{aligned}\mathbb{E}[\beta_k^{\epsilon,\delta}(r)] &= \frac{n^{k+1}}{(k+1)!}\mathbb{E}\left[g_r^{\epsilon,\delta}(\mathcal{Y}', \mathcal{Y}' \cup \mathcal{P}_n)\right] \\ &= \frac{n^{k+1}}{(k+1)!}\int_{DM^{k+1}} p_{\epsilon,\varphi}(y) h_r^{\epsilon,\delta}(y) e^{-nV(B(c(y),r_2)\cap M)} |dvol_g(y)|,\end{aligned}$$

where $p_{\epsilon,\varphi}(\mathbf{y}) := \mathbb{P}\left(A_\epsilon^{(\varphi)} \subset \bigcup_{x\in\mathcal{P}_n} B(x,\rho(\mathcal{Y})) \mid \mathcal{Y}' = \mathbf{y},\ \mathcal{P}_n \cap B(c(\mathbf{y}), r_2) = \mathcal{Y}'\right)$.

We first establish as in [10] that for $\varphi \sim (\log n)^{-1}$, $p_{\epsilon,\varphi}(\mathbf{y}) \to \mathbf{1}$ uniformly over $\mathbf{y}$. In [10] this is done by building a small net over the annulus A_ϵ, and then deriving the estimate

$$1 - Ce^{-Cn\epsilon^d r^d} \leq p_\epsilon(\mathbf{y}),$$

for some $C > 0$. Similarly here, we may build a small net over the partial annulus and derive, using that $\varphi \sim (\log n)^{-1}$, the estimate

$$1 - Ce^{-Cn\epsilon^d r^d} \leq p_{\epsilon,\varphi}(\mathbf{y}),$$

for some (other) $C > 0$.

Thus, we can write

$$\mathbb{E}[\beta_k^{\epsilon,\delta}(r)] \gtrsim \frac{n^{k+1}}{(k+1)!}\int_{DM^{k+1}} h_r^{\epsilon,\delta}(\mathbf{y}) e^{-nVol(B_{r_2}(c(\mathbf{y}))\cap M)} |dvol_g(\mathbf{y})|.$$

Define

$$\mathcal{W} := \{V \in Gr(k,d) \mid V = span\{w_i \mid i \in [k]\},\ \forall i \in [k],\ w_i \in W\}.$$

Note that by construction, for every $y_i \in \mathbf{y}$ counted above, we have $w_i := y_i - y_0 \in W$ hence $V \in \mathcal{W}$, where $V = span\{w_i |\ i \in [k]\}$. Using the Blaschke-Petkantschin formula for manifolds with non-empty boundary, we then find

$$\mathbb{E}[\beta_k^{\epsilon,\delta}(r)] \gtrsim \frac{n^{k+1}}{(k+1)!} \int_{\partial M_{[\delta,2\delta]}} |dvol_g(c)| \int_{r_1}^{r} du u^{dk-1} \int_{\mathcal{W}} d\mu_{k,d}(V)$$
$$\times \int_{(S_1(V))^{k+1}} \Upsilon_1^{d-k}(v) \prod_{i=1}^{k+1} \sqrt{|det(g_{\exp_c(v_i)})|} |dvol_{S_1^{k+1}(V)}(v)| f(\exp_c(v)),$$

where $f(\mathbf{y}) := h_r^{\epsilon,\delta}(\mathbf{y}) e^{-nV(B_{r_2}(c(\mathbf{y}))\cap M)}$.

By Lemma 7.2.4, we know that

$$\begin{aligned} V(B_{r_2}(c(\mathbf{y})) \cap M) &\leq \tfrac{1}{2}(1+q)\omega_d r_2^d \\ &\leq \tfrac{1}{2}\omega_d r_2^d (1 + q + O(r_2)), \end{aligned}$$

and $q = \frac{\delta}{r} \sim (\log n)^{-1}$ from our choice of δ.

Furthermore, $r_2 \sim r = o((\log n)^{-1})$, hence

$$\begin{aligned} V(B_{r_2}(c(\mathbf{y})) \cap M) &\leq \tfrac{1}{2}\omega_d r_2^d (1 + O((\log n)^{-1})) \\ &= \alpha \omega_d r_2^d, \end{aligned}$$

for some $\alpha := \frac{1}{2} + O((\log n)^{-1})$.

Thus

$$\mathbb{E}[\beta_k^{\epsilon,\delta}(r)] \gtrsim \frac{n^{k+1}}{(k+1)!} e^{-\alpha\Lambda_{r_2}} \Big(\int_{\partial M_{[\delta,2\delta]}} |dvol_g(c)| \int_{r_1}^{r} du u^{dk-1} \int_{\mathcal{W}} d\mu_{k,d}(V)$$
$$\times \int_{S_1^{k+1}(V)} \Upsilon_1^{d-k}(w) \prod_{i=1}^{k+1} \sqrt{|det(g_{\exp_c(w_i)})|} |dvol_{S_1^{k+1}(V)}(w)| h_r^{\epsilon,\delta}(\exp_c(uw)).$$

Let us estimate a lower bound for

$$D_k^{\epsilon} := \int_{\mathcal{W}} d\mu_{k,d}(V) \int_{S_1^{k+1}(w)} \Upsilon_1^{d-k}(w) \prod_{i=1}^{k+1} \sqrt{|det(g_{\exp_c(w_i)})|} |dvol_{S_1^{k+1}(V)}(w)| h_r^{\epsilon,\delta}(\exp_c(uw)).$$

From the choice of $\mathcal{W}$, we have $h_r^{\epsilon,\delta}(\exp_c(uw)) = 1$ for all $w \in S_1^{k+1}(V)$ and $V \in \mathcal{W}$. Furthermore for sufficiently small r, the determinant term tends to 1, and the volume $\Upsilon_1(w)$ is bounded below in terms of ϵ (as observed in [10]).

Hence it remains to estimate a lower bound for $\mu_{k,d}(\mathcal{W})$. Let us emphasise that the proposed argument below, bounding from below $\mu_{k,d}(\mathcal{W})$, is different

from the one provided in [42]. This geometric argument relies on principal angles and the covering number of a manifold and we find it elegant and concise. While different from that given in [42], it yields the exact same bounds, which is a good confirmation that this bound is fairly sharp.

Lemma 7.8.4. *We show that*

$$\mu_{k,d}(\mathcal{W}) \gtrsim \varphi^k.$$

Definition 40. *Let dist denote the geodesic distance on the Grassmannian $Gr(k,d)$, which is given by*

$$dist(V_0, V) = \sqrt{\sum_{i=1}^{k} \varphi_i^2},$$

where the φ_i's are the principal angles between the spaces V_0 and V.

Proof of Lemma 7.8.4. Let $V_0 \in Gr(k,W) \subset \mathcal{W}$ and let $V \in B(V_0, \varphi) \subset Gr(k,d)$ (ball centered at V_0 under the above geodesic distance on $Gr(k,d)$). Thus, we can write $V = span\{w_i \mid i \in [k]\}$ and $V_0 = span\{v_i \mid i \in [k]\}$ such that for all $i \in [k]$

$$\angle(w_i, v_i) \leq dist(V, V_0) \leq \varphi.$$

In particular, since $v_i \in W$, we must have

$$\angle(w_i, W) \leq \angle(w_i, v_i) \leq \varphi,$$

hence $V \in \mathcal{W}$ and $B(V_0, \varphi) \subset \mathcal{W}$.

Furthermore by Theorem 2.2 in [33], there exists $C > 0$ such that for all $m \in \mathbb{N}$, there exists $S_m = \{x_1, \ldots, x_m\} \subset Gr(k,W) = Gr(k, d-1)$, with $|S_m| = m$, and for every $x \neq y \in Gr(k, d-1)$

$$dist(x,y) \geq Cm^{-1/D},$$

where $D := dim(Gr(k, d-1)) = k(d-k-1)$.

Observe that the geodesic distance in $Gr(k, d-1)$ is the reduced distance from $G(k,d)$ to the subspace $Gr(k, d-1)$. Thus, the balls $\{B(x_i, (C/2)m^{-1/D}) \mid i \in [m]\}$ are disjoint.

Picking m such that $\varphi = \frac{C}{2} m^{-1/D}$, i.e., $m \sim \varphi^{-k(d-k-1)}$, we know from above that each of those balls is contained in $\mathcal{W}$ (because every basis vector in every subspace of every ball has angle at most φ with W).

Finally, each of the m distinct balls has volume $\mu_{k,d}(B(x_i,\varphi)) \sim \varphi^{k(d-k)}$, hence we have the following lower bound

$$\mu_{k,d}(\mathcal{W}) \gtrsim \varphi^{k(d-k)}\varphi^{-k(d-k-1)} = \varphi^k.$$

Theorem 2.2 in [33] being optimal (cf, the covering number of a manifold), we could not have found more distinct balls above in our volume estimation. This suggests that the above lower bound for $\mu_{k,d}(\mathcal{W})$ is sharp.

□

Wrapping up, we have

$$\mathbb{E}[\beta_k^{\epsilon,\delta}(r)] \gtrsim n^{k+1}\varphi^k e^{-\alpha\Lambda_{r_2}} \int_{\partial M_{[\delta,2\delta]}} |dvol_g(c)| \int_{r_1}^{r} du u^{dk-1}.$$

With $r_1 := r(1-\xi^2/(2c_g^2))$, $r_2 := r(1+\xi)$ and $\xi = \Lambda^{-1}$, we have as in [10], $e^{-\alpha\Lambda_{r_2}} = \Omega(e^{-\alpha\Lambda})$, and we find

$$\begin{aligned}\mathbb{E}[\beta_k^{\epsilon,\delta}] &\gtrsim \varphi^k n\Lambda^k e^{-\alpha\Lambda_{r_2}}\xi^2 \int_{\partial M_{[\delta,2\delta]}} |dvol_g(c)| \\ &\gtrsim n\Lambda^{k-2}e^{-\alpha\Lambda} r(\log n)^{-(k+1)}.\end{aligned}$$

□

7.9 Second moment for the lower threshold

In the upper threshold case, controlling the expected Betti numbers from above suffices to control the Betti numbers w.h.p. by Markov's inequality (a first moment inequality). In the lower threshold case we wish to establish that in the designated regime for Λ, w.h.p. the Betti numbers diverge to ∞ (in which case we do not recover the homology of the manifold). Yet, thus far we only have a lower bound for the expected Betti numbers via the expected number of Θ-cycles near the boundary counted in the previous section. We shall use Chebyshev's inequality (a second moment inequality):

$$\mathbb{P}(\beta_k^{\epsilon,\delta}(r) \leq \gamma\mathbb{E}[\beta_k^{\epsilon,\delta}(r)]) \leq \frac{Var(\beta_k^{\epsilon,\delta}(r))}{(1-\gamma)^2\mathbb{E}[\beta_k^{\epsilon,\delta}(r)]^2}.$$

We seek a lower bound for $\beta_k^{\epsilon,\delta}(r)$ from the lower bound on $\mathbb{E}[\beta_k^{\epsilon,\delta}(r)]$ obtained in the previous section, hence it suffices to show that the RHS above tends to 0, as $n \to \infty$.

The calculations of the different terms involved in the variance for the RHS above are fairly tedious. They require the use of the multivariable Blaschke-Petkantschin formula (proved above) and tedious integral calculations which are almost identical to those already done in the argument in [10]. Hence we

do not reproduce them here, as the calculations themselves do not constitute a significant conceptual step of the overall argument. For a detailed exposition of the calculations related to the second moment for the lower threshold, see Section 8 in [10] and Section 8 in [42].

The same calculations than in Section 8, [10] hold up to some suitable normalising factors, and we eventually obtain, similarly to [10], the following upper bound

$$\frac{Var(\beta_k^{\epsilon,\delta}(r))}{\mathbb{E}[\beta_k^{\epsilon,\delta}(r)]^2} \lesssim \frac{ne^{-\Lambda/2}\Lambda^{2k-3}r}{n^2\Lambda^{2k-4}e^{-2\alpha\Lambda}r^2(\log n)^{-2(k+1)}}(e^{-(\Lambda\epsilon\omega_d)/(2\omega_{d-1})}+\epsilon^d),$$

where $\alpha = \frac{1}{2} + O((\log n)^{-1})$ is as in Lemma 7.8.3.

Choose $\epsilon := \frac{2(2k-1)\omega_{d-1}}{\omega_d}\frac{\log\log n}{\log n}$ and

$$\Lambda := (2-2/d)\log n + 2(1/d-3)\log\log n - w(n).$$

Such a choice for ϵ is similar to the one made in [10] in their second moment section - Section 8 in [10], up to a suitable normalising factor. The choice for Λ comes from the value for the lower threshold which we derive when proving the lower threshold part of the main theorem (see proof of Theorem 7.1.1 below). It is thus natural to seek whether this candidate does satisfy the second moment requirement (in which case, it is valid).

With the above choices, we find indeed that the RHS above tends to 0 as $n \to \infty$, hence that w.h.p. $\beta_k(r) \geq \beta_k^{\epsilon,\delta}(r) \geq (1/2)\mathbb{E}[\beta_k^{\epsilon,\delta}(r)]$. This proves Proposition 7.1.4 claimed above.

Note carefully that the RHS above tends to 0 only due to the specific choices made above. In particular, this was only possible due to the error estimate of $O((\log n)^{-1})$ of α directly coming from our choice of $\delta \sim (\log n)^{-1}r$.

7.10 Proof of the main result

We now present the proof of the main result - Theorem 7.1.1, discussed in the outline section.

Proof of Theorem 7.1.1. **Upper threshold**

Letting $\Lambda := a_k \log n + b_k \log\log n + w(n)$, let us solve for the best possible coefficients a_k, b_k such that, using the upper bound estimates,

$$\mathbb{E}[C_{k+1}(r,r_0)], \mathbb{E}[C_k(r,r_0)] \leq n\Lambda^k e^{-\Lambda} \leq n^{1-1/d}\Lambda^k e^{-1/2\Lambda} \to 0, \text{ as } n \to \infty.$$

Indeed, this implies by Markov's inequality that

$$\mathbb{P}\left(C_k(r,r_0), C_{k+1}(r,r_0) > 0\right) \to 0,$$

hence by Morse theory

$$\mathbb{P}\left(\beta_k(r_0) > \beta_k(r)\right) \to 0,$$

which implies that

$$P\left(H_k(\mathcal{C}(n,r) \cong H_k(\mathcal{C}(n,r_0)\right) \to 1.$$

We have

$$n^{1-1/d}\Lambda^k e^{-1/2\Lambda} \sim \exp(k\log\Lambda - 1/2(a_k\log n + b_k\log\log n + w(n)) + (1-1/d)\log n).$$

Discarding eventual constant terms in the exponent above and seeing that $-w(n) \to -\infty$, it suffices to find a_k and b_k such that the $\log n$ and the $\log\log n$ coefficients vanish. Gathering the terms, we find the following condition for the $\log n$ coefficient to vanish:

$$-1/2a_k + 1 - 1/d = 0 \Rightarrow a_k = 2(1-1/d).$$

Similarly, using $\log\Lambda \sim \log\log n$, we find the following condition for the $\log\log n$ coefficient to vanish:

$$k - b_k/2 = 0 \Rightarrow b_k = 2k.$$

Let $r_0 := r(\omega_d/\kappa(1+|\log r|))^{1/d}$ be as in Proposition 7.1.2. For n sufficiently large, using the covering result stated in Theorem 7.3.2 (see [21]), Λ_{r_0} is beyond the upper threshold value, such that w.h.p. $M \subset \cup_{x\in\mathcal{P}_n} B(x,r_0)$ and by the Nerve Lemma, for n sufficiently large $H_k(\mathcal{C}(n,r_0)) \cong H_k(M)$ w.h.p..

Hence

$$\mathbb{P}\left(H_k(\mathcal{C}(n,r)) \cong H_k(M)\right) \to 1,$$

as required.

Lower threshold

Let

$$\Lambda = a_k\log n + b_k\log\log n - w(n),$$

and recall the obtained lower bound on $\beta_k(r)$ (w.h.p.)

$$\Omega\left(e^{-\alpha\Lambda} n\Lambda^{k-2} r(\log n)^{-(k+1)}\right).$$

Let us seek the best possible coefficients a_k, b_k such that this threshold diverges to ∞. Indeed, if such is the case, then in particular we cannot have $\beta_k(r) = \beta_k(M) < \infty$.

We can rewrite it as follows

$$\begin{aligned} re^{-\alpha\Lambda} n\Lambda^{k-2}(\log n)^{-(k+1)} &\sim re^{-\alpha\Lambda} e^{\log n} e^{(k-2)\log\Lambda}(\log n)^{-(k+1)} \\ &\sim \exp(\frac{1}{d}\log\log n - \frac{1}{d}\log n - (k+1)\log\log n - \log w(n) \\ &- \alpha(a_k\log n + b_k\log\log n - w(n)) + \log n + (k-2)\log(a_k\log n)) \\ &+ \alpha w(n) - \log w(n)); \end{aligned}$$

the third line on the RHS above, i.e., $\alpha w(n) - \log w(n)$ will diverge, hence it suffices to find a_k, b_k such that the rest of the exponent remains constant or converges to 0.
We first look at the coefficients of $\log n$. Taking $a_k = 2 - 2/d$ yields:

$$\begin{aligned} -1/d - \alpha a_k + 1 &= -1/d - (1/2 + \epsilon(n))(2 - 2/d) + 1 \\ &= -1/d - 1 - 2\epsilon(n) + 1/d - 2\epsilon(n)/d + 1 \\ &= (-2 - 2/d)\epsilon(n) \\ &= o((\log n)^{-1}), \end{aligned}$$

hence the $\log n$ term will converge to 0.
Since $a_k \geq 1$, we have the trivial lower bound

$$(k-2)\log(a_k \log n) \geq (k-2)\log\log n;$$

hence it suffices to solve for b_k such that (looking at the coefficients of $\log\log n$)

$$k - 2 - b_k/2 - k - 1 + 1/d = 0,$$

which yields $b_k = 2(-3 + 1/d)$.

We have thus shown that choosing

$$\Lambda = (2 - 2/d)\log n + 2(-3 + 1/d)\log\log n - w(n),$$

$\beta_k(r) \to \infty$ w.h.p.. In particular, for such Λ we do not recover the homology of M w.h.p..

□

7.11 Conclusion

In this chapter, we investigated homology of a Čech complex from a random sample on a compact Riemannian manifold with smooth non-empty boundary. This can be seen as a continuation of previous works by Bobrowski and Weinberger ([11]) and Bobrowski and Oliveira ([10]) which investigated the cases of, respectively, a torus and a closed (empty boundary) Riemannian boundary. As in these earlier works, we find a tight gap between the upper threshold value of the bandwidth beyond which we recover the homology of the manifold w.h.p., and the lower threshold value below which we do not recover the homology of the manifold w.h.p.. Indeed both thresholds occur for $\Lambda := n\omega_d r^d \sim 2(1-1/d)\log n$. Note that the coefficient is almost twice as much as the threshold values obtained in [10]. This is naturally understood by the fact that Riemannian balls intersecting the non-empty boundary of the manifold have volume about half of a ball of the same radius completely contained in the manifold (cf, the section on Riemannian approximations above).

Chapter 8

Conclusion and future work

We have defined and discussed several results related to geometric/topological methods in machine learning. As we saw in Chapters 2, 3, 4, and as noticed in [63], many problems of interest on graphs arise as optimisation problems on graph functionals of the form

$$\sum_{x \in X_n} \sum_{y \in X_n} \eta_r(x-y)(u(x)-u(y))^{\alpha}.$$

This observation motivated the authors of [63], in a series of works, to draw on techniques from the calculus of variations (e.g., Γ-convergence) and propose a *variational approach* to tackle the various problems on graphs arising as above. In particular, this variational approach proved successful in establishing various consistency results: e.g., spectral clustering, through the spectral convergence of the graph to the continuous Laplacian; Cheeger consistency; minimal bisection functional consistency. These results were presented in Chapter 2.

In Chapter 3, we noted how some basic results on the regularity of empirical measures with respect to the underlying measure allowed us to derive similar consistency results for k-NN graph constructions. In particular, we were able to establish asymptotic conditions on the choice of k as a function of n, to guarantee consistency of spectral clustering done with a graph Laplacian sparsified via a k-NN construction, and similarly Cheeger consistency and the consistency of minimal bisection functionals.

In Chapter 4, we focused on some of the key results of the succesful variational arguments previously discussed: the concentration of empirical measures (a discrepancy-type result) and the Γ-convergence of certain functionals. There, we proposed a setting motivated by the work of Owada and Adler in [52], in order to investigate possible generalisations of these key results to the case where the sampling domain is now all of $\mathbb{R}^d$ (more generally is unbounded). We established such extensions and also discussed some of the current limitations which

prevent one from carrying over the variational approach completely. In particular, there is still a need to understand how to establish better compactness properties of some functionals in $\mathbb{R}^d$.

From Chapter 5 onwards, we changed our paradigm and focused on *random geometric complexes*. These can be thought as generalisations of random geometric graphs, where we not only consider vertices and edges, but also triangles and higher order simplices. Chapter 5 serves as a brief introduction to random geometric complexes. We refer to the work of Kahle in [41] for a nice introduction to the types of results one generally is interested in proving.

In Chapter 6, we proposed an extension of the results in [1, 52], investigating the homology of noise when allowing the sampling density to be supported on $\mathbb{R}^d$ (instead of a bounded domain or more generally a compact manifold). We showed that some conditions on the choice of the bandwith parameter can be weakened with a well-chosen variable bandwidth construction, and how in some cases we may *decrackle the noise*.

Finally, in Chapter 7, we investigated the problem of homology recovery from a sampling on a compact Riemannian manifold with non-empty boundary, building from previous works of Bobrowski and Weinberger ([11]) and Bobrowski and Oliveira ([10]). This work follows the argument found in our joint work with Ulrike Tillmann and Oliver Vipond in [42].

There are many future directions one can investigate. In Chapter 7, while the upper and lower threshold values for Λ found in the case of a compact manifold with non-empty boundary (cf, [42]) have a tight gap, in the sense that they have the same leading term of $(2 - 2/d) \log n$, the transition is not shown to be sharp, unlike the case of a closed manifold (cf, [12]). In fact, since the obtained lower threshold is independent of k, one may wonder whether such a sharp transition can occur.

In Chapter 6, a natural next step is to continue the investigation of the homology of noise to that of persistent homology, much like the work of Owada and Bobrowski in [53] continues the work in [52].

With regards to random geometric complexes, one may be interested in investigating the *random connection model*. There, we are given a smooth kernel function on the graph, which we use for every two vertices x, y as a probability to connect the points. In other words, given $x, y \in X_n$ and a smooth kernel η_r, connect the points x, y with probability $\eta_r(x - y)$. This produces a random graph (where both the vertices and the edges are random), from which we may build a Vietoris-Rips complex (i.e., the clique complex of the graph). We may seek, as in [41], to investigate the expected topological features of this complex, such as its expected Betti numbers, as $n \to \infty$. Note in particular that if $\eta_r(x - y) = r^{-d} \mathbb{1}(|x - y| < r)$, we recover the usual Vietoris-Rips com-

plex construction. There are several advantages and disadvantages with this model, which as far as we know has never been investigated from a topological point of view (at the graph level on the other hand, the random connection model is fairly well known). First, note that if we take the expectation of the graph (of the affinity matrix inducing the graph) with respect to the edges (keeping the vertices random), the expected graph is just the weighted graph $\{\eta_r(x-y) \mid x, y \in X_n\}$. Hence this provides us with a way of investigating topological features of a weighted graph. Furthermore, this matrix is smooth under small noisy perturbations on the vertices, hence the various expected topological features to investigate, e.g., the expected Betti numbers, will be robust to noise, which is desirable and not generally the case (a motivation for studying persistent homology instead). There are on the other hand some drawbacks. First of all, while such a construction is natural on a Vietoris-Rips complex, since it is a purely combinatorial construction, it is not clear how to make sense of such a random connection model for a Čech complex. Hence such analysis would have limited geometric interpretations. Furthermore from a computational point of view, this model seems demanding.

For computational complexity motivations, we have investigated in Chapter 3 some consistency results for sparse graph representations, in particular k-NN constructions. Noting that these results rely on a discrepancy-type result, which was extended to various settings on $\mathbb{R}^d$ in Chapter 4, it is natural to try and combine the methods of Chapter 3 with the work of Chapter 4, and similarly deduce some consistency (or even just connectivity results) in the case of a k-NN graph sampled from an unbounded domain.

There are many more graph functionals which have slightly more complicated forms than the ones studied in this thesis. In some cases, little is known about them and it would be desirable to develop some approaches to similarly establish consistency results there (e.g., [27]).

Finally, many of the results described in this thesis are asymptotic results. Stronger results would control some error of convergence (of the various quantities of interest) as a function of n. Such results are of course more difficult to attain, but are highly desirable for practical points of view.

Bibliography

[1] Robert J. Adler, Omer Bobrowski, and S. Weinberger. Crackle: The Homology of Noise. *Discrete and Computational Geometry*, 52:680 – 704, 2014.

[2] G. Alberti and G. Bellettini. A non-local anisotropic model for phase transitions: asymptotic behavior or rescaled energies. *European Journal of Applied Mathematics*, 9.

[3] P.S. Alexandroff. Über den allgemeinen dimensionsbegriff und seine bezeihungen zur elementaren geometrischen anschauung. *Mathematische Annalen*, 98:617 – 635, 1928.

[4] S. Balakrishnan, A. Rinaldo, D. Sheehy, A. Singh, and L. Wasserman. Minimax rates for homology inference. *Proceedings of Machine Learning Research*, 22:64 – 72, 2012.

[5] Paul Balister, Bela Bollobas, Amites Sarkar, and Mark Walters. Connectivity of random k-nearest-neighbours graphs. *Advances in Applied Probability*, 37:1 – 24, 2005.

[6] Paul Balister, Bela Bollobas, Amites Sarkar, and Mark Walters. A critical constant for the k-nearest neighbour model. *Advances in Applied Probability*, 41(1):1 – 12, 2009.

[7] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in Neural Information Processing Systems*, 14:586 – 691, 2001.

[8] Tyrus Berry and John Harlim. Variable bandwidth diffusion kernels. *Applied and Computational Harmonic Analysis*, 40(1):68–96, 2016.

[9] O. Bobrowski and M. Khale. Topology of random geometric complexes: a survey. *J Appl. and Comput. Topology*, 1:331–364, 2018.

[10] O. Bobrowski and G. Oliveira. Random Čech Complexes on Riemannian Manifolds. *Random Structures and Algorithms*, 54(3):374 – 412, 2019.

[11] O. Bobrowski and S. Weinberger. On the Vanishing of Homology in Random Čech Complexes. *Random Structures and Algorithms*, 51:14–51, 2017.

[12] Omer Bobrowski. Homological connectivity in random Čech complexes. *arXiv:1906.04861*, 2019.

[13] Omer Bobrowski, Matthew Kahle, and Primoz Skraba. Maximally persistent cycles in random geometric complexes. *The Annals of Applied Probability*, 27(4), 2017.

[14] Karol Borsuk. On the imbedding of systems of compacta in simplicial complexes. *Fundamenta Mathematicae*, 35(1):217 – 234, 1948.

[15] M.R. Brito, E.L. Chavez, A.J. Quiroz, and J.E. Yukich. Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection.

[16] P. Bubenick and P. T. Kim. A Statistical Approach to Persistent Homology. *Homology, Homotopy and Applications*, 9(2):337 – 362, 2007.

[17] M. Buchet, F. Chazal, S. Y. Oudot, and D. R. Sheehy. Efficient and robust persistent homology for measures. *Computational Geometry*, 58:70 – 96, 2016.

[18] N. R. Campbell. The study of discontinuous phenomena. *Proc. Camb. Phil. Soc.*, 15:117 – 136, 1909.

[19] N. R. Campbell. Discontinuities in light emission. *Proc. Camb. Phil. Soc.*, 15:310 – 328, 1910.

[20] G. Carlsson. Topology and Data. *Bull. Amer. Math. Soc.*, 46(2):255 – 308, 2009.

[21] W. Chai. *Random Topological Structures.* PhD thesis. University of Chicago, 2018.

[22] F. Chazal, L. J. Guibas, S. Oudot, and P. Skraba. Scalar field analysis over point cloud data. *Discrete and Computational Geometry*, 46(743), 2011.

[23] Jeff Cheeger. *A lower bound for the smallest eigenvalue of the Laplacian.* Problems in analysis (Papers dedicated to Salomon Bochner, 1969. Princeton University Press.

[24] Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5 – 30, 2006.

[25] Vin de Silva and Robert Ghrist. Coverage in sensor networks via persistent homology. *Algebraic and Geometric Topology*, 7(1):339 – 358, 2007.

[26] R.M. Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40 – 50, 1969.

[27] Matthew M. Dunlop, Dejan Slepcev, Andrew M. Stuart, and Matthew Thorpe. Large data and zero noise limits of graph-based semi-supervised learning algorithms. *Applied and Computational Harmonic Analysis*, 2019.

[28] Herbert Edelsbrunner and John Harer. Persistent Homology – a Survey. *Contemp. Math.*, 453:257 – 282, 2008.

[29] L. C. Evans and R. F. Gariepy. *Measure Theory and Finite Properties of Functions*. CRC Press, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742.

[30] C. Fefferman. Fitting a C^m smooth function to data iii. *Annals of Mathematics*, 170(1), 2009.

[31] C. Fefferman and B. Klartag. Fitting a C^m smooth function to data i. *Annals of Mathematics*, 169:315 – 346, 2009.

[32] C. Fefferman and B. Klartag. Fitting a C^m smooth function to data ii. *Rev. Math. Iberoamericana*, 25:49 – 273, 2009.

[33] L. Flatto and Donald J. Newman. Random Coverings. *Acta Math.*, 138:241–264, 1977.

[34] Robert Forman. A User's Guide to Discrete Morse Theory. *Sém. Lothar. Combin.*, 2002.

[35] V. Gershkovich and H. Rubinstein. Morse theory for min-type functions. *Asian J. Math.*, 1(4):696 – 715, 1997.

[36] C. R. Givens and R. M. Shortt. A Class of Wasserstein Metrics for Probability Distributions, 1984.

[37] P. Hall. On the coverage of k-dimensional space by k-dimensional spheres. *Annals of Probability*, 1985.

[38] H. Hanche-Olsen and H. Holden. The Kolmogorov-Riesz compactness theorem. *Expositiones Mathematicae*, 28(4):385 – 394, 2010.

[39] M.C. Jones. Variable Kernel Density Estimates. *Austral. J. Statist.*, 32(3):361 – 371, 1990.

[40] H. Jung. Ueber die kleinste kugel, die eine räumliche figur einschliesst. *J. Reine Angew. Math.*, 123:241 – 257, 1901.

[41] M. Kahle. Random Geometric Complexes. *Discrete and Computational Geometry*, 45(3):553–573, 2011.

[42] Henry-Louis de Kergorlay, Ulrike Tillmann, and Oliver Vipond. Random Čech Complexes on Manifolds with Boundary. *arXiv:1906.07626*, 2019.

[43] Shoshichi Kobayashi and Katsumi Nomizu. *Foundations of Differential Geometry Volume* 1. Wiley Classics Library, 1996.

[44] John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, second edition edition, 2013.

[45] Markus Maier, Matthias Hein, and Ulrike von Luxburg. *Optimal construction of k-nearest-neighbours graphs for identifying noisy clusters*, 410(19):1749 – 1764, 2009.

[46] R. E. Miles. Isotropic random simplices. *Advances in Applied Probability*, 3:353 – 383, 1971.

[47] J. Milnor. *Lectures on the h-Cobordism Theorem*. Princeton University Press, 1965.

[48] P.A.P. Morgan and S. Fazekas de St Groth. Random circles on a sphere. *Biometrika*, 1962.

[49] T. Müller and M. D. Penrose. Optimal Cheeger Cuts and Bisections of Random Graphs. *arXiv:1805.08669*, 2018.

[50] P. Niyogi, S. Smale, and S. Weinberger. Finging the homology of submanifolds with high confidence from random samples. *Discrete and Computational Geometry*, 39(1), 2008.

[51] P. Niyogi, S. Smale, and S. Weinberger. A topological view of unsupervised learning from noisy data. *SIAM Journal on Computing*, 40(3):646 – 663, 2011.

[52] T. Owada and Robert J. Adler. Limit Theorems for Point Processes Under Geometric Constraints (and Topological Crackle). *Ann. Probab.*, 45(3):2004–2055, 2017.

[53] T. Owada and O. Bobrowski. Convergence of persistence diagram for topological crackle. *arXiv:1810.01602*, 2018.

[54] Mathew Penrose. A strong law for the longest edge of the minimal spanning tree. *Annals of Probability*, 27(1):246 – 260, 1999.

[55] Mathew D. Penrose. *Random Geometric Graphs*, volume 5 of *Oxford Studies in Probability*. Oxford University Press, Oxford, 2003.

[56] Peter Peterson. *Riemannian Geometry*, volume 171 of *Graduate Texts in Mathematics*. Springer, 2006.

[57] S.T. Roweis and L.K. Saul. Non-linear dimension reduction by locally linear embedding. *Science*, 290:2323 – 2326, 2000.

[58] Stephan R. Sain and David W. Scott. On Locally Adaptive Density Estimations. *Journal of the American Statistical Association*, 91(436), 1996.

[59] A. Singer. From graph to manifold Laplacian: the convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006.

[60] George R. Terrell and David W. Scott. Variable Kernel Density Estimations. *The Annals of Statistics*, 20(3):1236 – 1265, 1992.

[61] N. G. Trillos, M. Gerlach, M. Hein, and D. Slepcev. Error Estimates for Spectral Convergence of the Graph Laplacian on Random Geometric Graphs Towards the Laplace–Beltrami Operator. *Foundations of Computational Mathematics*, 2019.

[62] N. G. Trillos and D. Slepcev. On the rate of convergence of empirical measures in ∞-transportation distance. *Canadian Journal of Mathematics*, 67(6):1358 – 1383, 2015.

[63] N. G. Trillos and D. Slepcev. Continuum Limit of Total Variation on Point Clouds. *Archive for Rational Mechanics and Analysis*, 220(1):193–241, 2016.

[64] N. G. Trillos and D. Slepcev. A variational approach to the consistency of spectral clustering. *Applied and Computational Harmonic Analysis*, 45(2):239 – 281, 2018.

[65] N. G. Trillos, D. Slepcev, and J. von Brecht. Estimating Perimeter Using Graph Cuts. *Advances in Applied Probability*, 49(4):1067 – 1090, 2017.

[66] N. G. Trillos, D. Slepcev, J. von Brecht, T. Laurent, and X. Bresson. Consistency of Cheeger and Ratio Graph Cuts. *The Journal of Machine Learning Reasearch*, 17(1):6268 – 6313.

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

[68] Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consitency of spectral clustering. *The Annals of Statistics*, 36(2):555 – 586, 2008.

[69] Feng Xue and P.R. Kumar. The nmuber of neighbours needed for connectivity of wireless networks. *Wireless Netwroks*, 10(2):169 – 181, 2004.

[70] Lihi Zelnic-Manor and Pietro Perona. Self-tuning spectral clustering. *NIPS*, 2004.

[71] Afra Zomorodian and Gunnar Carlsson. Computing Persistent Homology. *Discrete Comput. Geom.*, 33(2):249 – 274, 2005.

Appendix A

Naive approach to noise decrackling (cf, Chapter 6)

A.1 Introduction

As discussed in Section 6.2, this appendix looks at the use of a naive scaling choice to study the classic argument of Kahle in [41] in the case of a variable bandwidth construction and an arbitrary radial density supported on $\mathbb{R}^d$. More useful scalings were considered in Chapter 6. Let us start with some notation.

A.2 Notation

Given two functions f, g, write

$$f(n) = O(g(n)),$$

if $\exists C > 0$, $\exists n_0 \in \mathbb{N}$, $\forall n \geq n_0$

$$|f(n)| \leq Cg(n);$$

write

$$f(n) = o(g(n)),$$

if $\forall C > 0$, $\exists n_0 \in \mathbb{N}$, $\forall n \geq n_0$

$$|f(n)| \leq Cg(n);$$

write

$$f(n) = \Omega(g(n)),$$

if $\exists C > 0$, $\exists n_0 \in \mathbb{N}$, $\forall n \geq n_0$

$$f(n) \geq Cg(n);$$

and write

$$f(n) = \omega(g(n)),$$

if $\forall C > 0$, $\exists n_0 \in \mathbb{N}$, $\forall n \geq n_0$

$$\big|f(n)\big| \geq C\big|g(n)\big|.$$

If the above constant C depends explicitly on a parameter k, we will write for instance $f(n) = \Omega_k(g(n))$.

If $\lim_n g(n) = 0$, we write

$$f(n) \sim g(n),$$

if $\exists k_1, k_2 > 0,\ \exists n_0 \in \mathbb{N},\ \forall n \geq n_0$

$$k_1 g(n) \leq f(n) \leq k_2 g(n).$$

If $\lim_n g(n) = a$, we write

$$f(n) \sim g(n),$$

if $(a - f(n)) \sim (a - g(n))$.

A.3 A naive approach

Consider a set of points $X_n = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ independently sampled from a radial density $q : \mathbb{R}^d \to \mathbb{R}^+$, where we assume that $\text{supp}(q) = \mathbb{R}^d$.
Denote by μ the probability distribution associated to the density q, i.e.,

$$\mu(A) := \int_A q(x)dx.$$

In particular let $y \in \mathbb{R}^d$, $r > 0$, and let

$$\begin{aligned} \rho : \mathbb{R}^d &\longrightarrow \mathbb{R}^+ \\ x &\longmapsto q(x)^{-1/d}; \end{aligned}$$

it is easy to see after a suitable change of variable, that

$$\mu(B(y, r\rho(y))) \sim r^d.$$

Thus, we can scale any ball of $\mathbb{R}^d$ by a function of its centre to make its volume roughly constant for a fixed radius r. This well known idea allows us to follow a similar approach to [41] where the distribution is uniform.

While ρ is defined on $\mathbb{R}^d$, we will sometimes also refer to the induced function $\rho : \mathbb{R}_+ \to \mathbb{R}_+$, where

$$\rho(R) := sup\{\rho(x) \mid |x| \leq R\}$$

for $R \in \mathbb{R}_+$. Similarly, we will sometimes refer to $q : \mathbb{R}_+ \to \mathbb{R}_+$ where

$$q(R) := \inf\{q(x) \mid |x| \leq R\}$$

for $R \in \mathbb{R}_+$. Note that $\rho(R) = q(R)^{-1/d}$.
Finally, we will make use of the function

$$\begin{aligned} \varphi : \mathbb{R}^+ &\longrightarrow \mathbb{R}_+ \\ R &\longmapsto \sup\left\{ \frac{|x|}{\rho(x)} \mid |x| \geq R \right\}. \end{aligned}$$

When clear from context, we may refer to $\rho(x_j)$ by ρ_j to simplify notation.

We wish to investigate topological properties of $R(X_n; r\rho)$ as $n \to \infty$.
In particular, we are interested in the asymptotic behavior of the expected ranks of the homology groups H_k, i.e., the expected Betti numbers $\mathbb{E}[\beta_k]$, $k \geq 0$, for various choices of r.

A.3.1 Discrete Morse theory

We briefly recall the necessary tools from Discrete Morse Theory used in [41], which are also used in our setting. A more thorough introduction can be found in [34].

Definition 41. *Let Δ be a simplicial complex, and let α and β be faces of Δ. We write $\alpha \prec \beta$ if α is a codimension 1 face of β.*

Definition 42. *Let Δ be a simplicial complex, and let $V := \{\alpha \prec \beta\}$ be a collection of pairs of faces of Δ.*
We say that V is a discrete vector field of Δ, if each face from V is in at most one pair.
We say that a simplex is critical (wrt V), if it is not contained in any pair in V.

Definition 43. *Let Δ be a simplicial complex, and let V be a discrete vector field of Δ. Let $n \in \mathbb{N}$, and let $\{\alpha_i \mid i \in \{0, \ldots, n\}\}$ and $\{\beta_i \mid i \in \{0, \ldots, n\}\}$ be faces of Δ, such that for all $i \in \{0, \ldots, n\}$, $\alpha_{i+1} \neq \alpha_i$ and*

$$\{\alpha_i \prec \beta_i\} \in V.$$

The sequence of faces

$$\alpha_0 \prec \beta_0 \succ \alpha_1 \prec \beta_1 \succ \cdots \prec \beta_n \succ \alpha_{n+1} = \alpha_0$$

is called a closed V-path.
We say that V is a discrete gradient field, if there are no closed V-paths.

We can now state the Fundamental Theorem of Discrete Morse Theory, cf, [41, 34].

Theorem A.3.1 ([34]). *Let Δ be a simplicial complex, and suppose that V is a discrete gradient field of Δ.*
Then Δ is homotopy equivalent to a CW complex with one cell of dimension k for each critical k-dimensional simplex.

Letting f_k be the number of cells of dimension k, we then have that $\beta_k \leq f_k = C_k$, where C_k is the number of critical k-dimensional faces of Δ. Thus, as observed in [41], to analyze the asymptotic behavior of the expected Betti numbers, it suffices to obtain bounds on $\mathbb{E}[C_k]$.

Recall the definition of k-connectivity for a topological space.

Definition 44 ([41]). *A topological space T is k-connected if for every $0 \leq i \leq k$, every map from an i-dimensional sphere S^i to T is homotopically trivial.*

In [41] it is shown for the VR complex on a smooth convex body of $\mathbb{R}^d$, that for every $k \geq 0$, there exists $c_k > 0$ such that if $r \geq c_k(\log n/n)^{1/d}$, then $R(X_n; r)$ is k-connected w.h.p..
Using the above observation that $\beta_k \leq C_k$, it suffices to show for $r \geq c_k(\log n/n)^{1/d}$, that we have

$$\mathbb{E}[C_k] = o(1).$$

A.3.2 Main result

The result obtained in [41] on the sublinearity of the expected Betti numbers and on k-connectivity threshold values for the VR complex can be summarized by the following theorem.

Theorem A.3.2 (Theorem 6.5 in [41]). *Let $R(X_n; r)$ be a random VR complex, where X_n is an i.i.d. sample from the uniform distribution on a smoothly convex body of $\mathbb{R}^d$, and let $\Lambda := nr^d$.*
There exists $c > 0$ independent of X_n, such that for all $k \geq 0$

$$\mathbb{E}[\beta_k] = O\left(\Lambda^k e^{-c\Lambda} n\right).$$

In particular for $k \geq 0$, if $r = \omega(n^{-1/d})$ then

$$\mathbb{E}[\beta_k] = o(n);$$

and if $r = \Omega_k\left(\left(\frac{\log n}{n}\right)^{1/d}\right)$ then

$$\mathbb{E}[\beta_k] = o(1).$$

Analogously, we obtain the following result for a variable bandwidth VR complex on $\mathbb{R}^d$ defined as above.

Theorem A.3.3. *Let $R(X_n; r\rho)$ be the random variable bandwidth VR complex defined above, where X_n is independently sampled from a radial density q, with $\mathrm{supp}(q) = \mathbb{R}^d$, and $\rho(x) := q(x)^{-1/d}$. Let $\alpha = 4$ if q is decreasing everywhere on $\mathbb{R}^d$; otherwise let $\alpha = 6$.*
For every $R(n) \to \infty$, if $r \geq 4\varphi(R)$ (note that φ decays to 0), then for all $k \geq 0$

$$\mathbb{E}[\beta_k] = O\left(\Lambda^k e^{-c(n)\Lambda} n\right),$$

where $c(n) \sim \rho(R)^{-\alpha d}$, and as before $\Lambda := nr^d$.
In particular, let $k \geq 0$ and let $w(n) \to \infty$ (growing arbitrarily slowly).
Let $R_0(n) \in \rho_0^{-1}\left(w(n)^{-1} n^{1/\alpha d}\right)$ (taking $w(n)$ sufficiently slow such that $R_0(n) \to \infty$), where

$$\rho_0(R) := \rho(R)\left(\log \rho(R)\right)^{1/\alpha d}.$$

If $r = \Omega_k\left(\rho_0(R_0)^\alpha n^{-1/d}\right)$ and $r \geq 4\varphi(R_0)$, then

$$\mathbb{E}[\beta_k] = o(n).$$

Let $R_1(n) \in \rho^{-1}\left(w(n)^{-1}\left(\frac{n}{\log n}\right)^{1/\alpha d}\right)$ (taking $w(n)$ sufficiently slow such that $R_1(n) \to \infty$).
If $r = \Omega_k\left(\rho(R_1)^\alpha \left(\frac{\log n}{n}\right)^{1/d}\right)$ and $r \geq 4\varphi(R_1)$, then

$$\mathbb{E}[\beta_k] = o(1).$$

Corollary A.3.4. *Note that the above theorem applied to the case $k = 0$, indicates that the random variable bandwidth graph $G(X_n; r\rho)$ on $\mathbb{R}^d$ is connected w.h.p. for*

$$r \gtrsim \max\left\{\varphi(R_1), \rho(R_1)^\alpha \left(\frac{\log n}{n}\right)^{1/d}\right\},$$

with R_1 chosen as above; i.e., that for such r we have

$$\lim_n \mathbb{P}\left(G(X_n; r\rho) \text{ is connected}\right) = 1.$$

As it will be evident from the proof below, if the sampling domain is bounded (rather than $\mathbb{R}^d$), then the condition $r \geq \varphi(R)$ can be dropped and R can instead be taken to be a sufficiently large constant, independent of n. In particular, we recover in that case the same connectivity threshold values than those obtained in [41, 55].

A.3.3 Proof of Theorem A.3.3

Throughout the proof, as $n \to \infty$, let $R(n) \to \infty$ and let $r(n) \to 0$ be such that $r \geq 4\varphi(R)$. Note that such a choice of $r \to 0$ is valid, since $\varphi \to 0$ as $R \to \infty$. Indeed, since $q(x) = o(|x|^{-d})$ (otherwise it wouldn't be a density over $\mathbb{R}^d$), we have $\rho(x) = q(x)^{-1/d} = \omega(|x|)$, and

$$\frac{|x|}{\rho(x)} = o(1).$$

Similarly to [41], we require the use of a geometric lemma.

Lemma A.3.5 (Geometric Lemma)**.** *Let $\{y_0, \ldots, y_\ell\} \subset X_n$ be such that*

$$|y_0| < \cdots < |y_\ell|,$$

where

$$|y_0 - y_1| > r\rho_1,$$
$$\frac{r\rho_1}{2} < |y_1|,$$

and such that for all $\{i, j\} \subset \{0, \ldots, \ell\}$, $\{i, j\} \neq \{0, 1\}$,

$$|y_i - y_j| < r \max\{\rho_i, \rho_j\}.$$

Finally, let $I := \bigcap_{i=1}^{\ell} B(y_i, r\rho_i) \cap B(0, |y_1|)$.
There exists $\epsilon_d(n) > 0$ (decreasing in n) independent of $\{y_0, \ldots, y_\ell\}$, such that

$$\mu(I) \geq \epsilon_d r^d.$$

If q is decreasing, we can choose

$$\epsilon_d \sim (1 - \delta)^d,$$

and more generally we can choose

$$\epsilon_d \sim q(R)^2 (1 - \delta)^d,$$

where $\delta \sim \sqrt{1 - \rho(R)^{-3}}$.

The lower bound on $\mu(I)$ obtained above indicates that there is enough room in I to find a ball of radius $\sim \epsilon_d^{1/d} r$ contained in this intersection. In [41] ϵ_d is independent of n and of the y_i's. In our current setting it remains independent of the y_i's, but the sampling set ($\mathbb{R}^d$) being unbounded, we must make ϵ_d dependent on n (as defined above). As we will see, this dependency on n is carried over to the proof of theorem, which explains why the obtained threshold values of r (for the sublinearity of the expected Betti numbers and the k-connectivity

of the complex) carry some extra functions of R (i.e., of n) compared to [41] .

In order to prove the Geometric Lemma, we start by proving the following proposition which finds the centre y_m of the ball contained in the intersection I.

Proposition A.3.6. *There exists $t \in (0,1)$ with $t \sim \rho^{-1}(R)$, and $\delta \in (0,1)$ with $\delta \sim \sqrt{1-\rho(R)^{-3}}$, such that for all $\{y_0, \dots, y_\ell\} \subset X_n$ satisfying the same conditions as in the Geometric Lemma,*

$$\forall i \in [\ell],\ |y_i - y_m| < r\delta\rho_i,$$

where $y_m := ty_0 + (1-t)y_1$.
Note that the choices of t and δ are independent of $\{y_0, \dots, y_\ell\}$.

We note a mild issue in [41], where a similar fact is proved and y_m is chosen to be $y_m = \frac{y_0}{2} + \frac{y_1}{2}$. Such a choice of y_m is not suitable, and one must instead take $y_m = ty_0 + (1-t)y_1$ with a well-chosen t, as it is done in the above proposition. This is because, as we will see, based on the conditions of the Geometric Lemma, we cannot deal with the case $i = 1$ similarly to the case $i \geq 2$.

Proof of Proposition A.3.6. Let $i \in [\ell]$, and suppose that $|y_i| > R$. Using $r \geq 4\varphi(R)$, we find for all $t \in (0,1)$

$$|y_i - y_m| \leq 2|y_i| \leq \frac{r}{2}\rho_i.$$

Suppose now that $|y_i| \leq R$ and suppose that $i \geq 2$.
We have

$$\begin{aligned} |y_i - y_m|^2 &= \left|t(y_i - y_0) + (1-t)(y_i - y_1)\right|^2 \\ &= t^2|y_i - y_0|^2 + (1-t)^2|y_i - y_1|^2 + 2t(1-t)\left\langle y_i - y_0, y_i - y_1\right\rangle, \end{aligned}$$

and

$$2\left\langle y_i - y_0, y_i - y_1\right\rangle = |y_i - y_0|^2 + |y_i - y_0|^2 - |y_0 - y_1|^2,$$

hence

$$\begin{aligned} |y_i - y_m|^2 &= t|y_i - y_0|^2 + (1-t)|y_i - y_1|^2 - t(1-t)|y_0 - y_1|^2 \\ |y_i - y_m|^2 &< r^2\left(\rho_i^2 - t(1-t)\rho_1^2\right). \end{aligned}$$

Thus, it suffices to find $\delta \in (0,1)$ such that

$$\rho_i^2 - t(1-t)\rho_1^2 \leq \delta^2\rho_i^2,$$

hence we want

$$\delta \geq \sqrt{1 - t(1-t)\frac{\rho_1^2}{\rho_i^2}}.$$

Since $|y_i| \leq R \Leftrightarrow \rho_i \leq \rho(R)$, we can choose

$$\delta := \sqrt{1 - t(1-t)\frac{\rho_{min}^2}{\rho(R)^2}},$$

where $\rho_{min} := \min\{\rho(x) | x \in \mathbb{R}^d\}$. This shows that for all $i \geq 2$

$$|y_i - y_m| < r\delta\rho_i;$$

it remains to find $t \in (0,1)$ such that also

$$|y_1 - y_m| < r\delta\rho_1.$$

We have $|y_m - y_1| = t|y_0 - y_1|$, hence it suffices to find $t \in (0,1)$ such that

$$t|y_0 - y_1| \leq 2t|y_1| < 2tR \leq r\delta\rho_1.$$

Using $\rho_{min} \leq \rho_1$ and $\dfrac{R}{r} \leq \rho(R)$, it suffices to find $t \in (0,1)$ such that

$$4t^2\rho(R)^2 \leq (1 - t(1-t)\gamma^2)\rho_{min}^2,$$

where $\gamma := \dfrac{\rho_{min}}{\rho(R)}$;
i.e., we want a quadratic in t with positive leading coefficient, to be non-positive. This quadratic can be showed to intersect the x-axis at

$$t = \frac{-\gamma^2 + \sqrt{\gamma^4 + 4(4\gamma^{-2} - \gamma^2)}}{2(4\gamma^{-2} - \gamma^2)}.$$

Therefore, we can choose t asymptotically in n as

$$t \sim \rho(R)^{-1}$$

and thus

$$\delta \sim \sqrt{1 - \rho(R)^{-3}}.$$

□

We can now prove the Geometric Lemma.

Proof of the Geometric Lemma. From Proposition A.3.6, there exists $\delta \sim \sqrt{1 - \rho(R)^{-3}}$ such that

$$\forall i \in [\ell],\ |y_i - y_m| \leq r\delta\rho_i.$$

For $x \in B(y_m, r(1-\delta)\rho_1)$ and $i \in [\ell]$

$$\begin{aligned} |x - y_i| &\leq |x - y_m| + |y_m - y_i| \\ &\leq r(1-\delta)\rho_1 + r\delta\rho_i; \end{aligned}$$

using $\rho_1 \leq \rho_i$ we have

$$|x - y_i| \leq r\rho_i,$$

i.e., $B(y_m, r(1-\delta)\rho_1) \subset \bigcap_{i=1}^{\ell} B(y_i, r\rho_i)$, and so

$$B(y_m, r(1-\delta)\rho_1) \cap B(0,|y_1|) \subset \bigcap_{i=1}^{\ell} B(y_i, r\rho_i) \cap B(0,|y_1|).$$

Suppose first that q is decreasing. Using $|y_m| \leq |y_1|$ we find

$$\mu\left(B(y_m, r(1-\delta)\rho_1) \cap B(0,|y_1|)\right) \geq \mu\left(B(y_1, r(1-\delta)\rho_1) \cap B(0,|y_1|)\right),$$

and by assumptions $|y_1| > \dfrac{r\rho_1}{2} \geq \dfrac{r\rho_{min}}{2}$, hence

$$\mu\left(B(y_1, r(1-\delta)\rho_1) \cap B(0,|y_1|)\right) \gtrsim (1-\delta)^d r^d,$$

and the Geometric Lemma holds with $\epsilon_d \sim (1-\delta)^d$.

Suppose now that q is not decreasing everywhere on $\mathbb{R}^d$. Then it does not necessarily hold that

$$\mu\left(B(y_1, r(1-\delta)\rho_1) \cap B(0,|y_1|)\right) \geq \mu\left(B(y_1, r(1-\delta)\rho_1) \cap B(0,|y_1|)^c\right).$$

If it did, we would get better estimates for the lower bound, as done in the case where q is decreasing, hence we may assume that

$$\mu\left(B(y_1, r(1-\delta)\rho_1) \cap B(0,|y_1|)^c\right) \geq \mu\left(B(y_1, r(1-\delta)\rho_1) \cap B(0,|y_1|)\right),$$

hence that

$$\mu\left(B(y_1, r(1-\delta)\rho_1) \cap B(0,|y_1|)^c\right) \geq \frac{1}{2}\mu\left(B(y_1, r(1-\delta)\rho_1)\right).$$

Denote the Lebesgue measure by m, and let $q_{\max} = \sup\{q(x) \mid x \in \mathbb{R}^d\}$.

Observe that

$$m\left(B(y_1, r(1-\delta)\rho_1) \cap B(0,|y_1|))\right) \gtrsim m\left(B(y_1, r(1-\delta)\rho_1) \cap B(0,|y_1|))^c\right),$$

hence that

$$\begin{aligned}\mu\left(B(y_1, r(1-\delta)\rho_1) \cap B(0,|y_1|))^c\right) &\leq q_{\max} m\left(B(y_1, r(1-\delta)\rho_1) \cap B(0,|y_1|))^c\right) \\ &\lesssim m\left(B(y_1, r(1-\delta)\rho_1) \cap B(0,|y_1|))\right).\end{aligned}$$

Combining the above, we have

$$\begin{aligned}\mu\left(B(y_1, r(1-\delta)\rho_1) \cap B(0,|y_1|))\right) &\geq q(R) m\left(B(y_1, r(1-\delta)\rho_1) \cap B(0,|y_1|))\right) \\ &\gtrsim q(R)\mu\left(B(y_1, r(1-\delta)\rho_1)\right) \\ &\gtrsim q(R)((1-\delta)r)^d.\end{aligned}$$

Therefore, we find

$$\begin{aligned}\mu\left(B(y_m, r(1-\delta)\rho_1) \cap B(0,|y_1|\right) &\geq q(R)m\left(B(y_m, r(1-\delta)\rho_1) \cap B(0,|y_1|\right)\\ &\geq q(R)m\left(B(y_1, r(1-\delta)\rho_1) \cap B(0,|y_1|\right)\\ &\gtrsim q(R)\mu\left(B(y_1, r(1-\delta)\rho_1) \cap B(0,|y_1|\right)\\ &\gtrsim q(R)^2(1-\delta)^d r^d;\end{aligned}$$

hence the lemma holds in this more general case with $\epsilon_d \sim q(R)^2(1-\delta)^d$.

□

Using the Geometric Lemma, we can prove Theorem A.3.3.

Proof of Theorem A.3.3. Up to relabelling, we have a.s. $X_n = \{x_1, \dots, x_n\}$ with $|x_1| < |x_2| \cdots < |x_n|$.

We build a discrete gradient field V on $R(X_n; r\rho)$ as done in [41]. We explain again the construction, as it will be helpful to have it in mind for the rest of the proof.

For every face $S = [x_{i_1}, \dots, x_{i_j}]$, pair it if possible with the face $[x_{i_0}, x_{i_1}, \dots, x_{i_j}]$, where $i_0 < i_1$ is chosen as small as possible. As verified in [41], each face is paired at most once, which guarantees that V is a well-defined discrete gradient vector field.

Let us bound the probability p_k that a set of $k+1$ vertices span a k-dimensional face in the Vietoris-Rips complex. Let $x_1, \dots, x_{k+1}$ be the $k+1$ vertices, and assume without loss of generality (up to relabelling) that $\rho(x_{k+1}) = \max \rho(x_i)$. Then for all $i \in [k+1]$ we must have $x_i \in B(x_{k+1}, r\rho(x_{k+1}))$, hence

$$p_k = O\left(\mu\left(B(x_{k+1}, r\rho(x_{k+1}))\right)^k\right) = O(r^{dk}).$$

We then estimate the probability that such a k-dimensional face is critical with respect to the discrete field V.

Using a similar argument to that in [41], if the face F is critical, i.e., unpaired, then there is no common neighbor x_a of the vertices $x_{i_1}, \dots, x_{i_{k+1}}$ with $a < i_1$. And for similar reasons, there must be some x_{i_0} common neightbor of $F \setminus \{x_{i_1}\}$ with $i_0 < i_1$.

Note that letting $y_j := x_{i_j}$, we satisfy the conditions of the Geometric Lemma, hence we can deduce that

$$\mu(I) \geq \epsilon_d r^d,$$

where

$$I := \bigcap_{j=1}^{k+1} B(x_{i_j}, r\rho_j) \cap B(0,|x_{i_1}|),$$

and $\epsilon_d > 0$ is independent of F but dependent on n, as above. We have thus found a lower bound estimate for the measure of I, which yields an estimate for the probability that a random point from $\mathbb{R}^d$ belongs to I. Therefore, by independence of the random points in $\mathbb{R}^d$, the probability that a k-face F is critical is

$$\begin{aligned} p_c &\leq \left(1 - \epsilon_d r^d\right)^{n-k-2} \\ &= O\left(\exp(-c(n)\Lambda)\right), \end{aligned}$$

where $\Lambda := nr^d$ and $c(n)$ is an arbitrary constant in $(0, \epsilon_d)$.

Thus, the expected number of critical k-faces is given by

$$\begin{aligned} \mathbb{E}\left[C_k\right] &\leq \binom{n}{k+1} p_k p_c \\ &= O\left(n\Lambda^k e^{-c(n)\Lambda}\right). \end{aligned}$$

Recall that if q is decreasing we may choose $\epsilon_d \sim (1-\delta)^d$, and more generally, we may choose $\epsilon_d \sim q(R)^2(1-\delta)^d$.

With $\delta \sim \sqrt{1 - \rho(R)^{-3}}$, it is easy to verify that $(1-\delta)^d \gtrsim \rho(R)^{-4d}$. Hence we can pick

$$c(n) \sim \rho(R)^{-4d}$$

if q is decreasing, and

$$c(n) \sim q(R)^2 \rho(R)^{-4d} = \rho(R)^{-6d}$$

more generally.

It remains to find a value of r beyond which the expected Betti numbers grow sublinearly, and a value of r beyond which the VR complex becomes k-connected.

Let $k \geq 0$, and let R_0 be defined as in Theorem 5.2. This guarantees that $\rho_0(R_0)^\alpha n^{-1/d} = o(1)$. Let $c_k > (\alpha k d)^{1/d}$ and suppose that $r \geq 4\varphi(R_0)$ and that

$$\begin{aligned} r &= c_k \rho(R_0)^\alpha \log\left(\rho(R_0)\right)^{1/d} n^{-1/d} \\ \Leftrightarrow \Lambda \rho(R_0)^{-4d} &= c_k^d \log\left(\rho(R_0)\right). \end{aligned}$$

We see that

$$\log \Lambda = d \log(c_k) + \log\log\left(\rho(R_0)\right) + \alpha d \log\left(\rho(R_0)\right),$$

hence from the choice of c_k, we have

$$\lim_n (k \log \Lambda - c(n)\Lambda) = -\infty,$$

i.e.,

$$\lim_n \exp\left(k \log \Lambda - c(n)\Lambda\right) = 0,$$

which guarantees that

$$\mathbb{E}\left[\beta_k\right] = o(n).$$

Similarly, let R_1 be defined as in Theorem 5.2, and let $c_k > 0$ be such that

$$\begin{aligned} r &= c_k \rho(R_1)^\alpha \left(\frac{\log n}{n}\right)^{1/d} \\ \Leftrightarrow nr^d &=: \Lambda = c_k^d \rho(R_1)^{\alpha d} \log n. \end{aligned}$$

Using the fact that $\rho(R_1)^\alpha = o\left(\left(\frac{n}{\log n}\right)^{1/d}\right) \Rightarrow \log(\rho(R_1)) = o(\log n)$, we have

$$\lim_n (k \log(\Lambda) - \rho(R_1)^{-\alpha d}\Lambda + \log n) = -\infty,$$

i.e.,

$$\mathbb{E}\left[\beta_k\right] = \Lambda^k \exp\left(-c(n)\Lambda\right) n = o(1).$$

□