Electronic Journal of Statistics
Vol. 18 (2024) 5659–5678
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2331

A note on estimating the dimension from a random geometric graph

Caelan Atamanchuk[1], Luc Devroye[1] and Gábor Lugosi[2,3,4]

[1] *School of Computer Science, McGill University, Montreal, Canada, e-mail:* caelan.atamanchuk@gmail.com; lucdevroye@gmail.com

[2] *Department of Economics and Business, Pompeu Fabra University, Barcelona, Spain*

[3] *ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain*

[4] *Barcelona Graduate School of Economics, e-mail:* gabor.lugosi@gmail.com

Abstract: Let G_n be a random geometric graph with vertex set $[n]$ based on n i.i.d. random vectors $X_1, \ldots, X_n$ drawn from an unknown density f on $\mathbb{R}^d$. An edge (i, j) is present when $\|X_i - X_j\| \leq r_n$, for a given threshold r_n possibly depending upon n, where $\|\cdot\|$ denotes Euclidean distance. We study the problem of estimating the dimension d of the underlying space when we have access to the adjacency matrix of the graph but do not know r_n or the vectors X_i. The main result of the paper is that there exists an estimator of d that converges to d in probability as $n \to \infty$ for all densities with $\int f^5 < \infty$ whenever $n^{3/2} r_n^d \to \infty$ and $r_n = o(1)$. The conditions allow very sparse graphs since when $n^{3/2} r_n^d \to 0$, the graph contains isolated edges only, with high probability. We also show that, without any condition on the density, a consistent estimator of d exists when $n r_n^d \to \infty$ and $r_n = o(1)$.

MSC2020 subject classifications: Primary 62G05; secondary 05C80.
Keywords and phrases: Multivariate densities, nonparametric estimation, random geometric graphs, estimating the dimension, absolute continuity.

Received June 2024.

In network science one often seeks geometric representations of an observed network that help interpret and predict connections and understand the structure of the network. Indeed, embedding a (weighted) graph in a low-dimensional Euclidean space—called *multidimensional scaling* in statistics—is a thoroughly studied problem; see the monographs of Borg and Groenen [6] and Borg, Groenen, and Mair [7] for a comprehensive treatment. For a sample of the literature on closely related approaches in computational geometry and machine learning, see Reiterman, Rödl, Šiňajová [34], Tenenbaum, Silva, and Langford [36], Shavitt and Tankel [35], Kleinberg [27], Kang and Müller [25], Verbeek and Suri [38]. A more basic question is to determine the dimension of the underlying geometric space. In this paper we consider the problem of estimating the dimension of the Euclidean space underlying a geometric graph, upon observing a (combinatorial) graph.

arXiv: 2311.13059

In order to set up a rigorous statistical problem, we model the graph as a random geometric graph. Indeed, random geometric graphs occur naturally in network models in a variety of areas, including bioinformatics or the analysis of social media. We refer the reader to Duchemin and De Castro [14] for a review and pointers to the literature. Let G_n be a random geometric graph with vertex set $[n]$ based on n i.i.d. random vectors $X_1, \ldots, X_n$ drawn from an unknown density f on $\mathbb{R}^d$. An edge (i, j) is present when $\|X_i - X_j\| \leq r_n$, for a given threshold r_n possibly depending upon n, where $\|\cdot\|$ denotes Euclidean distance. Introduced by Gilbert [18, 19], the properties of these graphs have been well studied when f is the uniform density on a convex set of $\mathbb{R}^d$ or the torus $[0, 1]^d$ in $\mathbb{R}^d$. Its properties are surveyed by Penrose [31]. Noteworthy are the precise results on connectivity (Appel and Russo [2]; Balister, Bollobás and Sarkar [4]; Balister, Bollobás, Sarkar, and Walters [5]), cover time (Cooper and Frieze [12]), coverage (Gilbert [17]; Hall [22], Janson [24]), chromatic number (McDiarmid and Müller [30]), and minimal spanning tree (Penrose [32]).

In the dimension-estimation problem considered here, we observe the adjacency matrix of the graph G_n but we do not know d, r_n or the vertex locations X_i. The question then is whether one can estimate the underlying dimension d. In other words, can one develop an estimate Δ_n of d, based only on knowledge of G_n, with the property that $\Delta_n \to d$ in probability as $n \to \infty$? When this convergence happens, we say that Δ_n is a *consistent* estimator of d.

Whether consistent estimators exist, may depend on the parameters of the model, that is, the density f and the sequence of radii $\{r_n\}$. For example, if the graph is too sparse, there is no hope to estimate d. Indeed, suppose that f is the uniform density on $[0, 1]^d$, and r_n is such that $n^{3/2} r_n^d \to 0$. Then G_n only contains isolated edges, with high probability. Indeed,

$$\begin{aligned}
&\mathbb{P}\left\{\exists \text{ distinct } i, j, k \in [n] : \|X_i - X_j\| \leq r_n \text{ and } \|X_i - X_k\| \leq r_n\right\} \\
&\quad \leq \quad n^3 \mathbb{E}\left[\mathbb{P}\left\{\|X_1 - X_2\| \leq r_n \text{ and } \|X_1 - X_3\| \leq r_n \middle| X_1\right\}\right] \leq n^3 V_d^2 r_n^{2d} \to 0 \ ,
\end{aligned}$$

where V_d denotes the volume of the unit ball in $\mathbb{R}^d$

For such graphs it is clearly impossible to infer anything about the underlying geometry. The main result of this paper shows that, as soon as r_n is such that $n^{3/2} r_n^d \to \infty$, it is possible to consistently estimate the dimension, for a large class of densities. More precisely, we prove the following.

Theorem 0.1. *Let the density f on $\mathbb{R}^d$ satisfy $\int f^5 < \infty$. Assume furthermore that*

$$\lim_{n \to \infty} n^{3/2} r_n^d = \infty \ ,$$

and $r_n = o(1)$. Then there exists an estimate Δ_n such that $\Delta_n \to d$ in probability.

The condition on the radius r_n allows extremely sparse graphs. It suffices to have $r_n^d \sim n^{-3/2} \omega_n$ for $\omega_n \to \infty$ arbitrarily slowly. Note that in that case the graph has merely $O_p(\sqrt{n}\omega_n)$ edges. Such graphs are extremely sparse as the great majority of vertices are isolated. Indeed, the expected degree of a

typical vertex is of the order of $n^{-1/2}\omega_n$. The condition $\int f^5 < \infty$ excludes densities with pronounced infinite peaks but it does not assume anything about the smoothness or tails of the distribution. We also prove a consistency result for arbitrary densities, though under more stringent conditions on the radii r_n:

Theorem 0.2. *Let the density f in $\mathbb{R}^d$ be arbitrary, and assume that*

$$\lim_{n\to\infty} nr_n^d = \infty \ , \quad \textit{and} \quad \lim_{n\to\infty} r_n = 0 \ .$$

Then there exists an estimate Δ_n such that $\Delta_n \to d$ in probability.

The condition $nr_n^d \to \infty$ implies that the expected degree of a typical vertex goes to infinity, albeit arbitrarily slowly. While this is significantly more restrictive than the condition of Theorem 0.1, it still allows the graph to be relatively sparse.

It is an interesting open question whether there exist dimension estimators that are consistent for all densities under the minimal assumption $n^{3/2}r_n^d \to \infty$. We conjecture that the estimator used in this paper to prove Theorem 0.1 is not consistent for all densities, though the condition $\int f^5 < \infty$ may possibly be relaxed to $\int f^3 < \infty$, as discussed below.

The paper is organized as follows. After reviewing some of the related literature, in Section 1 we introduce four simple estimators of the dimension whose analysis proves Theorems 0.1 and 0.2. We start analyzing the estimators in Section 2 by focusing on the special–but important–case of the uniform density on the torus. Finally, in Section 3 we state Theorems 0.1 and 0.2 for general densities. In Section 4 we establish a geometric lemma that is a key tool in our approach of defining estimators of the dimension. Most proofs are presented in the Appendix.

Related literature

Numerous heuristics have been proposed for choosing the embedding dimension in multidimensional scaling (Hout, Papesh, and Goldinger [23]). For a principled appoach, see Peterfreund and Gavish [33].

Granata and Carnevale [20] consider the dimension-estimation problem in a more general framework of estimating the intrinsic dimension of geometric graphs defined in general metric spaces. Instead of focusing on general conditions for consistency, [20] aim to construct accurate estimates from graph distances.

Bubeck, Ding, Eldan, and Rácz [9] show that, based on a dense random geometric graph drawn from the uniform distribution on the surface of the d-dimensional unit sphere, it is possible to estimate d as long as $n \gg d$.

Lichev, and Mitsche [29] and Casse [11] study properties of the online nearest neighbor tree based on uniformly distributed points in $[0,1]^d$ and observe that it is possible to consistently estimate the dimension upon observing the combinatorial tree.

Dimension-estimation from geometric graphs is closely related to the problem of estimating the intrinsic dimensionality of high-dimensional data, see Fukunaga and Olsen [16], Verveer and Duin [40], Bruske and Sommer [8], Tenenbaum, Silva, and Langford [36], Facco, d'Errico, Rodriguez, Laio [15]. Indeed, often the first step of computing such estimates is to construct a geometric graph from the data, see, e.g., Grassberger and Procaccia [21], Camastra and Vinciarelli [10], Kégl [26], Costa and Hero [13], Levina and Bickel [28]. However, all these approaches assume knowledge of the pairwise distances between data points, which is equivalent to having access to the geometric graph G_n *at all scales* $r > 0$. In this paper we only assume access to the graph at a single threshold value r_n.

Araya and De Castro [3] study estimating the Euclidean distances between the point locations upon observing the combinatorial graph for dense random geometric graphs, though their results also apply to relatively sparse graphs. von Luxburg and Alamgir [41] discuss the problem of estimating the density function of the data distribution based on observing the k-nearest neighbor graph generated by the data.

1. The proposed estimates

In this section we introduce four simple estimators of the dimension d.

Consider the unit ball $B(0,1)$ in $\mathbb{R}^d$, and let X and Y be independent and uniformly distributed in $B(0,1)$. We define the quantity

$$w_d = \mathbb{P}\{\|X - Y\| \le 1\} \ .$$

All four estimators studied in this paper are based on estimating w_d. It is shown in Lemma 4.1 below that the sequence w_d decreases strictly monotonically to 0 as $d \uparrow \infty$.

If W is a data-based estimate, then we set

$$\Delta_n = \operatorname*{argmin}_d |W - w_d|.$$

In view of Lemma 4.1, if $W \to w_d$ in probability, then $\Delta_n \to d$ in probability and therefore it suffices to construct consistent estimators of w_d.

By using binary search (first doubling the dimension until an overshoot occurs, and then applying classical binary search), one can find Δ_n using only $O(\log d)$ computations of the function w_s.

We propose simple local estimates W_1 and W_4 and more powerful global estimates W_2 and W_3. Randomly label the nodes of the graph such that all labelings are equally likely. Denote the degree of vertex i in G_n by D_i, and let δ_i be the number of edges between nodes in N_i, the set of neighbors of vertex i (excluding the vertex i). Let M be the smallest index among the vertices of maximal degree. Let ξ_{ij} be the indicator that i is connected to j. Our estimates are as follows:

$$W_1 \stackrel{\text{def.}}{=} \frac{\delta_M}{\binom{D_M}{2}} \ ,$$

$$W_2 \stackrel{\text{def.}}{=} \frac{\sum_{i<j<k} \xi_{ki}\xi_{kj}\xi_{ij}}{\sum_{i<j<k} \xi_{ki}\xi_{kj}} ,$$

$$W_3 \stackrel{\text{def.}}{=} \frac{\sum_{i=1}^{n} \frac{\delta_i}{\binom{D_i}{2}}}{\sum_{i=1}^{n} \mathbb{1}_{D_i \geq 2}} ,$$

and

$$W_4 \stackrel{\text{def.}}{=} \frac{\delta_1}{\binom{D_1}{2}} .$$

We interpret all these ratios as zero when the denominator equals zero. Note that W_1 is the edge density of the subgraph spanned by the neighborhood of a vertex of maximal degree, while W_4 is that of a randomly chosen vertex. W_2 counts the ratio of the number of triangles in the graph and the number of paths of length two. W_2 is often called the *transitivity* of the network. Finally, the estimator W_3 is identical to the *clustering coefficient* introduced by Watts and Strogatz [42].

2. Analysis for the uniform density on the torus

In this section we focus on the uniform density on $[0,1]^d$, and measure Euclidean distances as in the torus, that is, for $x, y \in [0,1]^d$,

$$\|x-y\| \stackrel{\text{def.}}{=} \min_{z \in Z^d} \|x-y+z\| ,$$

where Z^d is the collection of all integer-valued d-dimensional vectors. This allows us to present some of the ideas in a transparent manner.

Assume that $r_n \leq 1/2$ to avoid the wraparound effect in the torus. We begin by analyzing the statistic δ_1, assuming that $D_1 \geq 2$. Note that we can represent δ_1 as

$$\delta_1 \stackrel{\mathcal{L}}{=} \sum_{i,j \in \{1,\ldots,D_1\}: i<j} Y_{ij} ,$$

where $Y_{ij} = \mathbb{1}_{\|Z_i - Z_j\| \leq 1}$, and $Z_1, \ldots, Z_{D_1}$ are i.i.d. random vectors uniformly distributed in the unit ball of $\mathbb{R}^d$. This representation is valid since the X_i are uniformly distributed and by the assumption $r_n \leq 1/2$. Moreover, conditioned on a fixed number of points falling in the neighborhood of radius r_n of any point, their distribution is independent uniform in that neighborhood. Each random variable Y_{ij} is Bernoulli (w_d). Thus, still for $D_1 \geq 2$,

$$\mathbb{E}\left\{ \frac{\delta_1}{\binom{D_1}{2}} \mid D_1 \right\} = w_d ,$$

so that the estimator W_4 is unbiased. Then,

$$\operatorname{Var}\{\delta_1 | D_1\}$$

$$= \mathbb{E}\left\{ \left(\sum_{i<j\le D_1} (Y_{ij} - w_d) \right)^2 \mid D_1 \right\}$$
$$= \binom{D_1}{2} \mathbb{E}\left\{ (Y_{12} - w_d)^2 | D_1 \right\} + 3\binom{D_1}{3} \mathbb{E}\left\{ (Y_{12} - w_d)(Y_{13} - w_d) | D_1 \right\} .$$

Note however that, given D_1, Y_{12} and Y_{13} are conditionally independent, so that we can conclude that

$$\mathrm{Var}\{\delta_1 | D_1\} = \binom{D_1}{2} w_d(1 - w_d) .$$

Therefore, by the Chebyshev-Cantelli inequality, if $D_1 \ge 2$ and $t > 0$,

$$\begin{aligned} \mathbb{P}\left\{ \left| \frac{\delta_1}{\binom{D_1}{2}} - w_d \right| \ge t \mid D_1 \right\} &\le \frac{\mathrm{Var}\{\delta_1 | D_1\}}{\mathrm{Var}\{\delta_1 | D_1\} + \binom{D_1}{2}^2 t^2} \\ &= \frac{\binom{D_1}{2} w_d(1 - w_d)}{\binom{D_1}{2} w_d(1 - w_d) + \binom{D_1}{2}^2 t^2} \\ &= \frac{w_d(1 - w_d)}{w_d(1 - w_d) + \binom{D_1}{2} t^2} \\ &\le \frac{w_d}{w_d + \binom{D_1}{2} t^2} . \end{aligned}$$

We conclude that $W_4 = \delta_1 / \binom{D_1}{2} \to w_d$ in probability when $D_1 \to \infty$ in probability. As it is discussed in the proof of the next theorem, a sufficient condition for this is that $nr_n^d \to \infty$. This latter condition may be somewhat relaxed if one replaces the estimator W_4 by W_1. This is shown in the following theorem which summarizes the consistency properties of the estimators W_1 and W_2.

Theorem 2.1. *Let the density f be the uniform density on the unit torus $[0,1]^d$, and assume that $r_n \le 1/2$ for all n.*

(i) If

$$\lim_{n\to\infty} nr_n^{d(1-\epsilon)} = \infty$$

for all $\epsilon > 0$, then $W_1 \to w_d$ (and thus $\Delta_n \to d$) in probability as $n \to \infty$. A sufficient condition for this is that $nr_n^d \ge L(n)$ with $L(n)$ slowly varying.

(ii) If $r_n \to 0$ and $n^{3/2} r_n^d \to \infty$, then $W_2 \to w_d$ (and thus $\Delta_n \to d$) in probability as $n \to \infty$.

The computational complexity of the inferior estimate W_1 is less than that of W_2, so both estimates have their use. On the other hand, W_1 requires at least $n/L(n)$ edges, where $L(n)$ is slowly varying. For example, for constant k, $n/\log^k(n)$ edges will do. At the same time, one may reduce the computational cost of W_2 by considering Monte-Carlo subsampling to approximate W_2.

3. General densities

In the Appendix we prove the following theorem, which implies Theorem 0.2.

Theorem 3.1. *Let the density f be arbitrary. Assume that $r_n = o(1)$ and $nr_n^d \to \infty$. Then $W_4 \to w_d$ (and thus $\Delta_n \to d$) in probability as $n \to \infty$.*

Finally, the following result—proven in the Appendix–implies Theorem 0.1.

Theorem 3.2. *Let the density f have $\int f^5 < \infty$. Assume that $r_n = o(1)$ and $n^{3/2} r_n^d \to \infty$. Then $W_2 \to w_d$ (and thus $\Delta_n \to d$) in probability as $n \to \infty$.*

We suspect that $\int f^3 < \infty$ suffices in Theorem 3.2, but this would require a substantially longer proof. In any case, the restriction $\int f^5 < \infty$ would imply, for example, that for the univariate beta (a, b) density, we need to have $\min(a, b) > 4/5$. Nevertheless the theorem still covers most densities, including some that are nowhere continuous.

4. A geometric lemma

Recall that

$$w_d = \mathbb{P}\{\|X - Y\| \le 1\} \ ,$$

where X and Y are independent, uniformly distributed random vectors in the unit ball $B(0, 1)$ in $\mathbb{R}^d$. The following property shows that by consistently estimating w_d, one obtains a consistent estimator of d.

Lemma 4.1. *We have*

$$w_d = \frac{3}{2}\mathbb{P}\left\{\beta\left(\frac{1}{2}, \frac{d+1}{2}\right) \ge \frac{1}{4}\right\} \ ,$$

where $\beta(a, b)$ denotes a beta random variable with shape parameters a and b. The sequence w_d decreases strictly monotonically to 0 as $d \uparrow \infty$.

On the value of w_d

The explicit density of $\|X - Y\|$ was derived by Aharonyan and Khalatyan [1]. From it, one can deduce a formula for w_d as a function of some gamma functions. As Lemma 4.1 shows, the constant w_d is simply related to the upper tail of a beta random variable, so w_d is a constant times an incomplete beta integral. For general properties of random variables uniformly distributed in high-dimensional convex sets, we refer to Vershynin [39].

On the sample size needed

The representation of w_d given in Lemma 4.1 permits us to show that $w_d - w_{d+1} \ge d^{-(d+o(d))/2}$ (see Appendix E). Our proposed algorithms are all based

on estimates of w_d, and have errors that decline at polynomial rates in n, the sample size. Thus, while all estimates are consistent in the limit, there is no hope of a good performance when $d \gg \log n / \log\log n$. In [9] it is shown that, in the case of very dense graphs and the uniform density on the surface of the unit sphere, there exist estimators that work well as soon as $n \gg d$. It is a challenging problem for further research to determine the exact tradeoff between edge density and required sample size for accurately estimating d.

Appendix A: Proof of Theorem 2.1

Proof of (i). Replacing D_1 by D_M in the analysis of $\delta_1/\binom{D_1}{2}$ implies that if we can show that $D_M \to \infty$ in probability, then $W_1 \to w_d$ in probability, and thus $\Delta_n \to d$ in probability as $n \to \infty$. For a random geometric graph on the torus of $\mathbb{R}^d$, we have from simple considerations that $D_M \to \infty$ in probability if for all $\epsilon > 0$

$$\lim_{n\to\infty} n r_n^{d(1-\epsilon)} = \infty \ .$$

A sufficient condition for this is that $n r_n^d \geq L(n)$ with $L(n)$ slowly varying. See Bingham, Goldie and Teugels [37] for more on this topic. We have $D_M \geq D_1$, and thus $D_M \to \infty$ in probability when $D_1 \to \infty$ in probability. As D_1 is binomial $(n-1, V_d r_n^d)$, where V_d denotes the volume of the unit ball in $\mathbb{R}^d$, we have $D_1 \to \infty$ in probability when $n r_n^d \to \infty$. When $n r_n^d \to c > 0$ for a constant c, we know that $D_M \sim \log n / \log\log n$ in probability by a Poissonization argument. Thus, to show that $D_M \to \infty$, we just need to consider the case $n r_n^d \to 0$.

Fix an arbitrary large integer t. Then $D_M < t$ means that for each data point, the t-th nearest neighbor is at least distance r away. So, we grid the torus with cubes of side length $\rho \stackrel{\text{def.}}{=} r_n/(2\sqrt{d})$, which ensures that each cell in the grid can at most have t data points. As the cardinalities of the cells jointly form a multinomial random vector, and the multinomial components are negatively associated, we have

$$\begin{aligned}
\mathbb{P}\{D_M < t\} &\leq \mathbb{P}\{\text{all cells have } \leq t \text{ data points}\} \\
&\leq (\mathbb{P}\{\text{Binomial}(n, \rho^d) \leq t\})^{1/\rho^d} \\
&\leq \exp\left(-\frac{\mathbb{P}\{\text{Binomial}(n, \rho^d) > t\}}{\rho^d}\right) \\
&\leq \exp\left(-\frac{\binom{n}{t+1}\rho^{d(t+1)}(1-\rho^d)^{n-t-1}}{\rho^d}\right) \ .
\end{aligned}$$

The absolute value of the exponent is of asymptotic order

$$\frac{(n\rho^d)^{t+1}}{\rho^d} = \Theta\left(n^{t+1} r_n^{dt}\right) \ ,$$

and this tends to ∞ as $n \to \infty$ by our condition. □

Proof of (ii). We rewrite the estimate as

$$W_2 \stackrel{\text{def.}}{=} A_1/A_2 \ ,$$

where

$$A_1 = \frac{1}{\binom{n}{3}} \sum_{1\le i<j<k\le n} \xi_{ki}\xi_{kj}\xi_{ij}$$

and

$$A_2 = \frac{1}{\binom{n}{3}} \sum_{1\le i<j<k\le n} \xi_{ki}\xi_{kj} \ .$$

We observe that $\mathbb{E}\{A_1\} = w_d\pi_n^2$ and $\mathbb{E}\{A_2\} = \pi_n^2$, where $\pi_n \stackrel{\text{def.}}{=} V_d r_n^d$. The ratio of these means is w_d. By bounding the variances of both we shall show that $A_1/\mathbb{E}\{A_1\} \to 1$ and $A_2/\mathbb{E}\{A_2\} \to 1$ in probability, so that $A_1/A_2 \to w_d$ in probability, as required.

We begin with

$$\mathrm{Var}\left\{\sum_{i<j<k} \xi_{ki}\xi_{kj}\right\} = \mathbb{E}\left\{\left(\sum_{i<j<k} (\xi_{ki}-\pi_n)(\xi_{kj}-\pi_n)\right)^2\right\} \ .$$

Let s be the set $\{i,j,k\}$ and let s' be the set $\{i',j',k'\}$. Observe that if $|s\cap s'| \le 1$, then $\xi_{ki}, \xi_{kj}, \xi_{k'i'}, \xi_{k'j'}$ are independent. When expanding the squared expression, we are left with contributions coming from the cases when $|s\cap s'| \ge 2$. If the intersection is of size three, then the two ordered triples are identical. This yields a term equal to

$$\mathbb{E}\left\{\sum_{i<j<k} (\xi_{ki}-\pi_n)^2(\xi_{kj}-\pi_n)^2\right\} = \binom{n}{3}\pi_n^2(1-\pi_n)^2 \le n^3\pi_n^2.$$

When $|s\cap t| = 2$, the graph formed by $(k,i),(k,j),(k',i'),(k',j')$ is either a tree (in fact, a star on four vertices) or a 4-cycle. Only in the latter case do we have dependence and a non-vanishing contribution. To see this, note that, for example, if $k=k'$ and $i=i'$, then the corresponding term equals

$$\mathbb{E}\left\{(\xi_{ki}-\pi_n)^2(\xi_{kj}-\pi_n)(\xi_{kj'}-\pi_n)\right\} = 0 \ .$$

On the other hand, in the case of a 4-cycle, for example, when $i=i'$ and $j=j'$, then the corresponding term may be bounded as follows:

$$\begin{aligned}
&\mathbb{E}\left\{(\xi_{ki}-\pi_n)(\xi_{kj}-\pi_n)(\xi_{k'i}-\pi_n)(\xi_{k'j}-\pi_n)\right\} \\
&\quad\le \quad \mathbb{E}\left\{|\xi_{ki}-\pi_n||\xi_{kj}-\pi_n||\xi_{k'i}-\pi_n|\right\} \\
&\quad= \quad \mathbb{E}\left\{|\xi_{ki}-\pi_n|\right\}^3 \le \pi_n^3 \ .
\end{aligned}$$

Therefore, the contribution to the variance coming from the 4-cycles is at most $n^4\pi_n^3$. We conclude that

$$\mathrm{Var}\left\{\sum_{i<j<k}\xi_{ki}\xi_{kj}\right\} \le n^3\pi_n^2 + n^4\pi_n^3 \ .$$

As

$$\left(\mathbb{E}\left\{\sum_{i<j<k}\xi_{ki}\xi_{kj}\right\}\right)^2 = \binom{n}{3}^2\pi_n^4 \ ,$$

Chebyshev's inequality shows that $A_2/\mathbb{E}\{A_2\} \to 1$ in probability whenever

$$n^3\pi_n^2 \to \infty \ .$$

The reasoning for A_1 is similar. When expanding

$$\mathrm{Var}\left\{\sum_{i<j<k}\xi_{ki}\xi_{kj}\xi_{ij}\right\} = \mathbb{E}\left\{\left(\sum_{i<j<k}(\xi_{ki}-\pi_n)(\xi_{kj}-\pi_n)(\xi_{ij}-\pi_n)\right)^2\right\},$$

we once again only need to consider triples s and s' with $|s\cap s'| \ge 2$. When the intersection is of size 3, the contribution to the variance is $O(n^3\pi_n^2)$. When the intersection is of size 2, after breaking up the two cycles in the graph formed by the five edges involves, the contribution to the variance is easily seen to be $O(n^4\pi_n^3)$. Arguing as for A_2, we conclude that $A_1/\mathbb{E}\{A_1\} \to 1$ in probability whenever $n^3\pi_n^2 \to \infty$. □

Appendix B: Proof of Theorem 3.1

We condition on X_1 and D_1, and let $Y_1, \ldots, Y_{D_1}$ be i.i.d. random vectors drawn from f restricted to the ball $B(X_1, r_n)$. Set $\xi_{ij} = \mathbb{1}_{||Y_i - Y_j|| \le r_n}$. Let $Z_1, Z_2, \ldots$ be i.i.d. uniform random variables on $B(X_1, r_n)$. Then

$$\delta_1 = \sum_{1\le i<j\le D_1}\xi_{ij}$$

and therefore,

$$\mathbb{E}\left\{W_4 \mid X_1, D_1\right\} = \mathbb{1}_{D_1\ge 2} \times \mathbb{P}\{||Y_1 - Y_2|| \le r_n | X_1\} \ .$$

Set $\pi_n = V_d r_n^d$ and $\mu_n(x) = \int_{B(x,r_n)} f$, $x \in \mathbb{R}^d$. The density of Y_1 given X_1 is given by

$$\frac{f(y)}{\mu_n(X_1)}\mathbb{1}_{y\in B(X_1,r_n)} \ .$$

The total variation distance $\mathrm{TV}(Y_1, Z_1)$ given X_1 is

$$\begin{aligned}\frac{1}{2}\int_{B(X_1,r_n)}\left|\frac{f(y)}{\mu_n(X_1)}-\frac{1}{\pi_n}\right|dy &= \frac{1}{2\mu_n(X_1)}\int_{B(X_1,r_n)}\left|f(y)-\frac{\mu_n(X_1)}{\pi_n}\right|dy \\ &\stackrel{\text{def.}}{=} \frac{\psi_n(X_1)}{2\mu_n(X_1)}. \\ &= \frac{\psi_n(X_1)}{\pi_n}\times\frac{\pi_n}{2\mu_n(X_1)}.\end{aligned}$$

The Lebesgue density theorem (see, e.g., Wheeden and Zygmund [43]), implies that for almost all x,

$$\lim_{r\downarrow 0}\frac{1}{V_d r^d}\int_{B(x,r)} f(y)dy = f(x)\ .$$

Thus, as $r_n \downarrow 0$, $\psi_n(x)/\pi_n \to 0$ as $n\to\infty$ for almost all x. Similarly, $\mu_n(x)/\pi_n \to f(x)$ as $n\to\infty$ for almost all x. We can couple Y_1 and Z_1 such that, for any x with $f(x)>0$, conditional on $X_1 = x$,

$$\mathbb{P}\{Y_1 \neq Z_1\} = o(1)$$

for almost all x. Similarly, we can couple Y_2 with a uniform random vector Z_2. Thus,

$$\begin{aligned}|\mathbb{P}\{||Y_1-Y_2||\le r_n\}-w_d| &= |\mathbb{P}\{||Y_1-Y_2||\le r_n\}-\mathbb{P}\{||Z_1-Z_2||\le r_n\}| \\ &\le \mathbb{P}\{[Y_1\neq Z_1]\cup[Y_2\neq Z_2]\} \\ &= \int f(x)\mathbb{P}\{[Y_1\neq Z_1]\cup[Y_2\neq Z_2]|X_1=x\}\,dx \\ &= o(1)\end{aligned}$$

by the Lebesgue dominated convergence theorem.

From the discussion above, it helps to define the mean

$$\nu_n(x) = \mathbb{E}\{\xi_{12}|X_1 = x\}\ .$$

We have

$$\begin{aligned}\mathrm{Var}\{\delta_1|X_1,D_1\} &= \mathbb{1}_{D_1\ge 2}\mathbb{E}\left\{\left(\sum_{1\le i<j\le D_1}(\xi_{ij}-\nu_n(X_1))\right)^2|X_1,D_1\right\} \\ &= \mathbb{1}_{D_1\ge 2}\binom{D_1}{2}\mathbb{E}\left\{(\xi_{12}-\nu_n(X_1))^2|X_1\right\} \\ &\le \mathbb{1}_{D_1\ge 2}\binom{D_1}{2}\nu_n(X_1)\ .\end{aligned}$$

Finally, for arbitrary $\epsilon>0$,

$$\mathbb{P}\left\{|W_4-w_d|>2\epsilon \mid X_1, D_1\right\}$$

$$\begin{aligned} &\leq \mathbb{1}_{D_1\geq 2}\mathbb{P}\left\{|W_4-\nu_n(X_1)|>\epsilon \mid X_1, D_1\right\} \\ &\quad +\mathbb{1}_{D_1\geq 2}\mathbb{P}\left\{|\nu_n(X_1)-w_d|>\epsilon \mid X_1\right\}+\mathbb{P}\{D_1\leq 1\} \\ &\stackrel{\text{def.}}{=} I+II+III. \end{aligned}$$

We have

$$\mathbb{E}\{II\}\leq \int f(x)\mathbb{P}\left\{|\nu_n(x)-w_d|>\epsilon\right\}\,dx\to 0$$

as $n\to\infty$ by the Lebesgue dominated convergence theorem, since $\nu_n(x)\to w_d$ as $n\to\infty$ for almost all x. By Chebyshev's inequality,

$$\begin{aligned} \mathbb{P}\left\{|W_4-\nu_n(X_1)|>\epsilon \mid X_1, D_1\right\} &\leq \frac{1}{\epsilon^2}\mathbb{E}\left\{\left(W_4-\nu_n(X_1)\right)^2 \mid X_1, D_1\right\} \\ &\leq \frac{1}{\epsilon^2}\mathbb{1}_{D_1\geq 2}\frac{\nu_n(X_1)}{\binom{D_1}{2}}\,, \end{aligned}$$

and therefore

$$\mathbb{E}\{I\}\leq \frac{1}{\epsilon^2}\int f(x)\nu_n(x)dx\times\mathbb{E}\left\{\frac{\mathbb{1}_{D_1\geq 2}}{\binom{D_1}{2}}\right\}\leq \frac{1}{\epsilon^2}\mathbb{E}\left\{\frac{\mathbb{1}_{D_1\geq 2}}{\binom{D_1}{2}}\right\}$$

which is $o(1)$ if $D_1\to\infty$ in probability. Finally, $\mathbb{E}\{III\}\to 0$ under the same condition on D_1. We conclude by noting that $D_1\to\infty$ in probability if $nr_n^d\to\infty$, as D_1 is binomial $(n-1,\mu_n(X_1))$. So, for any fixed t,

$$\begin{aligned} \mathbb{P}\{D_1\leq t\} &\leq \mathbb{P}\{(n-1)\mu_n(X_1)\leq 2t\}+\mathbb{P}\{\text{binomial}((n-1),2t/(n-1))\leq t\} \\ &\stackrel{\text{def.}}{=} A+B. \end{aligned}$$

Clearly, $B\leq 2/t$ by Chebyshev's inequality. Noting that $\mu_n(x)/\pi_n\to f(x)$ at almost all x as $n\to\infty$, we have for arbitrary $\epsilon>0$,

$$\begin{aligned} A &= \mathbb{P}\{(n-1)\mu_n(X_1)\leq 2t\} \\ &\leq \mathbb{1}_{(n-1)\pi_n\leq 1/\epsilon}+\mathbb{P}\left\{\frac{\mu_n(X_1)}{\pi_n}\leq 2t\epsilon\right\} \\ &= \int f(x)\mathbb{1}_{\mu_n(x)/\pi_n\leq 2t\epsilon}+o(1) \\ &\leq \int f(x)\mathbb{1}_{f(x)\leq 3t\epsilon}+o(1) \end{aligned}$$

which can be made as small as desired by our choice of ϵ. Hence, for any fixed $t>0$, $\mathbb{P}\{D_1\leq t\}\leq 2/t+o(1)$, which implies that $D_1\to\infty$ in probability. □

Appendix C: Proof of Theorem 3.2

We use the notation $\mu_n(x)=\mathbb{P}\{X_1\in B(x,r_n)\}$ and $\nu_n(x)=\mathbb{P}\{[X_1,X_2\in B(x,r_n)]\cap[\|X_1-X_2\|\leq r_n]\}$. We have

$$\mathbb{E}\{\xi_{12}\}=\mathbb{E}\{\mu_n(X_1)\}=\int f(x)\mu_n(x)\,dx\,.$$

Let us introduce the maximal function

$$f^*(x) = \sup_{r>0} \frac{\int_{B(x,r)} f(y)\,dy}{V_d r^d} \ ,$$

and observe that $f \le f^*$ almost everywhere, and that $\int f^p < \infty$ for fixed $p > 1$ implies $\int (f^*)^p < \infty$ [see, e.g., Wheeden and Zygmund [43]]. Thus, as $\int f\mu_n \le V_d r_n^d \int (f^*)^2$ and $\mu_n(x)/(V_d r_n^d) \to f(x)$ at almost all x by the Lebesgue density theorem and $r_n \to 0$, the Lebesgue dominated convergence theorem implies that $\int f\mu_n \sim V_d r_n^d \int f^2$ as $n \to \infty$. In other words,

$$\mathbb{E}\{\xi_{12}\} = V_d r_n^d \left(\int f^2 + o(1) \right) \ .$$

Next,

$$\mathbb{E}\{\xi_{12}\xi_{13}\} \stackrel{\text{def.}}{=} M_n = \mathbb{E}\{\mu_n^2(X_1)\} = \int f(x)\mu_n^2(x)\,dx \ .$$

Extending the argument given above, we see that if $\int f^3 < \infty$, then

$$\mathbb{E}\{\xi_{12}\xi_{13}\} = (V_d r_n^d)^2 \left(\int f^3 + o(1) \right) \ .$$

Using the coupling argument of the proof of Theorem 3.1, we can verify that

$$\mathbb{E}\{\xi_{12}\xi_{13}\xi_{23}\} \stackrel{\text{def.}}{=} M_n' = \int f(x)\nu_n(x)\,dx = w_d (V_d r_n^d)^2 \left(\int f^3 + o(1) \right) \ .$$

We also need a general upper bound for

$$\mathbb{E}\left\{ \prod_{e \in E} \xi_e \right\}$$

where E is a fixed finite set of pairs of indices drawn from $\{1, 2, \ldots, n\}$. An example includes $\mathbb{E}\{\xi_{12}\xi_{13}\xi_{23}\xi_{24}\xi_{45}\}$. Let $v(E)$ denote the size of the set of vertices involved in the definition of E, and assume that the graph defined by E is connected. Since the graph is connected, all vertices are at most at graph distance $v(E) - 1$ from the node of smallest index. Thus,

$$\mathbb{E}\left\{ \prod_{e \in E} \xi_e \right\} \le \left(V_d (v(E)-1)^d r_n^d \right)^{v(E)-1} \int f(x) (\mu_n'(x))^{v(E)-1}\,dx$$

where

$$\mu_n'(x) = \frac{\int_{B(x,(v(E)-1)r_n)} f}{V_d (v(E)-1)^d r_n^d} \le f^*(x).$$

As $f \le f^*$, we have

$$\mathbb{E}\left\{ \cap_{e \in E} \xi_e \right\} \le O\left(r_n^{d(v(E)-1)} \right) \int (f^*)^{v(E)} \ .$$

Armed with this, we have

$$\begin{aligned}\mathbb{E}\left\{\sum_{1\le i<j<k\le n}\xi_{ki}\xi_{kj}\right\} &= \binom{n}{3}\int f(x)\mu_n^2(x)\,dx \\ &= \binom{n}{3}(V_d r_n^d)^2\left(\int f^3 + o(1)\right)\to\infty.\end{aligned}$$

Recalling that $M_n = \int f(x)\mu_n^2(x)\,dx$, we have

$$\mathrm{Var}\left\{\sum_{1\le i<j<k\le n}\xi_{ki}\xi_{kj}\right\} = \mathbb{E}\left\{\left(\sum_{1\le i<j<k\le n}(\xi_{ki}\xi_{kj}-M_n)\right)^2\right\} = A_0+A_1+A_2\,,$$

where

$$A_0 = \mathbb{E}\left\{\sum_{1\le i<j<k\le n}(\xi_{ki}\xi_{kj}-M_n)^2\right\} = \binom{n}{3}(M_n - M_n^2) \le n^3 V_d^2 r_n^{2d}\int (f^*)^3\,,$$

A_1

$$\begin{aligned}&= \mathbb{E}\left\{\sum_{1\le i<j<k\le n}\sum_{1\le i'<j'<k'\le n}\mathbb{1}_{|\{i,j,k,i',j',k'\}|=5}(\xi_{ki}\xi_{kj}-M_n)(\xi_{k'i'}\xi_{k'j'}-M_n)\right\}\\ &= \mathbb{E}\left\{\sum_{1\le i<j<k\le n}\sum_{1\le i'<j'<k'\le n}\mathbb{1}_{|\{i,j,k,i',j',k'\}|=5}(\xi_{ki}\xi_{kj}\xi_{k'i'}\xi_{k'j'}-M_n^2)\right\}\\ &\le \mathbb{E}\left\{\sum_{1\le i<j<k\le n}\sum_{1\le i'<j'<k'\le n}\mathbb{1}_{|\{i,j,k,i',j',k'\}|=5}\xi_{ki}\xi_{kj}\xi_{k'i'}\xi_{k'j'}\right\}\\ &\le O(n^5)\times O(r_n^{4d})\times\int (f^*)^5\,,\end{aligned}$$

and

A_2

$$\begin{aligned}&= \mathbb{E}\left\{\sum_{1\le i<j<k\le n}\sum_{1\le i'<j'<k'\le n}\mathbb{1}_{|\{i,j,k,i',j',k'\}|=4}(\xi_{ki}\xi_{kj}-M_n)(\xi_{k'i'}\xi_{k'j'}-M_n)\right\}\\ &\le \mathbb{E}\left\{\sum_{1\le i<j<k\le n}\sum_{1\le i'<j'<k'\le n}\mathbb{1}_{|\{i,j,k,i',j',k'\}|=4}\xi_{ki}\xi_{kj}\xi_{k'i'}\xi_{k'j'}\right\}\\ &\le O(n^4)\times O(r_n^{3d})\times\int (f^*)^4\,.\end{aligned}$$

By Chebyshev's inequality, we see that

$$\frac{\left\{\sum_{1\le i<j<k\le n}\xi_{ki}\xi_{kj}\right\}}{\binom{n}{3}(V_d r_n^d)^2\int f^3}\to 1$$

in probability if $A_0+A_1+A_2=o(n^6 r_n^{4d})$, which is easily verified.

Finally, we will show that

$$\frac{\left\{\sum_{1\le i<j<k\le n}\xi_{ki}\xi_{kj}\xi_{ij}\right\}}{\binom{n}{3}(V_d r_n^d)^2\int f^3}\to w_d$$

in probability, so that $W_2\to w_d$ in probability, as required. To see this, we note that the above ratio has expected value tending to one, while its variance tends to zero. The variance bound mimics the bound obtained for the variance of $\sum_{1\le i<j<k\le n}\xi_{ki}\xi_{kj}$. The troublesome terms involve upper bounds for $\mathbb{E}\{\xi_{ki}\xi_{kj}\xi_{ij}\xi_{k'i'}\xi_{k'j'}\xi_{i'j'}\}$ when $|\{i,j,k,i',j',k'\}|\in\{4,5\}$. But by bounding ξ_{ij} and $\xi_{i'j'}$ by one, we have an expression similar to that dealt with above, and thus, the variance tends to zero. □

Appendix D: Proof of Lemma 4.1

We first show the following identity

$$w_d=\mathbb{P}\left\{\beta\left(\frac{d+1}{2},\frac{d+1}{2}\right)\le\frac{1}{4}\right\}+\mathbb{P}\left\{\beta\left(\frac{1}{2},\frac{d+1}{2}\right)\ge\frac{1}{4}\right\}\ . \tag{1}$$

We recall the formula for the volume of $B(0,1)$ in $\mathbb{R}^d$:

$$V_d\stackrel{\text{def.}}{=}\frac{\pi^{d/2}}{\Gamma\left(\frac{d+2}{2}\right)}\ .$$

Let X and Y be defined as above. It is well-known that $R\stackrel{\text{def.}}{=}\|X\|$ is distributed as $U^{1/d}$, where U is uniform on $[0,1]$: it has density dx^{d-1} on $[0,1]$. Without loss of generality, we can assume that $X=(R,0,0,\ldots,0)$. Then $\|X-Y\|\le 1$ if $Y\in A\stackrel{\text{def.}}{=}B(0,1)\cap B(X,1)$. A is a loon-shaped region formed by two spherical caps of the same size. Call one of the two spherical caps S. Let $\lambda(\cdot)$ denote the volume of a set, and recall that $V_d=\lambda(B(0,1))$. We have

$$\mathbb{P}\{Y\in A\}=\frac{2\mathbb{E}\{\lambda(S)\}}{V_d}\ ,$$

where the volume of the spherical cap is a function of R. Standard spatial integration yields

$$\lambda(S)=\int_{R/2}^1 V_{d-1}(1-y^2)^{\frac{d-1}{2}}\,dy\ .$$

Thus,

$$\begin{aligned}
\mathbb{E}\{\lambda(S)\} &= \int_0^1 dr^{d-1} \int_{r/2}^1 V_{d-1}(1-y^2)^{\frac{d-1}{2}}\, dy\, dr \\
&= V_{d-1}\int_0^{1/2}(1-y^2)^{\frac{d-1}{2}} \int_0^{2y} dr^{d-1}\, dr\, dy + V_{d-1}\int_{1/2}^1 (1-y^2)^{\frac{d-1}{2}}\, dy \\
&= V_{d-1}\int_0^{1/2}(1-y^2)^{\frac{d-1}{2}}(2y)^d\, dy + V_{d-1}\int_{1/2}^1 (1-y^2)^{\frac{d-1}{2}}\, dy \\
&= 2^{d-1}V_{d-1}\int_0^{1/4}(y(1-y)^{\frac{d-1}{2}}\, dy + \frac{V_{d-1}}{2}\int_{1/4}^1 (1-y)^{\frac{d-1}{2}} y^{-1/2}\, dy \\
&= I + II\ .
\end{aligned}$$

Now,

$$I = \alpha \mathbb{P}\left\{\beta\left(\frac{d+1}{2}, \frac{d+1}{2}\right) \le \frac{1}{4}\right\}\ ,$$

where

$$\alpha = 2^{d-1}V_{d-1}\frac{\Gamma^2\left(\frac{d+1}{2}\right)}{\Gamma(d+1)}\ .$$

Furthermore,

$$II = \alpha' \mathbb{P}\left\{\beta\left(\frac{1}{2}, \frac{d+1}{2}\right) \ge \frac{1}{4}\right\}\ ,$$

where

$$\alpha' = \frac{V_{d-1}}{2}\frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d+2}{2}\right)}\ .$$

Combining all of the above, we obtain

$$\mathbb{P}\{Y \in A\} = \frac{2\alpha}{V_d}\mathbb{P}\left\{\beta\left(\frac{d+1}{2}, \frac{d+1}{2}\right) \le \frac{1}{4}\right\} + \frac{2\alpha'}{V_d}\mathbb{P}\left\{\beta\left(\frac{1}{2}, \frac{d+1}{2}\right) \ge \frac{1}{4}\right\}\ .$$

We verify that $\alpha = \alpha' = V_d/2$, to conclude (1). The formal verification is as follows:

$$\begin{aligned}
\frac{2\alpha}{V_d} &= \frac{2^d V_{d-1}}{V_d}\frac{\Gamma^2\left(\frac{d+1}{2}\right)}{\Gamma(d+1)} = \frac{2^d\Gamma\left(\frac{d+2}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d+1}{2}\right)}\frac{\Gamma^2\left(\frac{d+1}{2}\right)}{\Gamma(d+1)} \\
&= \frac{2^d\Gamma\left(\frac{d+2}{2}\right)\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma(d+1)} = 1
\end{aligned}$$

by the duplication formula for the gamma function (see, e.g., Whittaker and Watson, [44, p.240]). This is also immediate by induction on d. Next,

$$\frac{2\alpha'}{V_d} = \frac{V_{d-1}}{V_d}\frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d+2}{2}\right)} = \frac{\Gamma\left(\frac{d+2}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d+1}{2}\right)}\frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d+2}{2}\right)} = 1\ .$$

This proves (1). Next, we show that

$$\mathbb{P}\left\{\beta\left(\frac{d+1}{2},\frac{d+1}{2}\right)\le\frac{1}{4}\right\}=\frac{1}{2}\mathbb{P}\left\{\beta\left(\frac{1}{2},\frac{d+1}{2}\right)\ge\frac{1}{4}\right\}\ . \tag{2}$$

That would complete the beta representation in Lemma 4.1. Let $B=\beta\left(\frac{d+1}{2},\frac{d+1}{2}\right)$. Observe that

$$\mathbb{P}\{B\le 1/4\}=\frac{1}{2}\left(\mathbb{P}\{B\le 1/4\}+\mathbb{P}\{B\ge 3/4\}\right)=\frac{1}{2}\mathbb{P}\{|2B-1|\ge 1/2\}\ .$$

Now, $|2B-1|$ has a density proportional to $(1-x^2)^{\frac{d-1}{2}}$ on $[0,1]$, and $(2B-1)^2$ is beta $(1/2,(d+1)/2)$. Thus,

$$\mathbb{P}\{B\le 1/4\}=\frac{1}{2}\mathbb{P}\left\{\beta\left(\frac{1}{2},\frac{d+1}{2}\right)\ge\frac{1}{4}\right\}.$$

The monotonicity claim follows easily. Finally, $w_d\to 0$ since $\beta(1/2,d)\to 0$ in probability as $d\to\infty$. □

Appendix E: Proof of $w_d - w_{d+1} \ge d^{-(d+o(d))/2}$

Observe that

$$\beta\left(\frac{1}{2},\frac{d-1}{2}\right)\stackrel{\mathcal{L}}{=}\frac{G(1)}{\sum_{i=1}^{d}G(i)},$$

where $G(1),G(2),\dots$ are i.i.d. gamma (1/2) random variables. Thus, with this coupling,

$$\begin{aligned}\mathbb{P}\left\{\beta\left(\frac{1}{2},\frac{d}{2}\right)\ge\frac{1}{4}\right\}&=\mathbb{P}\left\{\beta\left(\frac{1}{2},\frac{d-1}{2}\right)\ge\frac{1}{4}\right\}\\&\quad-\mathbb{P}\left\{\beta\left(\frac{1}{2},\frac{d-1}{2}\right)\ge\frac{1}{4}>\beta\left(\frac{1}{2},\frac{d}{2}\right)\right\}\ .\end{aligned}$$

The last summand reduces to

$$\begin{aligned}&\mathbb{P}\left\{\sum_{i=2}^{d}G(i)\le 3G(1)<\sum_{i=2}^{d+1}G(i)\right\}\\&\quad=\quad\mathbb{P}\left\{3G(1)-G(d+1)<\sum_{i=2}^{d}G(i)\le 3G(1)\right\}\\&\quad\ge\quad\mathbb{P}\{G(d+1)\ge 6,G(1)\in[1,2]\}\mathbb{P}\left\{\sum_{i=2}^{d}G(i)\le 3\right\}\\&\quad\stackrel{\text{def.}}{=}\quad\rho\mathbb{P}\left\{\sum_{i=2}^{d}G(i)\le 3\right\}\ .\end{aligned}$$

As $\sum_{i=2}^{d}G(i)$ is gamma $((d-1)/2)$, we see that

$$w_{d-2}-w_{d-1}\ge\rho\int_0^3\frac{x^{\frac{d-3}{2}}e^{-x}}{\Gamma\left(\frac{d-1}{2}\right)}\,dx\ge\frac{\rho}{e^3}\frac{3^{\frac{d-1}{2}}}{\Gamma\left(\frac{d+1}{2}\right)}=d^{-\frac{d}{2}+o(d)}\ .\square$$

Acknowledgments

The authors thank Jakob Reznikov for his assistance. The authors also thank the reviewers for numerous relevant remarks and pointers to related literature.

Funding

Luc Devroye acknowledges support of NSERC grant A3450. Gábor Lugosi acknowledges the support of Ayudas Fundación BBVA a Proyectos de Investigación Científica 2021 and of the Spanish grant PID2022-138268NB-I00, financed by MCIN/AEI/10.13039/501100011033, FSE+MTM2015-67304-P, and FEDER, EU.

References

[1] Aharonyan, N. G. and Khalatyan, V. (2020). Distribution of the distance between two random points in a body from. *Journal of Contemporary Mathematical Analysis (Armenian Academy of Sciences)* **55** 329–334. MR4174821

[2] Appel, M. J. B. and Russo, R. P. (2002). The connectivity of a graph on uniform points on [0, 1] d. *Statistics & Probability Letters* **60** 351–357. MR1947174

[3] Araya Valdivia, E. and De Castro, Y. (2019). Latent distance estimation for random geometric graphs. In *Advances in Neural Information Processing Systems* **32** 8721–8731.

[4] Balister, P., Bollobás, B. and Sarkar, A. (2008). Percolation, connectivity, coverage and colouring of random geometric graphs. *Handbook of Large-Scale Random Networks* 117–142. MR2582387

[5] Balister, P., Bollobás, B., Sarkar, A. and Walters, M. (2009). Highly connected random geometric graphs. *Discrete Applied Mathematics* **157** 309–320. MR2479805

[6] Borg, I. and Groenen, P. J. (2007). *Modern Multidimensional Scaling: Theory and Applications*. Springer Science & Business Media. MR2158691

[7] Borg, I., Groenen, P. J. and Mair, P. (2012). *Applied multidimensional scaling*. Springer Science & Business Media. MR2987167

[8] Bruske, J. and Sommer, G. (1998). Intrinsic dimensionality estimation with optimally topology preserving maps. *IEEE Transactions on pattern analysis and machine intelligence* **20** 572–575.

[9] Bubeck, S., Ding, J., Eldan, R. and Rácz, M. Z. (2016). Testing for high-dimensional geometry in random graphs. *Random Structures & Algorithms* **49** 503–532. MR3545825

[10] Camastra, F. and Vinciarelli, A. (2002). Estimating the intrinsic dimension of data with a fractal-based method. *IEEE Transactions on pattern analysis and machine intelligence* **24** 1404–1407.

[11] Casse, J. (2023). Siblings in d-dimensional nearest neighbour trees. *arXiv preprint arXiv:2302.10795*.
[12] Cooper, C. and Frieze, A. (2011). The cover time of random geometric graphs. *Random Structures & Algorithms* **38** 324–349. MR2663732
[13] Costa, J. A. and Hero, A. O. (2004). Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing* **52** 2210–2221. MR2085582
[14] Duchemin, Q. and De Castro, Y. (2022). Markov random geometric graph, MRGG: A growth model for temporal dynamic networks. *Electronic Journal of Statistics* **16** 671–699. MR4364740
[15] Facco, E., d'Errico, M., Rodriguez, A. and Laio, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports* **7** 12140.
[16] Fukunaga, K. and Olsen, D. R. (1971). An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on computers* **100** 176–183.
[17] Gilbert, E. N. (1965). The probability of covering a sphere with N circular caps. *Biometrika* **52** 323–330. MR0207005
[18] Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics* **30** 1141–1144. MR0108839
[19] Gilbert, E. N. (1961). Random plane networks. *Journal of SIAM* **9** 533–543. MR0132566
[20] Granata, D. and Carnevale, V. (2016). Accurate estimation of the intrinsic dimension using graph distances: Unraveling the geometric complexity of datasets. *Scientific Reports* **6** 31377.
[21] Grassberger, P. and Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Physica D: nonlinear phenomena* **9** 189–208. MR0732572
[22] Hall, P. (1988). *Introduction to the Theory of Coverage Processes*. Wiley, New York. MR0973404
[23] Hout, M. C., Papesh, M. H. and Goldinger, S. D. (2013). Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science* **4** 93–103.
[24] Janson, S. (1986). Random coverings in several dimensions. *Acta Mathematica* **156** 83–118. MR0822331
[25] Kang, R. J. and Müller, T. (2011). Sphere and dot product representations of graphs. In *Proceedings of the Twenty-Seventh Annual Symposium on Computational Geometry* 308–314.
[26] Kégl, B. (2002). Intrinsic dimension estimation using packing numbers. In *Advances in Neural Information Processing Systems* **15**.
[27] Kleinberg, R. (2007). Geographic routing using hyperbolic space. In *IEEE INFOCOM 2007-26th IEEE International Conference on Computer Communications* 1902–1909.
[28] Levina, E. and Bickel, P. (2004). Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems* **17** 777–784.
[29] Lichev, L. and Mitsche, D. (2021). New results for the random near-

est neighbor tree. *Probability Theory and Related Fields* **189** 229–279. MR4750559

[30] McDiarmid, C. and Müller, T. (2011). On the chromatic number of random geometric graphs. *Combinatorica* **31** 423–488. MR2861238

[31] Penrose, M. (2003). *Random Geometric Graphs. Oxford Studies in Probability* **5**. Oxford University Press, Oxford. MR1986198

[32] Penrose, M. D. (1999). A strong law for the longest edge of the minimal spanning tree. *The Annals of Probability* **27** 246–260. MR1681114

[33] Peterfreund, E. and Gavish, M. (2021). Multidimensional scaling of noisy high dimensional data. *Applied and Computational Harmonic Analysis* **51** 333–373. MR4191903

[34] Reiterman, J., Rödl, V. and Šiňajová, E. (1989). Geometrical embeddings of graphs. *Discrete Mathematics* **74** 291–319. MR0992741

[35] Shavitt, Y. and Tankel, T. (2004). Big-bang simulation for embedding network distances in Euclidean space. *IEEE/ACM Transactions on Networking* **12** 993–1006.

[36] Tenenbaum, J. B., Silva, V. d. and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319–2323.

[37] Teugels, J. L., Bingham, N. H. and Goldie, C. M. (1987). *Regular Variations*. Cambridge University Press. MR0898871

[38] Verbeek, K. and Suri, S. (2014). Metric embedding, hyperbolic space, and social networks. In *Proceedings of the Thirtieth Annual Symposium on Computational Geometry* 501–510. MR3382332

[39] Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science* **47**. Cambridge University Press. MR3837109

[40] Verveer, P. J. and Duin, R. P. W. (1995). An evaluation of intrinsic dimensionality estimators. *IEEE Transactions on pattern analysis and machine intelligence* **17** 81–86.

[41] von Luxburg, U. and Alamgir, M. (2013). Density estimation from unweighted k-nearest neighbor graphs: a roadmap. In *Advances in Neural Information Processing Systems* **26** 225–233.

[42] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *Nature* **393** 440–442.

[43] Wheeden, A. and Zygmund, R. L. (1977). *Measure and Integral*. Marcel Dekker, New York. MR0492146

[44] Whittaker, E. T. and Watson, G. N. (1927). *A Course of Modern Analysis*. Cambridge University Press. MR1424469