Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2017, Vol. 53, No. 3, 1051–1068
DOI: 10.1214/16-AIHP747
© Association des Publications de l'Institut Henri Poincaré, 2017

ANNALES DE L'INSTITUT HENRI POINCARÉ
PROBABILITÉS ET STATISTIQUES
www.imstat.org/aihp

On estimating the perimeter using the alpha-shape

Ery Arias-Castro[a] and Alberto Rodríguez-Casal[b]

[a]*Department of Mathematics, University of California, San Diego, 9500 Gilman Drive, San Diego, CA 92093-0112, USA.*
url: http://math.ucsd.edu/~eariasca
[b]*Departamento de Estatística e Investigación Operativa, Facultade de Matemáticas, Universidade de Santiago de Compostela, Rúa Lope Gómez de Marzoa, s/n. Campus sur, 15782, Santiago de Compostela, A Coruña, Spain.*
url: http://eio.usc.es/pub/alberto/

Received 19 June 2015; revised 31 January 2016; accepted 13 February 2016

Abstract. We consider the problem of estimating the perimeter of a smooth domain in the plane based on a sample from the uniform distribution over the domain. We study the performance of the estimator defined as the perimeter of the alpha-shape of the sample. Some numerical experiments corroborate our theoretical findings.

Résumé. Nous considérons le problème de l'estimation du périmètre d'un domaine à bord lisse dans le plan basé sur un échantillon tiré de la loi uniforme ayant pour support le domaine en question. Nous étudions la performance de l'estimateur défini par le périmètre de la forme-alpha (« alpha-shape ») de l'échantillon. Des expériences numériques confirment notre théorie.

MSC: 62G99; 60D05

Keywords: Perimeter estimation; α-shape; r-convex hull; Rolling condition; Sets with positive reach

1. Introduction

The problem of recovering topological and geometric information about the support of a distribution based on a sample has received a considerable amount of attention in a number of fields, such as computational geometry, computer vision, image analysis, clustering or pattern recognition. This includes, for example, estimating of the number of connected components [2], the intrinsic dimensionality [17] and, more generally, the homology [5,6,20,30,37], the Minkowski content [7], as well as the perimeter and area [3,29]. The estimation of the support or, more generally, level sets of a density is itself a rich line of research [4,27,31,33–35]. A closely related topic is that of set estimation [8,18]. We refer the reader to the classic book of [15], which treats a number of these topics.

We focus here on the problem of estimating the perimeter of the support. Concretely, we are given a set of points $\mathcal{X}_n = \{X_1, \dots, X_n\}$, which we assume are independently sampled uniformly at random from an unknown compact set $S \subset \mathbb{R}^2$, and our goal is to estimate the perimeter of S, by which we mean the length of its boundary. Let ∂S denote the boundary of a set $S \subset \mathbb{R}^2$, namely $\partial S = \bar{S} \cap \overline{S^{\mathsf{c}}}$, where $\bar{S}$ denotes the closure of S and $S^{\mathsf{c}} = \mathbb{R}^2 \setminus S$ is the complement of S.

1.1. *Related work*

[29] address this problem under the assumption that S is convex and estimate its perimeter by the perimeter of the convex hull of the sample $\mathcal{X}_n$. They obtain the precise rate of convergence in expectation, which is of order $O(n^{-2/3})$ when the boundary ∂S has bounded curvature. They also obtain an analogous result for the problem of estimating the

area of S. [3] extend their results to other sampling distributions. See [28] for a review on more recent results on the convex hull of a random sample.

There is a series of papers that consider the problem of estimating the surface area of the boundary of a more general class of supports S but under a different sampling scheme where two samples are given, one from the uniform distribution on S and another from the uniform distribution on $G \setminus S$, where G is a bounded set containing S. We refer to this model as *binary images*. In that line, [7] aim at estimating the Minkowski content of ∂S, and introduce an estimator that is proved to be consistent under weak assumptions on the set S. They obtain a convergence rate of $O(n^{-1/4})$ in dimension 2 when ∂S has bounded curvature – in which case the Minkowski content coincides with the perimeter. [22,23] follow their work and propose a different estimator, which is very closely related to the one we study here, obtaining an improved rate convergence of $O(n^{-1/3})$ in dimension 2. Continuing this line of work, [13] propose an estimator of the perimeter of S based on a Delaunay triangulation, which is shown to be consistent under mild assumptions on S. Working with the same sampling scheme, but allowing for noise, [14] consider the estimation of the length of the boundary of a horizon of the form $\{(x, y) \in [0, 1]^2 : y \le g(x)\}$, where $g : [0, 1] \mapsto [0, 1]$ is a function with Hölder regularity. We further comment on this paper in Section 6 in our discussion of the minimax (sub)optimality of our estimator.

1.2. *The r-rolling condition*

A set S is said to fulfill the r-rolling condition if for any $x \in \partial S$ there is a open ball with radius r, B, such that $B \cap S = \varnothing$ and $x \in \partial B$. In this paper, we work under the assumption that S satisfies the following condition:

> *S is a compact subset of $\mathbb{R}^2$ such that both S and S^{c} satisfy the r-rolling condition.*

From a geometrical point of view, we are assuming that a ball of radius r can roll inside S and S^{c}. This rolling condition implies that, for any $x \in \partial S$, there are two open balls B^+ and B^- such that $x \in \partial B^+ \cap \partial B^-$, $B^+ \subset S$ and $B^- \subset S^{\mathsf{c}}$. In fact, it can be easily seen [21, Lemma A.0.1] that this is only possible if there is a (unique) unit vector η_x (the unit normal vector at x pointing outward) such that $B^+ = B(x - r\eta_x, r)$ and $B^- = B(x + r\eta_x, r)$, where $B(a, \alpha)$ denotes the open ball with radius α and center $a \in \mathbb{R}^2$. See [36] for a comprehensive discussion, including a relation to Serra's regular model and mathematical morphology. The r-rolling condition is closely linked to the notion of *r-convexity*. A set S is said to be r-convex if for any point $x \notin \bar{S}$ there is a open ball B of radius r such that $x \in B$ and $B \cap \bar{S} = \varnothing$ [26,35]. It is known that, if both S and S^{c} satisfy the r-rolling condition, then S and S^{c} are r-convex; see [21, Lemma A.0.8] and also [36].

The r-rolling condition is also connected with the idea of sets of positive reach introduced in the seminal paper [12]. For a nonempty set $S \subset \mathbb{R}^2$ and $x \in \mathbb{R}^2$, define

$$\operatorname{dist}(x, S) = \inf\bigl\{\|x - s\| : s \in S\bigr\},$$

where $\|\cdot\|$ stands for the Euclidean norm. The reach of a set S, denoted $\rho(S)$, is the supremum over $r > 0$ such that there is a unique point realizing $\inf\{\|x - s\| : s \in S\}$ on the set $\{x : \operatorname{dist}(x, S) < r\}$. For twice differentiable submanifolds (e.g., curves), the reach bounds the radius of curvature from above [12, Lem. 4.17]. Also, if S and S^{c} satisfy the r-rolling condition then $\rho(\partial S) \ge r$; see [21, Lemma A.0.6]. Conversely, using results in [9], it follows easily that the converse is true if, in addition, S is equal to the closure of its interior.

1.3. *The estimator*

Our estimator for the perimeter of S is the perimeter of the *α-shape* of $\mathcal{X}_n$, for some fixed $0 < \alpha < r$. The α-shape of $\mathcal{X}_n$ is the polygon, denoted $C_\alpha(\mathcal{X}_n)$, whose edges – which we call α-edges – are defined as follows [11]. A pair (X_i, X_j) forms an α-edge if there is an open ball B of radius α such that $X_i, X_j \in \partial B$ and $B \cap \mathcal{X}_n = \varnothing$. If α is large enough, the α-shape coincides with the convex hull of the sample. For a smaller α, the α-shape is not necessarily convex. See Figure 1 for an illustration. The α-shape is well known in the computational geometry literature for producing good global reconstructions if the sample points are (approximately) uniformly distributed in the set S. Moreover, it can be computed efficiently in time $O(n \log n)$. See [10] for a survey.

[9] estimate the perimeter of S by the outer Minskowski content of the *r-convex hull* of the sample, defined as the smallest r-convex set that contains the sample. Since the boundary of that set is smooth except at a finite number of

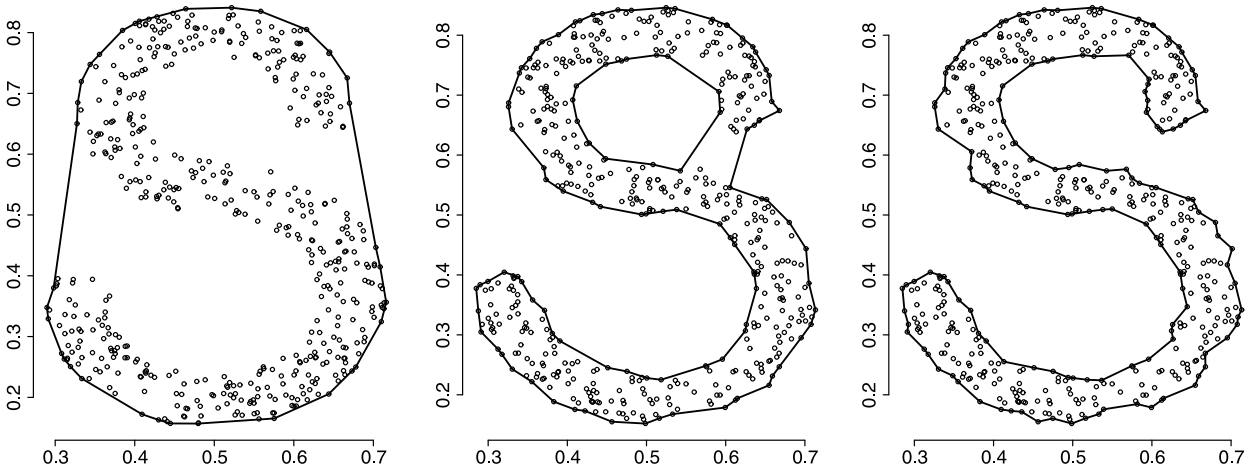

Fig. 1. The α-shape of a sample of size $n = 500$ from the uniform distribution of a thick S letter, for $\alpha = 1$ (left), $\alpha = 0.06$ (center) and $\alpha = 0.035$ (right). Note that in the second case the α-shape is made of two disconnected closed curves.

points, the outer Minskowski coincides with the perimeter. See [1] for a broader correspondence between these two quantities. [9] show that this estimator is consistent, but no convergence rate is provided. Note that, for large sample sizes, both estimators are quite similar; see Proposition 2 for a formal statement. From the computational point of view, the α-shape of the sample tends to be more stable with respect to the value of α, and is faster to compute over a range of values of α – the latter can be done in $O(n \log n)$ time, since the α-shape changes a finite number of times with α. The α-convex hull of the sample does not enjoy such properties.

1.4. *Main results*

Let λ denote the one-dimensional Hausdorff measure in $\mathbb{R}^2$, normalized so that it equals 1 for a line segment of length 1, and let $\operatorname{diam}(A) = \sup\{\|x - y\| : x, y \in A\}$ denote the diameter of a set $A \subset \mathbb{R}^2$.

Theorem 1. *Let $\mathcal{X}_n = (X_1, \ldots, X_n)$ be an independent sample from the uniform distribution on a compact set $S \subset \mathbb{R}^2$ such that S and S^c satisfy the r-rolling condition. Fix $\alpha \in (0, r)$. There is a constant A depending only on $(\alpha, r, \operatorname{diam}(S))$ and $t_0 > 0$ depending only on (α, r) such that, for all $0 \le t \le t_0$,*

$$\mathbb{P}\left(\left|\frac{\lambda(C_\alpha)}{\lambda(\partial S)} - 1\right| > t\right) \le An^2 \exp\bigl(-nt^{3/2}/A\bigr). \tag{1}$$

Remark 1. *In particular, defining $\varepsilon_n = (3A \log(n)/n)^{2/3}$, with probability one,*

$$(1 - \varepsilon_n)\lambda(\partial S) \le \lambda\bigl(C_\alpha(\mathcal{X}_n)\bigr) \le (1 + \varepsilon_n)\lambda(\partial S),$$

eventually, by applying the Borel–Cantelli lemma. So, the convergence rate of $\lambda(C_\alpha(\mathcal{X}_n))$ as an estimator of $\lambda(\partial S)$ is, up to a log factor, of order $n^{-2/3}$.

Remark 2. *We will argue later on that the same result holds also for the perimeter of the α-convex hull of the sample, refining, thus, the convergence established in* [9]. *See the discussion in Section 6.*

1.5. *Content*

The remaining of the paper is largely devoted to proving Theorem 1. In Section 2 we establish some auxiliary geometrical results. Section 3 is dedicated to the study of α-edges. Theorem 1 is proved in Section 4. Some numerical experiments are presented in Section 5. We discuss some extensions and open problems in Section 6.

1.6. *Notation and preliminaries*

We start by introducing some notation and some general concepts. Let $\mu(A)$ denote the Lebesgue measure of a measurable set $A \subset \mathbb{R}^2$. For a pair of distinct points $x_1, x_2 \in \mathbb{R}^2$, let (x_1x_2) denote the line passing through x_1 and x_2, and let $[x_1x_2]$ denote the line segment with endpoints x_1 and x_2. For a non empty set $A \subset \mathbb{R}^2$ and $\varepsilon > 0$, define

$$B(A, \varepsilon) = \big\{x \in \mathbb{R}^2 : \operatorname{dist}(x, A) < \varepsilon\big\}.$$

If $A = \{x\}$ is a singleton we use the notation $B(x, \varepsilon)$ (resp. $\bar{B}(x, \varepsilon)$) instead of $B(\{x\}, \varepsilon)$ for denoting the open (resp. closed) ball of radius $\varepsilon > 0$ and center $x \in \mathbb{R}^2$. Let P_A denote the metric projection onto a set A, i.e., $P_A(x) = \arg\min_{a \in A} \|x - a\|$, which is a singleton when $\operatorname{dist}(x, A) < \rho(A)$. For two nonempty sets $C, D \subset \mathbb{R}^2$, let $\mathcal{H}(C, D)$ denote their Hausdorff distance, defined as

$$\mathcal{H}(C, D) = \inf\big\{\varepsilon > 0 : C \subset B(D, \varepsilon) \text{ and } D \subset B(C, \varepsilon)\big\}.$$

For a curve $C \subset \mathbb{R}^2$ and $x \in C$, $\vec{C}_x$ denotes the tangent subspace of C at x when it exists. For two curves, C and D, respectively differentiable almost everywhere and differentiable, and such that $\rho(D) \geq r$ and $C \subset B(D, r)$, define the deviation angle of C with respect to D as

$$\angle(C, D) = \sup_{x \in C} \angle(\vec{C}_x, \vec{D}_{P_D(x)}),$$

where $\angle(\vec{C}_x, \vec{D}_{P_D(x)}) \in [0, \pi/2]$ denotes the angle between the tangent spaces of C and D at x and $P_D(x)$, respectively [19]. Note that it is not symmetric in C and D.

Where they appear, α and r are fixed. Everywhere in the proof, a constant only depends (at most) on α, r and the diameter of S. We will leave this dependence implicit most of the time.

We let n denote the sample size throughout. We say that an event holds with high probability if it happens with probability at least $1 - Ae^{-n/A}$ for some constant $A > 0$.

2. Some geometrical results

In this section we gather a few geometrical results that we will use later on in the paper.

Lemma 1. *Let $S \subset \mathbb{R}^2$ such that S and S^{c} satisfy the r-rolling condition. Any ball of radius $\alpha > 0$ with center in S contains a ball of radius $\frac{1}{2}\min\{\alpha, r\}$ included in S.*

Proof. Let Γ be a shorthand for ∂S. First, we will analyze the case $\alpha \leq r$. If $z \in S$ satisfies $\operatorname{dist}(z, \Gamma) \geq \alpha$, then $B(z, \alpha) \subset S$. Now, take $z \in S$ such that $\operatorname{dist}(z, \Gamma) < \alpha$ and let y be the metric projection of z onto Γ, which is well-defined since $\operatorname{dist}(z, \Gamma) < \rho(\Gamma)$. By the r-rolling property, there is an open ball B of radius r tangent to Γ at y that contains z and $B \subset S$. Therefore $B(z, \alpha) \cap B$ contains the ball of radius $\alpha/2$ tangent to Γ at y that contains z. See Figure 2 for an illustration. This concludes the proof for $\alpha \leq r$. If $\alpha > r$, the ball of radius α contains the ball of radius r with same center. By what we just did, that ball contains a ball of radius $r/2$ which belongs to S. □

Recall that μ denotes the Lebesgue measure on $\mathbb{R}^2$.

Lemma 2. *Let $S \subset \mathbb{R}^2$ be measurable and such that S and S^{c} satisfy the r-rolling condition. For any $\alpha \leq r$, there is a numeric constant $A > 0$ depending only on α such that, for any $z \notin S$,*

$$\mu\big(B(z, \alpha) \cap S\big) \geq A \max\big(0, \alpha - \operatorname{dist}(z, \partial S)\big)^{3/2}.$$

Proof. Let Γ be a shorthand for ∂S. It suffices to consider $z \notin S$ such that $h = \alpha - \operatorname{dist}(z, \Gamma) > 0$. Let y be the metric projection of z onto Γ, which is well-defined since $\operatorname{dist}(z, \Gamma) < \alpha \leq \rho(\Gamma)$, and let B be the open ball of radius α tangent to Γ at y and contained within S. It is clear that $\mu(B(z, \alpha) \cap S) \geq \mu(B(z, \alpha) \cap B)$. The intersection

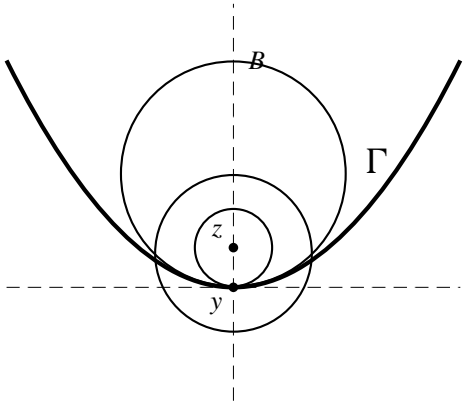

Fig. 2. Illustrates the proof of Lemma 1. The thick, parabolic line represents a portion of $\Gamma = \partial S$.

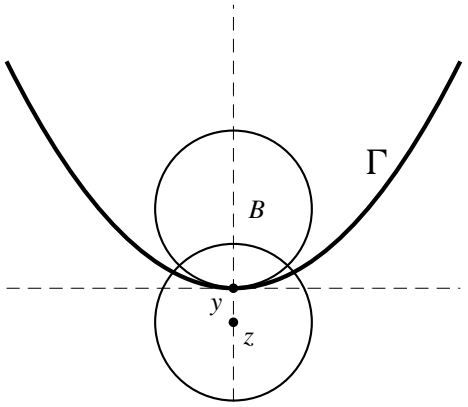

Fig. 3. Illustrates the proof of Lemma 2. The thick, parabolic line represents a portion of $\Gamma = \partial S$. The intersection of the two balls is the region of interest.

$B(z, \alpha) \cap B$ is the union of two spherical caps symmetric with respect to line joining the two points at the intersection $\partial B(z, \alpha) \cap \partial B$. See Figure 3 for an illustration. If C denotes one of them, we therefore have $\mu(B(z, \alpha) \cap B) = 2\mu(C)$, with C a spherical cap of radius α and height h. Its area is equal to

$$\mu(C) = 2\alpha^2 \int_0^{\operatorname{acos}(1-h/\alpha)} \sin^2(t)\, dt.$$

Using the bound $\sin(t) \geq 2t/\pi$, valid for $t \in [0, \pi/2]$, and the bound $\operatorname{acos}(1-t) \geq \sqrt{2t}$, valid for $t \in [0, 1]$, we obtain $2\mu(C) \geq Ah^{3/2}$ with $A = 32\sqrt{2\alpha}/(3\pi^2)$. □

For the following result, we use some heavy machinery from the seminal work of [12]. For a set $T \subset \mathbb{R}^2$, let $\mathcal{E}(T)$ denote its Euler–Poincaré characteristic, and recall that $\lambda(T)$ denotes its length.

Lemma 3. *Suppose $S \subset \mathbb{R}^2$ is compact, with both S and S^{c} satisfying the r-rolling condition. There are constants $A_0, A_1 > 0$ depending only on r and* $\operatorname{diam}(S)$ *such that $|\mathcal{E}(\partial S)| \leq A_0$ and $\lambda(\partial S) \leq A_1$.*

Proof. Let $\Gamma = \partial S$ and $d = \operatorname{diam}(S)$, and assume, without loss of generality, that $S \subset \bar{B}(0, d)$. For a given T such that $\rho(T) \geq r$, let Φ_k denote the kth curvature measure associated with T, $k \in \{0, 1, 2\}$, as defined in [12, Def. 5.7]. In [12, Rem. 5.10] we find that

$$\sup\big\{|\Phi_k|(T) : T \subset \bar{B}(0, d),\, \rho(T) \geq r\big\} < \infty, \tag{2}$$

where $|\Phi_k|(T)$ is the total variation of Φ_k over T. Now, by [12, Rem. 6.14], $\Phi_1(\Gamma)$ coincides with the one-dimensional Hausdorff measure, so that $|\Phi_1|(\Gamma) = \Phi_1(\Gamma) = \lambda(\Gamma)$. From this, we deduce the existence of A_1. By [12, Th. 5.19], $\Phi_0(\Gamma)$ coincides with $\mathcal{E}(\Gamma)$ and, by (2) for $k = 0$, we get that there is some constant A_0 such that $|\Phi_0(\Gamma)| \le |\Phi_0|(\Gamma) \le A_0$. □

We define an ε-net of a set S as any subset of points $x_1, \dots, x_m \in S$ such that $\|x_j - x_k\| \ge \varepsilon$ when $j \ne k$, and that, for any $x \in S$, $\|x - x_j\| < \varepsilon$ for some $j = 1, \dots, m$. Note that any bounded set $S \subset \mathbb{R}^2$ admits an ε-net of finite cardinality.

Lemma 4. *For any bounded $S \subset \mathbb{R}^2$, there is a constant A depending only on* $\operatorname{diam}(S)$ *such that, for any $0 < \varepsilon <$* $\operatorname{diam}(S)$*, any ε-net for S has cardinality bounded by $A\varepsilon^{-2}$. If, in addition, both S and S^{c} satisfy the r-rolling condition, then there is a constant A' depending only on r and* $\operatorname{diam}(S)$ *such that any ε-net for ∂S has cardinality bounded by $A'\varepsilon^{-1}$.*

Proof. Assume without loss of generality that $S \subset \bar{B}(0, d)$ where $d = \operatorname{diam}(S)$. Let $x_1, \dots, x_m$ be an ε-net of S. Since $B(x_j, \varepsilon/2) \cap B(x_k, \varepsilon/2) = \varnothing$ when $j \ne k$, we have

$$\pi d^2 \ge \sum_{j=1}^{m} \mu\big(\bar{B}(0, d) \cap B(x_j, \varepsilon/2)\big) \ge m\pi(\varepsilon/4)^2,$$

using Lemma 1 in the last inequality. We therefore have $m \le 16d^2/\varepsilon^2$. This proves the first part.

For the second part, let $\Gamma = \partial S$. It is enough to show the results for $\varepsilon \le 2r$. Note that $2r \le d$ by the r-rolling condition on S. Let $y_1, \dots, y_{m'}$ be an ε-net of Γ. Since $B(y_j, \varepsilon/2) \cap B(y_k, \varepsilon/2) = \varnothing$ when $j \ne k$, we have

$$m'\pi\left(\frac{\varepsilon}{2}\right)^2 = \mu\left(\bigcup_{j=1}^{m'} B\left(y_j, \frac{\varepsilon}{2}\right)\right) \le \mu\left(B\left(\Gamma, \frac{\varepsilon}{2}\right)\right). \tag{3}$$

By [12, Th. 5.6], we have

$$\mu\big(B(\Gamma, \varepsilon/2)\big) = \varepsilon\Phi_1(\Gamma) + \frac{\pi}{4}\varepsilon^2\Phi_0(\Gamma),$$

where $\Phi_1(\Gamma) = \lambda(\Gamma)$ [12, Rem. 6.14] and $\Phi_0(\Gamma)$ is the Euler–Poincaré characteristic of Γ [12, Th. 5.19]. By Lemma 3, there are positive constants A_0, A_1 depending only on r and d such that $\lambda(\Gamma) \le A_1$ and $|\Phi_0(\Gamma)| \le A_0$, yielding

$$\mu\big(B(\Gamma, \varepsilon/2)\big) \le A_1\varepsilon + A_0\frac{\pi}{4}\varepsilon^2 \le A_2\varepsilon,$$

where $A_2 = A_1 + A_0(\pi/4)d$, using the fact that $\varepsilon \le d$. Plugging this into (3), we conclude the proof of the second part. □

Next, we establish some basic properties of a line segment joining two points on a circle which barely intersects a set with smooth boundary.

Lemma 5. *Let $S \subset \mathbb{R}^2$ be such that both S and S^{c} satisfy the r-rolling condition. Fix $\alpha \in (0, r)$ and $0 < t \le \min\{\alpha, 2\alpha^2/r\}$. There is a constant $A > 0$ depending only on (r, α) such that, for any $z \notin S$ with $0 < \alpha - \operatorname{dist}(z, S) \le t/A$ and any $x_1, x_2 \in \partial B(z, \alpha) \cap S$, we have*

$$[x_1x_2] \subset B(\partial S, t), \tag{4}$$

$$\|x_1 - x_2\| \le \sqrt{t}, \tag{5}$$

$$\angle\big([x_1x_2], \partial S\big) \le \sqrt{t}. \tag{6}$$

(*The angle in (6) is well defined because of (4) and the bound $t \le \alpha < r$.*)

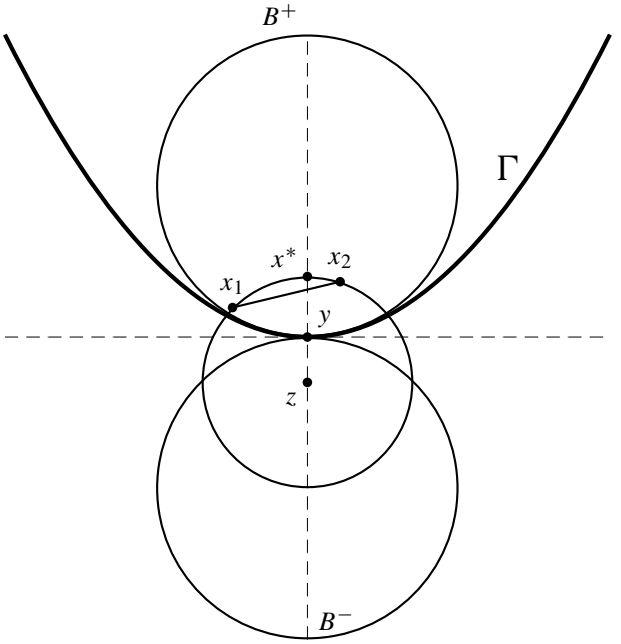

Fig. 4. Illustrates the proof of Lemma 5. The thick, parabolic line represents a portion of $\Gamma = \partial S$.

Proof. Let Γ be a shorthand for ∂S. Define $\delta = \alpha - \operatorname{dist}(z, S)$, and let e_1, e_2 denote the canonical basis vectors of $\mathbb{R}^2$. Since $p = \operatorname{dist}(z, \Gamma) = \operatorname{dist}(z, S) = \alpha - \delta < r$, $y = P_\Gamma(z)$ is well-defined. Without loss of generality, assume that y is the origin and that the tangent of Γ at y is the line spanned by e_1. Note that the line (yz) is perpendicular to the tangent at y, so that z is on the line defined by e_2 and without loss of generality we assume $z = -pe_2$. Let B be a shorthand for $B(z, \alpha)$ and let B^+ (resp. B^-) be the open ball centered at re_2 (resp. $-re_2$) with radius r. Since S and S^{c} satisfy the r-rolling condition, $B^+ \subset S$ and $B^- \subset S^{\mathsf{c}}$. Let $x^* = \delta e_2$. By construction x^* belongs to $(yz) \cap \partial B \cap B^+$. See Figure 4 for an illustration.

For any point $x \in B \cap S$,

$$\operatorname{dist}(x, \Gamma) = \operatorname{dist}\big(x, S^{\mathsf{c}}\big) \leq \operatorname{dist}\big(x, B^-\big) \leq \operatorname{dist}\big(x^*, B^-\big) = \delta.$$

Direct calculations show that $\partial B \cap \partial B^-$ is given by the points $\pm ae_1 - be_2$, where

$$\begin{cases} a^2 + (r-b)^2 = r^2, \\ a^2 + (p-b)^2 = \alpha^2. \end{cases}$$

So, using the fact that $p = \alpha - \delta$, we have

$$0 < b = \frac{\alpha^2 - p^2}{2(r-p)} = \frac{(\alpha - p)(\alpha + p)}{2(r-p)} \leq \frac{\alpha\delta}{r - \alpha}. \tag{7}$$

To prove (4), take $x \in [x_1 x_2]$. If $x \in S$, then $x \in B \cap S$ and we saw that $\operatorname{dist}(x, \Gamma) \leq \delta$. If $x \notin S$, let C be the closure of the intersection of B with the half-plane above the line $\mathbb{R}e_1 - be_2$. Since $B \cap C^{\mathsf{c}} \subset B^-$ and $B^- \cap S = \varnothing$, necessarily $x_1, x_2 \in C$, which in turn implies that $[x_1 x_2] \subset C$ since C is convex. In particular, $x \in C$, so that $\operatorname{dist}(x, [-ae_1, ae_1]) \leq \max\{b, \delta\}$. And since $\operatorname{dist}([-ae_1, ae_1], B^+) \leq b$ (by symmetry), we conclude with the triangle inequality that

$$\operatorname{dist}(x, \Gamma) = \operatorname{dist}(x, S) \leq \operatorname{dist}\big(x, B^+\big) \leq 2\max\{b, \delta\} \leq A_1 \delta, \tag{8}$$

for $A_1 = 2\max\{\alpha/(r-\alpha), 1\}$. This is valid for any $x \in [x_1 x_2]$, and proves (4) for any $A \geq A_1$.

To prove (5), we use the fact that $x_1, x_2 \in B \cap S \subset B \setminus B^-$, so that $\|x_1 - x_2\| \le \operatorname{diam}(B \setminus B^-)$, and $\operatorname{diam}(B \setminus B^-) = 2a$ when $b \le p$, which is the case since our assumptions that $\delta \le t/A$ and $t \le 2\alpha^2/r$ imply $\delta \le (r-\alpha)\alpha/r$, which forces $b \le p$ by (7). Continuing, we then have

$$a^2 = r^2 - (r-b)^2 = b(2r-b) \le 2br \le A_1 r \delta,$$

by (8). From this we get

$$\|x_1 - x_2\| \le \operatorname{diam}\bigl(B \setminus B^-\bigr) = 2a \le \sqrt{A_2 \delta}, \tag{9}$$

where $A_2 = 4A_1 r$. This proves (5) for any $A \ge \max\{A_1, A_2\}$.

We turn to proving (6). We first note that $\angle([x_1x_2], \Gamma)$ is well-defined. Indeed, by assumption $\delta \le t/A$, with $A \ge A_1 \ge 1$, and $t \le \alpha$, so that $B([x_1x_2], \Gamma) \le \alpha$ by (4), and we conclude with the fact that $\rho(\Gamma) \ge r > \alpha$. For any $x \in [x_1x_2]$ we can therefore compute the point $y' = P_\Gamma(x)$. Using the triangle inequality for angles, we have

$$\angle\bigl([x_1x_2], \vec{\Gamma}_{y'}\bigr) \le \angle\bigl([x_1x_2], \vec{\Gamma}_y\bigr) + \angle(\vec{\Gamma}_y, \vec{\Gamma}_{y'}) = \theta_1 + \theta_2. \tag{10}$$

We first bound θ_1. Direct trigonometric calculations show that

$$\sin(\theta_1) \le \frac{a}{\alpha} \le \frac{\sqrt{A_2\delta}}{2\alpha},$$

where the last inequality comes from (9). We use the fact that $\sin(\theta) \ge 2\theta/\pi$ for all $\theta \in [0, \pi/2]$, we get $\theta_1 \le A_3\sqrt{\delta}$, where $A_3 = \pi\sqrt{A_2}/(4\alpha)$. It remains to bound θ_2 in (10). We have $y = P_\Gamma(x^*)$ and $y' = P_\Gamma(x)$, and $\operatorname{dist}(x^*, \Gamma) = \delta < \alpha$ by construction, and also $\operatorname{dist}(x, \Gamma) \le t \le \alpha$ because of (4). Hence, by [12, Th. 4.8(8)], we get

$$\bigl\|y - y'\bigr\| \le \frac{r}{r-\alpha}\bigl\|x - x^*\bigr\|.$$

Using the fact that $x, x^* \in B \setminus B^-$, and then (9), we have $\|x - x^*\| \le \sqrt{A_2\delta}$. Now, if we denote by $\vec{\eta}_y$ and $\vec{\eta}_{y'}$ the outward pointing unit normal vector of Γ at y and y' respectively, [35, Th. 1] ensures that

$$\|\vec{\eta}_y - \vec{\eta}_{y'}\| \le \frac{1}{r}\bigl\|y - y'\bigr\|.$$

Since $\langle \vec{\eta}_y, \vec{\eta}_{y'} \rangle = \langle \vec{\Gamma}_y, \vec{\Gamma}_{y'} \rangle = \cos\theta_2$, we get

$$\|\vec{\eta}_y - \vec{\eta}_{y'}\| = \sqrt{2 - 2\cos\theta_2} = 2\sin(\theta_2/2).$$

We arrive at

$$\sin(\theta_2/2) \le \frac{\sqrt{A_2\delta}}{2(r-\alpha)}.$$

As before, this implies that $\theta_2 \le A_4\sqrt{\delta}$, where $A_4 = \pi\sqrt{A_2}/(4(r-\alpha))$. We conclude that

$$\angle\bigl([x_1x_2], \vec{\Gamma}_{y'}\bigr) \le (A_3 + A_4)\sqrt{\delta} = \sqrt{A_5\delta},$$

which proves (6) for any $A \ge \max\{A_1, A_2, A_5\}$. □

The following is a technical result involving two line segments, one on each of two intersecting circles of same radius, and a line passing through these line segments.

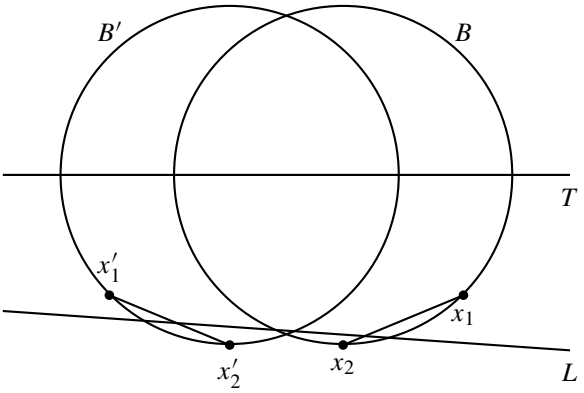

Fig. 5. Illustrating the proof of Lemma 6.

Lemma 6. *Let* $x_0, x_0' \in \mathbb{R}^2$ *such that* $0 < \|x_0 - x_0'\| < 2\alpha$, *and let* $x_1, x_2 \in \partial B(x_0, \alpha) \setminus B(x_0', \alpha)$ *and* $x_1', x_2' \in \partial B(x_0', \alpha) \setminus B(x_0, \alpha)$. *Let* L *be any line intersecting both* $[x_1x_2]$ *and* $[x_1'x_2']$. *Then there is a constant* $A > 0$ *depending only on* α *such that*

$$\max\{\angle((x_1x_2), L), \angle((x_1'x_2'), L)\} \le A\Big(\|x_0 - x_0'\| + \max_{i,j\in\{1,2\}} \|x_i - x_j'\|\Big).$$

Proof. Let B and B' be a shorthand for $B(x_0, \alpha)$ and $B(x_0', \alpha)$, respectively. Since the maximum above is bounded by $\pi/2$, it is enough to prove the inequality when

$$a = \|x_0 - x_0'\| + \max_{i,j\in\{1,2\}} \|x_i - x_j'\| < \alpha.$$

Let $T = (x_0x_0')$, and let H and $\tilde{H}$ denote the two half-spaces defined by T. Let t denote the intersection point $(\partial B \setminus B') \cap T$, and define t' analogously. Let m denote the intersection point $\partial B \cap \partial B' \cap H$, and define $\tilde{m}$ analogously. See Figure 5 for an illustration.

We claim that, when $a < \alpha$, the points x_1, x_2, x_1', x_2' are either all in H or all in $\tilde{H}$. Indeed, when x_i and x_j' are on opposite sides of T, then either $x_i \in \mathrm{arc}(mt)$ and $x_j' \in \mathrm{arc}(\tilde{m}t')$, or $x_i \in \mathrm{arc}(\tilde{m}t)$ and $x_j' \in \mathrm{arc}(mt')$. (For two points $s, t \in \partial B$, $\mathrm{arc}(st)$ denotes the shorter arc defined on ∂B by s and t.) The distance between a point in $\mathrm{arc}(mt)$ and a point in $\mathrm{arc}(\tilde{m}t')$ is not smaller than the minimum of $\|t - \tilde{m}\| \ge \sqrt{2}\alpha$ and $\|m - \tilde{m}\| \ge \sqrt{3}\alpha$, since $0 < \|x_0 - x_0'\| < \alpha$. Therefore, assume without loss of generality that $x_1, x_2, x_1', x_2' \in H$.

Let y be the point in $H \cap \partial B$ furthest from T, so the tangent of ∂B at y is parallel to T. Define y' similarly, with B' in place of B. We claim that $x_1, x_2 \in B(y, \sqrt{2}a)$ and $x_1', x_2' \in B(y', \sqrt{2}a)$. We prove this for x_1, without loss of generality, and consider the two possible cases:

- If $x_1 \in \mathrm{arc}(ym)$, then

$$\|y - x_1\| \le \|y - m\| \le \|y - y'\| = \|x_0 - x_0'\| \le a.$$

- If $x_1 \in \mathrm{arc}(ty)$, let us define $h = \|x_1 - y\|$, $d = \mathrm{dist}(x_1, (yx_0))$ and $z = P_{(yx_0)}(x_1)$. By the Pythagoras theorem,

$$d^2 + \|y - z\|^2 = h^2,$$
$$d^2 + \big(\alpha - \|y - z\|\big)^2 = \alpha^2.$$

From this we get $d^2 = h^2(1 - h^2/(4\alpha^2)) \ge h^2/2$, where the inequality is due to $s \le \mathrm{dist}(t, y) = \sqrt{2}\alpha$. But $d \le \max_{i,j\in\{1,2\}} \|x_i - x_j'\| \le a$. Hence, $h \le \sqrt{2}a$, as claimed.

By the fact that B is convex, the angle between (x_1x_2) and T is bounded from above by the maximum angle between T and the tangent of ∂B at any point in $\mathrm{arc}(x_1x_2)$. Moreover, by direct calculations, similar to that on

Lemma 5, for any point on $x \in \partial B$ such that $\|y - x\| \le \sqrt{2}\alpha$, the angle between T and the tangent of ∂B at x is bounded by $2\operatorname{asin}(\|y - x\|/(2\alpha)) \le \pi\|y - x\|/(2\alpha)$. Hence, by the fact that $x_1, x_2 \in B(y, \sqrt{2}a) \subset B(y, \sqrt{2}\alpha)$, we have

$$\angle\big((x_1x_2), T\big) \le \frac{\pi}{2\alpha} \max\big\{\|y - x_1\|, \|y - x_2\|\big\} \le \frac{\pi}{2\alpha}\sqrt{2}a = \frac{\pi a}{\sqrt{2}\alpha}.$$

Similarly,

$$\angle\big((x_1'x_2'), T\big) \le \frac{\pi a}{\sqrt{2}\alpha}.$$

By an analogous convexity argument, coupled with the fact that all the action is in half-space H, $\angle(L, T)$ is bounded from above by the maximum of any angle between T and a tangent of ∂B at any point in $\operatorname{arc}(x_1x_2)$, or any angle between T and a tangent of $\partial B'$ at any point in $\operatorname{arc}(x_1'x_2')$. Hence, as before, we get

$$\angle(L, T) \le \frac{\pi a}{\sqrt{2}\alpha}.$$

All the bounds combined, together with the triangle inequality, yield

$$\angle\big((x_1x_2), L\big) \le \angle\big((x_1x_2), T\big) + \angle(T, L) \le \frac{2\pi a}{\sqrt{2}\alpha},$$

and similarly for $(x_1'x_2')$. □

The following result is useful when comparing the length of two curves in terms of their Hausdorff distance and their deviation angle.

Lemma 7 (Th. 43 in [19]). *Let Γ be a compact curve in $\mathbb{R}^2$ such that $\rho(\Gamma) \ge r$ and let C be another curve in $\mathbb{R}^2$, differentiable almost everywhere, such that $C \subset B(\Gamma, r)$ and P_Γ is one-to-one on C. Then*

$$\frac{\cos\angle(C, \Gamma)}{1 + \frac{1}{r}\mathcal{H}(C, \Gamma)} \le \frac{\lambda(\Gamma)}{\lambda(C)} \le \frac{1}{1 - \frac{1}{r}\mathcal{H}(C, \Gamma)}.$$

Proof. The result is an immediate consequence of [19, Th. 43] and the fact that the reach bounds the radius of curvature from above [12, Lem. 4.17]. □

3. Some properties of α-edges

Our standing assumption in this section is the following:

($\star$) The data points $\mathcal{X}_n = \{X_1, \dots, X_n\}$ are independently sampled from a uniformly distribution with compact support $S \subset \mathbb{R}^2$ such that both S and S^{c} satisfy the r-rolling condition.

For any pair of distinct data points within distance 2α from each other, there are only two circles of radius α passing through them, symmetric with respect to the line joining the two points. In the special case of an α-edge, at least one of the two circles is empty of data points inside. The following result implies that, with probability tending to one, the center of such a circle lies outside of S.

Proposition 1. *Assume ($\star$). For any $\alpha > 0$, there is a constant $A > 0$ depending only on $(\alpha, r, \operatorname{diam}(S))$ such that, with probability at least $1 - Ae^{-n/A}$, there are no open balls of radius α with center in S empty of data points.*

Proof. Let $d = \mathrm{diam}(S)$ and assume without loss of generality that $S \subset \bar{B}(0,d)$. We will focus on the case $\alpha \leq r$. The case $\alpha > r$ can be analyzed similarly. By Lemma 1, if there is a ball of radius α with center in S empty of data points, then there is a ball of radius $\alpha/2$ included within S that is empty of data points. By Lemma 4, there is an $(\alpha/5)$-net of S, denoted $z_1, \ldots, z_m$, satisfying $m \leq A_1$, where A_1 depends only on d and α. By the triangle inequality any ball of radius $\alpha/2$ included within S contains a ball of the form $B(z_k, \alpha/5)$. Hence,

$$\begin{aligned}
\mathbb{P}\big(\exists z \in S : \mathcal{X}_n \cap B(z,\alpha) = \varnothing\big) &\leq \mathbb{P}\big(\exists k = 1, \ldots, m : \mathcal{X}_n \cap B(z_k, \alpha/5) = \varnothing\big) \\
&\leq \sum_{k=1}^{m} \mathbb{P}\big(\mathcal{X}_n \cap B(z_k, \alpha/5) = \varnothing\big) \\
&= \sum_{k=1}^{m} \left[1 - \frac{\mu(B(z_k, \alpha/5))}{\mu(S)}\right]^n \\
&\leq A_1 \big[1 - \big(\alpha/(5d)\big)^2\big]^n,
\end{aligned}$$

where in the second inequality we used the union bound and in the third we used the fact that $m \leq A_1$ and $S \subset \bar{B}(0,d)$. Therefore the result holds with $A = \max\{A_1, -1/\log[1 - (\alpha/(5d))^2]\}$. □

Remark 3. *We say that a data point is α-isolated if there are no other data points within distance 2α from it. Suppose that X_i is α-isolated so that $B(X_i, 2\alpha) \cap \mathcal{X}_n = \{X_i\}$. By the r-convexity of S^{c}, there is an open ball $B \subset S$ with radius α such that $X_i \in B$, which in particular satisfies $B \subset B(X_i, 2\alpha) \cap S$. Let $B' \subset B$ be an open ball of radius $\alpha/2$ such that $X_i \notin B'$. By construction, B' is included within S and is empty of data points. We conclude by Proposition 1 that, under ($\star$), with high probability, there are no α-isolated data points.*

Proposition 2. *Take $\alpha > 0$ and finite set of points $\mathcal{X} \subset \mathbb{R}^2$ such that there are no α-isolated points. Then the vertices of the α-shape of $\mathcal{X}$ and the vertices of the α-convex hull of $\mathcal{X}$ coincide.*

Proof. Let C and H denote the α-shape of $\mathcal{X}$ and the α-convex hull of $\mathcal{X}$, respectively. Note in particular that $H = \bigcap_{B \in \mathcal{B}} B^{\mathsf{c}}$ where $\mathcal{B}$ is the set of open balls of radius α that do not intersect $\mathcal{X}$. First, take $x \in \mathcal{X}$ such that $x \in \partial H$. By [9, Prop. 2], there is a open ball B of radius α such that $x \in \partial B$ but $B \cap \mathcal{X} = \varnothing$. Let B pivot on x. Since x is not α-isolated, the ball will eventually hit another data point, denoted x'. Then x and x' belong to the boundary of an open ball B' of radius α that does not contain any other data point by construction – for otherwise the ball would have hit that another data point before x' – so $[xx']$ forms an α-edge. This implies that x is a vertex of C. By definition of H above, $B' \subset H^{\mathsf{c}}$. Therefore $x \in \overline{B'} \subset \overline{H^{\mathsf{c}}}$, and since $x \in H$, we have $x \in H \cap \overline{H^{\mathsf{c}}} = \partial H$. □

The next proposition bounds the expected number of α-edges.

Proposition 3. *Assume ($\star$). For any $\alpha \in (0, r)$, there is a constant $A > 0$ depending only on $(\alpha, r, \mathrm{diam}(S))$ such that the expected number of α-edges is bounded by $An^{1/3}$.*

Proof. Let $N_\alpha^{\mathrm{shape}}$ and N_α^{hull} denote the number of vertices of the α-shape and α-convex hull, respectively, and let F denote the event that there are no α-isolated points. By Proposition 2, $N_\alpha^{\mathrm{shape}} = N_\alpha^{\mathrm{hull}}$ on F, so that $N_\alpha^{\mathrm{shape}} \leq N_\alpha^{\mathrm{hull}} \mathbb{1}_F + n \mathbb{1}_{F^{\mathsf{c}}}$, and consequently

$$\mathbb{E}\big(N_\alpha^{\mathrm{shape}}\big) \leq \mathbb{E}\big(N_\alpha^{\mathrm{hull}}\big) + n\,\mathbb{P}\big(F^{\mathsf{c}}\big).$$

On the one hand, $\mathbb{P}(F^{\mathsf{c}}) = 1 - \mathbb{P}(F) \leq A_1 e^{-n/A_1}$ for some constant A_1, by Proposition 1 and Remark 3. On the other hand, by [25, Th. 3], $\mathbb{E}(N_\alpha^{\mathrm{hull}}) \leq A_2 n^{1/3}$, for some constant A_2. From this, we conclude. □

Remark 4. *For $i < j$, let G_{ij} be the event that $[X_i X_j]$ forms an α-edge. By the fact that the points are iid, $\mathbb{P}(G_{ij})$ is independent of $i < j$. Hence, the expected number of α-edges $\binom{n}{2}\mathbb{P}(G_{ij})$ and Proposition 3 implies that $\mathbb{P}(G_{ij}) \leq An^{-5/3}$ for some constant A.*

The next result ensures that, with high probability, for each connected component of ∂S there is at least one α-edge within distance α.

Proposition 4. *Assume ($\star$). For any $\alpha \in (0, r)$, there is a constant $A > 0$ depending only on $(\alpha, r, \mathrm{diam}(S))$ such that, with probability at least $1 - Ae^{-n/A}$, for any connected component of ∂S, there is an α-edge with an endpoint within distance α of that component.*

Proof. Suppose that all the open balls of radius $\alpha/2$ centered at a point in S intersect the sample. By Proposition 1 this happens with probability at least $1 - Ae^{-n/A}$ for some constant $A > 0$. We saw in Remark 3 that this implies that there are no α-isolated data points. Let Γ_k be a connected component of $\Gamma = \partial S$. Fix $y \in \Gamma_k$ and let η denote the normal unit vector of Γ_k at y pointing away from S. For $s \geq 0$, define $y_s = y + s\eta$ and let $s^* = \inf\{s > 0 : B(y_s, \alpha) \cap \mathcal{X}_n = \varnothing\}$. Notice that $B(y_\alpha, \alpha) \subset S^{\mathsf{c}}$ and, therefore, it is empty of data points. Hence, $s^* < \alpha$. Moreover, we also have $s^* > 0$, since we are assuming that $B(y_0, \alpha/2)$ contains at least one data point (since $y_0 = y \in S$). By construction, there exists a data point $X_i \in \partial B(y_{s^*}, \alpha)$. Now, pivot the ball $B(y_{s^*}, \alpha)$ on X_i as we did in the proof of Proposition 2. Since X_i is not α-isolated, the ball will eventually hit another data point, denoted X_j, and $[X_i X_j]$ will form an α-edge. And, since $\|X_i - y_{s^*}\| = \alpha$ and $y_{s^*} \in S^{\mathsf{c}}$ (remember $0 < s^* < \alpha$), there is $z \in [X_i y_{s^*}]$ such that $z \in \Gamma$. We now use the fact that $B(y_{s^*}, \alpha) \cap \Gamma$ is contractible [12, Rem. 4.15], and since $B(y_s^*, \alpha) \cap \Gamma_k \neq \varnothing$, we must have $B(y_{s^*}, \alpha) \cap \Gamma = B(y_{s^*}, \alpha) \cap \Gamma_k$, which in turn implies that $z \in \Gamma_k$ and, therefore, $\mathrm{dist}(X_i, \Gamma_k) < \alpha$. □

Next, we prove some quantitative results about α-edges. In plain English, we show that, with probability tending to one, α-edges are near the boundary of S, have small length and their deviation angle with the boundary of S is small.

Proposition 5. *Assume ($\star$). For $i < j$, let G_{ij} denote the event that $[X_i X_j]$ is an α-edge, and for $t > 0$, let $H_{ij,t}$ denote the event that*

$$[X_i X_j] \subset B(\partial S, t), \qquad \|X_i - X_j\| \leq \sqrt{t} \quad \text{and} \quad \angle\big([X_i X_j], \partial S\big) \leq \sqrt{t}. \tag{11}$$

For any $\alpha \in (0, r)$, there is a constant $A > 0$ depending only on $(\alpha, r, \mathrm{diam}(S))$ such that, for any $0 < t \leq \min\{\alpha, 2\alpha^2/r\}$, $\mathbb{P}(G_{ij} \cap H_{ij,t}^{\mathsf{c}}) \leq Ae^{-nt^{3/2}/A}$.

Proof. Let Γ be a shorthand for ∂S. For any two distinct points $x, x' \in \mathbb{R}^2$ such that $\|x - x'\| < 2\alpha$, define

$$\zeta^{\pm}(x, x') = x + \alpha\, \Xi_{\pm\theta}\left(\frac{x' - x}{\|x' - x\|}\right),$$

where $\theta = \mathrm{acos}(\|x - x'\|/(2\alpha))$ and Ξ_θ denotes the rotation at angle θ. By construction, $x, x' \in \partial B(\zeta^{\pm}(x, x'), \alpha)$, and $\zeta^{\pm}(x, x')$ are the only two points with this property. Let $\zeta_{ij}^{\pm}$ be short for $\zeta^{\pm}(X_i, X_j)$, if $\|X_i - X_j\| < 2\alpha$, and $(\zeta_{ij}^{+}, \zeta_{ij}^{-}) = (X_i, X_j)$, otherwise.

Let E be the event that there are no open balls of radius α with center in S empty of data points. We studied this event in Proposition 1. With A_1 denoting the constant of Lemma 5, we have

$$H_{ij,t}^{\mathsf{c}} \cap G_{ij} \cap E \subset \big\{\exists \varepsilon \in \{-, +\} : \mathcal{X}_n \cap B\big(\zeta_{ij}^{\varepsilon}, \alpha\big) = \varnothing,\, \zeta_{ij}^{\varepsilon} \notin S \text{ and } \mathrm{dist}\big(\zeta_{ij}^{\varepsilon}, S\big) < \alpha - t/A_1\big\}.$$

Therefore, the union bound gives

$$\mathbb{P}\big(H_{ij,t}^{\mathsf{c}} \cap G_{ij} \cap E\big) \leq \sum_{\varepsilon = \pm} \mathbb{P}\big(\mathcal{X}_n \cap B\big(\zeta_{ij}^{\varepsilon}, \alpha\big) = \varnothing,\, \zeta_{ij}^{\varepsilon} \notin S \text{ and } \mathrm{dist}\big(\zeta_{ij}^{\varepsilon}, S\big) < \alpha - t/A_1\big).$$

With A_2 denoting the constant of Lemma 2, for any deterministic point $\zeta \notin S$ such that $\operatorname{dist}(\zeta, S) < \alpha - t/A_1$, we have

$$\begin{aligned}\mathbb{P}\big(\mathcal{X}_{n-2} \cap B(\zeta,\alpha) = \varnothing\big) &= \left(1 - \frac{\mu(S \cap B(\zeta,\alpha))}{\mu(S)}\right)^{n-2} \\ &\le \left(1 - \frac{A_2 t^{3/2}}{A_1^{3/2}\pi d^2}\right)^{n-2} \\ &\le A_3 e^{-nt^{3/2}/A_3},\end{aligned}$$

for some constant A_3 which depends only on α, r and $d := \operatorname{diam}(S)$. Hence, conditioning on (X_i, X_j), we have

$$\mathbb{P}\big(\mathcal{X}_n \cap B\big(\zeta_{ij}^{\varepsilon}, \alpha\big) = \varnothing, \zeta_{ij}^{\varepsilon} \notin S \text{ and } \operatorname{dist}\big(\zeta_{ij}^{\varepsilon}, S\big) < \alpha - t/A_1\big) \le A_3 e^{-nt^{3/2}/A_3}.$$

Together with Proposition 1, we arrive at

$$\mathbb{P}\big(H_{ij,t}^{\mathsf{c}} \cap G_{ij}\big) \le \mathbb{P}\big(H_{ij,t}^{\mathsf{c}} \cap G_{ij} \cap E\big) + \mathbb{P}\big(E^{\mathsf{c}}\big) \le A_4 e^{-nt^{3/2}/A_4},$$

for some constant A_4, again depending only on (α, r, d). □

The next two results combined imply that, with high probability, the α-edges form a simple polygon in one-to-one correspondence with ∂S. The first result shows that, with high probability, two distinct points in the union of all α-edges do not project on the same point on ∂S. We also show that α-edges are all one-sided in the sense that at least one of the two open balls of radius α that circumscribes an α-edge contains a data point.

Proposition 6. *Assume* ($\star$). *For any* $\alpha \in (0, r)$, *there is a constant* $A > 0$ *depending only on* $(\alpha, r, \operatorname{diam}(S))$ *such that, with probability at least* $1 - Ae^{-n/A}$: (i) *all* α*-edges are one-sided*; *and* (ii) *the metric projection onto* ∂S *is injective on the union of all* α*-edges*.

Proof. Let Γ be a shorthand for ∂S and $d = \operatorname{diam}(S)$. Assume there are no balls of radius α with center in S empty of data points and that, for t fixed (and chosen small enough in what follows), all the α-edges satisfy (11). Both events happen together with probability at least $1 - Ae^{-n/A}$, for some constant $A > 0$, by Propositions 1 and 5.

We first show that, if t is small enough, all α-edges are one-sided. Let $[x_1x_2]$ ($x_1 = X_{i_1}, x_2 = X_{i_2}$) be an arbitrary α-edge. Let $x_m = (x_1 + x_2)/2$ be the midpoint of that α-edge and $\rho = (\alpha^2 - \|x_1 - x_m\|^2)^{1/2}$. If there is a ball of radius α, B, such that $x_1, x_2 \in \partial B$, then the center of B is either $z_e = x_m + \rho u$ or $z_s = x_m - \rho u$, where u is the unit vector orthogonal to (x_1x_2) such that $\langle u, \eta\rangle > 0$, η being the outward pointing unit normal vector at $y_m = P_\Gamma(x_m)$, which is well-defined when $t < r$. Notice that the vector u is well defined when $\sqrt{t} < \pi/2$, since in that case (x_1x_2) is not orthogonal to Γ. We will prove that, for t even smaller, $z_s \in S$ and therefore $B(z_s, \alpha)$ is not empty of sample points. Define $c_s = y_m - \rho\eta$ and $c = y_m - r\eta$. By the r-rolling property, $B(c, r) \subset S$. By the triangle inequality and (11), we have

$$\|z_s - c_s\| \le \|x_m - y_m\| + \rho\|u - \eta\| \le t + \alpha\|u - \eta\|,$$

with, for t small enough,

$$\|u - \eta\|^2 = 2\big(1 - \langle u, \eta\rangle\big) \le 2\big(1 - \cos\angle\big([x_1x_2], \Gamma\big)\big) \le 2t,$$

using (11) (i.e., $\angle([x_1x_2], \Gamma) \le \sqrt{t}$) and the fact that $\cos(a) \ge 1 - a^2$ for any $a \in \mathbb{R}$. Using the triangle inequality and (11), again, we get

$$\|z_s - c\| \le \|z_s - c_s\| + \|c_s - c\| \le t + \alpha\sqrt{2t} + r - \sqrt{\alpha^2 - (\sqrt{t})^2} < r,$$

for t small enough, in which case $z_s \in B(c, r) \subset S$.

Now we prove that the metric projection onto Γ is injective on the union of all α-edges. Indeed, assume that this is not the case, so there are two distinct points belonging to some (necessarily distinct) α-edges, $x \in [X_{i_1} X_{i_2}]$ and $x' \in [X_{i'_1} X_{i'_2}]$, with the same metric projection onto Γ, denoted $y = P_\Gamma(x) = P_\Gamma(x')$. Let η be the outward pointing unit normal vector at y. For short, let $x_1 = X_{i_1}$, $x_2 = X_{i_2}$, $x'_1 = X_{i'_1}$, $x'_2 = X_{i'_2}$. By the triangle inequality and the fact that $\|x - x'\| \le \operatorname{dist}(x, \Gamma) + \operatorname{dist}(x', \Gamma)$, and then (11), we have

$$\max_{i,j \in \{1,2\}} \|x_i - x'_j\| \le \|x - x'\| + \|x_1 - x_2\| + \|x'_1 - x'_2\| \le 2t + 2\sqrt{t} \le 3\sqrt{t}, \tag{12}$$

when t is small enough. Also by the triangle inequality for angles and (11),

$$\angle\big((x_1 x_2), (x'_1 x'_2)\big) \le \angle\big((x_1 x_2), \vec{\Gamma}_y\big) + \angle\big(\vec{\Gamma}_y, (x'_1 x'_2)\big) \le 2\sqrt{t}. \tag{13}$$

Let B and B' denote the open balls of radius α circumscribing $[x_1 x_2]$ and $[x'_1 x'_2]$, respectively, and empty of data points. Since all α-edges are one-sided, these balls are uniquely defined. Also, define z'_e and z'_s analogously to z_e and z_s above, but based on x'_1 and x'_2, instead of x_1 and x_2. Using the same notation as above, we have $B = B(z_e, \alpha)$ and $B' = B(z'_e, \alpha)$ and

$$\|z_e - z'_e\| \le \|x_m - x'_m\| + \|\rho u - \rho' u'\|. \tag{14}$$

Reasoning as in (12) above, we have $\|x_m - x'_m\| \le 3\sqrt{t}$. Also,

$$\|\rho u - \rho' u'\|^2 = \rho^2 + (\rho')^2 - 2\rho\rho' \langle u, u' \rangle.$$

Using (11), $\rho^2 = \alpha^2 - \|x_1 - x_m\|^2 \ge \alpha^2 - t$ and, similarly, $(\rho')^2 \ge \alpha^2 - t$. Moreover, by (13) and using again the inequality $\cos(a) \ge 1 - a^2$ for any $a \in \mathbb{R}$, we get $\langle u, u' \rangle \ge 1 - 4t$. Hence,

$$\|\rho u - \rho' u'\|^2 \le 2\alpha^2 - 2(\alpha^2 - t)(1 - 4t) \le (8\alpha^2 + 2)t.$$

Hence, the bound in (14) leads to $\|z_e - z'_e\| \le 3\sqrt{t} + (8\alpha^2 + 2)^{1/2}\sqrt{t} = A_1\sqrt{t}$ when t is small enough, where A_1 is a constant. Combining this bound with that in (13), and applying Lemma 6, we obtain that

$$\max\big\{\angle\big((xx'), (x_1 x_2)\big), \angle\big((xx'), (x'_1 x'_2)\big)\big\} \le A_2\sqrt{t},$$

where A_2 is a constant. By the fact that (xx') is parallel to η [12, Th. 4.18(12)] and using (11), we also have

$$\max\big\{\angle\big((xx'), (x_1 x_2)\big), \angle\big((xx'), (x'_1 x'_2)\big)\big\} \ge \frac{\pi}{2} - \sqrt{t}.$$

We therefore have a contradiction when t is small enough that all the derivations above apply and, in addition, $\sqrt{t} < \pi/(2A_2 + 2)$. □

Remark 5. *Any one-sided α-edge shares each one of its endpoints with another α-edge. Indeed, suppose $[x_1 x_2]$ is an α-edge, so that there exists ζ such that $x_1, x_2 \in \partial B(\zeta, \alpha)$ and $\mathcal{X}_n \cap B(\zeta, \alpha) = \varnothing$. In that case, let $B(\zeta, \alpha)$ pivot on x_2, as we did in the proof of Proposition 4 away from x_1. Let x_3 denote the first data point that the ball hits. Then $[x_2 x_3]$ is an α-edge by construction. If x_2 is not shared with any other α-edge, then the ball pivots on x_2 away from x_1 until it touches x_1 from the other side. That (open) ball is empty of data points inside, and together with the ball we started with, makes $[x_1 x_2]$ two-sided.*

Proposition 7. *Assume ($\star$). For any $\alpha \in (0, r)$, there is a constant $A > 0$ depending only on $(\alpha, r, \operatorname{diam}(S))$ such that, with probability at least $1 - Ae^{-n/A}$, the union of all α-edges is in one-to-one correspondence with ∂S via the metric projection onto ∂S.*

Proof. Let Γ be a shorthand for ∂S and $d = \mathrm{diam}(S)$, and let C_α denote the union of all α-edges. Since Γ is a (compact) one dimensional manifold [36], it is well-known that each connected component of Γ is a closed curve homeomorphic to the unit circle, see [16, Thm. 5.27]. We prove that this is also the case for each connected component of C_α. We assume that the metric projection onto Γ, meaning P_Γ, is injective on C_α, that all α-edges are one-sided, that $C_\alpha \subset B(\Gamma, \alpha)$ – so that P_Γ is well-defined on C_α – and that $C_\alpha \cap B(\Gamma_k, \alpha) \neq \varnothing$ for any connected component Γ_k of Γ. This event happens with probability at least $1 - Ae^{-n/A}$ for some constant $A > 0$, by Propositions 5, 4 and 6. We prove that, under these circumstances, C_α is in one-to-one correspondence with Γ via P_Γ. Indeed, let Γ_k be a connected component of Γ. Let $[x_1x_2]$ be an α-edge such that $[x_1x_2] \cap B(\Gamma_k, \alpha) \neq \varnothing$. By assumption, there is a data point x_3 such that $[x_2x_3]$ is also an α-edge. Having constructed $[x_{a-1}x_a]$, let x_{a+1} be a data point such that $[x_ax_{a+1}]$ is an α-edge. Since $C_\alpha \subset B(\Gamma, \alpha) = \bigsqcup_\ell B(\Gamma_\ell, \alpha)$ – where the union is of disjoint sets by [12, Rem. 4.15, (1)] – and the polygon $\bigcup_a [x_ax_{a+1}]$ is connected, necessarily, $\bigcup_a [x_ax_{a+1}] \subset B(\Gamma_k, \alpha)$. Also, since the sequence $(x_a : a \geq 1)$ is made of finitely many data points, and $x_a \neq x_{a+1}$ for all a, there is $a, b \geq 1$ such that $x_a = x_{a+b+1}$, and we further may assume that $x_a, \ldots, x_{a+b}$ are all distinct. Therefore, by construction, $C = [x_ax_{a+1}] \cup \cdots \cup [x_{a+b-1}x_{a+b}]$ is a simple polygon made of α-edges such that $C \subset B(\Gamma_k, \alpha)$. In particular, the latter implies that $P_\Gamma(C) \subset \Gamma_k$, and since C is homeomorphic to the unit circle and P_Γ is continuous and injective on C, $P_\Gamma(C)$ is also homeomorphic to the unit circle. This forces $P_\Gamma(C) = \Gamma_k$, due to Γ_k being homeomorphic to the unit circle too. Since all this is true for any k, meaning any connected component of Γ, we conclude therefore that $P_\Gamma : C_\alpha \to \Gamma$ is not only injective, but also surjective. □

4. Proof of Theorem 1

We are now in a position to prove the main result, meaning, Theorem 1. Let Γ be a shorthand for ∂S and let C_α denote the union of all α-edges.

By Proposition 5 together with the union bound, and then Proposition 7, for any $0 < t \leq \min\{\alpha, 2\alpha^2/r\}$, with probability at least $1 - A_1 n^2 e^{-nt^{3/2}/A_1}$, for some constant $A_1 > 0$ depending only on $(\alpha, r, \mathrm{diam}(S))$, C_α is in one-to-one correspondence with Γ via the metric projection onto Γ, and satisfies $C_\alpha \subset B(\Gamma, t)$ and $\angle(C_\alpha, \Gamma) \leq \sqrt{t}$. Note that, because C_α and Γ are in one-to-one correspondence, $C_\alpha \subset B(\Gamma, t)$ implies that $\Gamma \subset B(C_\alpha, t)$, so that $\mathcal{H}(C_\alpha, \Gamma) \leq t$. We now apply Lemma 7, combined with the simple bounds $\cos a \geq 1 - a^2/2$, for $a > 0$, and $(1-a)^{-1} \leq 1 + 2a$, valid when $0 < a \leq 1/2$. Assuming $t \leq 1$, this yields

$$\frac{\lambda(C_\alpha)}{\lambda(\Gamma)} \leq \frac{1 + \frac{1}{r}\mathcal{H}(C_\alpha, \Gamma)}{\cos(\angle(C_\alpha, \Gamma))} \leq \frac{1 + t/r}{1 - t/2} \leq (1 + t/r)(1 + t) \leq 1 + (1 + 2/r)t$$

and

$$\frac{\lambda(C_\alpha)}{\lambda(\Gamma)} \geq 1 - \frac{1}{r}\mathcal{H}(C_\alpha, \Gamma) \geq 1 - t/r.$$

We get

$$\left| \frac{\lambda(C_\alpha)}{\lambda(\Gamma)} - 1 \right| \leq (1 + 2/r)t.$$

Hence, if $t \leq t_0 := \min\{\alpha/2, 2\alpha^2/r, 1\}$, we have

$$\mathbb{P}\left(\left| \frac{\lambda(C_\alpha)}{\lambda(\Gamma)} - 1 \right| > (1 + 2/r)t \right) \leq A_1 n^2 \exp\left(-nt^{3/2}/A_1\right).$$

Then a change of variable concludes the proof of Theorem 1.

5. Numerical experiments

In order to numerically check the conclusions of Theorem 1 we performed a small simulation study. For the set S we chose the corona $\{x \in \mathbb{R}^2 : 0.25 \leq \|x\| \leq 1\}$. In this case the value of r is equal to 0.25 (the radius of the hole)

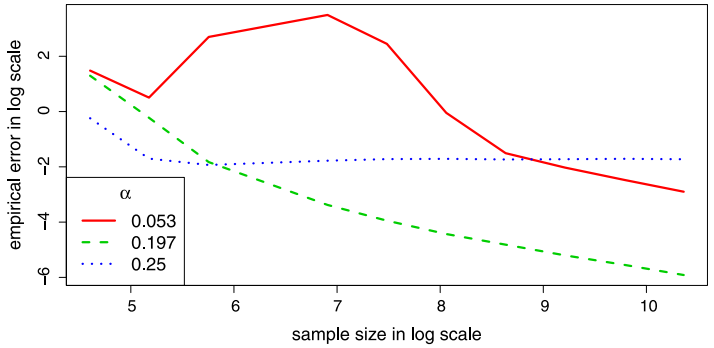

Fig. 6. Plot of error versus sample size, in log-log scale. The error corresponding to $\alpha = r = 0.25$ does not converge to zero. For values of $\alpha < r$, the plots show asymptotic slopes which are all very close to $-2/3$, as Theorem 1 predicts.

and $\lambda(\partial S) = 2\pi(0.25 + 1)$. The selected sample sizes were $n = 1000, 5000, 10{,}000, 30{,}000, 40{,}000, 50{,}000$. For each sample size n, we simulated $M = 1000$ samples from the uniform distribution on S and calculated the α-shape for each sample. The values of α were $0.05, 0.1, 0.15, 0.2, 0.24$, and the limit case $\alpha = r = 0.25$. Given n, α, and sample $m \in \{1, \dots, M\}$, we computed the sample α-shape, denoted $C_\alpha^{n,m}$, using the R-package `alphahull` of [24], and then its perimeter $\lambda(C_\alpha^{n,m})$. We estimated the expected error and bias by

$$e_\alpha(n) = \frac{1}{M}\sum_{m=1}^{M} \left|\lambda\big(C_\alpha^{n,m}\big) - \lambda(\partial S)\right| \quad \text{and} \quad b_\alpha(n) = \frac{1}{M}\sum_{m=1}^{M} \lambda\big(C_\alpha^{n,m}\big) - \lambda(\partial S),$$

respectively. Let $s_\alpha(n)$ denote the sample standard deviation of $\{\lambda(C_\alpha^{n,m}), m = 1, \dots, M\}$.

- Among the α's that we tried, the estimator performs best at $\alpha = 0.2$. It does not seem that, asymptotically, the best α converges to r. For instance, the ratio $e_{0.24}(n)/e_{0.2}(n)$ is around 6.7 for $n \geq 30{,}000$.
- Figure 6 shows the error versus sample size in log-log scale for $\alpha = 0.1, 0.2, 0.24, 0.25$. It can be seen that the error corresponding to $\alpha = r$ does no go to zero whereas $\alpha = 0.2$ always outperform the other considered values of α. The trend for large values of n is clearly linear and the slope is close to $-2/3$ as Theorem 1 predicts. This is particularly true when $\alpha = 0.2$ (our best choice), where fitting a line by least squares yields a slope of -0.67, with (Student) 95%-confidence interval of $(-0.73, -0.62)$, and an R-squared exceeding 0.99.
- For the limit case $\alpha = r$, the bias, $b_\alpha(n)$ does not go to zero as the sample size increases. The error $e_r(n)$ is approximately equal to 0.18; see Figure 6. This shows, from the numerical point of view, that the perimeter of the α-shape is not a consistent estimator of the $\lambda(\partial S)$ for $\alpha = r$. The main problem here is that the length of the α-edges does not go to zero, as Proposition 5 states for $\alpha < r$.
- The convergence rate of the standard deviation seems to be higher that $-2/3$. In fact, we have reasons to believe that the slope is of order $n^{-5/6}$. This is confirmed numerically. Indeed, if we fit a line to the log-log plot of $s_{0.2}(n)$, we get a slope with (Student) 95%-confidence interval of $(-0.86, -0.82)$. So, asymptotically, it seems that the error is dominated by the bias. This suggests that reducing the bias of the estimator could lead to improve the convergence rate of the method.
- The random variable $\lambda(C_\alpha)$ seems to be asymptotically normal. For the greatest considered $n = 50{,}000$, the sample $\{\lambda(C_\alpha^{n,m}), m = 1, \dots, M\}$ passes the Shapiro–Wilks normality test for several values of α. For instance, for $\alpha = 0.2$, we got a p-value of 0.82.

6. Discussion

We discuss a number of extensions and open problems.

Extensions. Our arguments extend more or less trivially to other sampling distributions. It is completely straightforward to see that Theorem 1 applies verbatim to a sampling distribution which has a density with respect to the

uniform distribution which is bounded away from zero near the boundary of S. A little less obvious is an extension to the case where this density converges to zero at some given rate near the boundary, which ends up impacting the rate of convergence of our estimator. In any case, our estimator remains consistent. The same results carry over to the case where ∂S has a finite number of 'kinks', i.e., points where the reach is infinite.

Choice of tuning parameter. The estimator depends on knowledge of r, or at least a lower bound on r, since any $\alpha \in (0, r)$ fixed appears to yield the convergence rate in $n^{-2/3}$. Choosing α automatically, therefore, requires an estimate on the size of r. This is done in recent work by [32]. Suppose we have an estimator $\hat{r}_n$ such that $r/2 \le \hat{r}_n \le 3r/2$ with high probability. We speculate that the convergence bound obtained in Theorem 1 with α chosen equal to $\hat{r}_n/4$ remains valid, albeit with a different multiplicative constant.

Finer asymptotics. [3] were able to compute the exact asymptotic expected value and variance of the perimeter of the convex hull of a sample, and also to show an asymptotic normal limit. An open problem would be to do the same here. Our numerical experiments lead us to speculate that our estimator is also normal in the large-sample limit.

Minimax rate. We conjecture that the rate that our estimator achieves, i.e., $\tilde{O}(n^{-2/3})$, is not minimax optimal, not even in the exponent. (The notation $\tilde{O}$ hides a poly-logarithmic factor.) Indeed, we learn in [15, Chapter 8] that for the problem of estimating the area under a Hölder-2 horizon in the context of binary images (see the Introduction), an estimator obtained from computing the area of an optimal set estimator (for the symmetric difference metric) only achieves the rate $\tilde{O}(n^{-2/3})$, while the optimal rate is $\tilde{O}(n^{-5/6})$. The same is true for the estimation of the perimeter as [14] prove. In fact, we speculate that the minimax rate they obtain for Hölder-2 periodic horizons, which is $\tilde{O}(n^{-5/6})$, is the same in our context.

Bias correction. [15] in the context of area estimation, and then [14] in the context of perimeter estimation, propose a plugin estimator followed by bias correction to achieve the minimax rate (within a polylog factor). The strategy is essentially the same and tailored to the setting of horizons in binary images. In particular, it relies heavily on function estimation is not easily adapted to our setting, which is rather based on set estimation.[1] Our estimator is also a form of plugin and we speculate that correcting for bias achieves the minimax rate. This is in line not only with the theory developed in [14,15] but also with our numerical experiments in Section 5. A bias correction based on sample splitting – as implemented in these previous works – seems viable, although we are still investigating this possibility.

Higher dimensions. Our setting is that of a set S in two dimensions. How about higher dimensions? The problem would be to estimate the $(d-1)$-volume of the boundary of a set $S \subset \mathbb{R}^d$, under the same conditions, and the estimator would be the $(d-1)$-volume of the α-shape of $\mathcal{X}_n$, which is the union of all the α-faces. We say that $X_{i_1}, \ldots, X_{i_d}$ form an α-face if they are affine-independent and there is an open ball B of radius α such that $X_{i_1}, \ldots, X_{i_d} \in \partial B$ and $B \cap \mathcal{X}_n = \varnothing$. Most of the auxiliary lemmas and propositions can be extended to the general framework. However, we have no idea how to extend Proposition 7.

The α-convex hull. Our results apply to the α-convex hull of the sample. This is because, with high probability, it shares the same vertices as the α-shape (by Proposition 2). When this is the case, the former is the union of arcs of radius α with base the α-edges. In particular, if an α-edge is of length ℓ, then the length of that arc is $2\alpha \sin^{-1}(\ell/(2\alpha)) = \ell + O(\ell^3)$. By Proposition 5 and an application of the union bound, the largest α-edge is of order $O_P(\log(n)/n)^{2/3}$. We conclude that the ratio between the perimeters of the α-convex hull and of the α-shape is of order $1 + O_P(\log(n)/n)^{4/3}$. We note, however, that the perimeter of the r-convex hull is consistent while the perimeter of the r-shape is not necessarily so. Our results require $\alpha < r$.

Acknowledgements

EAC was partially supported by grants from the US National Science Foundation (DMS-0915160 and DMS-1513465). ARC was partially supported by Projects MTM2008–03010 and MTM2013-41383-P from the Spanish Ministry of Science and Innovation, and by the IAP network StUDyS (Developing crucial Statistical methods for Understanding major complex Dynamic Systems in natural, biomedical and social sciences) of the Belgian Science Policy.

[1] Algorithmic complications that result from going from sets defined by horizons to more general sets is apparent in the treatment of [15].

References

[1] L. Ambrosio, A. Colesanti and E. Villa. Outer Minkowski content for some classes of closed sets. *Math. Ann.* **342** (4) (2008) 727–748. MR2443761

[2] G. Biau, B. Cadre and B. Pelletier. A graph-based estimator of the number of clusters. *ESAIM Probab. Stat.* **11** (2007) 272–280. MR2320821

[3] H. Bräker and T. Hsing. On the area and perimeter of a random convex hull in a bounded convex set. *Probab. Theory Related Fields* **111** (4) (1998) 517–550. MR1641826

[4] B. Cadre. Kernel estimation of density level sets. *J. Multivariate Anal.* **97** (4) (2006) 999–1023. MR2256570

[5] G. Carlsson. Topology and data. *Bull. Amer. Math. Soc. (N.S.)* **46** (2) (2009) 255–308. MR2476414

[6] F. Chazal and A. Lieutier. Weak feature size and persistant homology: Computing homology of solids in $\mathbb{R}^n$ from noisy data samples. In *Computational Geometry (SCG'05)* 255–262. ACM, New York, 2005. MR2460371

[7] A. Cuevas, R. Fraiman and A. Rodríguez-Casal. A nonparametric approach to the estimation of lengths and surface areas. *Ann. Statist.* **35** (3) (2007) 1031–1051. MR2341697

[8] A. Cuevas and R. Fraiman. Set estimation. In *New Perspectives in Stochastic Geometry* 374–397. Oxford Univ. Press, Oxford, 2010. MR2654684

[9] A. Cuevas, R. Fraiman and B. Pateiro-López. On statistical properties of sets fulfilling rolling-type conditions. *Adv. in Appl. Probab.* **44** (2) (2012) 311–329. MR2977397

[10] H. Edelsbrunner. Alpha shapes-a survey. In *Tessellations in the Sciences*, 2010.

[11] H. Edelsbrunner, D. G. Kirkpatrick and R. Seidel. On the shape of a set of points in the plane. *IEEE Trans. Inform. Theory* **29** (4) (1983) 551–559. MR0713690

[12] H. Federer. Curvature measures. *Trans. Amer. Math. Soc.* **93** (1959) 418–491. MR0110078

[13] R. Jiménez and J. E. Yukich. Nonparametric estimation of surface integrals. *Ann. Statist.* **39** (1) (2011) 232–260. MR2797845

[14] J.-C. Kim and A. Korostelëv. Estimation of smooth functionals in image models. *Math. Methods Statist.* **9** (2) (2000) 140–159. MR1780751

[15] A. P. Korostelëv and A. B. Tsybakov. *Minimax Theory of Image Reconstruction*. *Lecture Notes in Statistics* **82**. Springer-Verlag, New York, 1993. MR1226450

[16] J. M. Lee. *Introduction to Topological Manifolds*, 2nd edition. *Graduate Texts in Mathematics* **202**. Springer, New York, 2011. MR2766102

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

[18] E. Mammen and A. B. Tsybakov. Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist.* **23** (2) (1995) 502–524. MR1332579

[19] J.-M. Morvan. *Generalized Curvatures*. Springer, Berlin, 2008. MR2428231

[20] P. Niyogi, S. Smale and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.* **39** (1–3) (2008) 419–441. MR2383768

[21] B. Pateiro-Lopez. Set estimation under convexity type restrictions. Ph.D. thesis, Universidad de Santiago de Compostela, 2008.

[22] B. Pateiro-López and A. Rodríguez-Casal. Length and surface area estimation under smoothness restrictions. *Adv. in Appl. Probab.* **40** (2) (2008) 348–358. MR2431300

[23] B. Pateiro-López and A. Rodríguez-Casal. Surface area estimation under convexity type assumptions. *J. Nonparametr. Stat.* **21** (6) (2009) 729–741. MR2549435

[24] B. Pateiro-López and A. Rodrıguez-Casal. Generalizing the convex hull of a sample: The R package alphahull. *J. Stat. Softw.* **34** (5) (2010) 1–28.

[25] B. Pateiro-López and A. Rodríguez-Casal. Recovering the shape of a point cloud in the plane. *TEST* **22** (1) (2013) 19–45. MR3028242

[26] J. Perkal. Sur les ensembles ε-convexes. *Colloq. Math.* **4** (1956) 1–10. MR0077161

[27] W. Polonik. Measuring mass concentrations and estimating density contour clusters – an excess mass approach. *Ann. Statist.* **23** (3) (1995) 855–881. MR1345204

[28] M. Reitzner. Random polytopes. In *New Perspectives in Stochastic Geometry* 45–76. Oxford Univ. Press, Oxford, 2010. MR2654675

[29] A. Rényi and R. Sulanke. Über die konvexe Hülle von n zufällig gewählten Punkten. II. *Z. Wahrsch. Verw. Gebiete* **3** (1964) 138–147. MR0169139

[30] V. Robins. Towards computing homology from finite approximations. In *Proceedings of the 14th Summer Conference on General Topology and Its Applications (Brookville, NY, 1999)* 503–532, *Topology Proc.* **24**, 1999. MR1876386

[31] A. Rodríguez Casal. Set estimation under convexity type assumptions. *Ann. Inst. Henri Poincaré Probab. Stat.* **43** (6) (2007) 763–774. MR3252430

[32] A. Rodríguez-Casal and P. Saavedra-Nieves. A fully data-driven method for estimating the shape of a point cloud, 2014. Available at arXiv:1404.7397.

[33] A. Singh, C. Scott and R. Nowak. Adaptive Hausdorff estimation of density level sets. *Ann. Statist.* **37** (5B) (2009) 2760–2782. MR2541446

[34] A. B. Tsybakov. On nonparametric estimation of density level sets. *Ann. Statist.* **25** (3) (1997) 948–969. MR1447735

[35] G. Walther. Granulometric smoothing. *Ann. Statist.* **25** (6) (1997) 2273–2299. MR1604445

[36] G. Walther. On a generalization of Blaschke's rolling theorem and the smoothing of surfaces. *Math. Methods Appl. Sci.* **22** (4) (1999) 301–316. MR1671447

[37] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete Comput. Geom.* **33** (2) (2005) 249–274. MR2121296