# A note on estimating the dimension from a random geometric graph *

Caelan Atamanchuk
School of Computer Science
McGill University
Montreal, Canada

Luc Devroye
School of Computer Science
McGill University
Montreal, Canada

Gábor Lugosi
Department of Economics and Business,
Pompeu Fabra University, Barcelona, Spain
ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain
Barcelona Graduate School of Economics

November 21, 2023

**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 $\|X_i - X_j\| \le r_n$, 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 $r_n = o(1)$.

*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.

**Keywords:** Multivariate densities, nonparametric estimation, random geometric graphs, estimating the dimension, absolute continuity.

## 1 Introduction

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, significant effort has been devoted to embedding vertices of a graph in Euclidean or hyperbolic spaces, see Reiterman, Rödl, Šiňajová [22], Tenenbaum, Silva, and Langford [24], Shavitt and Tankel [23], Kleinberg [17], Kang and Müller [16], Verbeek and Suri [26] for a sample of the literature. 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.

In order to set up a rigorous statistical problem, we model the graph as a random geometric graph. Let $G_n$ be a random geometric graph with vertex set $[n]$ based on $n$ i.i.d. random vectors $X_1,\dots,X_n$ drawn from an unknown density $f$ on $\mathbb{R}^d$. An edge $(i,j)$ is present when $\|X_i - X_j\| \le r_n$, for a given threshold $r_n$ possibly depending upon $n$, where $\|\cdot\|$ denotes Euclidean distance. Introduced by Gilbert [10, 11], 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 [20]. 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 [8]), coverage (Gilbert [12]; Hall [14], Janson [15]), chromatic number (McDiarmid and Müller [19]), and minimal spanning tree (Penrose [21]).

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 $\Delta_n$ 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 $\Delta_n$ 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,

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

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 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 $\Delta_n$ 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. 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 2.** *Let the density $f$ in $\mathbb{R}^d$ be arbitrary, and assume that*

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

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

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 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 2 we establish a geometric lemma that is a key tool in our approach of defining estimators of the dimension. In Section 3 we introduce four simple estimators of the dimension whose analysis proves Theorems 1 and 2. We start analyzing the estimators in Section 4 by focusing on the special–but important–case of the uniform density on the torus. Finally, in Section 5 we prove Theorems 1 and 2 for general densities.

## Related literature

Granata and Carnevale [13] 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, [13] aim to construct accurate estimates from graph distances.

Bubeck, Ding, Eldan, and Rácz [6] 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 [18] and Casse [7] 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. Indeed, often the first step of computing such estimates is to construct a geometric graph from the data, see, e.g., Tenenbaum, Silva, and Langford [24], Facco, d'Errico, Rodriguez, Laio [9].

Araya and De Castro [3] study estimating the Euclidean distances between the point locations upon observing the combinatorial graph for dense random geometric graphs.

## 2 A geometric lemma

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

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

As all estimators studied in this paper are based on estimating $w_d$, the following property, proved in the Appendix, is a key ingredient of our arguments.

**Lemma 3.** *We have*

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

*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 computation 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 we showed in Lemma 3, 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 [27].

### On the sample size needed.

The representation of $w_d$ given in Lemma 3 permits us to show that $w_d - w_{d+1} \geq d^{-(d+o(d))/2}$ (see the Appendix). 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 [6] 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 $d$.

## 3 The proposed estimates

Here we introduce four simple estimators of the quantity $w_d$ defined above. If $W$ is a data-based estimate, then we set

$$\Delta_n = \underset{d}{\operatorname{argmin}} |W - w_d|.$$

In view of Lemma 3, 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 $\Delta_n$ 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 $\delta_i$ be the number of edges between nodes in $N_i$, the set of neighbors of vertex $i$. Let $M$ be the smallest index among the vertices of maximal degree. Let $\xi_{ij}$ 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

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

## 4 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 estimator $\delta_1$, assuming that $D_1 \geq 2$. Note that we can represent $\delta_1$ as

$$\delta_1 \stackrel{\mathcal{L}}{=} \sum_{i<j\leq D_1} Y_{ij} ,$$

where $Y_{ij} = \mathbb{1}_{\|X_i - X_j\| \leq 1}$, and $X_1, \ldots, X_{D_1}$ are i.i.d. random vectors uniformly distributed in the unit ball of $\mathbb{R}^d$. 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 $\delta_1$ is unbiased. Then,

$$\begin{aligned}
\mathrm{Var}\{\delta_1 | D_1\} &= \mathbb{E}\left\{ \left( \sum_{i<j\leq D_1} (Y_{ij} - w_d) \right)^2 \mid D_1 \right\} \\
&= \binom{D_1}{2} \mathbb{E}\left\{ (Y_{12} - w_d)^2 | D_1 \right\} + 3\binom{D_1}{3} \mathbb{E}\left\{ (Y_{12} - w_d)(Y_{13} - w_d) | D_1 \right\} .
\end{aligned}$$

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 $D_1 \geq 2$ and $t > 0$,

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

$$
\begin{aligned}
&= \frac{w_d(1-w_d)}{w_d(1-w_d)+\binom{D_1}{2}t^2} \\
&\leq \frac{w_d}{w_d+\binom{D_1}{2}t^2}\,.
\end{aligned}
$$

We conclude that $W_4 = \delta_1/\binom{D_1}{2} \to w_d$ in probability when $D_1 \to \infty$ in probability.

The following theorem summarizes the consistency properties of the estimators $W_1$ and $W_2$.

**Theorem 4.** *Let the density $f$ be the uniform density on the unit torus $[0,1]^d$, and assume that $r_n \leq 1/2$ for all $n$.*

*(i) If*

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

*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 $nr_n^d \geq L(n)$ 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.

**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 $nr_n^d \geq L(n)$ with $L(n)$ slowly varying. See Bingham, Goldie and Teugels [25] 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 $nr_n^d \to \infty$. When $nr_n^d \to c > 0$ 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 $nr_n^d \to 0$.

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

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

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 $\infty$ 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

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

and

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

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

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

Let $s$ be the set $\{i,j,k\}$ and let $s'$ be the set $\{i',j',k'\}$. Observe that if $|s \cap s'| \leq 1$, 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 $|s \cap s'| \geq 2$. If the intersection is of size three, then the two ordered triples are identical. This yields a term equal to

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

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:

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

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

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

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 $|s \cap s'| \geq 2$. 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 $n^3\pi_n^2 \to \infty$. $\square$

## 5 General densities

We prove the following theorem, which implies Theorem 2.

**Theorem 5.** *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 $n \to \infty$.*

**Proof.** We condition on $X_1$ and $D_1$, and let $Y_1, \dots, Y_{D_1}$ 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 $Z_1, Z_2, \dots$ 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,

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

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 [28]), 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 $X_1 = x$,

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

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

$$|\mathbb{P}\{\|Y_1 - Y_2\| \le r_n\} - w_d| \quad = \quad |\mathbb{P}\{\|Y_1 - Y_2\| \le r_n\} - \mathbb{P}\{\|Z_1 - Z_2\| \le r_n\}|$$

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

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

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

Finally, for arbitrary $\epsilon > 0$,

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

We have

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

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,

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

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 $t$,

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

Clearly, $B \leq 2/t$ by Chebyshev's inequality. Noting that $\mu_n(x) = \pi_n(\mu_n(x)/\pi_n)$ and $\mu_n(x)/\pi_n \to f(x)$ 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 $\epsilon$. 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. $\square$

Finally, we prove the following that implies Theorem 1.

**Theorem 6.** *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 $n \to \infty$.*

**Proof.** 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 $f \leq f^*$ almost everywhere, and that $\int f^p < \infty$ for fixed $p > 1$ implies $\int (f^*)^p < \infty$ [see, e.g., Wheeden and Zygmund [28]]. Thus, as $\int f\mu_n \leq V_d r_n^d \int (f^*)^2$ 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 5, 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

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

Armed with this, we have

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

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

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

$$\begin{aligned}
A_1 &= \mathbb{E}\left\{ \sum_{1\le i<j<k\le n} \sum_{1\le i'<j'<k'\le n} \mathbb{1}_{|\{i,j,k,i',j',k'\}|=5} (\xi_{ki}\xi_{kj} - M_n)(\xi_{k'i'}\xi_{k'j'} - M_n) \right\} \\
&= \mathbb{E}\left\{ \sum_{1\le i<j<k\le n} \sum_{1\le i'<j'<k'\le n} \mathbb{1}_{|\{i,j,k,i',j',k'\}|=5} (\xi_{ki}\xi_{kj}\xi_{k'i'}\xi_{k'j'} - M_n^2) \right\} \\
&\le \mathbb{E}\left\{ \sum_{1\le i<j<k\le n} \sum_{1\le i'<j'<k'\le n} \mathbb{1}_{|\{i,j,k,i',j',k'\}|=5} \xi_{ki}\xi_{kj}\xi_{k'i'}\xi_{k'j'} \right\} \\
&\le O(n^5) \times O(r_n^{4d}) \times \int (f^*)^5 \ ,
\end{aligned}$$

and

$$\begin{aligned}
A_2 &= \mathbb{E}\left\{ \sum_{1\le i<j<k\le n} \sum_{1\le i'<j'<k'\le n} \mathbb{1}_{|\{i,j,k,i',j',k'\}|=4} (\xi_{ki}\xi_{kj} - M_n)(\xi_{k'i'}\xi_{k'j'} - M_n) \right\} \\
&\le \mathbb{E}\left\{ \sum_{1\le i<j<k\le n} \sum_{1\le i'<j'<k'\le n} \mathbb{1}_{|\{i,j,k,i',j',k'\}|=4} \xi_{ki}\xi_{kj}\xi_{k'i'}\xi_{k'j'} \right\} \\
&\le O(n^4) \times O(r_n^{3d}) \times \int (f^*)^4 \ .
\end{aligned}$$

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 $\xi_{ij}$ and $\xi_{i'j'}$ by one, we have an expression similar to that dealt with above, and thus, the variance tends to zero. □

We suspect that $\int f^3 < \infty$ suffices in Theorem 6, 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.

## Appendix: proof of Lemma 3

We first show the following identity

$$w_d = \mathbb{P}\left\{\beta\left(\frac{d+1}{2}, \frac{d+1}{2}\right) \le \frac{1}{4}\right\} + \mathbb{P}\left\{\beta\left(\frac{1}{2}, \frac{d+1}{2}\right) \ge \frac{1}{4}\right\} . \tag{5.1}$$

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,

$$\mathbb{E}\{\lambda(S)\} \quad = \quad \int_0^1 dr^{d-1} \int_{r/2}^{1} V_{d-1}(1-y^2)^{\frac{d-1}{2}}\, dy\, dr$$

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

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 (5.1). 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, [29, 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 (5.1). Next, we show that

$$\mathbb{P}\left\{\beta\left(\frac{d+1}{2},\frac{d+1}{2}\right)\le\frac{1}{4}\right\}=\frac{1}{2}\mathbb{P}\left\{\beta\left(\frac{1}{2},\frac{d+1}{2}\right)\ge\frac{1}{4}\right\}. \tag{5.2}$$

That would complete the beta representation in Lemma 3. 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 $d\to\infty$. □

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

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

The last summand reduces to

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

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

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

## Acknowledgements

The authors thank Jakob Reznikov for his assistance.

## References

[1] N.G. Aharonyan and V. Khalatyan. Distribution of the distance between two random points in a body from. *Journal of Contemporary Mathematical Analysis (Armenian Academy of Sciences)*, 55(6):329–334, 2020.

[2] Martin J.B. Appel and Ralph P. Russo. The connectivity of a graph on uniform points on [0, 1] d. *Statistics & Probability Letters*, 60(4):351–357, 2002.

[3] Ernesto Araya Valdivia and Yohann De Castro. Latent distance estimation for random geometric graphs. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] Paul Balister, Béla Bollobás, and Amites Sarkar. Percolation, connectivity, coverage and colouring of random geometric graphs. *Handbook of Large-Scale Random Networks*, pages 117–142, 2008.

[5] Paul Balister, Béla Bollobás, Amites Sarkar, and Mark Walters. Highly connected random geometric graphs. *Discrete Applied Mathematics*, 157(2):309–320, 2009.

[6] Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures & Algorithms*, 49(3): 503–532, 2016.

[7] Jérôme Casse. Siblings in d-dimensional nearest neighbour trees. *arXiv preprint arXiv:2302.10795*, 2023.

[8] Colin Cooper and Alan Frieze. The cover time of random geometric graphs. *Random Structures & Algorithms*, 38(3):324–349, 2011.

[9] Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140, 2017.

[10] Edgar N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[11] Edgar N. Gilbert. Random plane networks. *Journal of SIAM*, 9(4):533–543, 1961.

[12] E.N. Gilbert. The probability of covering a sphere with n circular caps. *Biometrika*, 52(3/4):323–330, 1965.

[13] Daniele Granata and Vincenzo Carnevale. Accurate estimation of the intrinsic dimension using graph distances: Unraveling the geometric complexity of datasets. *Scientific Reports*, 6(1):31377, 2016.

[14] Peter Hall. *Introduction to the Theory of Coverage Processes*. Wiley, New York, 1988.

[15] Svante Janson. Random coverings in several dimensions. *Acta Mathematica*, 156:83–118, 1986.

[16] Ross J. Kang and Tobias Müller. Sphere and dot product representations of graphs. In *Proceedings of the Twenty-Seventh Annual Symposium on Computational Geometry*, pages 308–314, 2011.

[17] Robert Kleinberg. Geographic routing using hyperbolic space. In *IEEE INFOCOM 2007-26th IEEE International Conference on Computer Communications*, pages 1902–1909, 2007.

[18] Lyuben Lichev and Dieter Mitsche. New results for the random nearest neighbor tree. *arXiv preprint arXiv:2108.13014*, 2021.

[19] Colin McDiarmid and Tobias Müller. On the chromatic number of random geometric graphs. *Combinatorica*, 31(4):423–488, 2011.

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

[21] Mathew D Penrose. A strong law for the longest edge of the minimal spanning tree. *The Annals of Probability*, 27(1):246–260, 1999.

[22] J. Reiterman, V. Rödl, and E. Šiňajová. Geometrical embeddings of graphs. *Discrete Mathematics*, 74(3):291–319, 1989.

[23] Yuval Shavitt and Tomer Tankel. Big-bang simulation for embedding network distances in euclidean space. *IEEE/ACM Transactions on Networking*, 12(6):993–1006, 2004.

[24] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

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

[26] Kevin Verbeek and Subhash Suri. Metric embedding, hyperbolic space, and social networks. In *Proceedings of the Thirtieth Annual Symposium on Computational Geometry*, pages 501–510, 2014.

[27] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

[28] A. Wheeden and R.L. Zygmund. *Measure and Integral*. Marcel Dekker, New York, 1977.

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