Supplemental material to NIPS 2008 paper

Influence of graph construction on graph-based clustering measures

Markus Maier* Ulrike von Luxburg * Matthias Hein†

November 18, 2008

1 Definitions and assumptions (as in paper)

For the reader's convenience we repeat the corresponding section from the paper. Note, however, that in the rest of the supplemental material we omit the surface S in the expressions for the cut, the volume and NCut. This is possible because we always study these quantities for a fixed surface.
Given a graph $G = (V, E)$ with weights $w : E \to \mathbb{R}$ and a partition of the nodes V into $(C, V \setminus C)$ we define

$$\mathrm{Ncut}(C, V \setminus C) = \mathrm{cut}(C, V \setminus C)\left(\frac{1}{\mathrm{vol}(C)} + \frac{1}{\mathrm{vol}(V \setminus C)}\right),$$

where

$$\mathrm{cut}(C, V \setminus C) = \sum_{u \in C, v \in V \setminus C} w(u, v) \qquad \text{and} \qquad \mathrm{vol}(C) = \sum_{u \in C, v \in V} w(u, v).$$

In this paper we focus on Ncut, but the methods used here can easily be carried over to the study of other measures.
Given a finite sample $x_1, \ldots, x_n$ we can construct different types of neighborhood graphs:

- the r-neighborhood graph $G_{n,r}$: there is an edge from a point x_i to a point x_j if $\mathrm{dist}(x_i, x_j) \leq r$ for all $1 \leq i, j \leq n, i \neq j$. This graph is always undirected, as distances are symmetric.
- the directed k-nearest neighbor graph $G_{n,k}$: there is a directed edge from x_i to x_j if x_j is one of the k nearest neighbors of x_i for $1 \leq i, j \leq n, i \neq j$.

*Max Planck Insitute for Biological Cybernetics, Tübingen, Germany
†Saarland University, Saarbrücken, Germany

- the symmetric k-nearest neighbor graph: there is an undirected edge from x_i to x_j if the directed k-nearest neighbor graph contains a directed edge from x_i to x_j or vice versa.

In the following we work on the space $\mathbb{R}^d$ with Euclidean metric dist. We denote by η_d the volume of the d-dimensional unit ball in $\mathbb{R}^d$ and by $B(x,r)$ the ball with radius r centered at x. On the space $\mathbb{R}^d$ we will study partitions which are induced by some hypersurface S. Given a surface S which separates the data points in two non-empty parts C^+ and C^-, we denote by $\mathrm{cut}_{n,r}(S)$ the number of edges in $G_{n,r}$ that go from a sample point on one side of the surface to a sample point on the other side of the surface. The corresponding quantity for the directed k-nearest neighbor graph is denoted by $\mathrm{cut}_{n,k}(S)$. For a set $A \subseteq \mathbb{R}^d$ the volume of points in $\{x_1, \ldots, x_n\} \cap A$ in the graph $G_{n,r}$ is denoted by $\mathrm{vol}_{n,r}(A)$. Similarly, $\mathrm{vol}_{n,k}(A)$ denotes the corresponding volume in the graph $G_{n,k}$. In the rest of the paper we make the following

General assumptions: *The data points $x_1, ..., x_n$ are drawn independently from some density p on $\mathbb{R}^d$. This density is bounded from below and above, that is $0 < p_{\min} \leq p(x) \leq p_{\max}$. In particular, it has compact support C. We assume that the boundary ∂C of C is well-behaved, that means it is a set of Lebesgue measure 0 and we can find a constant $\gamma > 0$ such that for r sufficiently small, $\mathrm{vol}(B(x,r) \cap C) \geq \gamma \,\mathrm{vol}(B(x,r))$ for all $x \in C$. Furthermore we assume that p is twice differentiable in the interior of C and that the derivatives are bounded. The measure on $\mathbb{R}^d$ induced by p will be denoted by μ, that means, for a measurable set A we set $\mu(A) = \int_A p(x)\mathrm{d}x$. For the cut surface S, we assume that the volume of $S \cap \partial C$ with respect to the $(d-1)$-dimensional measure on S is a set of measure 0.*

In general, we study the setting where one wants to find two clusters which are induced by some hypersurface in $\mathbb{R}^d$. As cut surfaces we only consider hyperplanes in this work, that is a clustering on the data points is induced by some hyperplane in $\mathbb{R}^d$. Our results can be generalized to more general (smooth) surfaces, provided one makes a few assumptions on the regularity of the surface S. The proofs will be more technical, though.

2 Limit of $\mathbb{E}\,\mathrm{cut}_{n,r_n}$

Proposition 1 (Limit of $\mathbb{E}\,\mathrm{cut}_{n,r_n}$) *Under the assumptions:*

- *$p'_{\max}$ is the maximum of the absolute value of the directional derivative (over all directions) of p,*
- *r_n is sufficiently small, such that $p'_{\max} r_n \leq p_{\max}$, and*
- *$\mathcal{R}_n = \{x \in \mathbb{R}^d | \,\mathrm{dist}(x, \partial C) \leq 2r_n\}$.*

Then

$$\Big|\mathbb{E}\Big(\frac{\mathrm{cut}_{n,r_n}}{n(n-1)r_n^{d+1}}\Big) - \frac{2\eta_{d-1}}{d+1}\int_S p^2(s)\mathrm{d}s\Big| \leq \frac{16\eta_{d-1}p_{\max}p'_{\max}}{d+1}\,\mathrm{vol}_{d-1}(S\cap C)r_n + \frac{2\eta_{d-1}}{d+1}p_{\max}^2\,\mathrm{vol}_{d-1}(S\cap\mathcal{R}_n).$$

Proof. Let N_i denote the number of edges in the cut originating in point x_i. Then

$$\mathbb{E}(\mathrm{cut}_{n,r_n}) = \mathbb{E}(N_1) + \ldots + \mathbb{E}(N_n),$$

and thus, since the sample points are identically distributed,

$$\mathbb{E}(\mathrm{cut}_{n,r_n}) = n\mathbb{E}(N_1).$$

Conditioning on the position of the sample point x_1, we have

$$\mathbb{E}(\mathrm{cut}_{n,r_n}) = n\int_{\mathbb{R}^d}\mathbb{E}(N_1|X_1=x)p(x)\mathrm{d}x.$$

Setting for $r>0$

$$g(x,r) = \begin{cases}\mu(B(x,r)\cap C^+) & \text{if } x\in C^-\\ \mu(B(x,r)\cap C^-) & \text{if } x\in C^+,\end{cases}$$

we have

$$\mathbb{E}(N_1|X_1=x) = (n-1)g(x,r_n),$$

and thus

$$\mathbb{E}\Big(\frac{\mathrm{cut}_{n,r_n}}{n(n-1)r_n^{d+1}}\Big) = \frac{1}{r_n^{d+1}}\int_{\mathbb{R}^d} g(x,r_n)p(x)\mathrm{d}x.$$

For $s\in S$, i.e. on the hyperplane S let n_s be the normal through s (pointing towards C^+). Then

$$\mathbb{E}\Big(\frac{\mathrm{cut}_{n,r_n}}{n(n-1)r_n^{d+1}}\Big) = \int_S\int_{-\infty}^{\infty}\frac{1}{r^{d+1}}g(s+tn_s,r_n)p(s+tn_s)\;\mathrm{d}t\;\mathrm{d}s.$$

We set

$$h_n(s) = \frac{1}{r_n^{d+1}}\int_{-\infty}^{\infty} g(s+tn_s,r_n)p(s+tn_s)\mathrm{d}t,$$

and define

$$\begin{aligned}\mathcal{R}_n &= \{x\in\mathbb{R}^d|\,\mathrm{dist}(x,\partial C)\leq 2r_n\}\\ \mathcal{I}_n &= C\setminus\mathcal{R}_n\\ \mathcal{A}_n &= \mathbb{R}^d\setminus(\mathcal{I}_n\cup\mathcal{R}_n).\end{aligned}$$

Then we can decompose the integral into

$$\int_S h_n(s)\;\mathrm{d}s = \int_{S\cap\mathcal{I}_n} h_n(s)\;\mathrm{d}s + \int_{S\cap\mathcal{R}_n} h_n(s)\;\mathrm{d}s + \int_{S\cap\mathcal{A}_n} h_n(s)\;\mathrm{d}s$$

and bound the deviation from the limit for each part separately.
We start with the case $s \in S \cap \mathcal{A}_n$. Since $\operatorname{dist}(s+tn_s, S) = t$ we have $g(s+tn_s, r_n) = 0$ for $t \notin [-2r_n, 2r_n]$. By definition of $\mathcal{A}_n$ $p(x) = 0$ for $x \in B(s, 2r_n)$ and thus especially $p(s + tn_s) = 0$ for $t \in [-2r_n, 2r_n]$. Therefore $h(s) = 0$ for $s \in S \cap \mathcal{A}_n$. Since also $2\eta_{d-1}/(d+1) \int_{S\cap\mathcal{A}_n} p^2(s)\ \mathrm{d}s = 0$, we have

$$\Big| \int_{S\cap\mathcal{A}_n} h_n(s)\ \mathrm{d}s - \frac{2\eta_{d-1}}{d+1} \int_{S\cap\mathcal{A}_n} p^2(s)\mathrm{d}s \Big| = 0$$

Now let $s \in S \cap \mathcal{R}_n$ and define

$$A(t) = \operatorname{vol}(B(0,1) \cap \{x = (x^{(1)}, \ldots, x^{(d)}) | x^{(1)} \geq t\}).$$

In this case we have

$$\begin{aligned} h_n(s) &= \frac{1}{r_n^{d+1}} \int_{-\infty}^{\infty} g(s + tn_s, r_n) p(s + tn_s)\ \mathrm{d}t \\ &\leq \frac{1}{r_n^{d+1}} \int_{-r_n}^{r_n} r_n^d p_{\max} A(t/r_n) p_{\max}\ \mathrm{d}t \\ &= \frac{p_{\max}^2}{r_n} \int_{-r_n}^{r_n} A(t/r_n)\ \mathrm{d}t = \frac{2p_{\max}^2 \eta_{d-1}}{d+1}, \end{aligned}$$

where we use Lemma 2, a substitution in the integral and finally Lemma 3. Therefore

$$\begin{aligned} \int_{S\cap\mathcal{R}_n} h_n(s)\ \mathrm{d}s - \frac{2\eta_{d-1}}{d+1} \int_{S\cap\mathcal{R}_n} p^2(s)\mathrm{d}s &\leq \int_{S\cap\mathcal{R}_n} \frac{2p_{\max}^2\eta_{d-1}}{d+1}\mathrm{d}s - \frac{2\eta_{d-1}}{d+1} \int_{S\cap\mathcal{R}_n} p^2(s)\mathrm{d}s \\ &= \frac{2\eta_{d-1}}{d+1} \int_{S\cap\mathcal{R}_n} p_{\max}^2 - p^2(s)\mathrm{d}s \\ &\leq \frac{2\eta_{d-1}}{d+1} p_{\max}^2 \operatorname{vol}_{d-1}(S \cap \mathcal{R}_n). \end{aligned}$$

The other inequality holds trivially.
Finally, let $s \in S \cap \mathcal{I}_n$. By definition $B(s, 2r_n) \subseteq C$. Let $v \in \mathbb{R}^d$, $\|v\| = 1$ and $t \in [0, 2r]$. Applying Taylor's theorem in one dimension

$$p(s + tv) = p(s) + D_v p(s + \xi v)t,$$

where $\xi \in (0, t)$. Thus for all $y \in B(s, 2r_n)$

$$|p(y) - p(s)| \leq 2p'_{\max} r_n.$$

Thus for $s \in S \cap \mathcal{I}_n$,

$$h_n(s) = \frac{1}{r_n^{d+1}} \int_{-\infty}^{\infty} g(s + tn_s, r_n) p(s + tn_s) \, \mathrm{d}t \tag{1}$$

$$= \frac{1}{r_n^{d+1}} \int_{-r_n}^{r_n} g(s + tn_s, r_n) p(s + tn_s) \, \mathrm{d}t \tag{2}$$

$$\leq \frac{1}{r_n^{d+1}} \int_{-r_n}^{r_n} r_n^d (p(s) + 2p'_{\max} r_n) A(t/r_n)(p(s) + 2p'_{\max} r_n) \, \mathrm{d}t \tag{3}$$

$$= \frac{1}{r_n} \int_{-r_n}^{r_n} (p(s) + 2p'_{\max} r_n) A(t/r_n)(p(s) + 2p'_{\max} r_n) \, \mathrm{d}t \tag{4}$$

$$= \frac{1}{r_n} (p^2(s) + 4p(s)p'_{\max} r_n + 4(p'_{\max})^2 r_n^2) \int_{-r_n}^{r_n} A(t/r_n) \, \mathrm{d}t \tag{5}$$

$$= \frac{2\eta_{d-1}}{d+1} (p^2(s) + 4p'_{\max} r_n (p(s) + p'_{\max} r_n)) \tag{6}$$

$$\leq \frac{2\eta_{d-1}}{d+1} (p^2(s) + 4p'_{\max} r_n (p_{\max} + p'_{\max} r_n)) \tag{7}$$

On the other hand,

$$h_n(s) \geq \frac{2\eta_{d-1}}{d+1} (p(s) - 2p'_{\max} r_n)^2 \tag{8}$$

$$= \frac{2\eta_{d-1}}{d+1} (p^2(s) - 4p(s)p'_{\max} r_n + 4(p'_{\max})^2 r_n^2) \tag{9}$$

$$= \frac{2\eta_{d-1}}{d+1} (p^2(s) + 4p'_{\max} r_n (p'_{\max} r_n - p(s))) \tag{10}$$

$$\geq \frac{2\eta_{d-1}}{d+1} (p^2(s) + 4p'_{\max} r_n (p'_{\max} r_n - p_{\max})) \tag{11}$$

Therefore

$$\int_{S \cap \mathcal{I}_n} h_n(s) \, \mathrm{d}s - \frac{2\eta_{d-1}}{d+1} \int_{S \cap \mathcal{I}_n} p^2(s) \mathrm{d}s \leq \frac{8\eta_{d-1} p'_{\max}}{d+1} r_n (p_{\max} + p'_{\max} r_n) \operatorname{vol}_{d-1}(S \cap \mathcal{I}_n) \tag{12}$$

and

$$\frac{2\eta_{d-1}}{d+1} \int_{S \cap \mathcal{I}_n} p^2(s) \mathrm{d}s - \int_{S \cap \mathcal{I}_n} h_n(s) \, \mathrm{d}s \leq \frac{8\eta_{d-1} p'_{\max}}{d+1} r_n (p_{\max} - p'_{\max} r_n) \operatorname{vol}_{d-1}(S \cap \mathcal{I}_n) \tag{13}$$

$$\leq \frac{8\eta_{d-1} p_{\max} p'_{\max}}{d+1} \operatorname{vol}_{d-1}(S \cap \mathcal{I}_n) r_n. \tag{14}$$

Thus, if r_n is sufficiently small, such that $p'_{\max} r_n \leq p_{\max}$, we have

$$\Big| \int_{S \cap \mathcal{I}_n} h_n(s) \, \mathrm{d}s - \frac{2\eta_{d-1}}{d+1} \int_{S \cap \mathcal{I}_n} p^2(s) \mathrm{d}s \Big| \leq \frac{16\eta_{d-1} p_{\max} p'_{\max}}{d+1} \operatorname{vol}_{d-1}(S \cap \mathcal{I}_n) r_n \tag{15}$$

$$\leq \frac{16\eta_{d-1} p_{\max} p'_{\max}}{d+1} \operatorname{vol}_{d-1}(S \cap C) r_n. \tag{16}$$

□

Lemma 2 *Let S be a hyperplane with normal vector n_s and $x \in \mathbb{R}^d$ with $\operatorname{dist}(x, S) = t$. Set*

$$g(x,r) = \begin{cases} \mu(B(x,r) \cap C^+) & \text{if } x \in C^- \\ \mu(B(x,r) \cap C^-) & \text{if } x \in C^+ \end{cases}$$

and let $\tilde{p}_{\min} \leq p(y) \leq \tilde{p}_{\max}$ for all $y \in B(x,r)$. Then $g(x,r) = 0$ if $t \geq r_n$ and

$$\tilde{p}_{\min} r_n^d A\big(\frac{t}{r_n}\big) \leq g(x,r) \leq \tilde{p}_{\max} r_n^d A\big(\frac{t}{r_n}\big),$$

where

$$A(t) = \operatorname{vol}(B(0,1) \cap \{x = (x^{(1)}, \ldots, x^{(d)}) | x^{(1)} \geq t\}).$$

Proof. Suppose that $x \in C^+$ (the other case can be treated analoguously). We use that by a translation and a rotation of the coordinate system (such that the origin is at x and $-n_s$ is the direction of the first unit vector)

$$\operatorname{vol}(B(x, r_n) \cap C^-) = \operatorname{vol}\left(B(0, r_n) \cap \left\{x^{(1)} \geq t\right\}\right)$$

and the invariance of the Lebesgue measure with respect to linear transformations. By scaling we obtain

$$\operatorname{vol}\left(B(0, r_n) \cap \left\{x^{(1)} \geq t\right\}\right) = r_n^d \operatorname{vol}\left(B(0,1) \cap \left\{x^{(1)} \geq \frac{t}{r_n}\right\}\right).$$

□

Lemma 3 (Integral over cap volume) *With*

$$A(t) = \operatorname{vol}(B(0,1) \cap \{x = (x^{(1)}, \ldots, x^{(d)}) | x^{(1)} \geq t\})$$

we have

$$\int_0^1 A(t) \mathrm{d}t = \frac{\eta_{d-1}}{d+1}.$$

Proof. We have

$$\begin{aligned} \int_0^1 A(t) \mathrm{d}t &= \int_0^1 \operatorname{vol}(B(0,1) \cap \{x_1 \geq t\}) \mathrm{d}t = \int_0^1 \int_t^1 \eta_{d-1} \sqrt{1-r^2}^{d-1} \mathrm{d}r \mathrm{d}t \\ &= \int_0^1 \int_0^r \eta_{d-1} \sqrt{1-r^2}^{d-1} \mathrm{d}t \mathrm{d}r = \int_0^1 \eta_{d-1} \sqrt{1-r^2}^{d-1} \int_0^r \mathrm{d}t \mathrm{d}r \\ &= \int_0^1 \eta_{d-1} r \sqrt{1-r^2}^{d-1} \mathrm{d}r \end{aligned}$$

Substituting $r = \cos\theta$, we have to integrate from $\theta = \arccos(0) = \pi/2$ to $\theta = \arccos(1) = 0$, and have $\mathrm{d}r = -\sin\theta \mathrm{d}\theta$. Thus,

$$\int_0^1 A(t)\mathrm{d}t = \eta_{d-1}\int_0^{\pi/2} \cos\theta \sin^d\theta \mathrm{d}\theta = \eta_{d-1}\big[\frac{1}{d+1}\sin^{d+1}\theta\big]_0^{\pi/2} = \frac{\eta_{d-1}}{d+1}.$$

□

3 Limit of $\mathbb{E}\,\mathrm{cut}_{n,k_n}$

The following lemma is necessary for the proof of both cases, $k_n/n \to \alpha > 0$ and $k_n/n \to 0$. The result for the first case is in Proposition 5, the result for the case $k_n/n \to 0$ in Proposition 6.

Lemma 4 (Lower/upper bound on $\mathbb{E}\,\mathrm{cut}_{n,k_n}$) *Let the general assumptions above hold. For $q \in [0,1]$ we set*

$$\tilde{r}(x,q) = \min\{r|\mu(B(x,r)) = q\}.$$

Then for $\delta \in (0,1/2)$,

$$\mathbb{E}(\mathrm{cut}_{n,k_n}) \geq n(n-1)\int_{\mathbb{R}^d} p(x)g(x,\tilde{r}(x,(1-\delta)\alpha_n))\,\mathrm{d}x - 4\exp\Big(2\log n - \frac{1}{4}\delta^2 k_n\Big) \tag{17}$$

$$\mathbb{E}(\mathrm{cut}_{n,k_n}) \leq n(n-1)\int_{\mathbb{R}^d} p(x)g(x,\tilde{r}(x,(1+\delta)\alpha_n))\,\mathrm{d}x + 4\exp\Big(2\log n - \frac{1}{4}\delta^2 k_n\Big), \tag{18}$$

where $\alpha_n = k_n/(n-1)$.

Proof. Let $N_{n,k_n}^{(i)}$ be the number of edges originating in point x_i. Then, due to the independent and identical distribution of the sample points,

$$\mathbb{E}(\mathrm{cut}_{n,k_n}) = \sum_{i=1}^n \mathbb{E}(N_{n,k_n}^{(i)}) = n\mathbb{E}(N_{n,k_n}^{(1)}).$$

Let $R_x^{k_n}$ denote the random variable that measures the k_n-NN radius and let $F_{R_x^{k_n}}$ denote its conditional distribution given that the position of the point is x. Conditioning

on the position of the point and its k_n-NN radius, we obtain

$$\begin{aligned}
\mathbb{E}(\mathrm{cut}_{n,k_n}) &= n \int_{\mathbb{R}^d} \mathbb{E}(N^{(1)}_{n,k_n} | X = x) p(x) \, \mathrm{d}x \\
&= n \int_{\mathbb{R}^d} p(x) \int_0^\infty \mathbb{E}(N^{(1)}_{n,k_n} | X = x, R_x^{k_n} = r) \, \mathrm{d}F_{R_x^{k_n}}(r) \, \mathrm{d}x \\
&= n(n-1) \int_{\mathbb{R}^d} p(x) \int_0^\infty g(x,r) \, \mathrm{d}F_{R_x^{k_n}}(r) \, \mathrm{d}x \\
&= \int_{\mathbb{R}^d} p(x) n(n-1) \int_0^\infty g(x,r) \, \mathrm{d}F_{R_x^{k_n}}(r) \, \mathrm{d}x,
\end{aligned}$$

where $g(x,r)$ is defined as above.
Now we treat the inner integral. By the monotonicity in r of $g(x,r)$ and $0 \le g(x,r) \le 1$ for all x and r we have

$$\begin{aligned}
\int_0^\infty g(x,r) \, \mathrm{d}F_{R_x^{k_n}}(r) &\ge \int_{\tilde{r}(x,(1-\delta)\alpha_n)}^{\tilde{r}(x,(1+\delta)\alpha_n)} g(x,r) \, \mathrm{d}F_{R_x^{k_n}}(r) \\
&\ge g(x, \tilde{r}(x,(1-\delta)\alpha_n)) \Pr\left(\tilde{r}(x,(1-\delta)\alpha_n) \le R_x^{k_n} \le \tilde{r}(x,(1+\delta)\alpha_n)\right) \\
&\ge g(x, \tilde{r}(x,(1-\delta)\alpha_n)) - \Pr(R_x^{k_n} > \tilde{r}(x,(1+\delta)\alpha_n)) - \Pr(R_x^{k_n} < \tilde{r}(x,(1-\delta)\alpha_n)),
\end{aligned}$$

and

$$\begin{aligned}
\int_0^\infty g(x,r) \, \mathrm{d}F_{R_x^{k_n}}(r) &= \int_0^{\tilde{r}(x,(1-\delta)\alpha_n)} g(x,r) \, \mathrm{d}F_{R_x^{k_n}}(r) + \int_{\tilde{r}(x,(1-\delta)\alpha_n)}^{\tilde{r}(x,(1+\delta)\alpha_n)} g(x,r) \, \mathrm{d}F_{R_x^{k_n}}(r) \\
&\quad + \int_{\tilde{r}(x,(1+\delta)\alpha_n)}^\infty g(x,r) \, \mathrm{d}F_{R_x^{k_n}}(r) \\
&\le g(x, \tilde{r}(x,(1+\delta)\alpha_n)) + \Pr(R_x^{k_n} > \tilde{r}(x,(1+\delta)\alpha_n)) + \Pr(R_x^{k_n} < \tilde{r}(x,(1-\delta)\alpha_n)).
\end{aligned}$$

We first bound the terms $n(n-1)\Pr(R_x^{k_n} > \tilde{r}(x,(1+\delta)\alpha_n))$ and $n(n-1)\Pr(R_x^{k_n} < \tilde{r}(x,(1-\delta)\alpha_n))$. Setting $V \sim \mathrm{Bin}(n-1,(1-\delta)\alpha_n)$ we have

$$\begin{aligned}
n(n-1)\Pr(R_x^{k_n} < \tilde{r}(x,(1-\delta)\alpha_n)) &= n(n-1)\Pr(V > k_n - 1) \\
&= n(n-1)\Pr\Big(V > \frac{k_n - 1}{(n-1)(1-\delta)\alpha_n}(n-1)(1-\delta)\alpha_n\Big) \\
&= n(n-1)\Pr\Big(V > (1 + \frac{k_n - 1}{(n-1)(1-\delta)\alpha_n} - 1)(n-1)(1-\delta)\alpha_n\Big) \\
&= n(n-1)\Pr\Big(V > (1 + \frac{k_n - 1}{(1-\delta)k_n} - 1)(1-\delta)k_n\Big) \\
&\le n^2 \exp\Big(-\frac{1}{4}(1-\delta)k_n(\frac{k_n - 1}{(1-\delta)k_n} - 1)^2\Big) \\
&\le n^2 \exp\Big(-\frac{1}{4}(1-\delta)k_n\Big(\frac{\delta k_n - 1}{(1-\delta)k_n}\Big)^2\Big) \\
&\le \exp\Big(2\log n - \frac{1}{4}\frac{(\delta k_n - 1)^2}{(1-\delta)k_n}\Big)
\end{aligned}$$

by a Chernoff bound (given that $(k-1)/((1-\delta)k) < 2e$, which is the case if $1/(1-\delta) < 2e$). For $\delta < 1/2$ as stated above we have

$$\frac{(\delta k_n - 1)^2}{(1-\delta)k_n} = \frac{\delta^2 k_n^2 - 2\delta k_n + 1}{(1-\delta)k_n} = \frac{\delta^2}{1-\delta}k_n - \frac{2\delta}{1-\delta} + \frac{1}{(1-\delta)k_n} \tag{19}$$

$$\geq \delta^2 k_n - 2 \tag{20}$$

For the other side we have, with $U \sim \mathrm{Bin}(n-1, (1+\delta)\alpha_n)$),

$$\begin{aligned}
n(n-1)\Pr(R_x^{k_n} > \tilde{r}(x,(1+\delta)\alpha_n)) &= n(n-1)\Pr(U < k_n) \\
&= n(n-1)\Pr(U < \frac{k_n}{(n-1)(1+\delta)\alpha_n}(n-1)(1+\delta)\alpha_n) \\
&= n(n-1)\Pr(U < (1-(1-\frac{k_n}{(n-1)(1+\delta)\alpha_n})(n-1)(1+\delta)\alpha_n) \\
&= n(n-1)\Pr(U < (1-(1-\frac{1}{1+\delta})(1+\delta)k_n) \\
&\leq n^2 \exp(-\frac{1}{2}(1+\delta)k_n(1-\frac{1}{(1+\delta)})^2) \\
&= \exp(2\log n - \frac{1}{2}\frac{\delta^2}{1+\delta}k_n) \\
&\leq \exp(2\log n - \frac{\delta^2}{3}k_n).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
n(n-1)\Pr(R_x^{k_n} < \tilde{r}(x,(1-\delta)\alpha_n)) &< \exp\left(2\log n - \frac{1}{4}\delta^2 k_n + \frac{1}{2}\right) \\
&\leq 2\exp\left(2\log n - \frac{1}{4}\delta^2 k_n\right),
\end{aligned}$$

where we have used $\sqrt{e} \leq 2$, and

$$n(n-1)\Pr(R_x^{k_n} > \tilde{r}(x,(1+\delta)\alpha_n)) < \exp(2\log n - \frac{\delta^2}{3}k_n).$$

Finally,

$$\begin{aligned}
\mathbb{E}(\mathrm{cut}_{n,k_n}) &= \int_{\mathbb{R}^d} p(x)n(n-1)\int_0^\infty g(x,r)\mathrm{d}F_{R_x^{k_n}}(r)\ \mathrm{d}x \\
&\geq \int_{\mathbb{R}^d} p(x)n(n-1)(g(x,\tilde{r}(x,(1-\delta)\alpha_n)) - \Pr(R_x^{k_n} > \tilde{r}(x,(1+\delta)\alpha_n)) \\
&\quad - \Pr(R_x^{k_n} < \tilde{r}(x,(1-\delta)\alpha_n)))\ \mathrm{d}x \\
&\geq \int_{\mathbb{R}^d} p(x)(n(n-1)g(x,\tilde{r}(x,(1-\delta)\alpha_n)) - 4\exp\left(2\log n - \frac{1}{4}\delta^2 k_n\right))\ \mathrm{d}x \\
&= n(n-1)\int_{\mathbb{R}^d} p(x)g(x,\tilde{r}(x,(1-\delta)\alpha_n)\ \mathrm{d}x - 4\exp\left(2\log n - \frac{1}{4}\delta^2 k_n\right),
\end{aligned}$$

and analoguously for the other inequality. □

Proposition 5 (Limit of $\mathbb{E}\,\mathrm{cut}_{n,k_n}$ for k_n linear in n) *If $\alpha_n = k_n/n \to \alpha$ ($n \to \infty$), then*

$$\lim_{n\to\infty} \mathbb{E}(\frac{1}{n^2}\,\mathrm{cut}_{n,k_n}) = \int_{\mathbb{R}^d} g(x,\tilde{r}(x,\alpha))p(x)\,\mathrm{d}x.$$

In particular for any sequence $\delta_n \to 0$ we have

$$\Big|\mathbb{E}\Big(\frac{\mathrm{cut}_{n,k_n}}{n(n-1)}\Big) - \int_{\mathbb{R}^d} p(x)g(x,\tilde{r}(x,\alpha))\Big| \le \delta_n\alpha_n + |\alpha_n - \alpha| + 4\exp\Big(2\log n - \frac{1}{4}\delta_n^2 k_n\Big).$$

Proof. Suppose w.l.o.g. that $x \in C^+$, $\alpha_1, \alpha_2 > 0$ with $\alpha_1 \le \alpha_2$. Then

$$\begin{aligned}
g(x,\tilde{r}(x,\alpha_1)) &= g(x,\tilde{r}(x,\alpha_2)) - (g(x,\tilde{r}(x,\alpha_2)) - g(x,\tilde{r}(x,\alpha_1)))\\
&= g(x,\tilde{r}(x,\alpha_2)) - (\mu(B(x,\tilde{r}(x,\alpha_2))\cap C^-)\\
&\quad - \mu(B(x,\tilde{r}(x,\alpha_1))\cap C^-)\\
&= g(x,\tilde{r}(x,\alpha_2)) - \mu((B(x,\tilde{r}(x,\alpha_2))\setminus B(x,\tilde{r}(x,\alpha_1)))\cap C^-)\\
&\ge g(x,\tilde{r}(x,\alpha_2)) - \mu(B(x,\tilde{r}(x,\alpha_2))\setminus B(x,\tilde{r}(x,\alpha_1))\\
&= g(x,\tilde{r}(x,\alpha_2)) - \mu(B(x,\tilde{r}(x,\alpha_2)) + \mu(B(x,\tilde{r}(x,\alpha_1)))\\
&= g(x,\tilde{r}(x,\alpha_2)) - \alpha_2 + \alpha_1,
\end{aligned}$$

where we use that

$$\mu(B(x,\tilde{r}(x,\alpha_2))\setminus B(x,\tilde{r}(x,\alpha_1))) = \mu(B(x,\tilde{r}(x,\alpha_2))) - \mu(B(x,\tilde{r}(x,\alpha_1))),$$

since $B(x,\tilde{r}(x,\alpha_1)) \subseteq B(x,\tilde{r}(x,\alpha_2))$ due to $\alpha_1 \le \alpha_2$.
Using

$$(1+\delta_n)\alpha_n \le \alpha + \delta_n\alpha_n + |\alpha_n - \alpha|$$

and

$$(1-\delta_n)\alpha_n \ge \alpha - \delta_n\alpha_n - |\alpha_n - \alpha|,$$

we obtain

$$\begin{aligned}
g(x,\tilde{r}(x,(1-\delta_n)\alpha_n)) &\ge g(x,\tilde{r}(\alpha - \delta_n\alpha_n - |\alpha_n - \alpha|)\\
&\ge g(x,\tilde{r}(x,\alpha)) - \alpha + (\alpha - \delta_n\alpha_n - |\alpha_n - \alpha|)\\
&= g(x,\tilde{r}(x,\alpha)) - \delta_n\alpha_n - |\alpha_n - \alpha|
\end{aligned}$$

and

$$\begin{aligned}
g(x,\tilde{r}(x,(1+\delta_n)\alpha_n)) &\le g(x,\tilde{r}(\alpha + \delta_n\alpha_n + |\alpha_n - \alpha|)\\
&\le g(x,\tilde{r}(x,\alpha)) - \alpha + (\alpha + \delta_n\alpha_n + |\alpha_n - \alpha|)\\
&= g(x,\tilde{r}(x,\alpha)) + \delta_n\alpha_n + |\alpha_n - \alpha|.
\end{aligned}$$

Thus,

$$\int_{\mathbb{R}^d} p(x)g(x,\tilde{r}(x,(1-\delta)\alpha_n))\mathrm{d}x \geq \int_{\mathbb{R}^d} p(x)(g(x,\tilde{r}(x,\alpha)) - \delta_n\alpha_n - |\alpha_n - \alpha|)\mathrm{d}x$$
$$= \int_{\mathbb{R}^d} p(x)g(x,\tilde{r}(x,\alpha))\mathrm{d}x - \delta_n\alpha_n - |\alpha_n - \alpha|$$

Analoguously, we can show that

$$\int_{\mathbb{R}^d} p(x)g(x,\tilde{r}(x,(1+\delta)\alpha_n))\mathrm{d}x \leq \int_{\mathbb{R}^d} p(x)g(x,\tilde{r}(x,\alpha))\mathrm{d}x + \delta_n\alpha_n + |\alpha_n - \alpha|.$$

Thus, using the result from Lemma 4 we obtain

$$\Big|\mathbb{E}\Big(\frac{\mathrm{cut}_{n,k_n}}{n(n-1)}\Big) - \int_{\mathbb{R}^d} p(x)g(x,\tilde{r}(x,\alpha))\Big| \leq \delta_n\alpha_n + |\alpha_n - \alpha| + 4\exp\Big(2\log n - \frac{1}{4}\delta_n^2 k_n\Big).$$

□

Proposition 6 (Limit of $\mathbb{E}\,\mathrm{cut}_{n,k_n}$ for $k_n/n \to 0$) *Let p be bounded from above on $\mathbb{R}^d$ and from below away from 0 on C, that is $p(x) \leq p_{\max}$ for all $x \in \mathbb{R}^d$ and $p(x) \geq p_{\min} > 0$ for all $x \in C$. Furthermore let p be differentiable in the interior of C and the absolute value of the directional derivative bounded by $p'_{\max}$.*
Assume further that we can find constants r_γ, $\gamma > 0$ such that for all $x \in C$ and $r \leq r_\gamma$

$$\mathrm{vol}(B(x,r) \cap C) \geq \gamma\,\mathrm{vol}(B(x,r)). \tag{21}$$

Let $\alpha_n = k_n/(n-1) \to 0$ and $\delta_n \to 0$ with $\delta_n < 1/2$ for all $n \in \mathbb{N}$.
Let $n \geq 2$ be sufficiently large such that

$$\sqrt[d]{\frac{2\alpha_n}{\gamma p_{\min}\eta_d}} < r_\gamma,$$

and set

$$\nu_n = 2\frac{p'_{\max}}{p_{\min}}\sqrt[d]{\frac{2\alpha_n}{\gamma p_{\min}\eta_d}}.$$

Further define

$$\mathcal{R}_n = \{x \in \mathbb{R}^d | \,\mathrm{dist}(x,\partial C) \leq 2r_n^{\max}\}$$
$$\mathcal{I}_n = C \setminus \mathcal{R}_n$$

Then we have for $\nu_n \leq 1/2$

$$\mathbb{E}\Big(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}\Big) \geq (1 - 4\nu_n - 2\delta_n)\frac{2\eta_{d-1}}{d+1}\eta_d^{-1-1/d}\int_{S\cap\mathcal{I}_n} p^{1-1/d}(s)$$
$$- 4\exp\Big((1+1/d)(\log n - \log k_n) - \frac{1}{4}\delta_n^2 k_n\Big),$$

and

$$\mathbb{E}\Big(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}\Big) \le (1+64\nu_n+3\delta_n+72n^{-1})\frac{2\eta_{d-1}}{d+1}\eta_d^{-1-1/d}\int_{S\cap\mathcal{I}_n} p^{1-1/d}(s) + \frac{8p_{\max}^2}{\gamma^{1+1/d}p_{\min}^{1+1/d}\eta_d^{1/d}}\,\mathrm{vol}_{d-1}(S\cap\mathcal{R}_n) + 4\exp\Big((1+1/d)(\log n-\log k_n)-\frac{1}{4}\delta_n^2k_n\Big).$$

Proof. According to Lemma 4 we have for an arbitrary sequence δ_n with $\delta_n < 1/2$ for all $n \in \mathbb{N}$

$$\mathbb{E}\Big(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}\Big) \ge \frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}\int_{\mathbb{R}^d} p(x)g(x,\tilde{r}(x,(1-\delta_n)\alpha_n))\,\mathrm{d}x - 4\exp\Big((1+1/d)(\log n-\log k_n)-\frac{1}{4}\delta_n^2k_n\Big),$$

and

$$\mathbb{E}\Big(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}\Big) \le \frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}\int_{\mathbb{R}^d} p(x)g(x,\tilde{r}(x,(1+\delta_n)\alpha_n))\,\mathrm{d}x + 4\exp\Big((1+1/d)(\log n-\log k_n)-\frac{1}{4}\delta_n^2k_n\Big).$$

First we show, for a suitable sequence δ_n, an upper bound for the integral

$$\frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}\int_{\mathbb{R}^d} p(x)g(x,\tilde{r}(x,(1+\delta_n)\alpha_n))\,\mathrm{d}x.$$

If S denotes the separating hyperplane and for $s \in S$, n_s denotes the normal in S, then

$$\frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}\int_{\mathbb{R}^d} p(x)g(x,\tilde{r}(x,(1+\delta_n)\alpha_n))\,\mathrm{d}x \tag{22}$$

$$= \int_S \frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}\int_{-\infty}^{\infty} p(s+tn_s)g(s+tn_s,\tilde{r}(s+tn_s,(1+\delta_n)\alpha_n))\,\mathrm{d}t\,\mathrm{d}s \tag{23}$$

$$= \int_S h_n(s)\mathrm{d}s, \tag{24}$$

where we have set

$$h_n(s) = \frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}\int_{-\infty}^{\infty} p(s+tn_s)g(s+tn_s,\tilde{r}(s+tn_s,(1+\delta_n)\alpha_n))\,\mathrm{d}t. \tag{25}$$

We set

$$\tilde{r}_{\max}((1+\delta_n)\alpha_n) = \sqrt[d]{\frac{(1+\delta_n)\alpha_n}{\gamma p_{\min}\eta_d}},$$

and assume that $\tilde{r}_{\max}((1+\delta_n)\alpha_n) < r_\gamma$. Then, for $x \in C$,

$$\begin{aligned}
\mu\left(B\left(x, \sqrt[d]{\frac{(1+\delta_n)\alpha_n}{\gamma p_{\min}\eta_d}}\right)\right) &\geq p_{\min} \operatorname{vol}\left(B\left(x, \sqrt[d]{\frac{(1+\delta_n)\alpha_n}{\gamma p_{\min}\eta_d}}\right) \cap C\right) \\
&\geq p_{\min}\gamma \operatorname{vol}\left(B\left(x, \sqrt[d]{\frac{(1+\delta_n)\alpha_n}{\gamma p_{\min}\eta_d}}\right)\right) \\
&= p_{\min}\gamma \frac{(1+\delta_n)\alpha_n}{\gamma p_{\min}\eta_d}\eta_d \\
&= (1+\delta_n)\alpha_n
\end{aligned}$$

and therefore we have $\tilde{r}(x, (1+\delta_n)\alpha_n) \leq \tilde{r}_{\max}((1+\delta_n)\alpha_n)$ for all $x \in C$. Define $r_n^{\max} = \tilde{r}_{\max}((1+\delta_n)\alpha_n)$ and

$$\begin{aligned}
\mathcal{R}_n &= \{x \in \mathbb{R}^d | \operatorname{dist}(x, \partial C) \leq 2r_n^{\max}\} \\
\mathcal{I}_n &= C \setminus \mathcal{R}_n \\
\mathcal{A}_n &= \mathbb{R}^d \setminus (\mathcal{I}_n \cup \mathcal{R}_n).
\end{aligned}$$

Then we can decompose the integral into

$$\int_S h_n(s)\, \mathrm{d}s = \int_{S\cap\mathcal{I}_n} h_n(s)\, \mathrm{d}s + \int_{S\cap\mathcal{R}_n} h_n(s)\, \mathrm{d}s + \int_{S\cap\mathcal{A}_n} h_n(s)\, \mathrm{d}s.$$

Let $s \in S \cap \mathcal{A}_n$. Then $d(s, C) \geq r_n^{\max}$ and thus for $|t| \leq r_n^{\max}$ we have $p(s + tn_s) = 0$. If $|t| > r_n^{\max}$ and $s + tn_s \notin C$ then $p(s + tn_s) = 0$ as well. Otherwise, if $s + tn_s \in C$ we have $\tilde{r}(s + tn_s, (1+\delta_n)\alpha_n) \leq r_n^{\max}$ but $d(s + tn_s, S) \geq r_n^{\max}$ and thus $g(s + tn_s, \tilde{r}(s + tn_s, (1+\delta_n)\alpha_n)) = 0$. Therefore

$$\int_{S\cap\mathcal{A}_n} h_n(s)\, \mathrm{d}s = 0. \tag{26}$$

Now let $s \in S \cap \mathcal{R}_n$. We have for any $s \in S$ and $t \in \mathbb{R}$

$$p(s + tn_s)g(s + tn_s, \tilde{r}(s + tn_s, (1+\delta)\alpha_n)) \leq p_{\max} g(s + tn_s, \tilde{r}_{\max}((1+\delta)\alpha_n)),$$

since either $p(s+tn_s)=0$ or $p(s+tn_s)\le p_{\max}$ and the condition on $\tilde{r}$ above holds. Therefore we have (assume $\delta<1$)

$$\begin{aligned}
h_n(s) &\le \frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}\int_{-\tilde{r}_{\max}((1+\delta_n)\alpha_n)}^{+\tilde{r}_{\max}((1+\delta_n)\alpha_n)} p_{\max} g(s+tn_s, \tilde{r}_{\max}((1+\delta_n)\alpha_n))\,\mathrm{d}t \\
&\le \frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}\int_{-\tilde{r}_{\max}((1+\delta_n)\alpha_n)}^{+\tilde{r}_{\max}((1+\delta_n)\alpha_n)} p_{\max}^2\eta_d\tilde{r}_{\max}^d((1+\delta_n)\alpha_n)\,\mathrm{d}t \\
&= \frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}p_{\max}^2\eta_d\frac{(1+\delta_n)\alpha_n}{\gamma p_{\min}\eta_d}2\tilde{r}_{\max}((1+\delta_n)\alpha_n) \\
&= (1+\delta_n)\frac{n-1}{n}\sqrt[d]{\frac{n}{k_n}}\frac{2p_{\max}^2}{\gamma p_{\min}}\sqrt[d]{\frac{(1+\delta_n)\alpha_n}{\gamma p_{\min}\eta_d}} \\
&\le (1+\delta_n)^{1+1/d}\frac{2p_{\max}^2}{\gamma^{1+1/d}p_{\min}^{1+1/d}\eta_d^{1/d}} \\
&\le \frac{8p_{\max}^2}{\gamma^{1+1/d}p_{\min}^{1+1/d}\eta_d^{1/d}},
\end{aligned}$$

and thus

$$\int_{S\cap\mathcal{R}_n} h_n(s)\,\mathrm{d}s \le \frac{8p_{\max}^2}{\gamma^{1+1/d}p_{\min}^{1+1/d}\eta_d^{1/d}}\int_{S\cap\mathcal{R}_n} 1\,\mathrm{d}s \tag{27}$$

$$= \frac{8p_{\max}^2}{\gamma^{1+1/d}p_{\min}^{1+1/d}\eta_d^{1/d}}\,\mathrm{vol}_{d-1}(S\cap\mathcal{R}_n). \tag{28}$$

Finally, we consider the case $s\in S\cap\mathcal{I}_n$, that means $B(s,2r_n^{\max})\subseteq C$. For all $y\in B(s,2r_n^{\max})$ we have

$$p(s)-2p'_{\max}r_n^{\max}\le p(y)\le p(s)+2p'_{\max}r_n^{\max}, \tag{29}$$

or, written differently,

$$p(s)(1-2\frac{p'_{\max}}{p(s)}r_n^{\max})\le p(y)\le p(s)(1+2\frac{p'_{\max}}{p(s)}r_n^{\max}). \tag{30}$$

Setting

$$\nu_n = 2\frac{p'_{\max}}{p_{\min}}\sqrt[d]{\frac{2\alpha_n}{\gamma p_{\min}\eta_d}}, \tag{31}$$

under the assumption $\delta_n\le 1$

$$p(s)(1-\nu_n)\le p(y)\le p(s)(1+\nu_n), \tag{32}$$

and we have for all $x\in B(s,r_n^{\max})$,

$$\sqrt[d]{\frac{1}{1+\nu_n}}\sqrt[d]{\frac{(1+\delta_n)\alpha_n}{p(s)\eta}}\le\tilde{r}(x,(1+\delta_n)\alpha_n)\le\sqrt[d]{\frac{1}{1-\nu_n}}\sqrt[d]{\frac{(1+\delta_n)\alpha_n}{p(s)\eta}}.$$

We denote

$$\tilde{r}_+ = \sqrt[d]{\frac{1}{1-\nu_n}} \sqrt[d]{\frac{(1+\delta_n)\alpha_n}{p(s)\eta}}.$$

By the monotonicity of g, we have

$$\begin{aligned}
h_n(s) &= \frac{n-1}{k_n} \sqrt[d]{\frac{n}{k_n}} \int_{-\infty}^{\infty} p(s+tn_s) g(s+tn_s, \tilde{r}(s+tn_s, (1+\delta_n)\alpha_n)) \\
&\leq \frac{n-1}{k_n} \sqrt[d]{\frac{n}{k_n}} \int_{-\tilde{r}_+}^{\tilde{r}_+} (1+\nu_n) p(s) g(s+tn_s, \tilde{r}_+) \mathrm{d}t \\
&\leq \frac{n-1}{k_n} \sqrt[d]{\frac{n}{k_n}} \int_{-\tilde{r}_+}^{\tilde{r}_+} (1+\nu_n) p(s) (\tilde{r}_+)^d A(t/\tilde{r}_+) (1+\nu_n) p(s) \mathrm{d}t \\
&= \frac{n-1}{k_n} \sqrt[d]{\frac{n}{k_n}} (1+\nu_n)^2 p^2(s) (\tilde{r}_+)^d \int_{-\tilde{r}_+}^{\tilde{r}_+} A(t/\tilde{r}_+) \mathrm{d}t \\
&= \frac{n-1}{k_n} \sqrt[d]{\frac{n}{k_n}} (1+\nu_n)^2 p^2(s) (\tilde{r}_+)^d 2\tilde{r}_+ \int_0^1 A(u) \mathrm{d}u \\
&= \frac{n-1}{k_n} \sqrt[d]{\frac{n}{k_n}} \frac{2\eta_{d-1}}{d+1} (1+\nu_n)^2 p^2(s) \tilde{r}_+^{d+1} \\
&= \frac{n-1}{k_n} \sqrt[d]{\frac{n}{k_n}} \frac{2\eta_{d-1}}{d+1} (1+\nu_n)^2 p^2(s) \Big(\frac{1}{1-\nu_n}\Big)^{1+1/d} \alpha_n^{1+1/d} \Big(\frac{(1+\delta_n)}{p(s)\eta}\Big)^{1+1/d} \\
&= \frac{(1+\nu_n)^2}{(1-\nu_n)^{1+1/d}} (1+\delta_n)^{1+1/d} \sqrt[d]{\frac{n}{n-1}} \frac{2\eta_{d-1}}{d+1} \eta_d^{-1-1/d} p^{1-1/d}(s).
\end{aligned}$$

And thus

$$\int_{S\cap\mathcal{I}_n} h_n(s)\ \mathrm{d}s \leq \frac{(1+\nu_n)^2}{(1-\nu_n)^{1+1/d}} (1+\delta_n)^{1+1/d} \sqrt[d]{\frac{n}{n-1}} \frac{2\eta_{d-1}}{d+1} \eta_d^{-1-1/d} \int_{S\cap\mathcal{I}_n} p^{1-1/d}(s)\ \mathrm{d}s.$$

For $\nu_n \leq 1/2$ we have

$$\begin{aligned}
\frac{(1+\nu_n)^2}{(1-\nu_n)^{1+1/d}} &= \left(\frac{1+\nu_n}{1-\nu_n}\right)^{1+1/d} (1+\nu_n)^{1-1/d} = \left(1+\frac{2\nu_n}{1-\nu_n}\right)^{1+1/d} (1+\nu_n)^{1-1/d} \\
&\leq (1+4\nu_n)^{1+1/d} (1+4\nu_n)^{1-1/d} = (1+4\nu_n)^2 \\
&= 1+8\nu_n+16\nu_n^2 \leq 1+16\nu_n,
\end{aligned}$$

and, since $\delta_n < 1$ by assumption,

$$(1+\delta_n)^{1+1/d} \leq (1+\delta_n)^2 = 1+2\delta_n+\delta_n^2 \leq 1+3\delta_n.$$

Combining these, and

$$\sqrt[d]{\frac{n}{n-1}} = \sqrt[d]{1+\frac{1}{n-1}} \leq 1+\frac{2}{n}$$

for $n \geq 2$, we have for $\nu_n \leq 1/2$ and $\delta_n < 1$

$$\begin{aligned}\frac{(1+\nu_n)^2}{(1-\nu_n)^{1+1/d}}(1+\delta_n)^{1+1/d}\sqrt[d]{\frac{n}{n-1}} &\leq (1+16\nu_n+3\delta_n+48\nu_n\delta_n)(1+2n^{-1})\\ &\leq (1+64\nu_n+3\delta_n)(1+2n^{-1})\\ &\leq 1+64\nu_n+3\delta_n+2n^{-1}+128\nu_n n^{-1}+6\delta_n n^{-1}\\ &\leq 1+64\nu_n+3\delta_n+2n^{-1}+64n^{-1}+6n^{-1}\\ &= 1+64\nu_n+3\delta_n+72n^{-1}.\end{aligned}$$

Now we treat the other side, that means the integral

$$\frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}\int_{\mathbb{R}^d} p(x)g(x,\tilde{r}(x,(1-\delta_n)\alpha_n))\,\mathrm{d}x.$$

We set

$$h_n^-(s) = \frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}\int_{-\infty}^{\infty} p(s+tn_s)g(s+tn_s,\tilde{r}(s+tn_s,(1-\delta_n)\alpha_n))\,\mathrm{d}t \tag{33}$$

and can show similarly to above

$$\frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}\int_{\mathbb{R}^d} p(x)g(x,\tilde{r}(x,(1-\delta_n)\alpha_n))\,\mathrm{d}x = \int_S h_n^-(s)\mathrm{d}s.$$

With $\mathcal{R}_n$, $\mathcal{A}_n$, and $\mathcal{I}_n$ as above we certainly have

$$\begin{aligned}\int_S h_n^-(s)\mathrm{d}s &= \int_{S\cap\mathcal{I}_n} h_n^-(s)\,\mathrm{d}s + \int_{S\cap\mathcal{R}_n} h_n^-(s)\,\mathrm{d}s + \int_{S\cap\mathcal{A}_n} h_n^-(s)\,\mathrm{d}s &(34)\\ &\geq \int_{S\cap\mathcal{I}_n} h_n^-(s)\,\mathrm{d}s. &(35)\end{aligned}$$

Let $s \in S\cap\mathcal{I}_n$ and ν_n as above. For all $y \in B(s,2r_n^{\max})$, under the assumption $\delta_n < 1$, we have $p(y) \leq (1+\nu_n)p(s)$ and thus for all $x \in B(s,r_n^{\max})$,

$$\tilde{r}(x,(1-\delta_n)\alpha_n) \geq \sqrt[d]{\frac{1}{1+\nu_n}}\sqrt[d]{\frac{(1-\delta_n)\alpha_n}{p(s)\eta}}.$$

Setting

$$\tilde{r}_- = \sqrt[d]{\frac{1}{1+\nu_n}}\sqrt[d]{\frac{(1-\delta_n)\alpha_n}{p(s)\eta}},$$

we have by the monotonicity of g,

$$\begin{aligned}
h_n^-(s) &= \frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}\int_{-\infty}^{\infty} p(s+tn_s)g(s+tn_s,\tilde{r}(s+tn_s,(1-\delta_n)\alpha_n)) \\
&\geq \frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}\int_{-\tilde{r}_-}^{\tilde{r}_-}(1-\nu_n)p(s)g(s+tn_s,\tilde{r}_-)\mathrm{d}t \\
&\geq \frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}\int_{-\tilde{r}_-}^{\tilde{r}_-}(1-\nu_n)p(s)(\tilde{r}_-)^d A(t/\tilde{r}_-)(1-\nu_n)p(s)\mathrm{d}t \\
&= \frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}(1-\nu_n)^2p^2(s)(\tilde{r}_-)^d\int_{-\tilde{r}_-}^{\tilde{r}_-}A(t/\tilde{r}_-)\mathrm{d}t \\
&= \frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}(1-\nu_n)^2p^2(s)(\tilde{r}_-)^d 2\tilde{r}_-\int_0^1 A(u)\mathrm{d}u \\
&= \frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}\frac{2\eta_{d-1}}{d+1}(1-\nu_n)^2p^2(s)\tilde{r}_-^{d+1} \\
&= \frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}\frac{2\eta_{d-1}}{d+1}(1-\nu_n)^2p^2(s)\Big(\frac{1}{1+\nu_n}\Big)^{1+1/d}\alpha_n^{1+1/d}\Big(\frac{(1-\delta_n)}{p(s)\eta}\Big)^{1+1/d} \\
&= \frac{(1-\nu_n)^2}{(1+\nu_n)^{1+1/d}}(1-\delta_n)^{1+1/d}\sqrt[d]{\frac{n}{n-1}}\frac{2\eta_{d-1}}{d+1}\eta_d^{-1-1/d}p^{1-1/d}(s).
\end{aligned}$$

And thus

$$\int_{S\cap\mathcal{I}_n} h_n^-(s)\ \mathrm{d}s \geq \frac{(1-\nu_n)^2}{(1+\nu_n)^{1+1/d}}(1-\delta_n)^{1+1/d}\sqrt[d]{\frac{n}{n-1}}\frac{2\eta_{d-1}}{d+1}\eta_d^{-1-1/d}\int_{S\cap\mathcal{I}_n} p^{1-1/d}(s)\ \mathrm{d}s.$$

We have

$$\begin{aligned}
\frac{(1-\nu_n)^2}{(1+\nu_n)^{1+1/d}} &= \left(\frac{1-\nu_n}{1+\nu_n}\right)^{1+1/d}(1-\nu_n)^{1-1/d} = \left(1-\frac{2\nu_n}{1+\nu_n}\right)^{1+1/d}(1-\nu_n)^{1-1/d} \\
&\geq (1-2\nu_n)^{1+1/d}(1-2\nu_n)^{1-1/d} = (1-2\nu_n)^2 \\
&= 1-4\nu_n+4\nu_n^2 \geq 1-4\nu_n,
\end{aligned}$$

and

$$(1-\delta_n)^{1+1/d} \geq (1-\delta_n)^2 = 1-2\delta_n+\delta_n^2 \geq 1-2\delta_n.$$

Combining the above, we have

$$\frac{(1-\nu_n)^2}{(1+\nu_n)^{1+1/d}}(1-\delta_n)^{1+1/d}\sqrt[d]{\frac{n}{n-1}} \geq 1-4\nu_n-2\delta_n.$$

□

Corollary 1 *Under the general assumptions above and* $k_n \to 0$, $\log(n)/k_n \to 0$

$$\mathbb{E}\Big(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}\Big) \to \frac{2\eta_{d-1}}{d+1}\eta_d^{-1-1/d}\int_S p^{1-1/d}(s)\;\mathrm{d}s$$

for $n \to \infty$.

Proof. Clearly,

$$\int_{S\cap\mathcal{I}_n} p^{1-1/d}(s)\;\mathrm{d}s \le \int_{S\cap C} p^{1-1/d}(s)\;\mathrm{d}s = \int_S p^{1-1/d}(s)\;\mathrm{d}s.$$

On the other hand,

$$\begin{aligned}\int_S p^{1-1/d}(s)\;\mathrm{d}s - \int_{S\cap\mathcal{I}_n} p^{1-1/d}(s)\;\mathrm{d}s &= \int_{S\cap(C\setminus\mathcal{I}_n)} p^{1-1/d}(s)\;\mathrm{d}s\\ &\le p_{\max}^{1-1/d}\,\mathrm{vol}_{d-1}(S\cap(C\setminus\mathcal{I}_n))\end{aligned}$$

Using these simple observation, the result of Proposition 6 and the sandwich theorem we show convergence:
Setting $\delta_n = \sqrt[4]{\log(n)/k_n}$, according to our assumptions $\delta_n \to 0$ and

$$\delta_n^2\frac{k_n}{\log n} = \sqrt{\frac{\log n}{k_n}}\frac{k_n}{\log n} = \sqrt{\frac{k_n}{\log n}} \to \infty$$

for $n \to \infty$. Thus

$$\begin{aligned}\exp\Big((1+1/d)(\log n - \log k_n) - \frac{1}{4}\delta_n^2 k_n\Big) &\le \exp\Big((1+1/d)\log n - \frac{1}{4}\delta_n^2 k_n\Big)\\ &\le \exp\Big(\log n\Big((1+1/d) - \frac{1}{4}\delta_n^2\frac{k_n}{\log n}\Big)\Big) \to 0\end{aligned}$$

Since $S\cap\partial C$ is a set of $(d-1)$-dimensional measure 0 by assumption, we have $\mathrm{vol}_{d-1}(S\cap\mathcal{R}_n)\to 0$. Similarly, $\mathrm{vol}_{d-1}(S\cap(C\setminus\mathcal{I}_n))\to 0$. □

4 Variance term for cut_{n,k_n} and cut_{n,r_n}

Proposition 7 (Deviation from mean for scaled cut_{n,k_n}**)** *Under our general assumptions,* $k_n/\log n \to \infty$ *and* $k_n/n \to 0$,

$$\Pr\Big(\Big|\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n} - \mathbb{E}\Big(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}\Big)\Big| > \epsilon\Big) \le 2\exp\Big(-\frac{2\epsilon^2 n^{1-2/d}k_n^{2/d}}{(\tau_d+1)^2}\Big),$$

where τ_d *denotes the kissing number in* d *dimensions (i.e. the number of unit hyperspheres in* $\mathbb{R}^d$ *which can touch a unit hypersphere without any intersections, see [2]).*

Proof. Let $x_1, \dots, x_n$ be points drawn i.i.d. from our density p and let $\bar{x}_i \in \mathbb{R}^d$. Let $\mathrm{cut}^{(i)}_{n,k_n}$ denote the cut induced by S in the k_n-nearest neighbor graph that is constructed on the points $x_1, \dots, x_{i-1}, \bar{x}_i, x_{i+1}, \dots, x_n$. Changing the position of one point x_i to $\bar{x}_i$ at most $k_n + 2\tau_d k_n \leq 3\tau_d k_n$ (where τ_d is the kissing number in d dimensions) edges across the cut can change (because the number of outgoing edges is k_n and the number of incoming edges is bounded by $\tau_d k_n$, cf. [2]).
Thus

$$|\,\mathrm{cut}_{n,k_n} - \mathrm{cut}^{(i)}_{n,k_n}\,| \leq 3\tau_d k_n.$$

Hence,

$$\Big|\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n} - \frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}^{(i)}_{n,k_n}\Big| \leq \frac{3\tau_d k_n}{nk_n}\sqrt[d]{\frac{n}{k_n}} = \frac{3\tau_d}{n}\sqrt[d]{\frac{n}{k_n}}.$$

Thus by McDiarmid,

$$\begin{aligned}\Pr\Big(\Big|\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n} - \mathbb{E}\Big(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}\Big)\Big| > \epsilon\Big) &\leq 2\exp\Big(-\frac{2\epsilon^2}{n(\frac{3\tau_d}{n}\sqrt[d]{\frac{n}{k_n}})^2}\Big)\\ &= 2\exp(-\frac{2\epsilon^2 n^{1-2/d}k_n^{2/d}}{(3\tau_d)^2})\end{aligned}$$

Since $k_n/\log n \to \infty$ we have $n^{1-2/d}k_n^{2/d} \to \infty$ for $d \geq 2$ and $n \to \infty$. □

Corollary 2 (Limit of cut_{n,k_n}) *Let $d \geq 2$. For $n \to \infty$, $k_n/\log n \to \infty$ and $k_n/n \to 0$*

$$\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n} \xrightarrow{a.s.} \frac{2\eta_{d-1}}{d+1}\eta_d^{-1-1/d}\int_S p^{1-1/d}(s)\mathrm{d}s.$$

Proof. Let $\epsilon > 0$ and let

$$\mathrm{cut}_{\infty,\infty} := \frac{2\eta_{d-1}}{d+1}\eta_d^{-1-1/d}\int_S p^{1-1/d}(s).$$

Then

$$\begin{aligned}&\Pr\Big(\Big|\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n} - \mathrm{cut}_{\infty,\infty}\Big| > \epsilon\Big)\\ &\quad\leq \Pr\Big(\Big|\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n} - \mathbb{E}\Big(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}\Big)\Big| > \frac{\epsilon}{2}\Big)\\ &\quad+ \Pr\Big(\Big|\mathbb{E}\Big(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}\Big) - \mathrm{cut}_{\infty,\infty}\Big| > \frac{\epsilon}{2}\Big).\end{aligned}$$

The second term converges deterministically (as we have seen in Proposition 6), and for the first term we have

$$\Pr\Big(\Big|\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n} - \mathbb{E}\Big(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}\Big)\Big| > \frac{\epsilon}{2}\Big) \leq \exp\Big(-\frac{\epsilon^2 n^{1-2/d}k_n^{2/d}}{3(\tau_d)^2}\Big).$$

Thus, for $d \geq 2$ and $k_n/\log n \to \infty$ we have

$$\sum_{n=1}^{\infty} \Pr\Big(\Big|\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n} - \mathbb{E}\Big(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}\Big)\Big| > \frac{\epsilon}{2}\Big) < \infty,$$

which implies almost sure convergence by Borel-Cantelli. □

Proposition 8 (Deviation from mean of cut_{n,r_n}) *For every $\epsilon > 0$ there exists a constant $c_\epsilon > 0$ such that*

$$\Pr\Big(\Big|\frac{1}{n(n-1)r_n^{d+1}}\,\mathrm{cut}_{n,r_n} - \mathbb{E}\Big(\frac{1}{n(n-1)r_n^{d+1}}\,\mathrm{cut}_{n,r_n})\Big| > \epsilon\Big) \leq 2\exp(-c_\epsilon n r_n^{d+1})$$

Proof. For $i, j \in \{1, \ldots, n\}$, $i \neq j$ set

$$N_{i,j} = \begin{cases} 1 & \text{if } (X_i, X_j) \text{ edge in } G_{n,r_n} \text{ and } X_i \text{ and } X_j \text{ on different sides of } S, \\ 0 & \text{otherwise.} \end{cases}$$

Clearly,

$$\mathrm{cut}_{n,r_n} = \sum_{i=1}^{n}\sum_{j\neq i} N_{i,j},$$

and thus

$$\frac{1}{n(n-1)r_n^{d+1}}\,\mathrm{cut}_{n,r_n} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{r_n^{d+1}}N_{i,j}.$$

Setting

$$U = \frac{1}{n(n-1)r_n^{d+1}}\,\mathrm{cut}_{n,r_n}$$

and

$$g_{ij} = \frac{1}{r_n^{d+1}}N_{i,j},$$

we have

$$0 \leq g_{ij} \leq \frac{1}{r_n^{d+1}}.$$

With $b = 1/r_n^{d+1}$ we have

$$g_{ij} \leq \mathbb{E}U + b$$

and setting

$$\sigma^2 = \operatorname{Var} g_{ij} = \frac{1}{r_n^{2(d+1)}} \operatorname{Var}(N_{ij})$$

we have by a remark in [1]

$$\begin{aligned}\Pr(U - \mathbb{E}U \geq \epsilon) &\leq \exp\left(- \frac{n\epsilon}{2b}\left[\left(1 + \frac{\sigma^2}{b\epsilon}\right)\log\left(1 + \frac{b\epsilon}{\sigma^2}\right) - 1\right]\right) \\ &= \exp\left(- \frac{nr_n^{d+1}\epsilon}{2}\left[\left(1 + \frac{r_n^{d+1}\sigma^2}{\epsilon}\right)\log\left(1 + \frac{\epsilon}{r_n^{d+1}\sigma^2}\right) - 1\right]\right).\end{aligned}$$

Since $\log(1+x) > x/(1+x)$ for $x > 0$ we have $(1+x)\log(1+1/x) > 1$ and furthermore one can show that this function is decreasing. Thus the expression in the exponent can be bounded by using an upper bound for $r_n^{d+1}\sigma^2$.
We have

$$\begin{aligned}r_n^{d+1}\sigma^2 &= r_n^{d+1} \operatorname{Var} g_{ij} = r_n^{d+1} \frac{1}{r_n^{2(d+1)}} \operatorname{Var}(N_{ij}) \\ &= \frac{1}{r_n^{d+1}} \Pr(N_{ij} = 1)(1 - \Pr(N_{ij} = 1)) \leq \frac{1}{r_n^{d+1}} \Pr(N_{ij} = 1).\end{aligned}$$

Conditioning on the location of point X_i to the surface S we have $\Pr(N_{ij} = 1) = 0$ if $d(X_i, S) > r_n$. Otherwise $\Pr(N_{ij} = 1) \leq p_{\max} r_n^d \eta_d$. Therefore

$$\Pr(N_{ij} = 1) \leq p_{\max} r_n^d \eta_d \Pr(d(X_i, S) \leq r_n).$$

Under our assumptions there exists a constant c_1 (close to the (d-1)-dimensional measure of $S \cap C$) such that $\Pr(d(X_i, S) \leq r_n) \leq c_1 r_n$. Thus there exists a constant $c_2 > 0$ with

$$\Pr(N_{ij} = 1) \leq \tilde{c_2} r_n^{d+1}.$$

Combining the above, for each $\epsilon > 0$ we can find a constant c_ϵ such that

$$\Pr(U - \mathbb{E}U \geq \epsilon) \leq \exp(-c_\epsilon n r_n^{d+1}).$$

Finally, we use $U - \mathbb{E}U \leq -\epsilon \Leftrightarrow (-U) - \mathbb{E}(-U) \geq \epsilon$ and a similar argument to the one above in order to show the statement for the absolute deviation from the mean. □

Corollary 3 (Limit of $\operatorname{cut}_{n,r_n}$) *Let $nr_n^{d+1} \to \infty$ for $n \to \infty$. Then*

$$\frac{1}{n^2 r_n^{d+1}} \operatorname{cut}_{n,r_n} \xrightarrow{a.s.} \frac{2\eta_{d-1}}{d+1} \int_S p^2(s) \mathrm{d}s.$$

Proof. Similar to the proof for the k-NN graph and using that under the condition $nr_n^{d+1} \to \infty$ we have for every $\epsilon > 0$

$$\sum_{n=1}^{\infty} 2\exp(-c_\epsilon n r_n^{d+1}) < \infty.$$

□

5 Bounds for the volume vol

Here we derive the limit for the volume. Limits for other quantities used in the balancing terms, such as the number of points in one side of a cut, can be derived similarly.

Proposition 9 (Limit of vol_{n,k_n}) *Let $H \subseteq \mathbb{R}^d$ be a subset of $\mathbb{R}^d$ with $\mu(H) > 0$. Then*

$$\mathbb{E}\Big(\frac{1}{nk_n}\mathrm{vol}_{n,k_n}(H)\Big) = \mu(H),$$

and

$$\Pr\Big(\Big|\frac{1}{nk_n}\mathrm{vol}_{n,k_n}(H) - \mu(H)\Big| > \epsilon\Big) \le 2\exp\Big(-\frac{1}{2}\epsilon^2 n\Big).$$

Proof. The expected number of points in H is $n\mu(H)$, each of them has exactly k outgoing edges, thus

$$\mathbb{E}(\mathrm{vol}_{n,k_n}(H)) = nk_n\mu(H).$$

Changing the position of one point will change the volume by at most k_n, thus with McDiarmid's inequality

$$\Pr(|\frac{1}{nk_n}\mathrm{vol}_{n,k_n}(H) - \mu(H)| > \epsilon) < 2\exp(-\frac{2\epsilon^2}{n(\frac{k}{nk_n})^2}) = 2\exp(-2\epsilon^2 n).$$

□

Proposition 10 (Limit of $\mathbb{E}(\mathrm{vol}_{n,r_n})$) *Let $H \subseteq \mathbb{R}^d$ be a measurable subset of $\mathbb{R}^d$ with $\mu(H) > 0$ and let the following assumptions hold:*

- *$p'_{\max}$ is the supremum of the absolute value of the directional derivative (over all directions) of p, and*
- *we can find a constant $c_\partial > 0$ such that for all ϵ sufficiently small,*

$$\mathrm{vol}(\{x \in \mathbb{R}^d | \operatorname{dist}(x, \partial(C \cap H)) \le \epsilon\} \le 2c_\partial \epsilon \,\mathrm{vol}_{d-1}(\partial(C \cap H)).$$

Then if $r_n \to 0$ *and* $nr_n^{2d} \to \infty$ *for* $n \to \infty$

$$\Big|\mathbb{E}\big(\frac{1}{n(n-1)r_n^d}\operatorname{vol}_{n,r_n}(H)\big) - \eta_d \int_H p^2(x)\mathrm{d}x\Big| \leq r_n\eta_d\big(p'_{\max} + 2c_\partial p_{\max}^2 \operatorname{vol}_{d-1}(\partial(H\cap C))\big),$$

for n *sufficiently large (such that the condition above holds for* r_n*).*

Proof. Let $\mathcal{E}_{n,r_n}$ denote the edges of the graph G_{n,r_n}. With

$$M_i = \begin{cases} |\{(x_i, x_j) \in \mathcal{E}_{n,r_n}\}| & \text{if } x_i \in H \\ 0 & \text{otherwise,} \end{cases}$$

we have

$$\operatorname{vol}_{n,r_n}(H)) = M_1 + \ldots + M_n$$

and thus, due to the independent identical distribution of the sample point,

$$\mathbb{E}(\operatorname{vol}_{n,r_n}(H)) = n\mathbb{E}(M_1).$$

Conditioning on the realization of the random variable X_1, that is the position of x_1, we have

$$\begin{aligned}\mathbb{E}(\operatorname{vol}_{n,r_n}(H)) &= n\int_{\mathbb{R}^d} \mathbb{E}(M_1|X_1 = x)p(x)\ \mathrm{d}x \\ &= n\int_H (n-1)\mu(B(x,r_n))p(x)\ \mathrm{d}x \\ &= n\int_{H\cap C} (n-1)\mu(B(x,r_n))p(x)\ \mathrm{d}x\end{aligned}$$

and thus

$$\mathbb{E}\Big(\frac{1}{n(n-1)r_n^d}\operatorname{vol}_{n,r_n}(H)\Big) = \frac{1}{r_n^d}\int_{H\cap C}\mu(B(x,r_n))p(x)\ \mathrm{d}x.$$

Set

$$\mathcal{R}_n = \{x \in H\cap C | \operatorname{dist}(x, \partial(H\cap C)) \leq r_n\}$$

and

$$\mathcal{I}_n = (H\cap C)\setminus \mathcal{R}_n.$$

We have

$$\begin{aligned}&\frac{1}{r_n^d}\int_{H\cap C}\mu(B(x,r_n))p(x)\ \mathrm{d}x \\ &\qquad = \frac{1}{r_n^d}\int_{\mathcal{R}_n}\mu(B(x,r_n))p(x)\ \mathrm{d}x + \frac{1}{r_n^d}\int_{\mathcal{I}_n}\mu(B(x,r_n))p(x)\ \mathrm{d}x.\end{aligned}$$

Let $x \in \mathcal{I}_n$, $v \in \mathbb{R}^d$, $\|v\| = 1$ and $t > 0$ sufficiently small. Applying Taylor's theorem in one dimension

$$p(x+tv) = p(x) + D_v p(x+\xi v)t,$$

where $\xi \in (0,t)$. Thus for all $y \in B(x,r_n)$

$$|p(y)-p(x)| \le p'_{\max} r_n.$$

Hence, we can approximate the integral

$$\begin{aligned}
\frac{1}{r_n^d}\int_{\mathcal{I}_n} \mu(B(x,r_n))p(x)\,\mathrm{d}x &\le \frac{1}{r_n^d}\int_{\mathcal{I}_n} (p(x)+p'_{\max}r_n) r_n^d \eta_d p(x)\,\mathrm{d}x \\
&= \frac{1}{r_n^d}\int_{\mathcal{I}_n} p(x)^2 r_n^d \eta_d\,\mathrm{d}x + \frac{1}{r_n^d}\int_{\mathcal{I}_n} p'_{\max} r_n r_n^d \eta_d p(x)\,\mathrm{d}x \\
&\le \eta_d \int_{\mathcal{I}_n} p(x)^2\,\mathrm{d}x + \eta_d p'_{\max} r_n \int_{\mathcal{I}_n} p(x)\,\mathrm{d}x \\
&\le \eta_d \int_{\mathcal{I}_n} p(x)^2\,\mathrm{d}x + \eta_d p'_{\max} r_n.
\end{aligned}$$

Similarly we can show the lower bound, and thus

$$\Big|\frac{1}{r_n^d}\int_{\mathcal{I}_n} \mu(B(x,r_n))p(x)\,\mathrm{d}x - \eta_d \int_{\mathcal{I}_n} p(x)\,\mathrm{d}x\Big| \le \eta_d p'_{\max} r_n.$$

Now we turn to the border strip $\mathcal{R}_n$. We have

$$\begin{aligned}
&\frac{1}{r_n^d}\int_{\mathcal{R}_n} \mu(B(x,r_n))p(x)\,\mathrm{d}x - \eta_d \int_{\mathcal{R}_n} p^2(x)\,\mathrm{d}x \\
&\quad\le \frac{1}{r_n^d}\int_{\mathcal{R}_n} p_{\max}\eta_d r_n^d p(x)\,\mathrm{d}x - \eta_d \int_{\mathcal{R}_n} p^2(x)\,\mathrm{d}x \\
&\quad= \eta_d \int_{\mathcal{R}_n} p_{\max} p(x)\,\mathrm{d}x - \eta_d \int_{\mathcal{R}_n} p^2(x)\,\mathrm{d}x = \eta_d \int_{\mathcal{R}_n} (p_{\max}-p(x))p(x)\,\mathrm{d}x \\
&\quad\le \eta_d p_{\max} \int_{\mathcal{R}_n} p(x)\,\mathrm{d}x
\end{aligned}$$

On the other hand, clearly

$$\begin{aligned}
\eta_d \int_{\mathcal{R}_n} p^2(x)\,\mathrm{d}x - \frac{1}{r_n^d}\int_{\mathcal{R}_n} \mu(B(x,r_n))p(x)\,\mathrm{d}x &\le \eta_d \int_{\mathcal{R}_n} p^2(x)\,\mathrm{d}x \\
&\le \eta_d p_{\max} \int_{\mathcal{R}_n} p(x)\,\mathrm{d}x.
\end{aligned}$$

Thus,

$$\begin{aligned}
\Big|\frac{1}{r_n^d}\int_{\mathcal{R}_n} \mu(B(x,r_n))p(x)\,\mathrm{d}x - \eta_d \int_{\mathcal{R}_n} p(x)\,\mathrm{d}x\Big| &\le \eta_d p_{\max}^2 \operatorname{vol}(\mathcal{R}_n) \\
&\le 2c_\partial \eta_d p_{\max}^2 \operatorname{vol}_{d-1}(\partial(H\cap C)) r_n
\end{aligned}$$

Clearly,

$$\Big|\frac{1}{r_n^d}\int_{\mathcal{H}\cap C^c}\mu(B(x,r_n))p(x)\;\mathrm{d}x-\eta_d\int_{\mathcal{H}\cap C^c}p(x)\;\mathrm{d}x\Big|=0$$

since both terms are 0. □

Proposition 11 (Deviation from mean of vol_{n,r_n}) *For every $\epsilon>0$ there exists a constant $c_\epsilon>0$ such that*

$$\Pr\Big(\Big|\frac{1}{n(n-1)r_n^d}\,\mathrm{vol}_{n,r_n}-\mathbb{E}\Big(\frac{1}{n(n-1)r_n^d}\,\mathrm{vol}_{n,r_n})\Big|>\epsilon\Big)\le 2\exp(-c_\epsilon n r_n^d)$$

Proof. For $i,j\in\{1,\dots,n\}$, $i\neq j$ set

$$N_{i,j}=\begin{cases}1 & \text{if } (X_i,X_j) \text{ edge in } G_{n,r_n} \text{ and } X_i\in H\ ,\\ 0 & \text{otherwise.}\end{cases}$$

Clearly,

$$\mathrm{vol}_{n,r_n}=\sum_{i=1}^{n}\sum_{j\neq i}N_{i,j},$$

and thus

$$\frac{1}{n(n-1)r_n^d}\,\mathrm{vol}_{n,r_n}=\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{r_n^d}N_{i,j}.$$

Setting

$$U=\frac{1}{n(n-1)r_n^d}\,\mathrm{vol}_{n,r_n}$$

and

$$g_{ij}=\frac{1}{r_n^d}N_{i,j},$$

we have

$$0\le g_{ij}\le\frac{1}{r_n^d}.$$

With $b=1/r_n^d$ we have

$$g_{ij}\le\mathbb{E}U+b$$

and setting

$$\sigma^2 = \operatorname{Var} g_{ij} = \frac{1}{r_n^{2d}} \operatorname{Var}(N_{ij})$$

we have by a remark in [1]

$$\begin{aligned}\Pr(U - \mathbb{E}U \geq \epsilon) &\leq \exp\left(- \frac{n\epsilon}{2b} \left[\left(1 + \frac{\sigma^2}{b\epsilon}\right) \log\left(1 + \frac{b\epsilon}{\sigma^2}\right) - 1 \right] \right) \\ &= \exp\left(- \frac{nr_n^d \epsilon}{2} \left[\left(1 + \frac{r_n^d \sigma^2}{\epsilon}\right) \log\left(1 + \frac{\epsilon}{r_n^d \sigma^2}\right) - 1 \right] \right).\end{aligned}$$

Since $\log(1+x) > x/(1+x)$ for $x > 0$ we have $(1+x)\log(1+1/x) > 1$ and furthermore one can show that this function is decreasing. Thus the expression in the exponent can be bounded by using an upper bound for $r_n^d \sigma^2$.
We have

$$\begin{aligned}r_n^d \sigma^2 &= r_n^d \operatorname{Var} g_{ij} = r_n^d \frac{1}{r_n^{2d}} \operatorname{Var}(N_{ij}) \\ &= \frac{1}{r_n^d} \Pr(N_{ij} = 1)(1 - \Pr(N_{ij} = 1)) \leq \frac{1}{r_n^d} \Pr(N_{ij} = 1).\end{aligned}$$

Conditioning on the location of point X_i we have $\Pr(N_{ij} = 1) = 0$ if $X_i \notin H$. Otherwise $\Pr(N_{ij} = 1) \leq p_{\max} r_n^d \eta_d$. Therefore

$$\Pr(N_{ij} = 1) \leq p_{\max} r_n^d \eta_d \Pr(X_i \in H) = p_{\max} r_n^d \eta_d \mu(H).$$

Combining the above, for each $\epsilon > 0$ we can find a constant c_ϵ such that

$$\Pr(U - \mathbb{E}U \geq \epsilon) \leq \exp(-c_\epsilon n r_n^d).$$

Finally, we use $U - \mathbb{E}U \leq -\epsilon \Leftrightarrow (-U) - \mathbb{E}(-U) \geq \epsilon$ and a similar argument to the one above in order to show the statement for the absolute deviation from the mean. □

Corollary 4 (Limit of vol_{n,r_n}) *Let $nr_n^d \to \infty$ for $n \to \infty$. Then*

$$\frac{1}{n^2 r_n^d} \mathrm{vol}_{n,r_n} \xrightarrow{a.s.} \eta_d \int_H p^2(x)\, \mathrm{d}x.$$

Proof. Similar to the proof for the k-NN graph and using that under the condition $nr_n^d \to \infty$ we have for every $\epsilon > 0$

$$\sum_{n=1}^{\infty} 2 \exp(-c_\epsilon n r_n^d) < \infty.$$

□

References

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

[2] Gary L. Miller, Shang-Hua Teng, William Thurston, and Stephen A. Vavasis. Separators for sphere-packings and nearest neighbor graphs. *J. ACM*, 44(1):1–29, 1997. ISSN 0004-5411. doi: http://doi.acm.org/10.1145/256292.256294.