BOUNDARY ESTIMATION FROM POINT CLOUDS: ALGORITHMS, GUARANTEES AND APPLICATIONS

JEFF CALDER, SANGMIN PARK, AND DEJAN SLEPČEV

ABSTRACT. We investigate identifying the boundary of a domain from sample points in the domain. We introduce new estimators for the normal vector to the boundary, distance of a point to the boundary, and a test for whether a point lies within a boundary strip. The estimators can be efficiently computed and are more accurate than the ones present in the literature. We provide rigorous error estimates for the estimators. Furthermore we use the detected boundary points to solve boundary-value problems for PDE on point clouds. We prove error estimates for the Laplace and eikonal equations on point clouds. Finally we provide a range of numerical experiments illustrating the performance of our boundary estimators, applications to PDE on point clouds, and tests on image data sets.

Keywords: boundary detection, distance to boundary, PDE on point clouds, meshfree methods
MSC (2020): 65N75, 62G20, 65N12, 65N15, 65D99

Notation.

Ω:	bounded domain in $\mathbb{R}^d$. We denote the volume of Ω by $\|\Omega\|$.
R:	lower bound for the reach of $\partial\Omega$.
d_Ω:	the distance function $d_\Omega = \operatorname{dist}(x, \partial\Omega) : \Omega \to \mathbb{R}_+$.
$\partial_a\Omega$:	boundary region $\partial_a\Omega := \{x \in \Omega : \operatorname{dist}(x, \partial\Omega) \leq a\}$ for $a > 0$.
ω_d:	volume of the unit ball in $\mathbb{R}^d$.
ρ:	probability density function $\rho : \Omega \to [\rho_{\min}, \rho_{\max}]$ where $0 < \rho_{\min} \leq \rho_{\max} < \infty$.
L:	Upper bound for the Lipschitz constant of ρ.
$\mathcal{X}$:	$= \{x^1, \cdots, x^n\}$: set of *i.i.d.* sample points drawn from density ρ.
n:	total number of sample points considered.
r:	neighborhood radius.
ε:	thickness of the boundary region we seek to identify.
ν:	inward unit normal vector to $\partial\Omega$, extended to $\partial_R\Omega$ by (1.1).
$\bar{v}_r$, $\bar{\nu}_r$:	population-based estimator of the normal vector, and its unit normalization, (1.3).
$\hat{v}_r$, $\hat{\nu}_r$:	first-order empirical estimator of the normal vector, and its unit normalization, (1.2).
$\hat{v}_r^2$, $\hat{\nu}_r^2$:	second-order empirical estimator of the normal vector, and its unit normalization, (1.5).
$\hat{d}_r^1(x^0), \hat{d}_r^2(x^0)$	first and second-order estimators of the distance to boundary of Ω, (1.12) and (1.17).
C_x, C_y, C_r:	dimensionless constants explicitly stated in Appendix D.

J. CALDER: SCHOOL OF MATHEMATICS, UNIVERSITY OF MINNESOTA, 127 VINCENT HALL, 206 CHURCH ST. S.E., MINNEAPOLIS, MN 55455

S. PARK, D. SLEPČEV: DEPARTMENT OF MATHEMATICAL SCIENCES, CARNEGIE MELLON UNIVERSITY, 5000 FORBES AVE., PITTSBURGH, PA 15213

E-mail addresses: jwcalder@umn.edu, sangminp@andrew.cmu.edu, slepcev@math.cmu.edu.

Date: June 28, 2022.

Acknowledgments. JC was supported by NSF grant DMS 1944925, the Alfred P. Sloan Foundation, and a McKnight Presidential Fellowship. SP and DS were supported by NSF grant DMS 1814991. The authors would like to thank Eddie Aamari for valuable comments. The authors are grateful to CNA of CMU, IMA of Univ. of Minnesota, and Simons Institute at UC Berkeley for hospitality.

1. INTRODUCTION

We focus on determining the boundary of a domain given sample points in the domain. By determining the boundary we mean identifying the points which lie within an $\varepsilon > 0$ neighborhood of the boundary; see Figure 1 for illustration. Our aim is develop an algorithm that is efficient to compute, accurate (so that the boundary strip can be identified even for $\varepsilon > 0$ which is smaller than the typical distance between neighboring sample points), and guarantees that we identify a high percentage of points that are within distance ε, while misidentifying as few points as possible that are at distance greater than 2ε as boundary points. Having such a set is sufficient for imposing boundary values for computing solutions of PDE on point clouds.

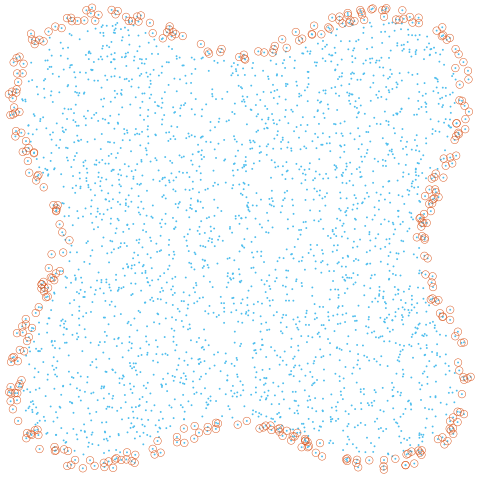


FIGURE 1. Boundary points identified using the proposed test (1.20).

Estimating the boundary of the support of an unknown distribution and the normal vector to the boundary are important and basic tasks with many applications. Identification of boundary points are crucial to solving partial differential equations (PDEs) on data clouds [24, 57, 69, 77], and have applications such as detecting anomalies in a point cloud [38] or assigning a notion of depth to each point (Section 6.3). Estimation of the distance of each point to the boundary is also used to improve the accuracy of kernel distance estimators near the boundary [11]. When the distribution is supported on a lower dimensional manifold, identifying points close to the boundary is important for estimation of the manifold itself. See [1] and references therein. While identifying the boundary of a point cloud is a basic problem, there are relatively few works that investigate the question in depth, see Section 1.5, and none satisfied the desired criteria above. In this work we introduce an approach that is simple, efficient, accurate and has the desired guarantees.

Our approach is to first estimate the approximate normal vector to the boundary using a kernel average. In fact, in Section 1.2 we develop two such estimators: a first-order estimator, given in (1.2), which estimates the normal vector to first-order with respect to the kernel bandwidth, and a second-order estimator, given in (1.5). We use these normal vector estimators in Section 1.3 to define estimators for the distance to the boundary, (1.12) and (1.17), which are, respectively, first and second-order accurate for points near the boundary. This allows us to define in Section 1.4 the statistical test for the boundary strip in (1.20). We implement our boundary test using MATLAB and Python, and make our code available on Github [1].

In this work we provide rigorous non-asymptotic error bounds of the first-order estimators and only asymptotic estimates for the second-order estimators. We focus on the first-order estimators in this paper, since nonasymptotic bounds for the second-order versions would be highly complicated, involving nontrivial

[1] https://github.com/sangmin-park0/BoundaryTest

dependence on a large number of parameters, including higher order derivatives of the density ρ and the boundary of Ω, which the first-order estimators do not require.

In Sections 1.2 and 1.3 we motivate and define the normal vector and distance-to-boundary estimators. The estimates on the normal vector estimators are provided in Section 2. Section 3 then establishes nonasymptotic estimates for the first-order test. In particular, the nonasymptotic error bounds on the distance estimator are provided in Theorem 3.3, and Corollary 3.5 establishes the nonasymptotic estimates for the first-order test. Asymptotic error estimates for the second-order distance test are given in Section 6.2.

In Section 5 we state our boundary tests in the form of a practical procedure, see Algorithm 1 and Algorithm 3. We conduct a number of experiments that illustrate the qualitative and quantitative performance of the algorithms. We also discuss the optimal selection of parameters, in particular the bandwidth of the kernel.

In Section 6 we turn to applications of the boundary test towards solving PDE boundary value problems using graph-based approximations, which is one of the problems that motivated our work. Since we estimate both the boundary points and the normal vector to the boundary, we are able to assign Dirichlet, Neumann, and Robin boundary conditions. In particular, we study the eikonal equation with Dirichlet boundary conditions and Poisson equations with Robin conditions on point clouds, and prove quantitative convergence rates to the solutions of the continuum PDEs. It is important to point out that not all methods for detecting boundary points will lead to convergent numerical approximations of PDEs. If too few points are identified, the boundary conditions may not be attained continuously as the mesh is refined [24]. Similar problems can occur if points far inside the interior of the domain are falsely identified as boundary points. The purpose of this section is to illustrate that our boundary detection method is compatible with setting boundary conditions for PDEs on point clouds. Our results cover only some preliminary examples, with much investigation left to future work.

Finally, in Sections 6.1.1 and 6.2.1 we implement numerical schemes for solving the eikonal and Robin equations on point clouds and conducted a number of experiments to both illustrate the solutions and numerically investigate the rate of convergence. Solving the eikonal equation enables us to estimate the distance to the boundary of any point in the dataset, which gives a notion of data depth on a point cloud. While our boundary test is not designed for working with manifolds in high dimensional spaces, Section 6.3 include experiments with notions of data depth based on the eikonal equation and Dirichlet eigenfunctions of the graph Laplacian on MNIST and FashionMNIST, using our boundary detection method to set the Dirichlet boundary conditions. The results are intriguing and agree with intuition; the boundary images are clearly outliers while the deepest images are good representatives of their class.

1.1. **Setting.** Consider a domain $\Omega \subset \mathbb{R}^d$ such that both Ω and $\mathbb{R}^d \setminus \Omega$ has reach at least $R > 0$, where reach is the maximal distance such that for all x with $\operatorname{dist}(x, \Omega) \leq R$ there exists a *unique* point $y \in \overline{\Omega}$ such that $|x - y| = \operatorname{dist}(x, \Omega)$. Denote by $\rho : \mathbb{R}^d \to [0, \infty)$ a probability density function, which we assume satisfies $\rho_{\min} \leq \rho \leq \rho_{\max}$ on Ω for some positive numbers $\rho_{\min} \leq \rho_{\max}$ and $\rho = 0$ outside of Ω. We assume that on Ω, the function ρ is Lipschitz continuous with Lipschitz constant L. Given a set of *i.i.d.* points $\mathcal{X}$ distributed according to ρ, our goal is to identify the points that are close to the boundary $\partial\Omega$ with high probability; namely, we aim to approximate the set

$$\partial_\varepsilon \Omega \cap \mathcal{X} = \{x \in \mathcal{X} : d_\Omega(x) \leq \varepsilon\}$$

of ε-boundary points, where $d_\Omega : \Omega \to \mathbb{R}_+$ is the distance function

$$d_\Omega(x) := \operatorname{dist}(x, \partial\Omega).$$

Our approach is as follows: we approximate inward normal vectors, use these to estimate the distance of each point to the boundary, and threshold the distance to obtain a boundary test. For $x \in \partial\Omega$ we denote by $\nu(x)$ the unit inward normal to $\partial\Omega$ at x. We extend the unit normal to a vector field on the set $\partial_R\Omega$ by setting

$$\nu(x) = \nu(x^*), \tag{1.1}$$

where $x^* \in \partial\Omega$ is the closest point to x on $\partial\Omega$. Note that x^* is uniquely defined on $\partial_R\Omega$. We can also equivalently set $\nu(x) = \nabla d_\Omega(x)$.

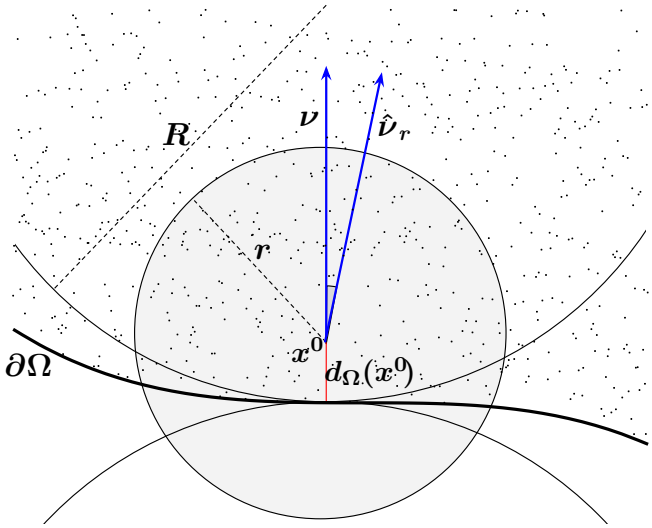


FIGURE 2. Illustration of the test setup: x^0 is the point tested.

1.2. **Estimation of the inward normal vector.** We now introduce the first and second-order estimator of $\nu(x^0)$. These estimators are accurate when x^0 is near the boundary. This is sufficient as our test does not require any accuracy of the estimated normal vectors in the interior. In fact, even in the continuum case the normal vectors are not necessarily well-defined for points outside of $\partial_R\Omega$.

First-order normal vector estimator. Let $r > 0$ and $\mathcal{X} = \{x^1, x^2, \cdots, x^n\}$ be the set of *i.i.d.* points distributed according to ρ. For each $x^0 \in \mathcal{X}$ we define the first-order normal vector estimator

$$\boxed{\hat{v}_r(x^0) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{B(x^0,r)}(x^i)(x^i - x^0), \qquad \hat{\nu}_r(x^0) = \frac{\hat{v}_r(x^0)}{|\hat{v}_r(x^0)|}}. \tag{1.2}$$

If $\hat{v}_r(x^0) = 0$ then we set $\hat{\nu}_r(x^0) = 0$. In this case, our test will identify x^0 as a boundary point. Note that this can happen with nonzero probability only when x^0 is an isolated point. We also define the corresponding population level estimator

$$\bar{v}_r(x^0) = \int_{\Omega\cap B(x^0,r)} (x - x^0)\rho(x)\,dx, \qquad \bar{\nu}_r(x^0) = \frac{\bar{v}_r(x^0)}{|\bar{v}_r(x^0)|}. \tag{1.3}$$

Theorem 2.6 establishes precise error bounds on the normal estimator, which in particular imply that

$$\mathbb{P}\left(|\hat{\nu}_r(x^0) - \nu(x^0)| > C\left(\frac{\log n}{n}\right)^{\frac{1}{d+2}}\right) \leq \frac{2d}{n^3} \tag{1.4}$$

for $r \sim (\log n/n)^{1/(d+2)}$, where $C > 0$ is a constant independent of n, with scaling $C \sim d^2$.

Second-order normal vector estimator. In addition to the assumptions for the first-order test, we now assume that ρ is a C^2 function and that the boundary of Ω is a C^3 manifold. To reduce the bias that arises

from the fact that ρ is not constant near x^0 we weight the points by the inverse of a kernel density estimate of ρ. For each $x^0 \in \mathcal{X}$ we define the second-order normal vector estimator

$$\boxed{\hat{v}_r^2(x^0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}_{B(x^0,r)}(x^i)}{\hat{\theta}(x^i)}(x^i - x^0), \qquad \hat{\nu}_r^2(x^0) = \frac{\hat{v}_r^2(x^0)}{|\hat{v}_r^2(x^0)|},} \tag{1.5}$$

where

$$\boxed{\hat{\theta}(x) = \frac{1}{\omega_d n}\left(\frac{2}{r}\right)^d \sum_{j=1}^{n} \mathbb{1}_{B(x,r/2)}(x^j).} \tag{1.6}$$

Similarly, we set $\hat{\nu}_r^2(x^0) = 0$ if $\hat{v}_r^2(x^0) = 0$. We note that the radius for estimating θ, namely $\frac{r}{2}$ is somewhat arbitrary. Using r instead of $\frac{r}{2}$ results in the error of the same order, however in practice using $r/2$ resulted in smaller error than using r.

At the population level our estimator takes the form

$$\bar{v}_r^2(x^0) = \int_{B(x^0,r)\cap\Omega} \frac{\rho(x)}{\theta(x)}(x - x^0)dx, \qquad \bar{\nu}_r^2(x^0) = \frac{\bar{v}_r^2(x^0)}{|\bar{v}_r^2(x^0)|}, \tag{1.7}$$

where

$$\theta(x) = \frac{2^d}{\omega_d r^d}\int_{B(x,r/2)\cap\Omega} \rho(z)dz. \tag{1.8}$$

In Section 2.1 we provide a proof that the error is indeed of size r^2 when $r \gtrsim (\log n/n)^{1/(d+4)}$, for n large enough. In contrast to our results for the first-order test (Theorem 2.6) we did not carry out a careful analysis of the second-order estimator to determine the exact constants appearing in the error bounds, and only determined the asymptotic scaling law. A more careful analysis of the second-order estimator is a nontrivial undertaking that we leave to future work.

We note that in addition to its use for distance estimation and the boundary test, the estimation of normal vectors is itself important to PDEs on graphs. It allows for the solution of PDEs on point cloud with not only Dirichlet boundary conditions but also Neumann, oblique, and Robin boundary conditions, which we study in Section 6.

1.3. **Estimation of the distance to the boundary.** The distance to $\partial\Omega$, $d_\Omega : \Omega \to \mathbb{R}$, is differentiable in $\partial_R\Omega$; see for example Lemma 2.21 in [12]. Furthermore, the gradient of the distance function conicides with the extension of the inward normal vector, that is, for $x \in \partial_R\Omega$ we have

$$\nabla d_\Omega(x) = \nu(x). \tag{1.9}$$

We exploit this relationship to approximate the distance function using the normal vectors near the boundary. First, we observe that d_Ω satisfies

$$d_\Omega(x) = \max_{y\in B(x,r)\cap\Omega} \big\{d_\Omega(x) - d_\Omega(y)\big\} \tag{1.10}$$

provided $B(x,r)\cap\partial\Omega$ is not empty. Indeed, the maximum is attained at $y \in \partial\Omega$ where $d_\Omega(y) = 0$. Suppose $d_\Omega \in C^2$ near the boundary. Then we can use the Taylor expansion

$$d_\Omega(y) = d_\Omega(x) + \nabla d_\Omega(x)\cdot(y - x) + O(r^2)$$

in (1.10), along with (1.9), to obtain

$$d_\Omega(x) = \max_{y\in B(x,r)\cap\Omega} \big\{\nu(x)\cdot(x - y)\big\} + O(r^2). \tag{1.11}$$

Replacing the true normal $\nu(x)$ in (1.11) with our first-order normal estimator $\hat{\nu}_r(x^0)$, and restricting the maximum to the point cloud, leads to our first-order estimator of the distance to the boundary.

First-order estimator for the distance to the boundary of Ω. Let $r > 0$ and $\mathcal{X} = \{x^1, x^2, \cdots, x^n\} \subset \Omega$. We define the first-order distance function estimator $\hat{d}_r^1 : \mathcal{X} \to \mathbb{R}$ by

$$\boxed{\hat{d}_r^1(x^0) = \max_{x^i \in B(x^0,r)\cap\mathcal{X}} (x^0 - x^i) \cdot \hat{\nu}_r(x^0).} \tag{1.12}$$

In Sections 2 and 3, we show that the assumption that $\partial\Omega$ has positive reach guarantees the error rate $O(r^2)$ of the first-order distance estimator near the boundary.

The associated population based estimator $\bar{d}_r^1$ defined by

$$\bar{d}_r^1(x^0) = \max_{x \in \overline{B(x^0,r)\cap\Omega}} (x^0 - x) \cdot \bar{\nu}_r(x^0). \tag{1.13}$$

Note that the population based estimator has a positive bias, meaning $d_\Omega(x^0) \leq \bar{d}_r(x^0)$. In Lemma 2.4 we obtain explicit bounds on the bias which establish that $\bar{d}_r(x^0) - d_\Omega(x^0) = O(r^2)$ as $r \to 0$. We combine this with variance bounds on $\bar{\nu}_r - \hat{\nu}_r$ established in Lemma 2.5 to show, in Theorem 3.3 that when $r \gtrsim (\log n/n)^{1/(d+2)}$ we have $|\hat{d}_r^1(x^0) - d_\Omega(x^0)| = O(r^2)$, with high probability, for x^0 sufficiently close to the boundary. The dependence of the error bounds on the parameters is explicitly stated.

Second-order estimator for the distance to the boundary of Ω. If the boundary of Ω is C^3, and thus d_Ω is C^3 within the a sufficiently small tubular neighborhood of the boundary [46], then we can use the second-order estimator $\hat{\nu}_r^n$ of the unit normal vector to obtain a second-order accurate estimator for the distance.

To derive a second-order distance function estimation near the boundary, we proceed from (1.10), as before, except now we use the higher order Taylor expansion

$$d_\Omega(y) = d_\Omega(x) + \nabla d_\Omega(x) \cdot (y - x) + \frac{1}{2}(y - x) \cdot \nabla^2 d_\Omega(x)(y - x) + O(r^3). \tag{1.14}$$

To handle the second-order terms, which cannot be easily estimated from the point cloud, we use the Taylor expansion

$$\nabla d_\Omega(y) = \nabla d_\Omega(x) + \nabla^2 d_\Omega(x)(y - x) + O(r^2).$$

Taking dot products of both sides with $y - x$ yields

$$(y - x) \cdot \nabla^2 d_\Omega(x)(y - x) = (\nabla d_\Omega(y) - \nabla d_\Omega(x)) \cdot (y - x) + O(r^3).$$

Combining this with the first expansion (1.14) yields

$$d_\Omega(y) = d_\Omega(x) + \frac{1}{2}(\nabla d_\Omega(x) + \nabla d_\Omega(y)) \cdot (y - x) + O(r^3).$$

Inserting this into (1.10) and using that $\nabla d_\Omega(x) = \nu(x)$ we obtain

$$d_\Omega(x) = \max_{y \in B(x,r)\cap\Omega} \left\{ (x - y) \cdot \frac{1}{2}(\nu(x) + \nu(y)) \right\} + O(r^3). \tag{1.15}$$

Hence, the second-order distance estimator simply involves averaging the normals at x and y. When discretizing to the point cloud, this yields the distance function estimation

$$\max_{x^i \in B(x^0,r)\cap X_n} (x^0 - x^i) \cdot \frac{1}{2}(\hat{\nu}_r(x^0) + \hat{\nu}_r(x^i)) \tag{1.16}$$

The above test is second-order accurate when applied to points that are closer to boundary than $\frac{r}{2}$, however at far away points, in particular those further than r, $\hat{\nu}_r^2(x^0)$ and $\hat{\nu}_r^2(x^i)$ are to large extent random and can be almost opposite to each other. This can lead to the distance being severely underestimated by the test above.

To avoid this problem, we define the second-order estimator with cutoff

$$\hat{d}^2_r(x^0) = \max_{x^i \in B(x^0,r)\cap\mathcal{X}} (x^0 - x^i)\cdot\left[\hat{\nu}^2_r(x^0) + \frac{\hat{\nu}^2_r(x^i) - \hat{\nu}^2_r(x^0)}{2}\mathbb{1}_{\mathbb{R}_+}(\hat{\nu}^2_r(x^i)\cdot\hat{\nu}^2_r(x^0))\right]. \tag{1.17}$$

The rationale for the particular cutoff function is as follows. We need a highly accurate estimate of the distance, for example, to determine the points in a boundary strip, only when $d_\Omega(x^0) < \frac{1}{2}r \ll R$. The point where the right-hand side of (1.16) is maximized is on the boundary. Thus the point where (1.17) is maximized, provided the normals are accurate, are close to the boundary. Points far away from the boundary can only maximize the right hand side if there is cancellation between the normal vector estimates. So we just need to discard the points where the normal is very poorly estimated, or rather, where the normal estimation is irrelevant as $B(x^0, r) \cap \partial\Omega = \varnothing$. Selecting the points where $\hat{\nu}_r(x^i)\cdot\hat{\nu}_r(x^0) > 0$ provides a convenient way to do so. We note that instead of discarding such points, we simply resort back to the first-order test, which provides another layer of robustness, in the case that the assumptions under which the second-order test was derived do not hold.

Henceforth, by the second-order estimator we refer to the estimator with cutoff (1.17), unless stated otherwise. In practice, we recommend the use of the second-order estimator. The estimates of Section 2.1 imply that for $r \gtrsim (\log n/n)^{1/(d+4)}$, the test (1.17) provides a second-order estimator of the normal vector. We note that unlike for the first-order test, our analysis for the second-order test is in the asymptotic regime, without precise estimates in the non-asymptotic regime. Developing the full error analysis of the second-order estimators remains a future task.

1.3.1. *Extension to manifolds.* We can generalize both the first and the second-order distance estimators to the case where ρ is supported on an m-dimensional manifold $\mathcal{M}$ with $m < d$. We simply replace the normal vectors by their projection onto the relevant tangent spaces approximated using PCA locally. Using such projections in boundary estimation for manifolds has been exploited in [1]. Let us denote by $\hat{T}^j$ the m-dimensional subspace spanned by the largest m eigenvectors of the sample covariance matrix from the observations $x^i - x^j$ for $x^i \in B(x^j, r)$, and Π^j the projection onto such a subspace. Thus we may define the first-order distance estimator in the manifold case as

$$\hat{d}^1_r(x^0) = \max_{x^i \in B(x^0,r)\cap\mathcal{X}} \Pi^0((x^0 - x^i))\cdot\hat{\nu}_r(x^0), \tag{1.18}$$

and the corresponding second-order estimator as

$$\hat{d}^2_{r,\mathcal{M}}(x^0) = \max_{x^i \in B(x^0,r)\cap\mathcal{X}} \left(\Pi^0(x^0 - x^i)\right)\cdot\left[\hat{\nu}^2_r(x^0) + \frac{\hat{\nu}^2_r(x^i) - \hat{\nu}^2_r(x^0)}{2}\mathbb{1}_{\mathbb{R}_+}(\Pi^0(\hat{\nu}^2_r(x^i))\cdot\Pi^0(\hat{\nu}^2_r(x^0)))\right]. \tag{1.19}$$

Note we have the equivalent distance estimators when we replace every vector w that appear in the above definitions with $\Pi^0 w$, which we avoid to keep notation simple. When $\mathcal{M}$ itself has positive reach, Π^j approximates the projection onto the true tangent plane at x^j with an error of $O(r)$ in the operator norm with high probability; when $\mathcal{M}$ is a C^3 manifold, the error is of order $O(r^2)$ (see Theorem 2 of [2]). In fact, this is also true in the presence of small additive noise. Further, Aamri and Levrard [2] suggest the same order of accuracy in the presence of small additive, possibly non-random noise of order $O(r^2)$. This means that the error rates for the estimated normal vector carry over, hence we can expect similar bounds on the distance estimators. Figure 9 shows experiments for 2 dimensional surfaces. However, the analysis required in this case is more intricate. One would need to bound the additional errors due to curvature and empirical estimation of the tangent plane. Thus we do not include the analysis in the current paper, and instead leave it to future work.

1.4. **The new boundary test.** Now we are ready to present our boundary test. Our aim is to create a test such that given $\varepsilon > 0$ small the test would recognize as boundary points all of the points within the distance ε from the true boundary of Ω and none of the points which are further than 2ε from $\partial\Omega$.

The boundary test we introduce depends on the empirical estimator of the distance to the boundary.

Boundary region test. Let $\mathcal{X} = \{x^1, x^2, \cdots, x^n\} \subset \Omega$ be an *i.i.d.* random sample of the density ρ. Let $\varepsilon, r > 0$ and $x^0 \in \mathcal{X}$. Given an empirical estimator of the distance to the boundary $\hat{d}_r$ we define the test $\widehat{T}_{\varepsilon,r} : \mathcal{X} \to \{0, 1\}$ by

$$\boxed{\widehat{T}_{\varepsilon,r}(x^0) = \begin{cases} 1 & \text{if } \hat{d}_r(x^0) < \frac{3\varepsilon}{2} \\ 0 & \text{otherwise.} \end{cases}} \tag{1.20}$$

We denote by $\widehat{T}^1_{\varepsilon,r}$ the estimator that uses the first-order estimator for the distance $\hat{d}^1_r(x^0)$ defined in (1.12) and by $\widehat{T}^2_{\varepsilon,r}$ the estimator that uses the second-order estimator for the distance $\hat{d}^2_r(x^0)$ defined in (1.17).

Our theoretical guarantees focus on $\widehat{T}^1_{\varepsilon,r}$. In particular we show that $\widehat{T}^1_{\varepsilon,r}$ identifies the *ε-boundary points* with high probability, even when ε is much smaller than the typical distance between nearby points. In particular Theorem 3.3 shows that, for $\varepsilon \gtrsim (\log n/n)^{2/(d+2)}$, under appropriate assumptions,

$$\mathbb{P}(\widehat{T}^1_{\varepsilon,r}(x^0) = 0 \mid d_\Omega(x^0) \le \varepsilon) + \mathbb{P}(\widehat{T}^1_{\varepsilon,r}(x^0) = 1 \mid d_\Omega(x^0) \ge 2\varepsilon) \le (2d+1)n^{-3}. \tag{1.21}$$

The assumptions we make on the geometric parameters are as follows.

Assumption 1.1. $\frac{\varepsilon}{r} \le \frac{1}{3\sqrt{d}}$.

Assumption 1.2. $r^2 \le R\varepsilon$.

Assumption 1.1 assures that r is sufficiently large so that distances to boundary of size ε can be detected. In particular it ensures that there are points $x \in B(x^0, r)$ for which $\big(x - x^0\big) \cdot \hat{\nu}_r(x^0) < -\frac{3\varepsilon}{2}$. Assumptions 1.1 and 1.2 together imply

$$\left(\frac{\varepsilon}{r} - \frac{r}{R}\right)^2 \le \frac{1}{d+1}, \tag{1.22}$$

which bounds the rate of growth of constant C in Lemma 2.2 in d. Assumption 1.2 is needed in Lemma 3.1 to ensure that $\hat{d}^1_r(x^0)$ does not underestimate the distance for positively curved domains. Assumptions 1.1 and 1.2 imply

$$r \le R\frac{\varepsilon}{r} \le \frac{R}{3\sqrt{d}}. \tag{1.23}$$

This guarantees that at least one third of $B(x^0, r)$ is in Ω, which is crucial for establishing the lower bound in Lemma 2.1. Finally, $r \le \frac{R}{2}$ follows easily from the assumptions. This implies the estimate

$$R - \sqrt{R^2 - x^2} \le \frac{x^2}{R} \quad \text{for} \quad |x| \le r,$$

which is used in the proof of Lemmas 2.1 and 2.2.

Now we summarize our result on the accuracy of the boundary test. Corollary 3.8 states that under suitable conditions $\partial_{\varepsilon,r}\mathcal{X} = \{x \in \mathcal{X} : \widehat{T}^1_{\varepsilon,r}(x) = 1\}$ satisfies

$$\partial_\varepsilon\Omega \subset \partial_{\varepsilon,r}\mathcal{X} \subset \partial_{2\varepsilon}\Omega$$

with probability at least $1 - 2dn^{-3}$, if

$$\varepsilon \ge C\left(\frac{\log n}{n}\right)^{\frac{2}{d+2}} \tag{1.24}$$

for some constant $C = C(d, R, L, \rho_{\min}, \rho_{\max})$. For our second-order boundary test, our analysis in the asymptotic regime suggest that we can identify ε-boundary points with $\varepsilon \gtrsim (\log n/n)^{3/(d+4)}$ with high probability. Please see Sections 2.1 and 4 for precise statements.

We can compare the above result with that from Cuevas and Rodríguez-Casal [36], which gives the best available theroetical guarantee the authors are aware of. Theorem 4 of [36] states that with probability one, the estimated set of boundary points $\partial\Omega_n$ based on the Devroye-Wise estimator [38] satisfies

$$d_H(\partial\Omega_n, \partial\Omega) \leq (2s^{-1}\omega_d^{-1})^{\frac{1}{d}} \left(\frac{\log n}{n}\right)^{\frac{1}{d}} \quad \text{eventually.} \tag{1.25}$$

Here, s denotes the *standardness constant*, which in our case is at least $\frac{1}{3}$. Further, Theorem 5 of [36] states that the rate in n in (1.25) is optimal for the Devroye-Wise estimator. Let us temporarily denote the right hand side of (1.25) by ε_n. Note that this allows identifying all points within ε_n of the boundary and none farther than 2ε via taking the points within ε_n of $\partial\Omega_n$.

Note that our test satisfies, under suitable choices of ε, r,

$$d_H(\partial_{\varepsilon,r}\mathcal{X}, \partial\Omega) \leq 2\varepsilon = O\left(\frac{\log n}{n}\right)^{\frac{2}{d+2}} \text{ with probability at least } 1 - 2dn^{-3},$$

provided we choose ε at the lower bound in (1.24). Thus for $d \geq 3$ our rate in n compares favorably to the optimal rate of the Devroye-Wise estimator (1.25). However, the constant in (1.24) is of order $C \sim O(d^{5/2})$, while the constant $(2s^{-1}\omega_d^{-1})^{1/d}$ in (1.25) is of order $O(d^{1/2})$. Details on the dependence of the constants on d can be found in Remark 3.4.

Another notable difference is that identifying the boundary points through [36] does not seem computationally tractable in higher dimensions. The points corresponding x^i whose balls $B(x^i, r)$ contribute to the boundary correspond exactly to points on the boundary of the α-shape [41] of $\mathcal{X}$. However, computing this involves Delaunay triangulation and may be difficult in dimensions higher than 3. See Section 1.5 for more details.

In contrast, our proposed boundary test is easy to implement and computationally efficient, as can be seen in Algorithms 1 and 3. The range search task of identifying $B(x^0, r) \cap \mathcal{X}$ for each $x^0 \in \mathcal{X}$ is the computational bottleneck of our test. This is computationally equivalent to performing a k-nearest neighbor search for each point in $\mathcal{X}$ (all-kNN) for suitable k. Empirically, k-nearest neighbor search (kNN) can be done in almost linear time with high accuracy [10, 39]. For further details, we refer the reader to the discussions in Section 5.

Finally, our test does not require the knowledge of the intrinsic dimension of $\operatorname{supp}\rho$. For instance, if Ω is an m-dimensional disc, the proposed boundary test will perform exactly the same when Ω is embedded in $\mathbb{R}^d$ for any $d \geq m$, besides the slightly higher computational cost of performing range search or kNN in higher dimensions. This is because our test is based on estimation of the distance d_Ω, which is intrinsic.

1.5. **Related works.** One of most studied approaches to boundary and support estimation is via the Devroye-Wise estimator, which approximates the support of ρ by a union of balls:

$$\Omega_n := \bigcup_{i=1}^{n} B\left(x^i, r_n\right). \tag{1.26}$$

Devroye and Wise [38] establish the convergence of Ω_n to $\Omega := \operatorname{supp}\rho$ as $n \to \infty$ and $r_n \to 0$, at a suitable rate, in the following sense: $\rho(\Omega \Delta \Omega_n) \to 0$ in probability if $r_n \gg n^{-1/d}$, while $r_n \gg (\log n/n)^{1/d}$ implies almost sure convergence.

Cuevas and Rodriguez-Casal, [36], established that, under certain smoothness assumptions, the Hausdorff distances $d_H(\Omega_n, \Omega), d_H(\partial\Omega_n, \partial\Omega) \sim (\log n/n)^{1/d}$, and that the rate is *optimal*. Furthermore, it is possible to compute the points x^i contributing to the boundary $\partial\Omega_n$ using α-shapes, introduced in [41]. However, α-shapes are a union of a certain subset of simplicies of the Delaunay triangulation. This poses challenges as the Delaunay triangulation in $d > 3$ dimensions is itself not an easy computational problem,

as the number of simplices can be large, up to $O(n^{\lceil d/2 \rceil})$ [59]. Thus, while efficient $O(n^2)$ algorithms are established for $d \leq 3$ [42], less is known for higher dimensions.

We also note that the Devroye-Wise boundary estimators have been used to estimate the Minkowski content of the boundary of S, which for sufficiently regular sets approximates the surface area ($(d-1)$-dimensional Hausdorff measure). This is shown to be L_2-consistent for general dimensions in [34] and convergent at $O(n^{-1/(2d)})$ for $d = 2, 3$ in [35].

Casal [67] defines an estimator called r-convex hull, based on the Minkowski sum and differences of sets and closely related to α-shapes, to approximate the support Ω with improved rate of $(\log n/n)^{2/(d+1)}$ in the Hausdorff distance with high probability.

We note that the while the works of Devroye-Wise and Casal propose different estimators for the boundary of the set, the data points x^i which are identified as being near the boundary are the same for both estimators, see Section 5.1 for explanation and Figure 8 for illustration.

Another family of approaches are associated with the kernel density estimators (KDE). Estimating the density level set via the kernel density estimator is well-studied [29] [65]. Cuevas and Fraiman [33] approximate the support by the super-level sets $\{\hat{f} > \alpha_n\}$ of the KDE $\hat{f}$, where tuning parameter $\alpha_n \to 0$ as $n \to \infty$, and establish d_H almost at the aforementioned optimal rate.

On the other hand, Berry and Sauer [11] approximates the distance d_Ω of points to the boundary of the manifold to improve accuracy of KDE near the boundary. To do so, they use the graph Laplacian to estimate the normal vectors, and compute d_Ω by solving an expression it satisfies in relation to the expectation of the said graph Laplacian.

For self-similar but possibly non-smooth $\partial\Omega$, such as the von Koch snowflake, Lachièze-Rey and Vega [52] use Voronoi cells to define an estimator that converges to Ω at the optimal rate in d_H when ρ is uniform.

Several further works, [1, 3, 30, 66, 74], have focused on identifying the boundary when ρ is supported on a lower dimensional manifold $\mathcal{M}$. Aamari, Aaron, and Levrard [1] generalize the result of Casal [67] to the manifold setting. They project the relevant geometric quantities onto the approximate tangent space estimated using principal component analysis (PCA) to identify the set $\mathcal{Y} \subset \mathcal{X}$ of points such that with high probability, for all $y^i \in \mathcal{Y}$ we have $d_H(y^i, \partial\mathcal{M}) \lesssim (\log n/n)^{2/(d+1)}$. Based on $\mathcal{Y}$, they use the weighted Tangential Delaunay Complex to provide an estimator approximating $\partial\mathcal{M}$ with rate $(\log n/n)^{2/(d+1)}$ in the Hausdorff distance with high probability. Further, they establish that this rate is minimax over the class of convex submanifolds (i.e. those diffeomorphic to a convex subset of $\mathbb{R}^d$), thus showing not only that their upper bound is tight, but also that estimation of boundary under the assumption of positive reach is not more difficult than that in the convex case.

Our first-order test identifies the set of boundary points such that with high probability each point is at most $(\log n/n)^{2/(d+2)}$. While our theoretical results are established for flat domains, we believe the same rate would apply to the generalized first-order estimator (1.18) in the manifold case. Through the same boundary reconstruction process as stated in [1], we may construct boundary estimators with the same rate, which is slightly slower than the minimax rate proven by [1]. However, we note that our test identifies w.h.p. *all points* within such tubular neighborhood of the boundary, which is stronger than obtaining the same bound in the Hausdorff distance, and is important for application to PDEs on graphs.

It is also interesting to note that the asymptotic error rate for our second-order test (1.20) based on distance estimator (1.17) in the Euclidean case is $(\log n/n)^{3/(d+4)}$, see Sections 2.1 and 4. This estimator however requires that manifolds are of class C^3 and that ρ is C^2, while the rates in [1] hold for manifolds which are merely C^2 and bounded densities. Determining minimax rates for estimators for C^3, and more regular manifolds and densities, remains an open problem.

Aaron and Cholaquidis [3] devise a statistical test to determine whether a random sample supported on a manifold has a boundary, along with heuristics to identify some of the points closer to the boundary. While their test uses k-nearest neighbor search instead of range search, the suggested test statistic for each point x^0 is similar to the size of the projection of $\hat{v}_r(x^0)$ onto the approximate tangent space at x^0. Thus, loosely speaking, this statistic exploits that the normal vector is of order $O(r)$ near the boundary, while $O(r^2)$ in the

interior. We note that this approaches only use the size of the estimated normal, while we utilize the normal vector itself.

Wu and Wu [74] use the behavior of the locally-linear embedding (LLE) near the boundary to identify boundary points. Interestingly, their test statistic is a quadratic function of a kNN-analogue of our normal vector $\hat{v}_r$, where the coefficients take into account the curvature of $\partial\Omega$ and density fluctuations. Further, they provide theoretical guarantees for their test statistic (see Proposition 5.1 of [74]).

A couple other methods try to use the normal vectors, but approximated in a different way. BORDER algorithm [30] uses that, given a fixed $k \in \mathbb{N}$ and sufficiently many points, the number of points of which x^0 is a k-neighbor of will be roughly half when x^0 is near the boundary, compared to that when x^0 is in the interior. BRIM algorithm introduced in [66], exploits the fact that given a suitable approximation of the inward normal at x^0, say $\nu(x^0)$, the number of points x^i such that $(x^i - x^0) \cdot \nu(x^0)$ is positive is greater than the number of points for which the inner product is negative, when x^0 is near the boundary. BRIM approximates the inward normal by identifying the point $y \in B(x^0, r) \cap \mathcal{X}$ such that $|B(y, r) \cap \mathcal{X}|$ is largest, then using $y - x^0$ as the estimator. However, for both approaches, such difference is of the same order as the statistic, which is weaker than the dichotomy used in [74]. Moreover, none of the approaches above use the normal vector to measure the distance to the boundary, which is one of the key elements for the improved accuracy.

Our convergence proofs for the solutions of PDEs on point clouds in Section 6 utilize the maximum principle, building upon previous related works in the field [15, 17, 44, 50, 80]. We also expect that recent advances in the studies of PDEs on point clouds [22, 23, 49] can also be applied in this setting, to obtain, for example, spectral convergence for the Dirichlet graph Laplacian. There are many methods in the numerical analysis literature for solving PDEs on unstructured meshes or point clouds. Methods with rigorous convergence results include the wide stencil schemes for Hamilton-Jacobi equations and elliptic PDEs [62], which were originally defined on regular grids and have subsequently been extended to unstructured point clouds [43, 47], and the point integral method [55]. Other works without convergence guarantees include upwind schemes for Hamilton-Jacobi equations on unstructured meshes [68], mesh-free generalized finite difference methods [71, 72], least squares manifold approximation methods [56, 75, 78], the local mesh method [53], radial basis function methods [45, 48, 63, 64], and a recent approach using graph Laplacians and deep learning [57]. A general survey of meshfree methods in PDEs is given in [28].

Regarding data depth, the ordering of multivariate data is an old problem in statistics [6, 58]. The goal is generally to extend robust statistical notions, like quantiles and the median, to multivariate data. For point clouds, there are notions of depth like the Tukey halfspace depth [76], which has been extended to graphs [70] and metric spaces [27], and the Monge-Kantorovich depth [31]. There are also notions of depth for curves [37] It was recently shown in [61] that the Tukey depth satisfies a non-standard eikonal equation in the viscosity sense, at the population level. To the best of our knowledge, the eikonal equation on a graph has not been used for data depth previously. Two forthcoming papers will study the graph eikonal depth in more detail [21, 60]. Other examples of connections between data depth and PDEs include convex hull peeling [25], non-dominated sorting [20], and Pareto envelope peeling [13].

Outline. The remainder of this paper is organized as follows. In Section 2 we establish preliminary estimates and error estimates on normal vectors estimators that will be useful in proving the main results, which are presented in Sections 3 and 4. Section 3 rigorously establishes nonasymptotic error bounds for the first-order test, which is the theoretical basis for applications to PDEs on graphs presented later in the paper. Section 4, under some additional regularity assumptions, establishes asymptotic error bounds for the second-order test, which we recommend for practical use. Then we present the algorithm and discuss the computational aspects of the boundary test in Section 5. Turning to applications, in Section 6 we will apply the boundary test to solving PDEs on graphs with various boundary conditions. Particular attention is paid to computing data-depth using PDEs in two ways: by solving the graph eikonal equation, and considering the first eigenfunction of the graph Laplacian. We also demonstrate these to MNIST and FashionMNIST data sets; see Section 6.3.

2. Preliminary results and error bounds for normal vector estimators

In this section we establish several results on the geometry of the empirical estimates we use, most importantly the error bounds for the normal vector estimators. Nonasymptotic $O(r)$ error bound for the first-order normal vector estimator is given in Theorem 2.6, and Section 2.1 establishes asymptotic $O(r^2)$ error bound for the second-order normal vector estimator . All the constants introduced in this and the following sections can also be found in Appendix D, and are non-dimensional. That is, they are invariant under the change of length-scale.

First we derive useful bounds on $\int_{B(x^0,r)} \rho(x)\, dx$ from the assumptions. We note that the following lemma is closely related to the 'standardness constant' in [36], which denotes the constant $s > 0$ in such that for all $x^0 \in \Omega$

$$\frac{|B(x^0, r) \cap \Omega|}{|B\left(x^0, r\right)|} \geq s. \tag{2.1}$$

This constant is of importance as it gives a lower bound on the number of points in $B(x^0, r) \cap \Omega$ with high probability. Our first lemma asserts that the Assumptions 1.1, 1.2 imply that $s \geq \frac{1}{3}$.

Lemma 2.1. *Let $r > 0$. Then*

$$\frac{\rho_{\min}\omega_d r^d}{3} \leq \int_{B(x^0,r)} \rho(x)\, dx \leq \rho_{\max}\omega_d r^d \tag{2.2}$$

Proof. As the upper bound is obvious, we focus on the lower bound, which easily follows from $s \geq 1/3$. We claim that (2.1) holds for $s = \frac{1}{2}\left(1 - \frac{\sqrt{d}r}{R}\right)$. Note that $B(x^0, r) \cap \Omega$ at least consists of the hemisphere minus the area between the tangent hyperplane at x^0. As the assumption $r \leq \frac{R}{3}$ implies that the height of the region between the tangent hyperplane and Ω with reach R is bounded above by $\frac{r^2}{R}$. Therefore, we may upper bound the area of the region by considering the cylinder with base $(d-1)$-dimensional hypersphere of radius r and height $\frac{r^2}{R}$. Thus its area is $\omega_{d-1} r^{d-1} \frac{r^2}{R} = \frac{\omega_{d-1} r^{d+1}}{R}$. Therefore

$$s \geq \frac{1}{2} - \frac{\omega_{d-1} r^{d+1} R^{-1}}{\omega_d r^d} = \frac{\omega_{d-1}}{\omega_d}\frac{r}{R} \tag{2.3}$$

We introduce the notation

$$\kappa_d = \frac{\omega_{d-1}}{\omega_d} \tag{2.4}$$

and claim that $\kappa_d \leq \sqrt{d}$. Note that since Γ is a logarithmically convex function

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

Therefore, $\omega_d^2 \geq \omega_{d-1}\omega_{d+1}$, and $\kappa_{d+1} \geq \kappa_d$. On the other hand,

$$\kappa_d\kappa_{d+1} = \frac{\omega_{d-1}}{\omega_{d+1}} = \frac{\Gamma\left(\frac{d+1}{2} + 1\right)}{\pi\Gamma\left(\frac{d-1}{2} + 1\right)} = \frac{d+3}{2\pi}.$$

Combining with $\kappa_{d+1} \geq \kappa_d$, we get $\kappa_d \leq \frac{\sqrt{d+3}}{2\pi} \leq \sqrt{d}$ as $d + 3 \leq 4\pi d$. Similarly, we have a lower bound $\kappa_{d+1} \geq \sqrt{\frac{d+3}{2\pi}} \geq \frac{1}{3}\sqrt{d+1}$, which will be of use later. Hence

$$\frac{\sqrt{d}}{3} \leq \frac{\omega_{d-1}}{\omega_d} \leq \sqrt{d}. \tag{2.5}$$

Combining the upper bound of (2.5) with (2.3), we have $s \geq \frac{1}{2}\left(1 - \frac{\sqrt{d}r}{R}\right)$. This, along with (1.23), implies that $s \geq \frac{1}{3}$. □

In the following two lemmas we examine the bias of the population-based estimators.

Lemma 2.2 (Bias of the estimated normal)**.** *For every $x^0 \in \Omega$ with $d_\Omega(x^0) \le r/2$ we have*

$$\left|\bar{v}_r(x^0) - C_y(x^0)\rho(x^0)r^{d+1}\nu(x^0)\right| \le \frac{C_x\rho(x^0)}{R}r^{d+2}, \tag{2.6}$$

provided $|\frac{\alpha}{r} - \frac{r}{R}| \le 1$, where

$$\begin{aligned} C_x &= 2\omega_{d-1} + \frac{LR\omega_d}{\rho_{min}} \\ C_y(x^0) &= \frac{\omega_{d-1}\left(1 - \left(\frac{d_\Omega(x^0)}{r} - \frac{r}{R}\right)^2\right)^{\frac{d+1}{2}}}{(d+1)}. \end{aligned} \tag{2.7}$$

In particular, whenever $d_\Omega(x^0) \le 2/(3\sqrt{d})$, we have $C_y(x^0) \ge \frac{\omega_{d-1}}{2(d+1)}$.

Remark 2.3 (Lower bound on C_y)**.** Suppose $d_\Omega(x^0) \le 2r/(3\sqrt{d})$. Then $\left(\frac{d_\Omega(x^0)}{r} - \frac{r}{R}\right)^2 \le \frac{d_\Omega(x^0)^2}{r^2} \le \frac{4}{9d} \le \frac{1}{d+1}$

$$\left(1 - \left(\frac{d_\Omega(x^0)}{r} - \frac{r}{R}\right)^2\right)^{\frac{d+1}{2}} \ge 1 - \frac{d+1}{2}\left(\frac{d_\Omega(x^0)}{r} - \frac{r}{R}\right)^2 \ge \frac{1}{2} \tag{2.8}$$

and so

$$C_y(x^0) \ge \frac{\omega_{d-1}}{2(d+1)}. \tag{2.9}$$

This lower bound will be important for results to follow. Observe that $d_\Omega(x^0) \lesssim r/\sqrt{d}$ allows a similar bound $C_y(x^0) \gtrsim \omega_{d-1}/d$.

Note also by Assumption 1.1, $d_\Omega(x^0) \le 2\varepsilon$ is a sufficient condition. As this is more intuitive and sufficient for theoretical results on the boundary test, we henceforth state the condition as $d_\Omega(x^0) \le 2\varepsilon$, but note here that all such conditions can be replaced by $d_\Omega(x^0) \le 2r/(3\sqrt{d})$. △

Proof of Lemma 2.2. We write

$$\bar{v}_r(x^0) = E_1 + \rho(x^0)E_2, \tag{2.10}$$

where

$$E_1 = \int_{\Omega\cap B(x^0,r)} (x - x^0)(\rho(x) - \rho(x^0))\,dx, \tag{2.11}$$

and

$$E_2 = \int_{\Omega\cap B(x^0,r)} (x - x^0)\,dx. \tag{2.12}$$

Since ρ is Lipschitz with constant L, the term E_1 is bounded by

$$|E_1| \le L\int_{B(x^0,r)} |x - x^0|^2\,dx = L\int_0^r \int_{\partial B(x^0,t)} t^2\,dS\,dt\,dx = L\int_0^r d\omega_d t^{d+1}\,dt = \frac{Ld\omega_d}{d+2}r^{d+2}. \tag{2.13}$$

We now estimate E_2. Without loss of generality, we may assume $x^0 = (0, 0, \ldots, 0, \alpha)$ for $\alpha = \text{dist}(x^0, \partial\Omega)$. By the assumption that the reach of $\partial\Omega$ is greater than $R > 0$, we have

$$\partial\Omega \cap B(x^0, r) \subset \left\{x \in B(x^0, r) \,:\, |x_d| \le \frac{r^2}{R}\right\},$$

provided $r \leq R/2$. Therefore

$$\left| E_2 - \int_{B(x^0,r)\cap\{x_d \geq \frac{r^2}{R}\}} (x - x^0)\, dx \right| \leq \int_{B(x^0,r)\cap\{|x_d| \leq \frac{r^2}{R}\}} |x^0 - x|\, dx \leq \frac{2\omega_{d-1} r^{d+2}}{R}. \tag{2.14}$$

We now change variables $z = (x - x^0)/r$ and write

$$\begin{aligned}\int_{B(x^0,r)\cap\{x_d \geq \frac{r^2}{R}\}} (x_d - x_d^0)\, dx &= r^{d+1} \int_{B(0,1)\cap\{z_d \geq \frac{r}{R} - \frac{\alpha}{r}\}} z_d\, dz \\ &= r^{d+1} \int_{B(0,1)\cap\{z_d \geq \left|\frac{\alpha}{r} - \frac{r}{R}\right|\}} z_d\, dz,\end{aligned}$$

where the last inequality comes from symmetry of the integrand. We now compute for any $0 \leq t \leq 1$

$$\begin{aligned}\int_{B(0,1)\cap\{z_d \geq t\}} z_d\, dz &= \omega_{d-1} \int_t^1 z_d (1 - z_d^2)^{\frac{d-1}{2}}\, dz \\ &= \frac{\omega_{d-1}}{2} \int_{t^2}^1 (1-s)^{\frac{d-1}{2}}\, ds \\ &= \frac{\omega_{d-1}}{d+1} (1-t^2)^{\frac{d+1}{2}}.\end{aligned}$$

Due to symmetry of the integrand, we have

$$\int_{B(x^0,r)\cap\{x_d \geq \frac{r^2}{R}\}} (x_j - x_j^0)\, dx = 0$$

for all $j = 1, \ldots, d-1$. Combining this with (2.14) we find that

$$\left| E_2 - \frac{\omega_{d-1}}{d+1} \left(1 - \left(\tfrac{\alpha}{r} - \tfrac{r}{R}\right)^2\right)^{\frac{d+1}{2}} r^{d+1} \nu(x^0) \right| \leq \frac{2\omega_{d-1} r^{d+2}}{R}, \tag{2.15}$$

provided $|\frac{\alpha}{r} - \frac{r}{R}| \leq 1$, since $\nu(x^0) = e_d$. Thus

$$\left| \bar{\nu}_r(x^0) - \frac{\omega_{d-1}}{d+1} \left(1 - \left(\tfrac{\alpha}{r} - \tfrac{r}{R}\right)^2\right)^{\frac{d+1}{2}} r^{d+1} \nu(x^0) \right| \leq \left(\frac{2\omega_{d-1}}{R} \rho(x^0) + L\omega_d \right) r^{d+2}.$$

We complete the proof by noting

$$\frac{2\omega_{d-1}}{R}\rho(x^0) + L\omega_d = \frac{\rho(x^0)}{R}\left(2\omega_{d-1} + \frac{LR}{\rho(x^0)}\right) \leq \frac{\rho(x^0)}{R}\left(2\omega_{d-1} + \frac{LR}{\rho_{\min}}\right) =: \frac{C_x \rho(x^0)}{R}.$$

□

Based on the bias of the estimated normal, we can approximate the bias of the distance estimator.

Lemma 2.4 (Bias of the distance estimator)**.** *Let $x^0 \in \Omega$ with $d_\Omega(x^0) \leq 2\varepsilon$. If*

$$r \leq \frac{RC_y}{2C_x} \tag{2.16}$$

then

$$d_\Omega(x^0) \leq \bar{d}_r^1(x^0) \leq d_\Omega(x^0) + \left(\frac{7C_x}{RC_y} + \frac{1}{R}\right) r^2. \tag{2.17}$$

Proof. (1) Recall

$$\hat{v}_r(x^0) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{B(x^0,r)}(x^i)(x^i - x^0)$$

and

$$\mathbb{E}\hat{v}_r(x^0) = \int_{B(x^0,r)} (x - x^0)\rho(x)\, dx = \bar{v}_r(x^0).$$

We consider the population based statistic

$$\bar{d}^1_\Omega(x^0) = \max_{x\in\Omega\cap B(x^0,r)} \{(x^0 - x)\cdot \bar{v}_r\},$$

where $\bar{\nu}_\varepsilon(x^0) := \frac{\bar{v}_r(x^0)}{\|\bar{v}_r(x^0)\|}$.

(2) By Lemma 2.2 we have

$$\bar{v}_r(x^0) = C_y\rho(x^0)r^{d+1}\nu(x^0) + \frac{1}{R}\mathcal{O}\left(C_x\rho(x^0)r^{d+2}\right).$$

Here, we can use the big-Oh notation very precisely, to mean that $f \in \mathcal{O}(g)$ if $|f| \le g$ (without any implicit constant). Therefore

$$|\bar{v}_r(x^0)| = C_y\rho(x^0)r^{d+1} + \frac{1}{R}\mathcal{O}\left(C_x\rho(x^0)r^{d+2}\right).$$

We also have

$$(x^0 - x)\cdot \bar{v}_r(x^0) = C_y\rho(x^0)r^{d+1}(x^0 - x)\cdot\nu(x^0) + \frac{1}{R}\mathcal{O}(C_x\rho(x^0)r^{d+3}). \tag{2.18}$$

We now write

$$\begin{aligned}\frac{1}{|\bar{v}_r(x^0)|} &= \frac{1}{C_y\rho(x^0)r^{d+1} + \frac{1}{R}\mathcal{O}(C_x\rho(x^0)r^{d+2})}\\ &= \frac{1}{C_y\rho(x^0)r^{d+1}\left(1 + \frac{1}{R}\mathcal{O}\left(\frac{C_x r}{C_y}\right)\right)}.\end{aligned}$$

We now use that

$$\frac{1}{1+t} = 1 + \mathcal{O}(4|t|) \ \text{ for } |x| \le \frac{1}{2}.$$

Hence, if

$$r \le \frac{RC_y}{2C_x}, \tag{2.19}$$

which implies that $\frac{C_x r}{RC_y} \le \frac{1}{2}$, then we have

$$\frac{1}{|\bar{v}_r(x^0)|} = \frac{1}{C_y\rho(x^0)r^{d+1}}\left(1 + \mathcal{O}\left(\frac{4C_x r}{RC_y}\right)\right). \tag{2.20}$$

Recall from (2.9) that $C_y > \frac{\omega_{d-1}}{2(d+1)}$. Thus

$$\frac{C_x}{C_y} \le 2(d+1)\left(2 + \frac{LR\kappa_d}{\rho_{\min}}\right).$$

(3) Inserting (2.20) into (2.18) we have

$$\begin{aligned}(x^0-x)\cdot\bar{\nu}_r(x^0) &= \frac{(x^0-x)\cdot\bar{v}_r(x^0)}{|\bar{v}_r(x^0)|}\\ &= \left((x^0-x)\cdot\nu(x^0)+\mathcal{O}\left(\frac{C_x r^2}{RC_y}\right)\right)\left(1+\mathcal{O}\left(\frac{4C_x r}{RC_y}\right)\right)\\ &= (x^0-x)\cdot\nu(x^0)+\mathcal{O}\left(\frac{C_x r^2}{RC_y}+\frac{4C_x r^2}{RC_y}+\frac{4C_x^2 r^3}{R^2C_y^2}\right)\\ &= (x^0-x)\cdot\nu(x^0)+\mathcal{O}\left(\frac{5C_x r^2}{RC_y}+\frac{4C_x^2 r^3}{R^2C_y^2}\right), \qquad (2.21)\end{aligned}$$

where $x\in\Omega\cap B(x^0,r)$.

(4) To obtain the lower bound we simply observe that $\max_{|x^i-x^0|\le r}(x^0-x^i)\cdot v$ is smallest when $v=\nu(x^0)$, in which case $\max_{|x^i-x^0|\le r}(x^0-x^i)\cdot\nu(x^0)=d_\Omega(x^0)$. Thus

$$d_\Omega(x^0)\le\bar{d}^1_\Omega(x^0) \qquad (2.22)$$

(5) For the other direction, by the assumption that the reach of $\partial\Omega$ is greater than R, we have

$$\Omega\cap B(x^0,r)\subset\left\{x\in B(x^0,r)\,:\,(x^0-x)\cdot\nu(x^0)\le d_\Omega(x^0)+\frac{r^2}{R}\right\}, \qquad (2.23)$$

provided $r\le R/2$. It follows that

$$\bar{d}^1_\Omega(x^0)\le d_\Omega(x^0)+\left(\frac{5C_x}{RC_y}+\frac{1}{R}\right)r^2+\frac{4C_x^2r^3}{R^2C_y^2}. \qquad (2.24)$$

(6) Now combining (2.22) and (2.24)

we have

$$d_\Omega(x^0)\le\bar{d}^1_\Omega(x^0)\le\left(\frac{5C_x}{RC_y}+\frac{1}{R}\right)r^2+\frac{4C_x^2}{R^2C_y^2}r^3\le\left(\frac{7C_x}{RC_y}+\frac{1}{R}\right)r^2$$

as desired, where the last inequality follows from the condition $r\le\frac{RC_y}{2C_x}$. Finally, as $\kappa_d\sim\sqrt{d}$ by (2.5), $\frac{C_x}{C_y}\sim d^{\frac{3}{2}}$. □

Next, we bound the variance of $\hat{\nu}_r$, the empirical estimator of the normal vector.

Lemma 2.5 (Bound on the variance)**.** *Let $\gamma>0$ and $c\le\frac{6d^3C_x\rho_{\max}\omega_d}{RC_y}$. If $d_\Omega(x^0)\le 2\varepsilon$ and r satisfies*

$$\left(\frac{3\gamma\rho_{\max}d^2\omega_d}{c^2}\frac{\log n}{n}\right)^{\frac{1}{d+2}}\le r\le\frac{RC_y}{2C_x} \qquad (2.25)$$

then

$$\mathbb{P}\left(|\hat{\nu}_r(x^0)-\bar{\nu}_r(x^0)|>\frac{6cr}{C_y\rho(x^0)}\right)\le 2dn^{-\gamma} \qquad (2.26)$$

Proof. Let us first fix $x^0\in\mathcal{X}$. For each $j=1,2,\cdots,d$ let

$$S_n^j=\sum_{i=1}^n\mathbb{1}_{B(x^0,r)}(x^i)(x_j^i-x_j^0).$$

Note

$$\sigma^2 = \mathrm{Var}\left(\mathbb{1}_{B(x^0,r)}(x^i)(x^i_j - x^0_j)\right) \leq \int_{B(x^0,r)} |x^i - x^0|^2 \rho(x)\, dx \leq \rho_{\max}\omega_d r^{d+2}.$$

By Bernstein's Inequality (C.3), we have

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{1}{n}S_n - \bar{v}_r(x^0)\right| > cr^{d+2}\right) &\leq \sum_{j=1}^{d} \mathbb{P}\left(\left|\frac{1}{n}S^j_n - \bar{v}_r(x^0)_j\right| > \frac{cr^{d+2}}{d}\right) \\
&\leq \sum_{j=1}^{d} 2\exp\left[-\frac{-nc^2r^{2d+4}}{2d^2\rho_{\max}\omega_d r^{d+2} + \frac{c}{3d}r^{d+3}}\right] \\
&\leq 2\sum_{j=1}^{d}\exp\left[-\frac{nc^2r^{d+2}}{2d^2\rho_{\max}\omega_d + \frac{cRC_y}{6dC_x}}\right] \leq 2d\exp\left[-\frac{nc^2r^{d+2}}{3d^2\rho_{\max}\omega_d}\right]
\end{aligned}$$

where the second last inequality follows from (2.16), and the last inequality from the condition

$$c \leq \frac{6d^3C_x\rho_{\max}\omega_d}{RC_y}.$$

The exponent is smaller than $-\gamma\log n$ when

$$r \geq \left(\frac{3\gamma\rho_{\max}d^2\omega_d}{c^2}\frac{\log n}{n}\right)^{\frac{1}{d+2}}$$

which is (2.25). Thus

$$\mathbb{P}\left(|\hat{v}_r(x^0) - \bar{v}_r(x^0)| > cr^{d+2}\right) \leq 2dn^{-\gamma} \tag{2.27}$$

Now, note that

$$|\hat{\nu}_r(x^0) - \bar{\nu}_r(x^0)| = \left|\frac{\hat{v}_r(x^0)}{|\hat{v}_r(x^0)|} - \frac{\bar{v}_r(x^0)}{|\bar{v}_r(x^0)|}\right| \leq \left|\hat{v}_r(x^0)\left(\frac{1}{|\hat{v}_r(x^0)|} - \frac{1}{|\bar{v}_r(x^0)|}\right)\right| + \frac{|\hat{v}_r(x^0) - \bar{v}_r(x^0)|}{|\bar{v}_r(x^0)|}.$$

Then (2.27) implies

$$\begin{aligned}
\left|\hat{v}_r(x^0)\left(\frac{1}{|\hat{v}_r(x^0)|} - \frac{1}{|\bar{v}_r(x^0)|}\right)\right| &= \frac{1}{|\bar{v}_r(x^0)|}|\hat{\nu}_r(x^0)(|\bar{v}_r(x^0)| - |\hat{v}_r(x^0)|)| \\
&\leq \frac{1}{|\bar{v}_r(x^0)|}|\bar{v}_r(x^0) - \hat{v}_r(x^0)| \leq \frac{cr^{d+2}}{|\bar{v}_r(x^0)|}
\end{aligned}$$

and

$$\frac{1}{|\bar{v}_r(x^0)|}|\hat{v}_r(x^0) - \bar{v}_r(x^0)| \leq \frac{cr^{d+2}}{|\bar{v}_r(x^0)|}.$$

Therefore, we have

$$\mathbb{P}\left(|\hat{\nu}_r(x^0) - \bar{\nu}_r(x^0)| > \frac{2cr^{d+2}}{|\bar{v}_r(x^0)|}\right) \leq 2dn^{-\gamma}. \tag{2.28}$$

Finally, from (2.20) and the condition $r \leq \frac{RC_y}{2C_x}$ we can deduce (2.26) as

$$\frac{2cr^{d+2}}{|\bar{v}_r(x^0)|} \leq \frac{2cr^{d+2}}{C_y\rho(x^0)r^{d+1}}\left(1 + \mathcal{O}\left(\frac{4C_xr}{RC_y}\right)\right) \leq \frac{6cr}{C_y\rho(x^0)}.$$

□

Theorem 2.6. *(Error estimates for the estimated normal vector)*
Let $x^0 \in \mathcal{X}$ with $d_\Omega(x^0) \leq 2\varepsilon$. Let $\gamma > 2$ and $\varepsilon, r > 0$ satisfy Assumption 1.2. Let r and n satisfy

$$\left(\frac{3\gamma\rho_{max}d^2\omega_d R^2}{C_x^2\rho_{\min}^2}\frac{\log n}{n}\right)^{\frac{1}{d+2}} \leq r \leq \frac{RC_y}{2C_x}. \tag{2.29}$$

Then

$$\mathbb{P}\left(|\hat{\nu}_r(x^0) - \nu(x^0)| \geq \frac{13C_x}{RC_y}r\right) \leq 2dn^{-\gamma} \tag{2.30}$$

Remark 2.7. Observe that if r satisfies (2.29), then we may choose $r = \left(\frac{3\gamma\rho_{max}d^2\omega_d R^2}{C_x^2\rho_{\min}^2}\frac{\log n}{n}\right)^{\frac{1}{d+2}}$, which means

$$\mathbb{P}\left(|\hat{\nu}_r(x^0) - \nu(x^0)| \geq C\left(\frac{\log n}{n}\right)^{\frac{1}{d+2}}\right) \leq 2dn^{-\gamma}$$

with

$$C := \frac{C_x}{C_y}\left(\frac{3\gamma\rho_{max}d^2\omega_d R^2}{C_x^2\rho_{\min}^2}\right) \sim d^2,$$

where the asymptotics in d can be derived using Stirling's formula. For a more detailed analysis of the how the constants scale with dimension, please see Remarks 3.4 and 3.6.

Further, we note that the above result holds for $x^0 \in \Omega_{2\varepsilon}$ – i.e. the reference point need not be one of the samples. The same applies to following results on the distance estimator. △

Proof. The upper bound of (3.4) allows us to apply Lemma 2.4, which we will combine with Lemma 2.5. The lower bound in (2.5) implies that

$$\frac{6d^3\omega_d}{C_y} \geq \frac{12d^3(d+1)}{\kappa_d} \geq 12$$

from which easily follows $\frac{C_x\rho(x^0)}{R} \leq \frac{6d^3\rho_{\max}\omega_d C_x}{RC_y}$. Thus we may set $c = \frac{C_x\rho(x^0)}{R}$. Then Lemma 2.5 implies that if

$$r \geq \left(\frac{3\gamma\rho_{max}d^2\omega_d R^2}{C_x^2\rho_{\min}^2}\frac{\log n}{n}\right)^{\frac{1}{d+2}}$$

then, by (2.20),

$$|\hat{\nu}_r(x^0) - \bar{\nu}_r(x^0)| \leq \frac{2C_x}{RC_y}\left(1 + \mathcal{O}\left(\frac{4C_x r}{RC_y}\right)\right)r \leq \frac{6C_x}{RC_y}r \tag{2.31}$$

with probability at least $1 - 2dn^{-\gamma}$, where the last inequality follows from the condition $r \leq \frac{RC_y}{2C_x}$.

Next we bound $|\bar{\nu}_r(x^0) - \nu(x^0)|$. Again by (2.20)

$$\begin{aligned}|\bar{\nu}_r(x^0) - \nu(x^0)| &= \left|\frac{\bar{v}_r(x^0)}{|\bar{v}_r(x^0)|} - \nu(x^0)\right| \\ &= \frac{1}{C_y\rho(x^0)r^{d+1}}\left|\bar{v}_r(x^0)\left(1 + \mathcal{O}\left(\frac{4C_x r}{RC_y}\right)\right) - C_y\rho(x^0)r^{d+1}\nu(x^0)\right|.\end{aligned}$$

By Lemma 2.2

$$\left|\bar{v}_r(x^0)\left(1+\mathcal{O}\left(\frac{4C_x r}{RC_y}\right)\right) - C_y\rho(x^0)r^{d+1}\nu(x^0)\right| \le |\bar{v}_r(x^0) - C_y\rho(x^0)r^{d+1}\nu(x^0)| + \frac{4C_x|\bar{v}_r(x^0)|r}{RC_y}.$$

Thus

$$|\bar{\nu}_r(x^0) - \nu(x^0)| \le \frac{C_x}{RC_y}r + \frac{4C_x r}{RC_y} + \frac{4C_x^2 r^2}{R^2C_y^2} \le \frac{7C_x}{RC_y}r \tag{2.32}$$

where the last inequality follows from (2.19). Combining (2.31) and (2.32) we have

$$|\hat{\nu}_r(x^0) - \nu(x^0)| \le \frac{13C_x}{RC_y}r$$

with probability at least $1 - 2dn^{-\gamma}$. □

2.1. **Second-order estimators: asymptotic error scaling.** Here we analyze the asymptotic error of the "second-order" estimator of the normal vector, $\hat{\nu}_r^2(x^0)$, defined in (1.5), and show that the error is indeed second-order in r, for points x^0 sufficiently close to the boundary, namely $d_\Omega(x^0) \lesssim r/\sqrt{d}$, which allows us to use (2.6) with a reasonable lower bound on $C_y(x^0)$ (see Remark 2.3). We note that in this section, in order to simplify expressions we use radius r for estimating θ, instead of the radius $r/2$ as in (1.6) and (1.8). However, a similar argument works when we set the radius to be $r/2$.

For simplicity, we first assume the boundary is the graph of a quadratic function near x^0. That is that near $x^0 = |x^0|e_d$ and the boundary is given by

$$x_d = H(x)^T A H(x)$$

where A is a $(d-1)\times(d-1)$ symmetric matrix and

$$H(x) = (x_1, \dots, x_{d-1})^T$$

We also introduce the symbols for projection of a vector to the e_d direction and for central symmetry with respect to the first $d-1$ variables

$$N(x) = x \cdot e_d\, e_d \quad \text{and} \quad S(x) = (-H(x), x_d).$$

Furthermore let $U(x) = B(x,r) \cap \Omega$.

Since $\bar{v}_r^2(x^0)\cdot e_d > Cr^{d+1}$ by estimate (2.6) it suffices to show that $|H(\bar{v}_r^2(x^0))| \le Cr^{d+3}$. We start by noting that due to symmetry of the quadratic function near x^0

$$\begin{aligned} H(\bar{v}_r^2(x^0)) &= \frac{1}{2}\int_{U(x^0)} H\left(\frac{\rho(x)}{\theta(x)}(x-x^0) + \frac{\rho(S(x))}{\theta(S(x))}(S(x)-x^0)\right)dx \\ &\le \frac{1}{2}\int_{U(x^0)} \frac{|\rho(x)\theta(S(x)) - \rho(S(x))\theta(x)|}{\theta(S(x))\theta(x)}|H(x)|dx \\ &\le \frac{8}{\rho_{min}^2} r^{d+1} \sup_{x\in U(x^0)} |\rho(x)\theta(S(x)) - \rho(S(x))\theta(x)|. \end{aligned}$$

For $x \in U(x^0)$ we now estimate, assuming $4r < R$ and using that S is isometry between $U(x)$ and $U(S(x))$

$$\begin{aligned}
|\rho(x)\theta(S(x)) - \rho(S(x))\theta(x)| &= \frac{1}{\omega_d r^d}\left|\rho(x)\int_{U(S(x))}\rho(z)-\rho(S(x))dz - \rho(S(x))\int_{U(x)}\rho(z)-\rho(x)dz\right| \\
&\leq \frac{1}{\omega_d r^d}\left|\rho(x)\int_{U(S(x))}\nabla\rho(N(0))\cdot(z-S(x))dz - \rho(S(x))\int_{U(x)}\nabla\rho(N(0))(z-x)dz\right| \\
&\quad + 4\|\rho\|_{L^\infty}\|D^2\rho\|_{L^\infty}r^2 \\
&= \frac{1}{\omega_d r^d}\left|(\rho(S(x))-\rho(x))\int_{U(x)}\nabla\rho(N(0))(z-x)dz\right| + 4\|\rho\|_{L^\infty}\|D^2\rho\|_{L^\infty}r^2 \\
&\leq 4\left(\|\nabla\rho\|^2_{\mathcal{L}^\infty} + \|\rho\|_{L^\infty}\|D^2\rho\|_{L^\infty}\right)r^2
\end{aligned}$$

Combining with the estimate above we obtain

$$|H(\bar{v}_r^2(x^0))| \leq Cr^{d+3}$$

where C depends on ρ alone.

We now relax the assumption that the boundary of Ω is a graph of a quadratic function. Namely note that since the boundary of Ω is C^3 there exists $C^r > 0$ such that near x^0 the boundary of Ω is between the graphs of $x_d = H(x)^T AH(x) - C_r|H(x)|^3$ and $x_d = H(x)^T AH(x) + C_r|H(x)|^3$. Note that neglecting the part of Ω between the graphs produces an error of size r^{d+3} and that all of the estimates above carry over to the part of Ω where $x_d > H(x)^T AH(x) + C_r|H(x)|^3$. Thus it still holds that $|T(\bar{v}_r^n(x^0))| \leq Cr^{d+3}$, only that C depends both of ρ and Ω.

We now outline the argument at the level of the sample. One can use standard concentration inequalities to control the variance and obtain the regime in which the empirical estimator $\hat{v}_r^n$ is within Cr^3 of the population based estimate $\bar{v}_r^n$.

Applying Bernstein's inequality to the random variables $Y^j = \frac{1}{\omega_d r^d}\mathbb{1}_{|x^j - x|\leq r/2}$ one obtains

$$|\hat{\theta}(x^i) - \theta(x^i)| \lesssim r^2 \tag{2.33}$$

with high probability provided that $r \gtrsim (\log n/n)^{1/(d+4)}$. Using the union bound the estimate holds uniformly for all i. Thus

$$\left|\hat{v}_r^2(x^0) - \frac{1}{n}\sum_{i=1}^n \frac{\mathbb{1}_{B(x^0,r)}(x^i)}{\theta(x^i)}(x^i - x^0)\right| \lesssim r^{d+3}$$

Using the Bernstein inequality once more one obtains that

$$\left|\frac{1}{n}\sum_{i=1}^n \frac{\mathbb{1}_{B(x^0,r)}(x^i)}{\theta(x^i)}(x^i - x^0) - \bar{v}_r^2(x^0)\right| \lesssim r^{d+3}$$

with high probability if $r \gtrsim (\log n/n)^{1/(d+4)}$

Combining with $\bar{v}_r^2(x^0)\cdot e_d \gtrsim r^{d+1}$ and $|H(\bar{v}_r^2(x^0))| \lesssim r^{d+3}$ we conclude that $|\hat{\nu}_r^2(x^0) - \nu(x^0)| \lesssim r^2$, as desired.

3. Nonasymptotic error bounds for first-order distance and boundary estimators

In this section we establish the main results. Namely in Theorem 3.3 we show that the estimator $\hat{d}_r^1(x^0)$ has $O(r^2)$ error, provided that $r \gtrsim (\log n/n)^{1/(d+2)}$. We then use this estimate to show that when $\frac{r^2}{R} \lesssim \varepsilon \lesssim r$ then we can accurately identify the ε-boundary points.

We start with establishing a lower bound on error of the distance estimator $\hat{d}_r^1$.

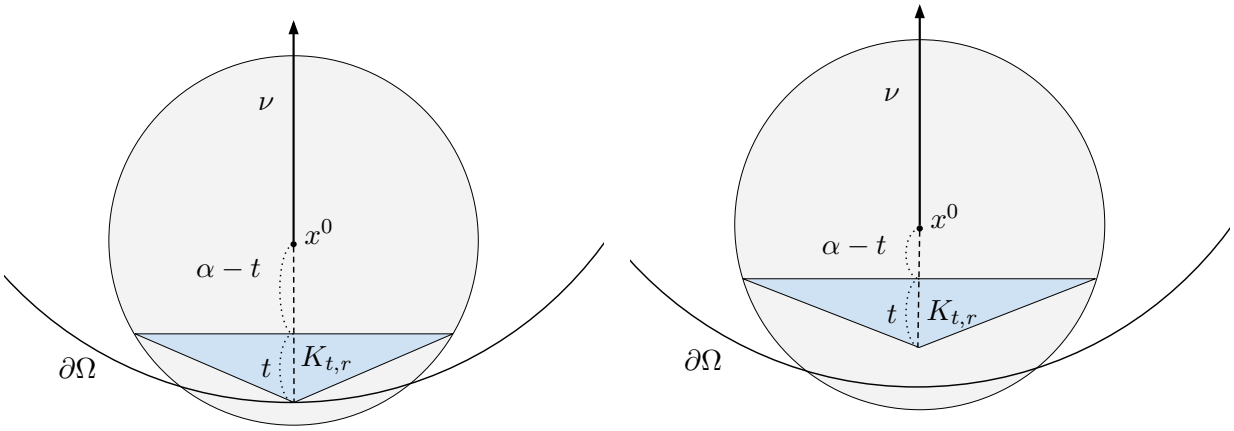


FIGURE 3. Geometry relevant to the lower bound on $\hat{d}_r^1(x^0)$. $\alpha = d_\Omega(x^0) \wedge \frac{r}{2}$. (Left) Case where $d_\Omega(x^0) < \frac{r}{2}$; (Right) case where $d_\Omega(x^0) > \frac{r}{2}$.

Lemma 3.1 (Lower bound on the distance estimator)**.** *Let $\gamma > 2$, $0 < t \le d_\Omega(x^0)$, and suppose Assumption 1.1 holds. If n and $\lambda > 0$ satisfy*

$$n \ge d \vee (1 + 4\lambda^{-1})$$

and t, r satisfy

$$tr^{d-1} \ge \frac{\gamma d^2 2^{(d-1)/2}}{\rho_{\min}\omega_{d-1}} \left(\frac{\log n}{n}\right), \tag{3.1}$$

then

$$\hat{d}_r^1(x^0) \ge (1-\lambda)(d_\Omega(x^0) \wedge \frac{r}{2}) - t \tag{3.2}$$

with probability at least $1 - n^{-\gamma}$.

Remark 3.2. In fact, the lemma holds for any unit vector $\hat{u}$ that may depend on $\mathcal{X}$. Recall that the second-order distance estimator $\hat{d}_r^2$ defined in (1.17) is of the form

$$\hat{d}_r^2(x^0) = \max_{x \in B(x^0,r)\cap\mathcal{X}} (x^0 - x^i) \cdot \hat{u},$$

where $|\hat{u}|$ can be as small as $\frac{1}{\sqrt{2}}$ in the interior, when $\hat{u}$ is an average of orthogonal unit vectors. Thus a slight modification allows us to obtain a similar result to the second-order distance estimator $\hat{d}_r^2$. △

Sketch of Proof. As the proof involves lengthy elementary calculations, we delay the full proof to Appendix A, and only present the main ideas here. The idea is to ensure that for any unit vector $u \in \mathbb{S}^{d-1}$, possibly depending on the samples $\mathcal{X}$, there is a point in the spherical segment $S^u \cap \Omega$ that contains points at least $(1-\lambda)(d_\Omega(x^0) \wedge \frac{r}{2}) - t$ away in the opposite direction of u. See Figure 3 for the illustration in the case $u = \nu$. As there are infinitely many choices of u, we shrink the spherical segment slightly so that we have a finite family $\{\tilde{S}^1, \cdots, \tilde{S}^N\}$ such that for any $u \in \mathbb{S}^{d-1}$ we can find $\tilde{S}^i \subset S^u$. This means it suffices to show that each $\tilde{S}^i$ is nonempty for $i = 1, \cdots, N$, and

$$\begin{aligned}\mathbb{P}(\hat{d}_r^1(x^0) \le (1-\lambda)(d_\Omega(x^0) \wedge \frac{r}{2} - t) &\le \mathbb{P}(S^u \cap \Omega \text{ is nonempty for all } u \in \mathbb{S}^{d-1}) \\ &\le \sum_{i=1}^N \mathbb{P}(\tilde{S}^i \cap \Omega \text{ is nonempty for all } i = 1, \cdots, N).\end{aligned}$$

For suitably chosen spherical segments, we may observe that $\tilde{S}^i \cap \Omega$ contains a cone K with the same base and height as the spherical segment. Thus the proof comes down to obtaining a lower bound for the volume of this cone, and an upper bound on the number N. △

We now state the nonasymptotic error bounds on the first-order distance estimator.

Theorem 3.3 (Error bounds for the distance estimator)**.** *Let $\varepsilon, r > 0$ satisfy Assumptions 1.1 and 1.2. Let constants C_x and C_y be as in (2.7), and*

$$C_r = \frac{1}{R} \max \left[\left(\frac{3\gamma \rho_{\max} d^2 \omega_d R^2}{{C_x}^2 \rho_{\min}^2} \right)^{\frac{1}{d+2}}, \left(\frac{4\gamma C_y d^2 2^{(d-1)/2}}{13 \rho_{\min} \omega_{d-1} C_x} \right)^{\frac{1}{d+1}} \right]$$

Suppose $\gamma > 2$, and n, r satisfy

$$n \geq d \vee \left(1 + \frac{RC_y}{13C_x} r^{-1} \right) \tag{3.3}$$

and

$$RC_r \left(\frac{\log n}{n} \right)^{\frac{1}{d+2}} \leq r \leq \frac{RC_y}{2C_x} \tag{3.4}$$

Then, for $x^0 \in \mathcal{X}$ we have

$$d_\Omega(x^0) \wedge \frac{r}{2} - \frac{13C_x}{RC_y} r^2 \leq \hat{d}_r^1(x^0) \tag{3.5}$$

with probability at least $1 - n^\gamma$. Moreover, if $d_\Omega(x^0) \leq 2\varepsilon \leq r$,

$$\hat{d}_r^1(x^0) \leq d_\Omega(x^0) + \left(\frac{13C_x}{RC_y} + \frac{1}{R} \right) r^2 \tag{3.6}$$

with probability at least $1 - 2dn^{-\gamma}$.

Remark 3.4. We make two brief remarks. Firstly, (3.3) is a much weaker condition than the lower bound of (3.4), as $r \geq RC_r \left(\frac{\log n}{n} \right)^{\frac{1}{d+2}}$ implies

$$\frac{RC_y}{13C_x} r^{-1} \leq \frac{RC_y}{13C_x} (RC_r)^{-\frac{1}{d+2}} \left(\frac{n}{\log n} \right)^{\frac{1}{d+2}},$$

which is much smaller than n for reasonably large n.

Secondly, we note that $C_r \sim \omega_d^{-1/d}$. Using Stirling's formula $d! \sim \sqrt{2\pi d}(d/e)^d$ one obtains $\omega_d \sim (1/\sqrt{\pi d})(2\pi e/d)^{d/2}$. Therefore $C_r \sim \omega_d^{-1/d} = O(\sqrt{d})$ △

Proof. We first prove the upper bound (3.6). Suppose $d_\Omega(x^0) \leq 2\varepsilon \leq r$. Condition (3.4) allows us to apply Theorem 2.6 to obtain (2.31) –i.e.

$$|\hat{\nu}_r(x^0) - \nu(x^0)| \leq \frac{13C_x}{RC_y} r$$

with probability at least $1 - 2dn^{-\gamma}$. Thus

$$\begin{aligned}\hat{d}_r^1(x^0) &= \max_{x^i \in B(x_0,r)\cap\mathcal{X}} \left\{(x^0 - x^i)\cdot(\hat{\nu}_r(x^0) - \nu(x^0) + \nu(x^0))\right\} \\ &\le \max_{x^i \in B(x_0,r)\cap\mathcal{X}} (x^0 - x^i)\cdot(\hat{\nu}_r(x^0) - \nu(x^0)) + \max_{x^i \in B(x_0,r)\cap\mathcal{X}} (x^0 - x^i)\cdot\nu(x^0) \\ &\le \frac{13C_x}{RC_y} r^2 + d_\Omega(x^0) + \frac{1}{R} r^2\end{aligned}$$

with the same probability. The last inequality uses the bound on $|\hat{\nu}_r(x^0) - \nu(x^0)|$ and that positive reach condition implies (2.23). Thus we have the upper bound (3.6).

Next, suppose $x^0 \in \mathcal{X}$, not necessarily close to the boundary. Letting $t = \frac{13C_x}{2C_y} r^2$ in Lemma 3.1, if r satisfies

$$r^{d+1} \ge \frac{4\gamma C_y d^2 2^{(d-1)/2}}{13\rho_{\min}\omega_{d-1}C_x} \frac{\log n}{n}$$

then Lemma 3.1 implies that

$$\hat{d}_r^1(x^0) \ge (1-\lambda)(d_\Omega(x^0) \wedge \frac{r}{2}) - \frac{13C_x}{2C_y} r^2 \tag{3.7}$$

with probability at least $1 - n^{-\gamma}$, given $n \ge d \vee (1 + 4\lambda^{-1})$. Further, choose $\lambda = \frac{13C_x}{RC_y} r$, so that by Assumption 1.1

$$\lambda(d_\Omega(x^0) \wedge \frac{r}{2}) \le \frac{\lambda r}{2} = \frac{13C_x}{2RC_y} r^2.$$

Then (3.3) implies

$$\hat{d}_r^1(x^0) \ge d_\Omega(x^0) \wedge 2\varepsilon - \lambda(d_\Omega(x^0) \wedge \frac{r}{2}) - t \ge d_\Omega(x^0) - \frac{13C_x}{RC_y} r^2,$$

hence we obtain (3.5). □

Corollary 3.5 (Accuracy of the boundary test)**.** *Let $x^0 \in \mathcal{X}$, $\gamma > 2$ and $\varepsilon, r > 0$ satisfy Assumptions 1.1 and 1.2. Let C_r be as in (3.3). If $n \ge d \vee 33$ and r, n satisfy*

$$RC_r \left(\frac{\log n}{n}\right)^{\frac{1}{d+2}} \le r \le \frac{RC_y}{2C_x} \tag{3.8}$$

and ε satisfies

$$\frac{1}{R}\left(\frac{26C_x}{C_y} + 2\right) r^2 < \varepsilon \tag{3.9}$$

then

$$\mathbb{P}(\widehat{T}^1_{\varepsilon,r}(x^0) = 1 \mid d_\Omega(x^0) \ge 2\varepsilon) + \mathbb{P}(\widehat{T}^1_{\varepsilon,r}(x^0) = 0 \mid d_\Omega(x^0) \le \varepsilon) \le (2d+1)n^{-\gamma}. \tag{3.10}$$

In particular, choosing the optimal r, ε

$$\varepsilon = \frac{1}{R}\left(\frac{26C_x}{C_y} + 2\right) r^2 = RC_r^2 \left(\frac{26C_x}{C_y} + 2\right) \left(\frac{\log n}{n}\right)^{\frac{2}{d+2}}, \tag{3.11}$$

the test identifies the ε-boundary with probability at least $1 - (2d+1)n^{-\gamma}$.

Remark 3.6. Recall that (1.22) implies

$$\frac{C_x}{C_y} \leq 2(d+1)\left(1+\frac{RL}{\rho_{\min}}\kappa_d\right) = O(d^{\frac{3}{2}})$$

as $\kappa_d \sim \sqrt{d}$ by (2.5). Also, recall from Remark (3.4) that $C_r = O(\sqrt{d})$. Therefore the constant for the optimal choice $\varepsilon = C(\log n/n)^{2/(d+2)}$ in (3.11) satisfies $C \sim C_r^2 C_x/C_y \sim d^{5/2}$. △

Proof. Suppose $n \geq d \vee (1+4\cdot 8) = d \vee 33$ and $d_\Omega(x^0) \geq 2\varepsilon$. Then we may choose $\lambda = \frac{1}{8}$ in (3.7) and apply Lemma 3.1 to deduce

$$\hat{d}_r^1(x^0) \geq \frac{7}{8}(d_\Omega(x^0) \wedge \frac{r}{2}) - \frac{13C_x}{2RC_y}r^2 \geq \frac{7}{8}(d_\Omega(x^0)\wedge 2\varepsilon) - \frac{13C_x}{2RC_y}r^2 \geq \frac{7\varepsilon}{4} - \frac{13C_x}{2RC_y}r^2 > \frac{3\varepsilon}{2}$$

with probability at least $1-n^{-\gamma}$, where last inequality follows from the condition (3.9). Note that we have used that Assumption 1.1 implies $2\varepsilon \leq \frac{3\sqrt{d}\varepsilon}{2} \leq \frac{r}{2}$.Thus we deduce

$$\mathbb{P}(\widehat{T}^1_{\varepsilon,r}(x^0) = 1 \,|\, d_\Omega(x^0) \geq 2\varepsilon) \leq n^{-\gamma}.$$

On the other hand, if $d_\Omega(x^0) \leq \varepsilon$, then the upper bound in (3.6) applies. Thus, again using (3.9)

$$\hat{d}_r^1(x^0) \leq d_\Omega(x^0) + \left(\frac{13C_x}{RC_y} + \frac{1}{R}\right)r^2 \leq \frac{3\varepsilon}{2},$$

with probability at least $1-2dn^{-\gamma}$. Hence

$$\mathbb{P}(\widehat{T}^1_{\varepsilon,r}(x^0) = 0 \,|\, d_\Omega(x^0) \leq \varepsilon) \leq n^{-\gamma}.$$

Combining this with the bound for the probability of false positive occurring, we obtain (3.10). □

For application to solving boundary value problems on graphs [24], it is crucial to limit the number of false positives, while the false negatives are not as detrimental. If we are only interested in bounding the probability of false positives, we may obtain the improved rate $\varepsilon \geq C\left(\frac{\log n}{n}\right)^{\frac{1}{d+1}}$ with $C \sim d$.

Theorem 3.7 (One-sided accuracy of the boundary test)**.** *Let $\gamma > 2$, and $x^0 \in \mathcal{X}$. Suppose $\varepsilon, r > 0$ satisfy Assumptions 1.1 and 1.2. If $n \geq d \vee 33$ and ε, r satisfy*

$$\left(\frac{\gamma d^2 2^{(d-1)/2}}{\rho_{\min}\omega_{d-1}}\frac{\log n}{n}\right)^{\frac{1}{d+1}} \leq r^2 < \frac{\varepsilon}{4}$$

then

$$\mathbb{P}(\widehat{T}^1_{\varepsilon,r}(x^0) = 1 \,|\, d_\Omega(x^0) > 2\varepsilon) \leq n^{-\gamma}.$$

Proof. Again, recall that Assumption 1.1 implies $2\varepsilon \leq \frac{r}{2}$. Applying Lemma 3.1 with $t = r^2$ and $\lambda = \frac{1}{8}$,we have $\hat{d}_r^1(x^0) \geq \frac{7}{8}(d_\Omega(x^0)\wedge r/2) - r^2$ with probability at least $n^{-\gamma}$. Thus if

$$4r^2 \leq \varepsilon$$

then, with probability at least $1-n^{-\gamma}$

$$\hat{d}_r^1(x^0) \geq \frac{7}{8}(d_\Omega(x^0)\wedge\frac{r}{2}) - r^2 \geq \frac{7}{8}(d_\Omega(x^0)\wedge 2\varepsilon) - r^2 > \frac{7\varepsilon}{4} - \frac{\varepsilon}{4} \geq \frac{3\varepsilon}{2}.$$

This implies that $\widehat{T}^1_{\varepsilon,r}(x^0) = 0$ by (1.20). □

Corollary 3.8. *Let $x^0 \in \mathcal{X}$, $\gamma > 2$. Let $n \geq d \vee 33$ and be sufficiently large such that*

$$\varepsilon = RC_\varepsilon \left(\frac{\log n}{n}\right)^{\frac{2}{d+2}}, \; r = C_r \left(\frac{\log n}{n}\right)^{\frac{1}{d+2}} \tag{3.12}$$

satisfy Assumptions 1.1 and 1.2. Recall the definitions

$$\partial_a \Omega = \{x^0 \in \mathcal{X} : d_\Omega(x^0) \leq a\}$$
$$\partial_{\varepsilon,r}\mathcal{X} = \{x^0 \in \mathcal{X} : \widehat{T}^1_{\varepsilon,r}\left(x^0\right) = 1\}.$$

Then, with probability at least $1 - (2d+1)n^{1-\gamma}$.

$$\partial_\varepsilon \Omega \subset \partial_{\varepsilon,r}\mathcal{X} \subset \partial_{2\varepsilon}\Omega. \tag{3.13}$$

In particular, by the Borel-Cantelli lemma, the test identifies a set between $\partial_\varepsilon\Omega$ and $\partial_{2\varepsilon}\Omega$ eventually with probability 1.

Proof. By Corollary 3.5, applying the test to all n points we have (3.13) hold with probability at least $1 - (2d+1)n^{-\gamma} \cdot n = 1 - (2d+1)n^{1-\gamma}$. □

Remark 3.9 (Reconstruction of boundary from boundary points)**.** Based on the set $\partial_{\varepsilon r}\mathcal{X}$ of boundary points we can reconstruct the boundary strip that approximates $\partial\Omega$ in the Hausdorff distance. See for instance Theorem 3.11 of [1] and the comment preceding it on the reconstruction process using Delaunay Complex, and [2] for further details. △

4. Asymptotic error bounds for second-order distance and boundary estimators

In this section, we use the $O(r^2)$ bound on the second-order normal estimator $\hat{\nu}^2_r$ from Section 2.1 to obtain $O(r^3)$ error bound on the second-order distance estimator $\hat{d}^2_r$ in the asymptotic regime, additionally assuming $\partial\Omega$ is of class C^3 and $\rho \in C^2_b(\Omega)$. Namely, we show that we can find some constant $C > 0$ independent of r such that

$$\begin{aligned} &\hat{d}^2_r(x^0) \geq d_\Omega(x^0) \wedge \frac{r}{2} - Cr^3, \text{ and} \\ &\hat{d}^2_r(x^0) \leq d_\Omega(x^0) + Cr^3 \text{ if } d_\Omega(x^0) \leq 2\varepsilon \end{aligned} \tag{4.1}$$

with high probability under the scaling $r \gtrsim (\log n/n)^{1/(d+4)}$. Note that the lower bound holds for general $x^0 \in \mathcal{X}$, not just those close to the boundary. Given the estimates above, we may set $\varepsilon = Cr^3/2 \sim (\log n/n)^{3/(d+4)}$ to see that our test (1.20) will identify the ε-boundary points with high probability. For a detailed argument deducing accuracy of the boundary estimator from that of the distance estimator, please see the the proof of Corollary 3.5; while the corollary applies to the first-order estimator, the same argument carries over to the second-order estimator.

For simplicity, we will show (4.1) for a slight modification of the estimator (1.17). Namely, instead of the cutoff $\mathbb{1}_{\mathbb{R}^+}(\hat{\nu}^2_r(x^i) \cdot \hat{\nu}^2_r(x^0))$, we use $\mathbb{1}_{\{x:\, x \leq cr\}}(|\hat{\nu}^2_r(x^i) - \hat{\nu}^2_r(x^0)|)$ for suitably large c, say, twice the Lipschitz constant of $d_\Omega(\cdot)$. Note that this is a reasonable cutoff, as

$$|\hat{\nu}^2_r(x^i) - \hat{\nu}^2_r(x^0)| \leq |\hat{\nu}^2_r(x^i) - \nu(x^i)| + |\nu(x^i) - \nu(x^0)| + |\nu(x^0) - \hat{\nu}^2_r(x^0)|.$$

From Section 2.1 we know that the first and third terms are small are of order $O(r^2)$ when $r \gtrsim (\log n/n)^{1/(d+4)}$; the second term is of order $O(r)$ as $\nu(x) = \nabla d_\Omega(x)$ near the boundary, which is a C^2 function as we assumed $\partial\Omega$ to be of class C^3. Thus, for sufficiently small r we have

$$|\hat{\nu}^2_r(x^i) - \hat{\nu}^2_r(x^0)| \leq \frac{c}{2}|x^i - x^0| + O(r^2) \leq cr.$$

Upper bound. For the upper bound, suppose $d_\Omega(x^0) \le 2\varepsilon$. Fix $c', C' > 0$ and $r > 0$, and denote by E_0 the event

$$E_0 := \{|\hat{\nu}_r^2(x^i) - \nu(x^i)| \le C' r^2 \ \text{ for all } x_i \in B(x^0, r) \cap \mathcal{X} \text{ such that } d_\Omega(x^i) \le r/\sqrt{d}\}.$$

Recall from Section 2.1 that E_0 occurs with high probability when $r \gtrsim (\log n/n)^{1/(d+4)}$ and $C' > 0$ is chosen suitably large.

For simplified notation, let us temporarily define $\hat{u}^i(x^0)$ for each $i = 1, \cdots, n$ by

$$\hat{u}^i(x^0) := \left[\hat{\nu}_r^2(x^0) + \frac{\hat{\nu}_r^2(x^i) - \hat{\nu}_r^2(x^0)}{2} \mathbb{1}_{\{x:\, x \le cr\}}(|\hat{\nu}_r^2(x^i) - \hat{\nu}_r^2(x^0)|)\right], \tag{4.2}$$

so that $\hat{d}_r^2(x^0) = \max_{x^i \in B(x^0,r)\cap\mathcal{X}} (x^0 - x^i) \cdot \hat{u}^i(x^0)$. Define the set $\hat{\mathcal{X}}$ by

$$\hat{\mathcal{X}} := \{x^i \in \mathcal{X} : (x^0 - x^i) \cdot \hat{u}^i(x^0) \ge 0\}.$$

Then we may write

$$\hat{d}_r^2(x^0) = \max_{x^i \in B(x^0,r)\cap\mathcal{X}} (x^0 - x^i) \cdot \hat{u}^i(x^0) = \max_{x^i \in B(x^0,r)\cap\hat{\mathcal{X}}} (x^0 - x^i) \cdot \hat{u}^i(x^0).$$

Indeed the right-hand side is the nonnegative part of $\hat{d}_r^2(x^0)$, while $\hat{d}_r^2(x^0) \ge 0$ due to that $x^0 \in B(x^0, r)\cap\hat{\mathcal{X}}$. Thus the above equality holds.

Due to the cutoff, note

$$\left|\frac{\hat{u}^i(x^0)}{|\hat{u}^i(x^0)|} - \nu(x^0)\right| \le \left|\frac{\hat{u}^i(x^0)}{|\hat{u}^i(x^0)|} - \hat{\nu}_r^2(x^0)\right| + |\hat{\nu}_r^2(x^0) - \nu(x^0)| \le cr + O(r^2) \le 2cr$$

for sufficiently small r. Thus, if $x^i \in \hat{\mathcal{X}}$, it is in the half plane opposite of $\hat{u}^i(x^0)$, which is closely approximated by the half plane opposite of $\nu(x^0)$. As $d_\Omega(x^0) \le 2\varepsilon$, collecting the errors due to curvature of the boundary and the difference between $\hat{u}^i(x^0)/|\hat{u}^i(x^0)|$ and $\nu(x^0)$, we see

$$d_\Omega(x^i) \le 2\varepsilon + \frac{r^2}{R} + 2cr^2 \le \frac{r}{\sqrt{d}}$$

when $\varepsilon \ll r$ and r is sufficiently small. Thus, by E_0 we have $|\hat{\nu}_r^2(x^i) - \nu(x^i)| \le C' r^2$ for all $x^i \in \hat{\mathcal{X}}$, and

$$\hat{d}_r^2(x^0) = \max_{x^i \in B(x^0,r)\cap\hat{\mathcal{X}}} (x^0 - x^i) \cdot \frac{\hat{\nu}_r^2(x^i) + \hat{\nu}_r^2(x^0)}{2}.$$

Now, when $\partial\Omega$ is of class C^3, recall (1.15) holds. Thus, we have

$$d_\Omega(x^0) \ge \max_{x^j \in B(x^0,r)} \left\{\frac{1}{2}(\nu(x^0) + \nu(x^j)) \cdot (x^0 - x^j)\right\} + O(r^3).$$

Then we have the upper bound on $\hat{d}_r^2$

$$\hat{d}_r^2(x^0) - d_\Omega(x^0) \le \max_{x^i \in B(x^0,r)\cap\hat{\mathcal{X}}} \left\{\frac{1}{2}\left(\hat{\nu}_r^2(x^i) + \hat{\nu}_r(x^0) - \nu(x^0) - \nu(x^i)\right) \cdot (x^0 - x^i)\right\} + O(r^3) = O(r^3),$$

as $|x^0 - x^i| \le r$ and $|\hat{\nu}_r^2(x^i) - \nu(x^i)| + |\hat{\nu}_r^2(x^0) - \nu(x^0)| \lesssim r^2$.

Lower bound. Recall the elementary equality $\frac{|u+w|^2}{4} = 1 - \frac{|u-w|^2}{4}$ that holds when $|u| = |w| = 1$. This implies the following lower bound on the magnitude of $\hat{u}^i(x^0)$ defined in (4.2)

$$|\hat{u}^i(x^0)| \ge (1 - c' r^2)^{1/2}. \tag{4.3}$$

Writing $\alpha = d_\Omega(x^0) \wedge \frac{r}{2}$, under the assumptions of Lemma 3.1, we have

$$\begin{aligned}\mathbb{P}(\hat{d}_r^2(x^0) \leq (1-\lambda)\alpha - t) &= \mathbb{P}\left(\max_{x^i \in B(x^0,r)\cap\mathcal{X}} (x^0 - x^i)\cdot \hat{u}^i \leq (1-\lambda)\alpha - t\right) \\ &= \mathbb{P}\left(\max_{x^i \in B(x^0,r)\cap\mathcal{X}} (x^0 - x^i)\cdot \frac{\hat{u}^i}{|\hat{u}^i|} \leq \frac{1}{|\hat{u}^i|}((1-\lambda)\alpha - t)\right).\end{aligned}$$

By (4.3), we can fix $C > 0$ such that $\frac{1}{|\hat{u}^i|} \leq \frac{1}{\sqrt{1-c'r^2}} \leq 1 + Cr^2$ when r is sufficiently small. As $t < \alpha \leq r$, we have

$$\begin{aligned}\mathbb{P}(\hat{d}_r^2(x^0) \leq (1-\lambda)\alpha - t) &\leq \mathbb{P}\left(\max_{x^i \in B(x^0,r)\cap\mathcal{X}} (x^0 - x^i)\cdot \frac{\hat{u}^i}{|\hat{u}^i|} \leq (1-\lambda)\alpha - t + Cr^2((1-\lambda)\alpha - t)\right) \\ &\leq \mathbb{P}\left(\max_{x^i \in B(x^0,r)\cap\mathcal{X}} (x^0 - x^i)\cdot \frac{\hat{u}^i}{|\hat{u}^i|} \leq (1-\lambda)\alpha - t + Cr^3\right) \leq n^{-\gamma}.\end{aligned}$$

The last inequality follows when $t > Cr^3$ by Lemma 3.1, as its proof only uses that $|\hat{\nu}_r(x^0)| = 1$. Choosing $t = 2Cr^3$ and $\lambda \leq Cr^2$ for instance, we obtain that $\hat{d}_r^2(x^0) \geq d_\Omega(x^0) - 3Cr^3$ with high probability, and the condition (3.1) becomes $r \gtrsim (\log n/n)^{1/(d+2)}$. Note that this is less restrictive than the scaling $r \gtrsim (\log n/n)^{1/(d+4)}$, required for the upper bound. While Lemma 3.1 also requires $n \geq d \wedge 4\lambda^{-1}$, but this is a much milder condition when $\lambda \sim r^2$. Thus we deduce that (4.1) holds with high probability, when $r \gtrsim (\log n/n)^{1/(d+4)}$.

5. Algorithms and Experiments

We now turn to the algorithms for our boundary tests and related numerical experiments. After presenting the pseudocode for the boundary tests and briefly commenting on the computational complexity, we demonstrate the efficiency and accuracy of our results, focusing on domains with constant positive or negative curvatures. Again we stress that, while the rigorous theoretical results in Section 3 are established for the first-order test, we recommend the second-order test for practical purposes. As we will see, the second-order test takes into account the curvature, hence performs much better than the first-order test.

To begin, we present the pseudocodes for the first- and second-order boundary tests, and the generalization of the second-order test to point clouds supported on manifolds.

Algorithm 1 First-order boundary test

Input: The set of points $\mathcal{X} = \{x^1, \cdots, x^n\}$, and parameters $r, \varepsilon > 0$
Output: $T(x_k) = 1$ if x_k is a ε-boundary point, 0 if an ε-interior point

1: **for** $i = 1 \cdots n$ **do**
2: $\quad T(i) \leftarrow 1$
3: $\quad \hat{v}_r(x^i) \leftarrow \sum_{y \in B(x^i,r)\cap\mathcal{X}} (y - x^i)$
4: $\quad \hat{\nu}_r(x^i) \leftarrow \hat{v}_r(x^i)/|\hat{v}_r(x^i)|$
5: $\quad$ **if** $\max_{x^j \in B(x^i,r)\cap\mathcal{X}} (x^i - x^j)\cdot\hat{\nu}_r > \frac{3\varepsilon}{2}$ **then** $T(i) = 0$
6: $\quad$ **end if**
7: **end for**

We add that the algorithms can take a percentile $p\%$ as an input instead of ε, so that it outputs the top $p\%$ of points with smallest estimated distance. This may be easier to implement in practice than choosing ε, as the lower bound for ε depends not only on n but also on R, ρ and d. Theoretically, $p\%$ and ε are interchangeable; we may set the largest estimated distance within the $p\%$ percentile to equal to the threshold, $\frac{3\varepsilon}{2}$.

Algorithm 2 Second-order boundary test

Input: The set of points $\mathcal{X} = \{x^1, \cdots, x^n\}$, and parameters $r, \varepsilon > 0$
Output: $T(x_k) = 1$ if x_k is a ε-boundary point, 0 if an ε-interior point

for $i = 1 \cdots n$ **do**
  $\hat{\theta}(x^i) \leftarrow \sum_{j=1}^{n} \mathbb{1}_{B(x^i, r/2)}(x^j)$
  $\hat{v}_r^2(x^i) \leftarrow \sum_{x^j \in B(x^i, r) \cap \mathcal{X}} \frac{(x^j - x^i)}{\hat{\theta}(x^j)}$
  $\hat{\nu}_r^2(x^i) \leftarrow \hat{v}_r^2(x^i) / |\hat{v}_r^2(x^i)|$
end for
for $i = 1 \cdots n$ **do**
  for $j = 1 \cdots n$ **do**
    $\hat{\nu}_{r,test}^{ij} = \hat{\nu}_r^2(x^i) + \dfrac{\hat{\nu}_r^2(x^j) - \hat{\nu}_r^2(x^i)}{2} \mathbb{1}_{\mathbb{R}_+}(\hat{\nu}_r^2(x^i) \cdot \hat{\nu}_r^2(x^i))$
  end for
  if $\max_{x^j \in B(x^i, r) \cap \mathcal{X}} (x^i - x^j) \cdot \hat{\nu}_{r,test}^{ij} > \frac{3\varepsilon}{2}$ **then** $T(i) = 0$
  end if
end for

Algorithm 3 Second-order boundary test for point clouds supported on manifolds

Input: The set of points $\mathcal{X} = \{x^1, \cdots, x^n\}$, parameters $r, \varepsilon > 0$, and the dimension of the manifold m
Output: $T(x_k) = 1$ if x_k is a ε-boundary point, 0 if an ε-interior point

for $i = 1 \cdots n$ **do**
  $\hat{\theta}(x^i) \leftarrow \sum_{j=1}^{n} \mathbb{1}_{B(x^i, r/2)}(x^j)$
  $\hat{v}_r^2(x^i) \leftarrow \sum_{x^j \in B(x^i, r) \cap \mathcal{X}} \frac{(x^j - x^i)}{\hat{\theta}(x^j)}$
  $\hat{\nu}_r^2(x^i) \leftarrow \hat{v}_r^2(x^i) / |\hat{v}_r^2(x^i)|$
  $Y^i \leftarrow$ rangesearch(x^i, r)
  $Y_i \leftarrow Y_i - \overline{Y_i}$
  $\{v_1, \cdots, v_m\} \leftarrow$ eigenvectors associated to m largest eigenvalues of $(Y^i - x^i)^T (Y^i - x^i)$
  $T^i \leftarrow \text{Span}\{v_1, \cdots, v_m\}$
end for
for $i = 1 \cdots n$ **do**
  for $j = 1 \cdots n$ **do**
    $\hat{\nu}_{r,test}^{ij} = \hat{\nu}_r^2(x^i) + \dfrac{\hat{\nu}_r^2(x^j) - \hat{\nu}_r^2(x^i)}{2} \mathbb{1}_{\mathbb{R}_+}(\Pi^i(\hat{\nu}_r^2(x^i)) \cdot \Pi^i(\hat{\nu}_r^2(x^i)))$
  end for
  if $\max_{x^j \in B(x^i, r) \cap \mathcal{X}} \Pi^i[(x^i - x^j)] \cdot \hat{\nu}_{r,test}^{ij} > \frac{3\varepsilon}{2}$ **then** $T(i) = 0$
  end if
end for

Remark 5.1 (Computational complexity)**.** Noting that range search task is essentially equivalent to k-nearest neighbor search for suitable k, we briefly remark on the computational expense. The best rigorous upper bounds for computing all-kNN for n points in $\mathbb{R}^d$ known to us, without number of parallel processors growing with n, are $O(n(\log n)^{d-1})$ [8] and $O(kd^d n \log n)$ [9]. Note that the the suitable choice of k for us is $k \sim \omega_d r^d n$, which, under the optimal choice of the test radius $r = RC_r(\log n/n)^{\frac{1}{d+2}} \leq C\omega_d^{-1/d}(\log n/n)^{\frac{1}{d+2}}$

for our first-order test, has the following scaling in n and d

$$k \lesssim (\log n)^{\frac{d}{d+2}} n^{\frac{2}{d+2}}.$$

Please see Remark 3.4 for further details.

While the computational cost of exact all-kNN is not cheap, approximate all-kNN can be performed at nearly linear time in n. For instance, the algorithm suggested in [39] reports that empirical cost scales like $n^{1.14}$ on average with above 90 percent accuracy. Python GraphLearning [19] package the Approximate Nearest Neighbors algorithm (ANNOY) [10], which also provides close to linear scaling in n. △

Remark 5.2 (Intrinsic dimension of $\mathcal{M}$)**.** In practice the intrinsic dimension m of $\mathcal{M}$ often unknown. However there are many ways to recover this from the eigenvalues $\lambda_1 \leq \cdots \leq \lambda_d$ of the sample covariance matrix $(Y^i - x^i)(Y^i - x^i)^T$. There are two big drops in the eigenvalue distribution. Near the boundary, eigenvectors sufficiently parallel to the normal direction have smaller eigenvalues due to the absence of points one one side of $\partial\mathcal{M}$. However, this gap should not reduce the eigenvalues much more than halving. On the other hand, $\lambda_{m+1}, \cdots, \lambda_d$ are due to curvature, and thus are much smaller compared to the first m when curvature is bounded. Thus we may recover the dimension m by for instance, counting the number of eigenvalues before the steepest drop in ratio $\frac{\lambda_{i+1}}{\lambda_i}$. △

We now describe the setting of our numerical experiments. In Figures 5, 6, and 7 we consider two types of domains: a ball, and an annulus, both with reach $R = 0.5$. Recall that this means the ball has radius R and the annulus has inner and outer radii $R_1 = R$, $R_2 = 1.6R$. By the boundary of the ball mean the sphere, and by that of the annulus we refer only to the inner boundary $\{x : |x| = R_1\}$, so we can observe how the test performs when the curvature is negative. Thus we test only the points satisfying $|x| \in [R_1, R_2 - r]$.

We consider the density function ρ parametrized by the Lipschitz constant L. The sinusoidal density has the form

$$\rho(x) = \frac{1}{|\Omega|}\left(1 + \frac{1}{2}\sin(L|\Omega|x_1)\right), \tag{5.1}$$

so that $\sup_{x\in\Omega} |\partial_1 \rho(x)| = L$. Note that our theory in Section 3 applies to Lipschitz functions that are not necessarily of class C^1. Indeed, we note that results obtained using the triangular wave density were similar.

The boundary tests are as described in (1.20), where the first-order test ('1st') uses the distance estimator (1.12), and the second-order test ('2nd') uses the estimator (1.17).

Measuring the test error. Let $\varepsilon, r > 0$ be the boundary width and the test radius. Given a test we are considering, let the set of *tested boundary points* be the set of points in $\mathcal{X}$ where the test $\widehat{T}_{\varepsilon,r}$ defined in (1.20). The *tested interior points* is the complement of the tested ε-boundary points in $\mathcal{X}$. Let P be the number of tested boundary points and N the number of tested interior points:

$$P = \sharp\{x^i \in \mathcal{X} \,:\, \widehat{T}_{\varepsilon,r}(x^i) = 1\} \quad \text{and} \quad N = \sharp\{x^i \in \mathcal{X} \,:\, \widehat{T}_{\varepsilon,r}(x^i) = 0\}.$$

We measure the error rate in a different way than is standard in hypothesis testing. We do it in a way that measures better whether we succeeded in our stated goal to create a test that would identify a large percentage of points near the boundary and would not misidentify as boundary points almost any points deep in the interior. This is important to be able to accurately set boundary conditions for PDE.

Thus we refer to $\partial_\varepsilon\Omega = \{x \in \mathcal{X} \,:\, \mathrm{dist}\,(x,\Omega) \leq \varepsilon\}$ and $\Omega^\circ_{2\varepsilon} = \{x \in \mathcal{X} \,:\, \mathrm{dist}\,(x,\Omega) > 2\varepsilon\}$ as true boundary and true interior points, respectively. We refer to tested boundary points which lie in $\Omega^\circ_{2\varepsilon}$ as false positives and tested interior points which lie in $\partial_\varepsilon\Omega$ as false negatives. We denote the number of false positives and false negatives by

$$FP = \sharp\{x \in \mathcal{X} \cap \Omega^\circ_{2\varepsilon} \,:\, \widehat{T}_{\varepsilon,r}(x^i) = 1\} \quad \text{and} \quad FN = \sharp\{x \in \mathcal{X} \cap \partial_\varepsilon\Omega \,:\, \widehat{T}_{\varepsilon,r}(x^i) = 0\}.$$

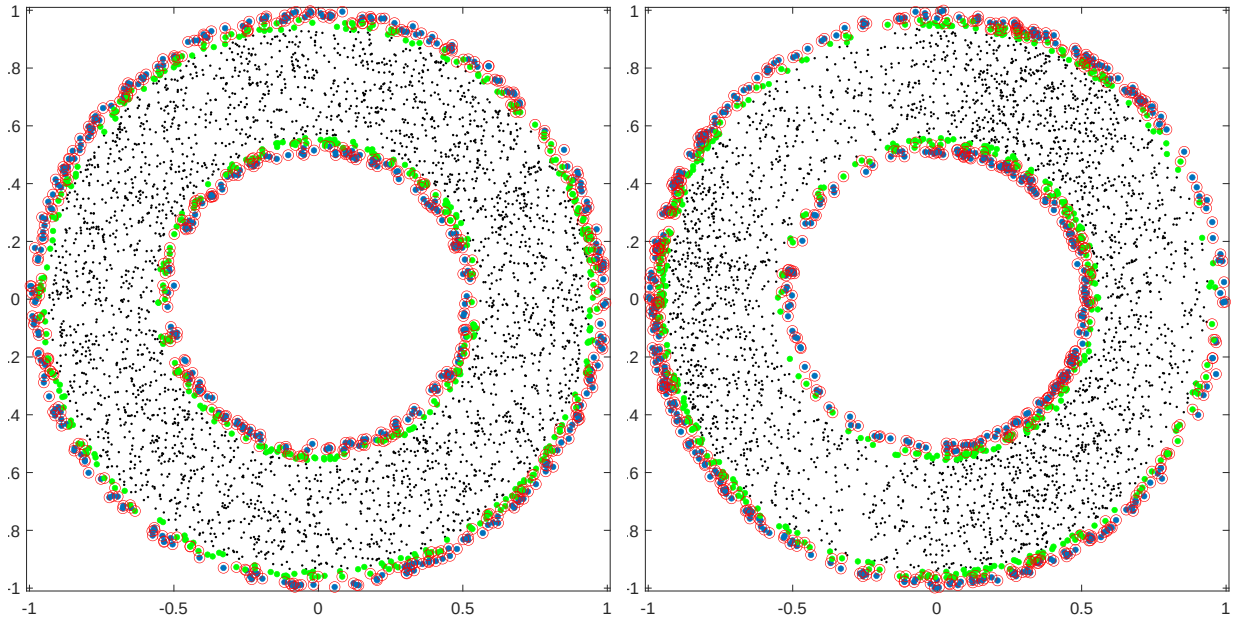


FIGURE 4. Boundary test on an annulus with inner and outer radii 0.5 and 0.8, respectively. $n = 2000$ points are drawn from uniform density on the left and sinusoidal density with $L = 2$ on the right. The point cloud is represented by black dots, while blue and green dots are the points whose true distance to the boundary are in $[0, \varepsilon)$ and $[\varepsilon, 2\varepsilon)$ respectively, for $\varepsilon = 0.03$. The red circles show the points identified by the 2nd order test, with $r = 0.18$, as boundary points. Observe that most blue dots are indeed correctly identified, and almost all points identified by the test are either blue or green dots.

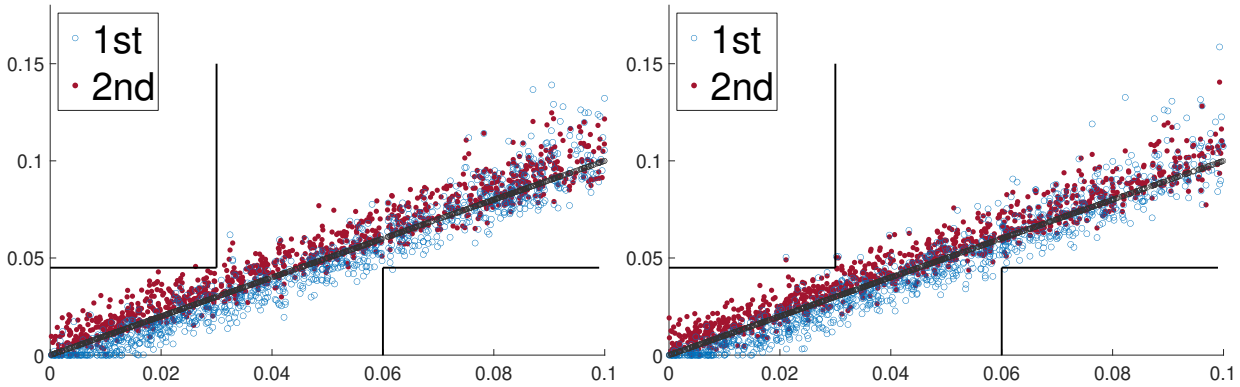
We denote by BP the number of true boundary points $BP = \sharp(\mathcal{X} \cap \partial_\varepsilon \Omega)$. We define false negative rate (FNR) and false positive rate (FPR) by

$$FNR = \frac{FN}{BP} \quad \text{and} \quad FPR = \frac{FP}{BP}.$$

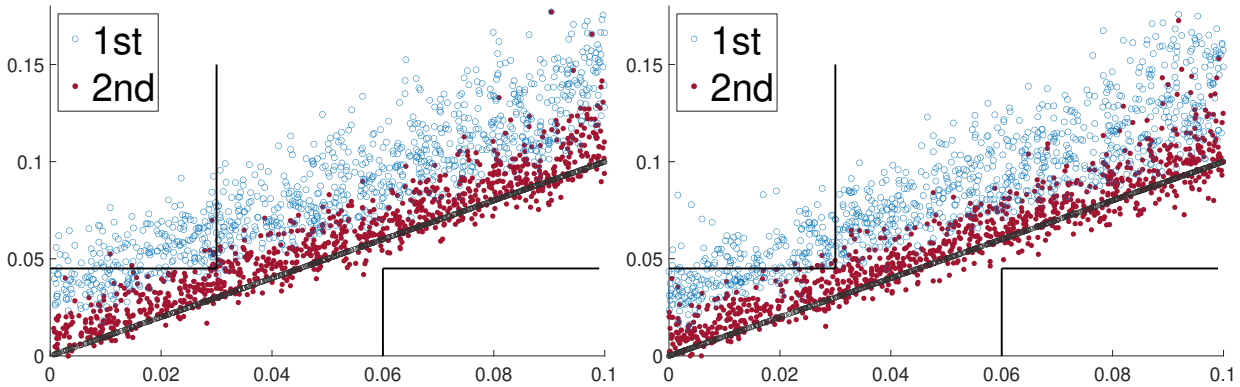
By the test failure rate (TFR) we mean the sum of FNR and FPR. Note the unusual definition of FPR. From the point of view hypothesis testing FPR would be the ratio of FP and true interior points. Given the large number of true interior points such measure of error would be small even if there is a significant number of points that were misidentified as boundary points. For our purposes it is important that the impact of false positives is small to the impact of the true positives. Thus we measure the error much more stringently and compare the number of the false positives to the number of true boundary points.

Remark 5.3 (Smoothing the estimated normals)**.** We observed that it is possible to further improve the accuracy of the estimated normals, thus of the test, if we smooth the normals in a small neighborhood using a suitable kernel. This reduces the variance, and tends to work well in combination with the second-order normal vector estimator (1.7), which limits the bias even in the presence of fluctuations in the density. However, when the second derivatives of the density ρ are large there can be a large bias in the estimated normal. In such cases we found that smoothing may worsen accuracy as errors accumulate. △

In Figure 7 we see that the first-order test for the ball shows $n \sim \varepsilon^{-2.5}$, corresponding almost exactly to the optimal theoretical scaling $\varepsilon \sim r^2$, $\varepsilon \sim (\log n/n)^{2/(d+2)}$ established in Corollary (3.5). We see similar trends with the second-order test for the ball. However, the first-order test shows extremely poor performance for the annulus, due to the negative curvature. For it to work, we need n large and ε, r small



(A) 3D ball with $R = 0.5$, and $n = 4000$. Left panel $L = 0$ and right panel $L = 2$.



(B) 3D annulus with $R_1 = 0.5$, $R_2 = 0.8$, and $n = 12000$. Left panel $L = 0$ and right panel $L = 2$.

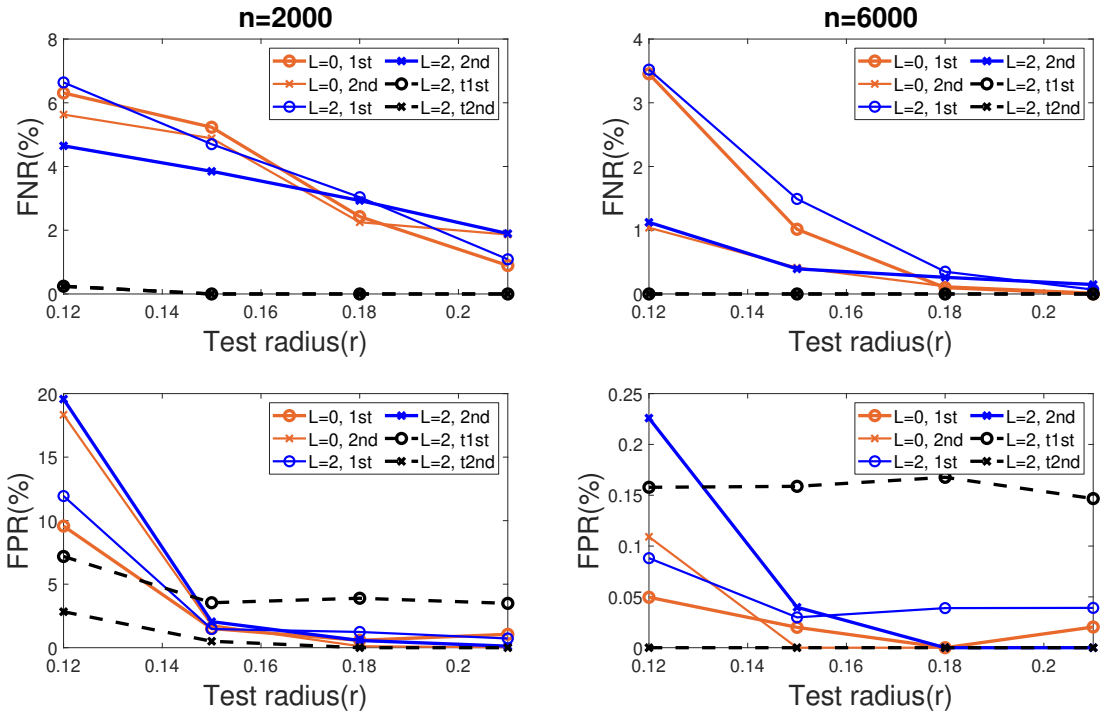
FIGURE 5. Plot of distance to boundary with $\varepsilon = 0.03$, $r = 0.18$. x- and y-axes each represents the true and the estimated distances respectively. 1st and 2nd refer to the order of the algorithm used. The boxes in the upper left and lower right corners specify the region for false negatives and false positives respectively. Only 1000 relevant points are plotted for improved visibility. Clear trend of 1st underestimating (resp. overestimating) the distance in a domain of positive (resp. negative) curvature is observed.

enough so that the curvature is negligible. On the other hand, the normalized second-order test shows exponential relationship between n and ε, although the exponent is worse than its counterpart for the ball.

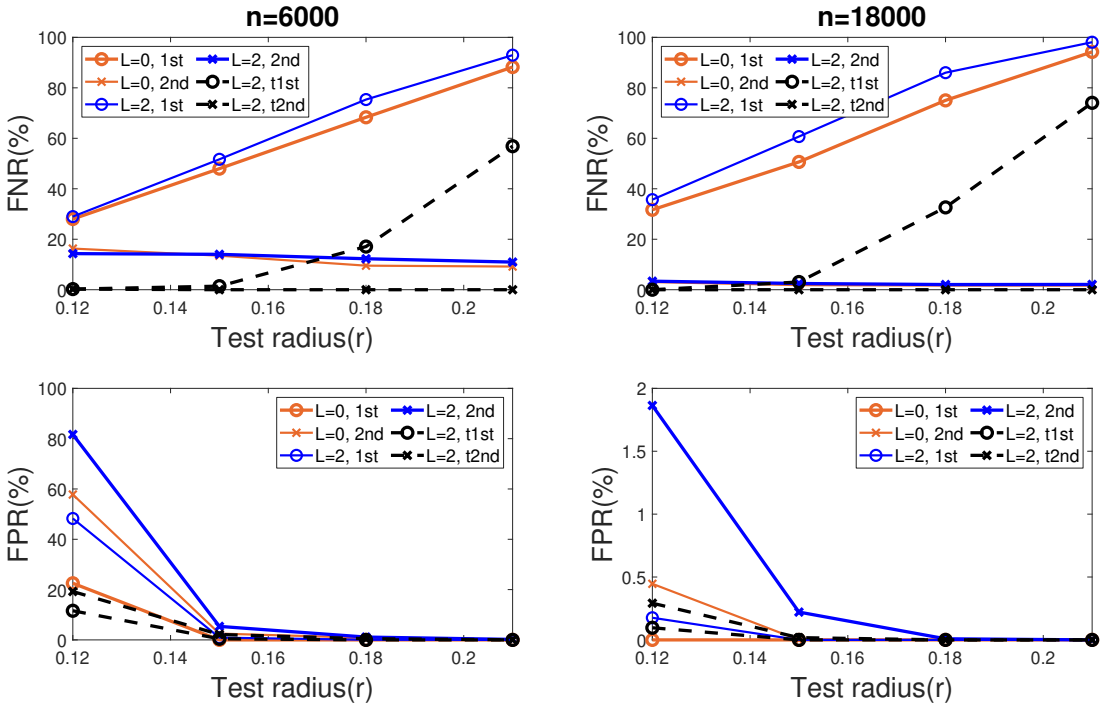
Remark 5.4 (Choice of parameters ε, r)**.** We have established in Theorem 3.3 that the optimal scaling for the first-order test is $r \sim (\log n/n)^{1/(d+2)}$ and $\varepsilon \sim r^2$ as $n \to \infty$. However, in practical situations, often n is not sufficiently large to guarantee that such scaling is realistic. Then how should we choose ε and r?

We observe from Figure 6 that the 2nd order test with the true normal vectors (t2nd) gives close to perfect results for both domains. This suggests that the 2nd order test for the most part resolves the challenge posed by curvature, which 1nd order test suffers from, and accurate estimation of normal vectors is key to boosting performance of the boundary test.

There are trade-offs in choosing r: clearly, when r is too small, the estimated normal is inaccurate due to high variance. On the other hand, large r leads to larger bias caused by curvature or fluctuations in the density. However, in Section 2.1 we have showed that the normalization by degree in the 2nd order estimator for the normal vector limits the bias to $O(r^2)$ even when ρ is non-uniform. Indeed, we see in Figure 6 (b) that FNR of 2nd is close to that of t2nd even in the presence of nontrivial fluctuation with $L = 2$ and relatively large r.



(A) Summary for the ball with sinusoidal density. $\varepsilon = 0.03$.



(B) Summary for the annulus with sinusoidal density. $\varepsilon = 0.03$.

FIGURE 6. Test failure rates depending on the test radius, number of points, sign of curvature, and the type of tests. $\varepsilon = 0.03$, $R = 0.5$. 1st, and 2nd are as in the previous experiments, while t1st and t2nd denote the first and second-order tests using the true normal vectors. Results have been averaged over 10 independent runs.

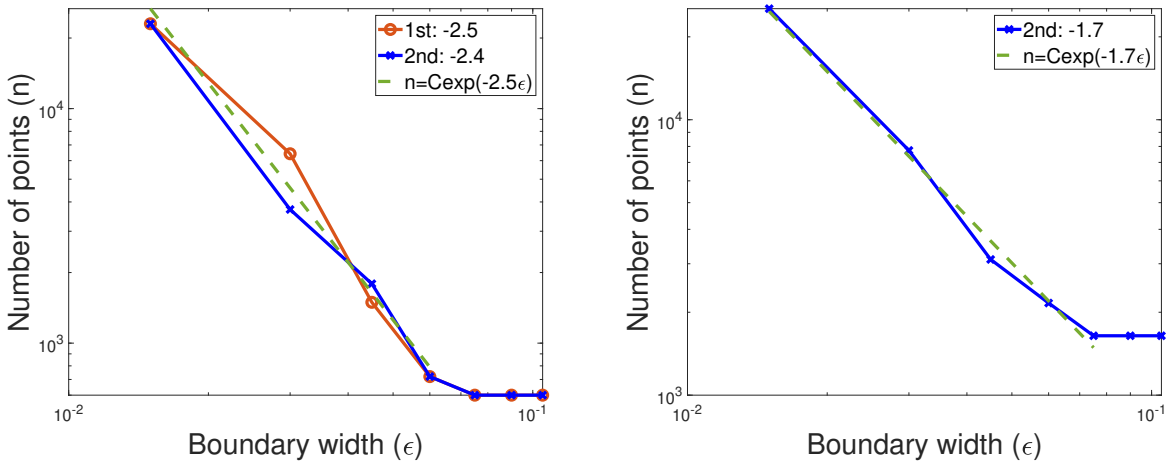


FIGURE 7. The plot shows the smallest number of points n for which TFR$\leq$ threshold, for given boundary width ε. (Left) Ball, threshold$= 0.5\%$, (Right) Annulus, threshold$= 10\%$. Maximal n considered was 20000 for the ball and 25000 for the annulus. We considered density with $L = 2$, and $r = \sqrt{\varepsilon}$. Number in the legend indicate the slope until n becomes stable. 1st order test applied to negatively curved domains have high false negatives, hence the TFR never went below the threshold. Hence the results from the 1st order test is not included for the annulus. Results have been averaged over 10 independent runs.

Thus, using the 2nd order test, it suffices to choose r in a reasonable range, so that $B(x^0, r)$ contains sufficiently many points, and r is not too close to the reach R, when a rough estimate of R is known. When the reach is completely unknown, then we recommend that r is taken to be the smallest so that each ball of radius r contains sufficient number of points.

Given r, ε should be chosen so that the ratio $\frac{|B(x^i, \frac{3\sqrt{2}\varepsilon}{2})|}{|B(x^i, r)|}$ of the volume of the balls is no larger than, say, $\frac{1}{2}$, to limit the number of false positives. The particular coefficient $\frac{3}{2}\sqrt{2}$ is is chosen as the threshold of our test is at $\frac{3\varepsilon}{2}$, and the $\frac{\hat{\nu}(x^i)+\hat{\nu}(x^j)}{2}$ can have magnitude as small as $\frac{1}{\sqrt{2}}$ when the sharp cutoff function is used. Note that for fixed r, ε, the ratio of the volumes decreases in dimension, as volume concentrates near the boundary of the ball in high dimensions. On the other hand, ε should be large enough so that the strips of height $\frac{\varepsilon}{2}$ and width around r contain enough points; this limits the possibility that points y with $d_\Omega(y)$ around 2ε are falsely tested positive. See Figure 3 and Lemma 3.1 for details. △

5.1. **Comparison with other approaches.** We limit our comparisons with other border detection algorithms to a couple of visual illustrations and remarks. The reason for this is that other algorithms were not designed to identify a boundary layer of desired width, ε, that our algorithm is designed for. Furthermore in most cases there is no straightforward way to adapt other algorithms to do detect a boundary layer of fixed width.

We compare our 2nd order boundary test with, tests based on the Devroye-Wise estimator (1.26) (DW), BRIM [66], and the statistic of Wu and Wu (WuWu) [74]. Recall that the Devroye-Wise estimator Ω_n approximates $\operatorname{supp}\rho$, and by boundary points we mean the points which contribute to the boundary $\partial\Omega_n$ – i.e. $x^i \in \mathcal{X}$ such that $\overline{B(x^i, \varepsilon)} \cap \partial\Omega_n \neq \emptyset$. We note that these are also exactly the data points that lie on the boundary estimator of Casal [67]. As discussed in Section 1.5, such points are precisely the boundary points of the α-shape [40, 41], a generalization of convex hull, with $\alpha = 1/\varepsilon$. In dimensions $d = 2, 3$, efficient algorithms for α-shapes exist, and we used the built-in function in MATLAB [73] to compute the contributing boundary points. For BRIM and WuWu, we implemented in MATLAB the algorithms described in [66] and [74] respectively.

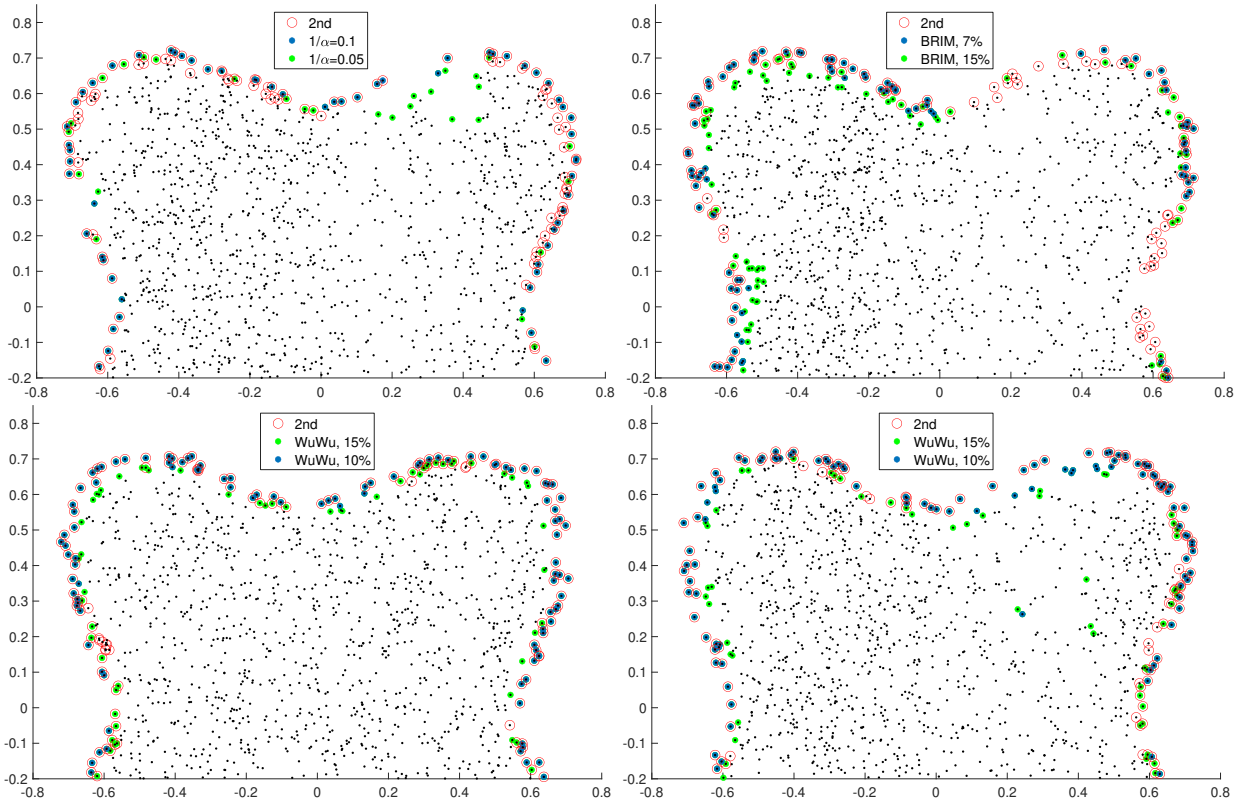


FIGURE 8. Comparison of tests for $n = 2000$ points drawn out of density ρ defined in (5.1), with (top, and bottom right) $L = 3$, and (bottom left) $L = 1$. The second-order test with $\varepsilon = 0.03$, $r = 0.18$ is compared with (top left) the Devroye-Wise estimator with radius α^{-1}, (top right) BRIM, and (bottom) WuWu. For BRIM and WuWu, the colored points are in the indicated top percentile according to the test statistic.

In Figure 8 we see that the Devroye-Wise estimator via α-shape effectively finds a thin boundary when a suitable α is used. The choice of appropriate α depends heavily on the density of the set of points considered. Smaller α^{-1} identifies more points, and in particular allows recognizing those where boundary has negative curvature. On the other hand, choosing α^{-1} too small increases the risk of falsely identifying interior points, lying in an area of low density, as boundary points. Indeed, the top plot of Figure 8 exhibits such a trade-off: the test with $\alpha^{-1} = 0.1$ misses boundary points around the concave indents, while choosing $\alpha^{-1} = 0.05$ results in false positives deep inside the interior. In the context of solving PDEs on graphs, such false positives can be catastrophic. As pointed out in Section 1.5, computing α-shapes becomes expensive when $d > 3$. We tested a commonly used alpha shapes package in Python [7] on a high performance computer with a 4.5GHz CPU, and found that the computational complexity in dimension for $n = 1000$ points independently and uniformly distributed on the unit ball in dimensions $d = 2$ up to $d = 9$ followed very closely to the exponential complexity $O(n^{0.23d})$. In terms of raw computational times, the alpha shape for $n = 1000$ points in dimension $d = 9$ took 110 minutes, and $d = 10$ and $d = 11$ would have taken roughly 12 and 77 hours, respectively. The memory requirements seem to grow very quickly as well, with $d = 8$ taking 13 GB and $d = 9$ requiring roughly 45 GB.

In contrast, BRIM easily generalizes to dimensions higher than 3. BRIM uses a similar basic idea as our approach: it approximate the inward normal direction. It does so by identifying the point $x^i \in B(x^0, r)$ maximizing $|B(x^i, r) \cap \mathcal{X}|$. To detect the boundary it compares the number of points in the normal direction and those opposite of it. The test is sensitive to variations in the density. Indeed the bottom plot of Figure

8 shows that BRIM identifies significantly more points on the left boundary, near which the density is high, than it does on the sparsely populated right.

WuWu also generalizes well to arbitrary dimension. Furthermore, it takes into account the curvature of the boundary by using spectral information of the 'sample covariance matrix' (see Section 1.5). We can see in Figure 8 that WuWu consistently detects points near negatively curved parts of the boundary. However, it is not as robust under fluctuations in density. Observe WuWu classifies considerably more points on the left side of the boundary, where points are densely distributed, compared to the right. Further, some interior points are in the top 15% according to the test statistic; this can be resolved by increasing k for kNN, but at the cost of successfully identifying fewer points close to the boundary.

We also ran experiments using the test statistic suggested by Aaron and Cholaquidis [3], but it did not perform well, as their statistic is designed to decide whether the manifold has a boundary or not, rather than to identify boundary points.

We stress again that all the other algorithms we compared were not designed for the task considered. We note that our method is as fast as any of the other methods and provides the best quality boundary for the task considered. Furthermore there is no error analysis that would suggest that any of the other methods are second-order accurate.

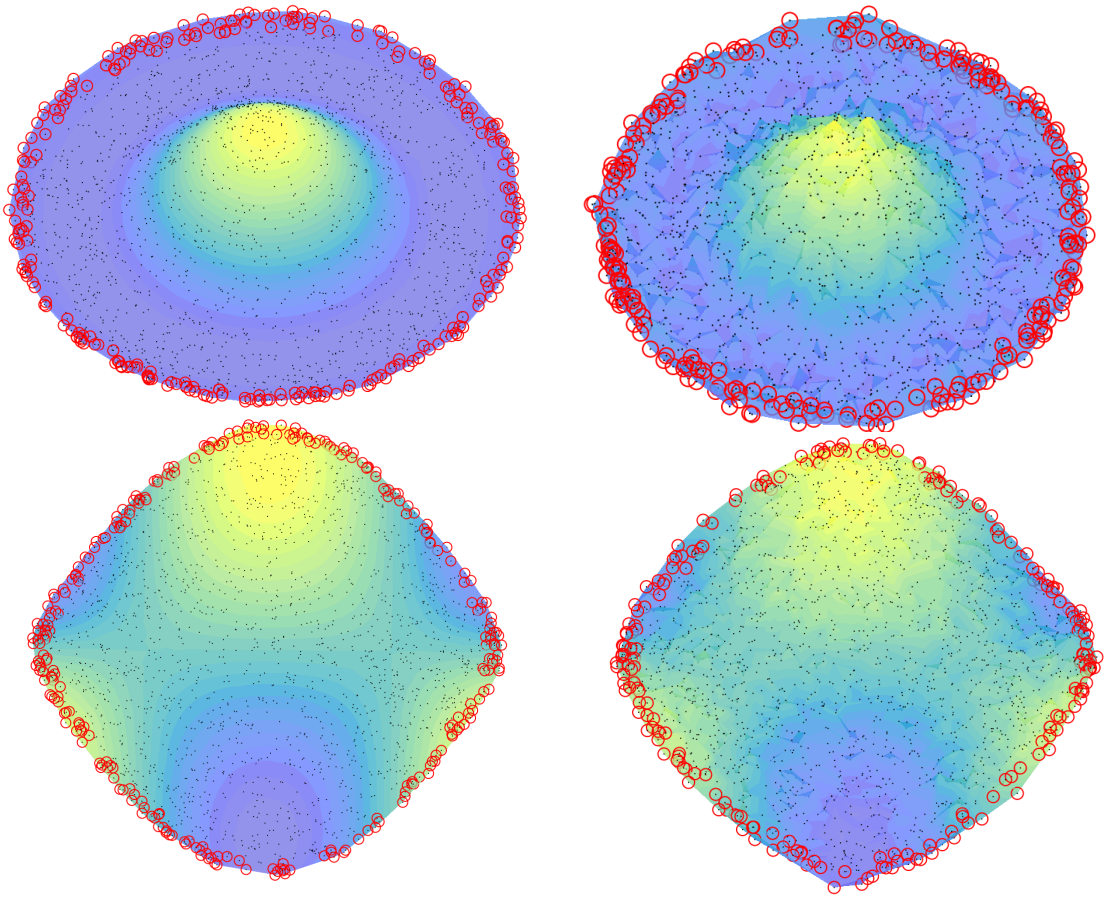


FIGURE 9. Boundary points of point clouds supported on 2-dimensional surfaces, identified using Algorithm 2. $n = 2000$, $r = 0.21$, $\varepsilon = 0.05$. Point clouds are marked in black, and the boundary points are circled in red. (Left) No additive noise. (Right) Additive Gaussian noise with standard deviation set as 1% of the diameter of the surface. Surfaces appear irregular as they are reconstructed from noisy samples.

6. Solving PDEs on data clouds

One immediate application of boundary detection is the ability to solve PDEs on point clouds with flexibility in the choice of boundary condition. All of the present approaches to solving PDEs on data clouds, where the boundary is not known in advance, rely on a variational description of the problem and thus result in natural variational boundary conditions. For the graph Laplacian this always yields homogeneous Neumann boundary conditions (see [24] for discussion of the graph Laplacian near the boundary). In this section, we show how we can use our boundary detection method, which includes an estimation of the normal vector to the boundary, to solve PDEs on point clouds with various boundary conditions, including Dirichlet, Neumann, oblique, and Robin problems. We then give applications to computing data-depth and medians on real datasets, and present intriguing numerical experiments on MNIST and FashionMNIST.

Throughout this section, we fix some additional notation. For $\varepsilon > 0$ we define

$$\partial_\varepsilon \Omega = \{x \in \Omega \,:\, \operatorname{dist}(x, \partial\Omega) \leq \varepsilon\}$$

and set $\Omega_\varepsilon = \Omega \setminus \partial_\varepsilon \Omega$. We recall that $\mathcal{X} = \{x^1, \dots, x^n\}$ is our point cloud, which is assumed to consist of independent and identically distributed random variables with density $\rho : \Omega \to \mathbb{R}$. We will place various assumptions on ρ throughout the section. We also assume we have an accurate estimation of the points from $\mathcal{X}$ that fall in the boundary tube $\partial_\varepsilon \Omega$. This is provided by our main results on boundary detection in Theorem 3.3 and Corollary 3.5. In order to make the results in this section as general as possible, we simply assume that we have computed a boundary set $\partial_\varepsilon \mathcal{X} \subset \mathcal{X}$ that satisfies

$$\mathcal{X}_\varepsilon \subset \Omega_\varepsilon \quad \text{and} \quad \partial_\varepsilon \mathcal{X} \subset \partial_{2\varepsilon} \Omega, \tag{6.1}$$

where $\mathcal{X}_\varepsilon = \mathcal{X} \setminus \partial_\varepsilon \mathcal{X}$.

6.1. The eikonal equation.

First, we consider extending Theorem 3.3 to estimate the distance function

$$d_\Omega(x) := \operatorname{dist}(x, \partial\Omega) \tag{6.2}$$

on the whole point cloud $\mathcal{X}$. We can do this by solving the graph eikonal equation

$$\left.\begin{aligned} \min_{y \in B_0(x^i, \varepsilon) \cap \mathcal{X}} \left\{ u_\varepsilon(y) - u_\varepsilon(x^i) + |y - x^i| \right\} = 0, \quad & \text{if } x^i \in \mathcal{X}_\varepsilon \\ u_\varepsilon(x^i) = 0, \quad & \text{if } x^i \in \partial_\varepsilon \mathcal{X}, \end{aligned}\right\} \tag{6.3}$$

where we write $B_0(x, \varepsilon) := B(x, \varepsilon) \setminus \{x\}$ for the punctured ball. The solution u_ε of the graph eikonal equation (6.3) is exactly the distance function on the graph with vertices $\mathcal{X}$ and edge weights $w_{ij} = |x^i - x^j|$ if $|x^i - x^j| \leq \varepsilon$, and $w_{ij} = \infty$ otherwise. When this graph is connected, the solution of (6.3) is unique. The solution of (6.3) can be computed with Dijkstra's algorithm in $O(nk \log(n))$ time, where k is an upper bound for the number of points in $B(x^i, \varepsilon) \cap \mathcal{X}$ over all i. We expect the solution u_ε converges to the distance function d_Ω as $\varepsilon \to 0$. Indeed this section is focused on proving this convergence with a quantitative $O(\varepsilon)$ error rate.

For (6.3) to be well-defined, we require the set $B_0(x^i, \varepsilon) \cap \mathcal{X}$ to be nonempty for all $x^i \in \mathcal{X}_\varepsilon$.

Proposition 6.1. *Let $n \geq 2$. The event that $B_0(x^i, \varepsilon) \cap \mathcal{X}$ is nonempty for all $x^i \in \mathcal{X}_\varepsilon$ has probability at least $1 - n \exp\left(-\frac{1}{2}\omega_d \rho_{min} n \varepsilon^d\right)$.*

Proof. By the *i.i.d.* law, the probability that $B_0(x^i, \varepsilon) \cap \mathcal{X}$ is empty conditioned on $x^i \in \mathcal{X}_\varepsilon$ is

$$\left(1 - \int_{B(x^i, \varepsilon)} \rho(x)\, dx\right)^{n-1} \leq \left(1 - \rho_{min} \omega_d \varepsilon^d\right)^{n-1} \leq \exp\left(-\rho_{min} \omega_d (n-1) \varepsilon^d\right).$$

The proof is completed by union bounding over $\mathcal{X}$, and using that $n - 1 \geq \frac{1}{2} n$ for $n \geq 2$. □

We briefly review some basic properties of the distance function. We recall a function $u : \Omega \to \mathbb{R}$ is *semiconcave* with constant C if $u - C|x|^2$ is concave. The distance function d_Ω is 1-Lipschitz and semiconcave with constant $1/R$ (see, e.g., [26]). By the Alexandrov theorem, a semiconcave function is twice differentiable almost everywhere in Ω. The distance function also satisfies the *dynamic programming principle*

$$d_\Omega(x) = \min_{y \in B(x,\varepsilon)} \big\{ d_\Omega(y) + |y - x| \big\}$$

for all balls $B(x, \varepsilon) \subset \Omega$. This can be rearranged into the form

$$\min_{y \in B(x,\varepsilon)} \big\{ d_\Omega(y) - d_\Omega(x) + |y - x| \big\} = 0. \tag{6.4}$$

Thus, the graph eikonal equation (6.3) is merely a discretization of the dynamic programming principle (6.4) to the point cloud $\mathcal{X}$. At any point $x \in \Omega$ where d_Ω is differentiable, we can Taylor expand d_Ω in (6.4) and compute the minimum explicitly to find that $|\nabla d_\Omega(x)| = 1$. If Ω is bounded, the distance function d_Ω always has points of nondifferentiability (for example at its maximum).

The equation $|\nabla u| = 1$ is referred to as the *eikonal* equation (more generally $|\nabla u| = f$). The distance function d_Ω can be interpreted as the unique *viscosity solution* of the eikonal equation. The viscosity solution is a type of weak solution to a partial differential equation (PDE) that allows non-differentiable functions to be solutions of first and second-order PDEs. In the case of the eikonal equation, and other first-order convex Hamilton-Jacobi equations, the viscosity solution coincides with the unique Lipschitz and *semiconcave* function that satisfies the PDE almost everywhere. We use the semiconcave interpretation here and do not discuss viscosity solutions directly. We refer the reader to [5, 16] for more details on viscosity solutions.

We now turn to convergence of the solution of the graph eikonal equation (6.3) to the distance function d_Ω. For this, we require a notion of asymptotic consistency.

Lemma 6.2. *Let $0 < t \le \frac{1}{d}$. The event that*

$$\min_{x \in B_0(x^i,\varepsilon) \cap \mathcal{X}} \Big\{ \lambda d_\Omega(x) - \lambda d_\Omega(x^i) + |x - x^i| \Big\} \le t\lambda\varepsilon + \frac{4\lambda\varepsilon^2}{R} - (\lambda - 1)\varepsilon \tag{6.5}$$

holds for all $\lambda \ge 1$ and $x^i \in \mathcal{X} \cap \Omega_\varepsilon$ has probability at least $1 - n \exp\left(-\frac{\omega_{d-1}}{4(d+1)} \rho_{min} n \varepsilon^d (2t)^{\frac{d+1}{2}} \right)$.

The proof of Lemma 6.2 requires some well-known properties of the distance function, which we summarize in the following Proposition, whose proof is postponed to the appendix.

Proposition 6.3. *Let $\varepsilon > 0$ and $x^0 \in \Omega_\varepsilon$. Let $x_* \in B(x^0, \varepsilon)$ such that*

$$d_\Omega(x_*) = \min_{B(x^0,\varepsilon)} d_\Omega. \tag{6.6}$$

Then $x_* \in \partial B(x^0, \varepsilon)$, $d_\Omega(x_*) = d_\Omega(x^0) - \varepsilon$, and for all $x \in \Omega$ we have

$$d_\Omega(x) - d_\Omega(x_*) \le p \cdot (x - x_*) + \frac{1}{R}|x - x_*|^2, \quad \text{where } p = \frac{x^0 - x_*}{\varepsilon}. \tag{6.7}$$

Proof of Lemma 6.2. Let $\lambda \ge 1$ and let $x^i_* \in B(x^i, \varepsilon)$ such that $d_\Omega(x^i_*) = \min_{B(x^i,\varepsilon)} d_\Omega$. For $x^i \in \mathcal{X} \cap \Omega_\varepsilon$ we can apply Proposition 6.3 to obtain

$$\begin{aligned} \lambda d_\Omega(x) - \lambda d_\Omega(x^i) + |x - x^i| &= \lambda d_\Omega(x) - \lambda d_\Omega(x^i_*) - \lambda\varepsilon + |x - x^i| \\ &\le \lambda\, p \cdot (x - x^i_*) + \frac{\lambda}{R}|x - x^i_*|^2 - \lambda\varepsilon + |x - x^i| \end{aligned}$$

for any $x \in B(x^i, \varepsilon)$, where $p = (x^i - x^i_*)/\varepsilon$. Since $|x - x^i_*| \le 2\varepsilon$ and $|x - x^i| \le \varepsilon$ we obtain

$$\lambda d_\Omega(x) - \lambda d_\Omega(x^i) + |x - x^i| \le \lambda\, p \cdot (x - x^i_*) + \frac{4\lambda\varepsilon^2}{R} - (\lambda - 1)\varepsilon. \tag{6.8}$$

For $0 \leq t \leq 1$ define the set

$$A_t^i = \left\{ x \in B(x^i, \varepsilon) \,:\, p \cdot (x - x_*^i) \leq t\,\varepsilon \right\}.$$

If (6.5) fails to hold, then it follows from (6.8) that the set $\mathcal{X} \cap A_t^i$ is empty. The remainder of the proof is focused on estimating the volume $|A_t^i|$ in order to control the probability that $\mathcal{X} \cap A_t^i$ is empty.

The measure of A_t^i is unchanged by taking $x^i = 0$, $x_*^i = \varepsilon e_d$, and $p = -e_d$, which gives

$$|A_t^i| = \left| B(0,\varepsilon) \cap \{x_d \geq (1-t)\varepsilon\} \right| = \varepsilon^d \left| B(0,1) \cap \{x_d \geq 1 - t\} \right|.$$

We lower bound the volume of the spherical cap by integrating

$$\begin{aligned} \left| B(0,1) \cap \{x_d \geq 1-t\} \right| &= \int_{1-t}^1 \omega_{d-1} (1 - x_d^2)^{\frac{d-1}{2}}\, dx_d \\ &\geq \int_{1-t}^1 \omega_{d-1} (1 - x_d^2)^{\frac{d-1}{2}} x_d\, dx_d \\ &= \frac{\omega_{d-1} (2t)^{\frac{d+1}{2}}}{d+1} \left(1 - \tfrac{t}{2}\right)^{\frac{d+1}{2}}. \end{aligned}$$

Now, since $t \mapsto \left(1 - \frac{t}{2}\right)^{\frac{d+1}{2}}$ is convex we have

$$\left(1 - \tfrac{t}{2}\right)^{\frac{d+1}{2}} \geq 1 - \left(\tfrac{d+1}{4}\right) t \geq \frac{1}{2},$$

provided $t \leq \frac{2}{d+1}$, which is satisfied when $t \leq \frac{1}{d}$. This yields

$$|A_t^i| \geq \frac{\omega_{d-1} \varepsilon^d (2t)^{\frac{d+1}{2}}}{2(d+1)} =: \Lambda.$$

Hence, the event that $\mathcal{X} \cap A_t^i$ is empty has probability bounded by

$$(1 - \rho_{min}\Lambda)^{n-1} \leq \exp\left(-\rho_{min}(n-1)\Lambda\right) \leq \exp\left(-\frac{1}{2}\rho_{min} n \Lambda\right),$$

since $n \geq 2$ so $n - 1 \geq \frac{1}{2}n$. The proof is completed by union bounding over $\mathcal{X}$. □

We now prove convergence of u_ε to the distance function d_Ω as $\varepsilon \to 0$ and $n \to \infty$.

Theorem 6.4. *Assume $\varepsilon \leq \frac{R}{8}$ and (6.1) holds. Let u_ε solve (6.3) and let $0 < t \leq \min\{\frac{1}{d}, \frac{1}{2} - \frac{4\varepsilon}{R}\}$. Then*

$$-2\varepsilon \leq u_\varepsilon - d_\Omega \leq 2 d_\Omega \left(t + \frac{4\varepsilon}{R}\right) \quad \text{on } \mathcal{X} \tag{6.9}$$

holds with probability at least $1 - 2n \exp\left(-\frac{\omega_{d-1}}{4(d+1)} \rho_{min} n \varepsilon^d (2t)^{\frac{d+1}{2}}\right)$.

Proof. The proof is split into three steps.

1. Let $0 < t \leq \frac{1}{d}$ and assume the results of Lemma 6.2 hold. Let $\lambda \geq 1$ and let $x^i \in \mathcal{X}$ such that $u_\varepsilon - \lambda d_\Omega$ attains its maximum over $\mathcal{X}$ at x^i. Then we have that

$$u_\varepsilon(x^j) - u_\varepsilon(x^i) \leq \lambda d_\Omega(x^j) - \lambda d_\Omega(x^i)$$

for all j. If $x^i \in \mathcal{X}_\varepsilon$, then since u_ε satisfies (6.3) we have

$$0 = \min_{y \in B_0(x^i,\varepsilon) \cap \mathcal{X}} \left\{ u_\varepsilon(y) - u_\varepsilon(x^i) + |y - x^i| \right\} \leq \min_{y \in B_0(x^i,\varepsilon) \cap \mathcal{X}} \left\{ \lambda d_\Omega(y) - \lambda d_\Omega(x^i) + |y - x^i| \right\}.$$

By (6.1) we have $x^i \in \Omega_\varepsilon$, which allows us to apply Lemma 6.2 to obtain that

$$0 \leq t\lambda\varepsilon + \frac{4\lambda\varepsilon^2}{R} - (\lambda - 1)\varepsilon.$$

This cannot hold when when $\lambda > \left(1 - t - \frac{4\varepsilon}{R}\right)^{-1}$ and $t + \frac{4\varepsilon}{R} < 1$. For any such λ we must have $x^i \in \partial_\varepsilon \mathcal{X}$ and so

$$\max_{\mathcal{X}}(u_\varepsilon - \lambda d_\Omega) = \max_{\partial_\varepsilon \mathcal{X}}(u_\varepsilon - \lambda d_\Omega) \leq 0.$$

It follows that $u_\varepsilon - d_\Omega \leq (\lambda - 1)d_\Omega$ on $\mathcal{X}$. Sending $\lambda \to \left(1 - t - \frac{4\varepsilon}{R}\right)^{-1}$ we obtain

$$u_\varepsilon - d_\Omega \leq d_\Omega \left[\left(1 - t - \frac{4\varepsilon}{R}\right)^{-1} - 1\right] \text{ on } \mathcal{X}.$$

The proof of this direction is completed by using the inequality

$$(1 - x)^{-1} - 1 \leq 2x \text{ for } 0 \leq x \leq \tfrac{1}{2}$$

and imposing the additional restriction that $t + \frac{4\varepsilon}{R} \leq \frac{1}{2}$ to simplify the right hand side.

2. For the other direction, let $0 < \lambda < 1$. Since d_Ω is 1-Lipschitz we have

$$\min_{y \in B_0(x^i, \varepsilon) \cap \mathcal{X}} \left\{\lambda d_\Omega(y) - \lambda d_\Omega(x^i) + |y - x^i|\right\} \geq (1 - \lambda) \min_{y \in B_0(x^i, \varepsilon) \cap \mathcal{X}} \left\{|y - x^i|\right\} > 0, \tag{6.10}$$

provided $B_0(x^i, \varepsilon) \cap \mathcal{X}$ is not empty. Thus, by (6.1) and Proposition 6.1, (6.10) holds for all $x^i \in \mathcal{X}_\varepsilon$ with probability at least $1 - n \exp\left(-\frac{1}{2}\omega_d \rho_{min} n \varepsilon^d\right)$. Let $x^i \in \mathcal{X}$ such that $u_\varepsilon - \lambda d_\Omega$ attains its minimum over $\mathcal{X}$ at x^i. By an argument similar to the first part of the proof, (6.3) and (6.10) imply that $x^i \in \partial_\varepsilon \mathcal{X}$. Therefore $u_\varepsilon(x^i) = 0$ and by (6.1) we have $x^i \in \partial_{2\varepsilon}\Omega$. It follows that

$$\min_{x \in \mathcal{X}}(u_\varepsilon(x) - \lambda d_\Omega(x)) = -\lambda d_\Omega(x^i) \geq -2\lambda\varepsilon.$$

Sending $\lambda \to 1^-$ completes the proof.

3. Union bounding over the events in steps 1 and 2 above, the results of the theorem hold with probability at least

$$1 - n \exp\left(-\frac{\omega_{d-1}}{4(d+1)} \rho_{min} n \varepsilon^d (2t)^{\frac{d+1}{2}}\right) - n \exp\left(-\frac{1}{2}\omega_d \rho_{min} n \varepsilon^d\right).$$

The first exponential is larger, provided

$$\frac{\omega_{d-1}}{2(d+1)}(2t)^{\frac{d+1}{2}} \leq \omega_d.$$

Recalling $\omega_{d-1}/\omega_d \leq \sqrt{d}$, this is true when $2t \leq 1$, which is implied by the assumption that $t \leq \frac{1}{d}$ and $d \geq 2$. Therefore, (6.9) holds with probability at least $1 - 2n \exp\left(-\frac{\omega_{d-1}}{4(d+1)} \rho_{min} n \varepsilon^d (2t)^{\frac{d+1}{2}}\right)$. □

Remark 6.5. We now provide an interpretation of the result of Theorem 6.4. To obtain the conditions under which the error rate is linear in ε we take $t = \varepsilon$ and obtain

$$-2\varepsilon \leq u_\varepsilon - d_\Omega \leq 2d_\Omega \left(1 + \frac{4}{R}\right)\varepsilon$$

holds with probability at least $1 - 2n^{-2}$ provided that the length scale ε satisfies:

$$\varepsilon \geq \left(\frac{6(d+1)\log(n)}{2^{\frac{d+1}{2}} \omega_{d-1} \rho_{min} n}\right)^{\frac{2}{3d+1}}. \tag{6.11}$$

Taking the smallest allowable ε above, we obtain that u_ε converges to the distance function d_Ω at a convergence rate of $\mathcal{O}(n^{-2/(3d+1)})$, up to logarithmic factors. We mention that we have numerically seen convergence rates closer to $\mathcal{O}(\varepsilon^2)$ for ε much larger than the lower bound in (6.11). This may indicate that, in practice, a sharper convergence rate, as a function of n, could be obtained by choosing larger value for ε.

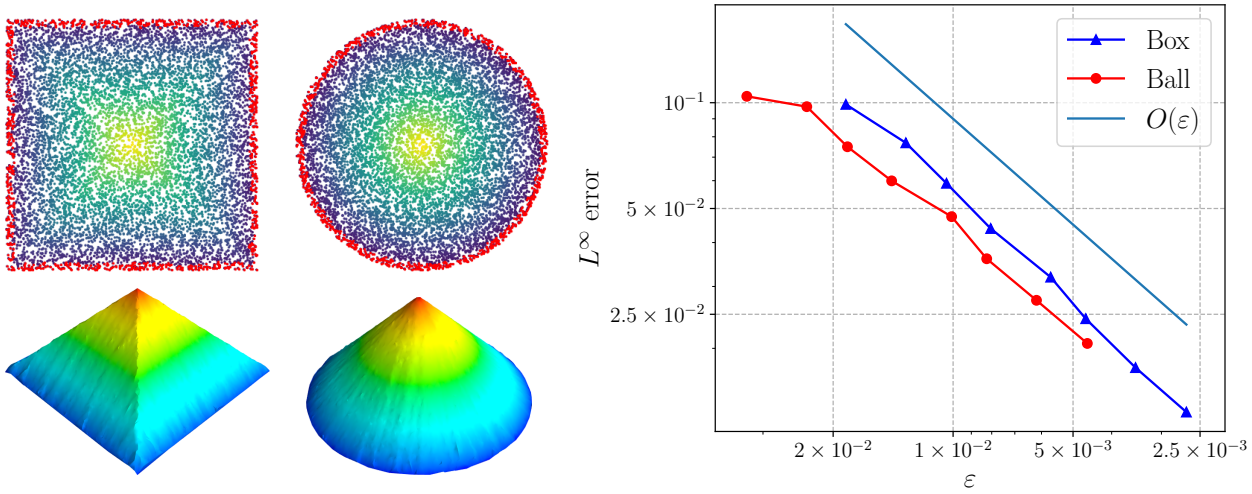


FIGURE 10. Plots of the solution to the graph eikonal equation (6.3) for $n = 10^4$ for both the box and ball domains, and error plots for varying ε averaged over 100 trials. The red points indicate the detected boundary points used in solving (6.3). We see convergence rates better than the linear $O(\varepsilon)$ rate guaranteed by Theorem 6.4.

To obtain a sufficient condition for uniform convergence alone we need conditions under which we can take $t_n \to 0$ as $n \to \infty$ and $\varepsilon_n \to 0$ for the estimate in Theorem 6.4 to hold with high probability. We see that this is possible whenever

$$\lim_{n\to\infty} \frac{n\varepsilon_n^d}{\log(n)} = \infty. \tag{6.12}$$

Then by the Borel-Cantelli lemma we have that $u_{\varepsilon_n} \to d_\Omega$ uniformly on $\mathcal{X}$ as $n \to \infty$ with probability one. △

6.1.1. *Numerical results.* We tested the $O(\varepsilon)$ convergence rate from Theorem 6.4 on a box $\Omega = [0,1]^2$ and ball $\Omega = B(0,1)$ domain. We used $n = 2^{10}$ up to $n = 2^{17} = 131,072$ *i.i.d.* random variables uniformly distributed on the domain, and chose ε adaptively based on the distance to the k^{th} nearest neighbor, where $k = 10n^{\frac{1}{5}}$. This is equivalent to the scaling $\varepsilon \sim n^{-\frac{2}{5}}$, since $k \sim n\varepsilon^2$. We detected the boundary by thresholding $\hat{d}_r(x)$ at $\frac{3\varepsilon}{2}$, where r is the distance from x to its k^{th} nearest neighbor, and ε satisfies $36\pi\rho n\varepsilon^2 = k$. In Figure 10 we show the solution of (6.3) for $n = 10^4$ as both a colored point cloud, and visualized as a surface, computed by constructing a triangulated mesh over the point cloud. In the plot in Figure 10 we show the L^∞ error $|u_\varepsilon - d_\Omega|$ versus ε averaged over 100 trials. Both domains track very closely to the theoretical $O(\varepsilon)$ convergence rates.

6.2. **Second-order equations.** We now turn to second-order equations on point clouds with general boundary conditions. In particular, we show how our estimation $\hat{\nu}_\varepsilon$ of the inward unit normal vector ν can be used to set general boundary conditions involving normal derivatives. We recall that Theorem 2.6 shows that $\hat{\nu}_\varepsilon$ is an $O(\varepsilon)$ approximation of ν with high probability. In order to state the results in the most general setting, we simply assume there exists a constant C_ν such that

$$|\hat{\nu}_\varepsilon(x^i) - \nu(x^i)| \leq C_\nu \varepsilon \tag{6.13}$$

for all $x^i \in \mathcal{X} \cap \partial_{2\varepsilon}\Omega$. We recall that Theorem 2.6 shows that the bound (6.13) holds with high probability as long as $\varepsilon \geq C(\log n/n)^{1/(d+2)}$. This lower bound on ε is also required for all the results in this section to hold with high probability. Indeed, Theorems 6.8 and 6.9 both require $n\varepsilon^{d+2} \geq C \log n$ for a sufficently large constant C, which amounts to the same lower bound on ε up to constants.

The graph PDEs we solve will involve the graph Laplacian $\mathcal{L}_\varepsilon$, which is defined by

$$\mathcal{L}_\varepsilon u(x^i) = \frac{2}{\sigma_\eta n \varepsilon^{d+2}} \sum_{j=1}^{n} \eta\left(\frac{|x^i - x^j|}{\varepsilon}\right)(u(x^j) - u(x^i)), \tag{6.14}$$

where $\sigma_\eta = \int_{\mathbb{R}^d} \eta(|z|) z_1^2 \, dz$, and η is smooth, compactly supported on $[0,1]$, and satisfies $\int_{\mathbb{R}^d} \eta(|z|)\, dz = 1$. We define the normal derivative $\nabla_\nu u(x) = \nabla u(x) \cdot \nu$ and the approximate normal derivative $\widehat{\nabla}_\nu$ by

$$\widehat{\nabla}_\nu u(x^i) = \frac{u(p_n(x^i + \varepsilon \hat{\nu}_\varepsilon(x^i))) - u(x^i)}{\varepsilon}, \tag{6.15}$$

where $p_n : \Omega \to \mathcal{X}$ is the closest point map. We consider the following graph Poisson equation with Robin-type boundary conditions

$$\left.\begin{aligned} \mathcal{L}_\varepsilon u(x^i) &= f(x^i), \quad \text{if } x^i \in \mathcal{X}_\varepsilon \\ \gamma u(x^i) - (1-\gamma)\widehat{\nabla}_\nu u(x^i) &= g(x^i), \quad \text{if } x^i \in \partial_\varepsilon \mathcal{X}. \end{aligned}\right\} \tag{6.16}$$

Here, $\gamma \in (0,1]$ and f and g are given smooth functions. In this section, we show that the solution of (6.16) converges as $n \to \infty$ and $\varepsilon \to 0$ to the solution of the Robin problem

$$\begin{cases} -\rho^{-1}\mathrm{div}(\rho^2 \nabla u) = f, & \text{in } \Omega \\ \gamma u - (1-\gamma)\nabla_\nu u = g, & \text{on } \partial\Omega. \end{cases} \tag{6.17}$$

Remark 6.6. We note that in order to solve the graph PDE (6.16) given a *nonconstant* boundary condition $g : \partial\Omega \to \mathbb{R}$, we need a way to define an extension $g_\varepsilon : \partial_{2\varepsilon}\Omega \to \mathbb{R}$ that is uniformly close to g within the boundary tube $\partial_{2\varepsilon}\Omega$. One way to do this is to define the closest point extension $g_\varepsilon(x) = g(x_*)$ where $x_* = \mathrm{argmin}_{y \in \partial\Omega}|x - y|$. The closest point x_* is unique for $x \in \partial_{2\varepsilon}\Omega$ when $2\varepsilon < R$ and if g is Lipschitz then $|g_\varepsilon(x) - g(x_*)| \le C\varepsilon$ for $x \in \partial_{2\varepsilon}\Omega$. It is important to note, however, that the closest point extension requires knowledge of the boundary $\partial\Omega$. In applications where the boundary $\partial\Omega$ is not known *a priori*, and is instead estimated from the point cloud, such as in data depth in machine learning, we can only handle constant boundary conditions (i.e., $g = 0$ on $\partial\Omega$ for data depth). △

Throughout this section we assume $\partial\Omega$ and ρ are smooth. By elliptic regularity, the solution u of (6.17) is smooth. The constants in this section will be denoted by $C, C_1, C_2, \cdots > 0$, and may depend on $\gamma, u, d, f, g, \rho, \Omega$ and $\partial\Omega$, and can change from line to line.

The proof of convergence is based on a maximum principle for (6.16).

Lemma 6.7. *If u satisfies*

$$\left.\begin{aligned} -\mathcal{L}_\varepsilon u(x^i) &< 0, \quad \textit{if } x^i \in X_\varepsilon \\ \gamma u(x^i) - (1-\gamma)\widehat{\nabla}_\nu u(x^i) &\le 0, \quad \textit{if } x^i \in \partial_\varepsilon \mathcal{X} \end{aligned}\right\} \tag{6.18}$$

then $u \le 0$ on $\mathcal{X}$.

Proof. Let us write $w_{ij} = \eta\left(\frac{|x^i - x^j|}{\varepsilon}\right)$ and $d_i = \sum_{j=1}^n w_{ij}$. Then by (6.18) we have

$$d_i u(x^i) - \sum_{j=1}^{n} w_{ij} u(x^j) = \sum_{j=1}^{n} w_{ij}(u(x^i) - u(x^j)) < 0$$

for all $x^i \in X_\varepsilon$. It follows that $d_i > 0$, and so $u(x^i) < \frac{1}{d_i}\sum_{j=1}^n w_{ij} u(x^j)$. Therefore, u attains its maximum over $\mathcal{X}$ at some $x^i \in \partial_\varepsilon \mathcal{X}$, and so

$$\gamma u(x^i) \le (1-\gamma)\frac{u(p_n(x^i + \varepsilon\hat{\nu}_\varepsilon(x^i))) - u(x^i)}{\varepsilon} \le 0.$$

Since $\gamma > 0$ we have $u(x^i) \le 0$. □

The convergence proof also requires pointwise consistency for the graph Laplacian. We refer to [18, Remark 5.26] for the following result.

Theorem 6.8. *Let $u \in C^4(\Omega)$, $\varepsilon > 0$ and $0 < \lambda \leq \varepsilon^{-1}$. Then*

$$\max_{x^i \in \Omega_\varepsilon \cap \mathcal{X}} \left| \mathcal{L}_\varepsilon u(x^i) - \rho(x^i)^{-1} div(\rho^2 \nabla u)|_{x_i} \right| \leq C_1 \|u\|_{C^4(\Omega)} (\varepsilon^2 + \lambda) \tag{6.19}$$

holds with probability at least $1 - 2n \exp\left(-C_2 n \varepsilon^{d+2} \lambda^2\right)$.

We now establish our main convergence result in this section.

Theorem 6.9. *Assume (6.1) and (6.13). Let $\varepsilon > 0$ and assume $C_\nu \varepsilon \leq 1$. Let u be the solution of (6.17) with $\gamma > 0$, and let u_ε satisfy (6.16). Then for any $0 < \lambda \leq \varepsilon^{-1}$ and $t > 0$, the event that*

$$|u(x^i) - u_\varepsilon(x^i)| \leq C \left(\|\gamma u - (1-\gamma) \nabla_\nu u - g\|_{L^\infty(\partial_{2\varepsilon}\Omega)} + (1-\gamma)(t + C_\nu \varepsilon + \varepsilon) + \varepsilon^2 + \lambda \right) \tag{6.20}$$

holds for all $x^i \in \mathcal{X}$ has probability at least $1 - n \exp\left(-\frac{1}{6}\omega_d \rho_{min} n \varepsilon^d t^d\right) - 2n \exp\left(-Cn\varepsilon^{d+2}\lambda^2\right)$.

Proof. The proof is split into three steps.

1. Note that $x^i + \varepsilon\nu \in \Omega_\varepsilon$. By (6.13) we have

$$|x^i + \varepsilon \hat{\nu}_\varepsilon(x^i) - (x^i + \varepsilon \nu)| = \varepsilon |\hat{\nu}_\varepsilon - \nu| \leq C_\nu \varepsilon^2.$$

Since $C_\nu \varepsilon \leq 1$ we have $x^i + \varepsilon \hat{\nu}_\varepsilon(x^i) \in \Omega$. Therefore, we can compute

$$\begin{aligned}
\widehat{\nabla}_\nu u(x^i) &= \frac{u(p_n(x^i + \varepsilon \hat{\nu}_\varepsilon(x^i))) - u(x^i)}{\varepsilon} \\
&= \frac{u(x^i + \varepsilon \nu(x^i)) - u(x^i)}{\varepsilon} + \mathcal{O}\left(\varepsilon^{-1} |p_n(x^i + \varepsilon \hat{\nu}_\varepsilon(x^i)) - (x^i + \varepsilon \hat{\nu}_\varepsilon(x^i))| + C_\nu \varepsilon \right) \\
&= \nabla_\nu u(x^i) + \mathcal{O}\left(\varepsilon^{-1} |p_n(x^i + \varepsilon \hat{\nu}_\varepsilon(x^i)) - (x^i + \varepsilon \hat{\nu}_\varepsilon(x^i))| + C_\nu \varepsilon + \varepsilon \right).
\end{aligned}$$

Let $t \geq 0$. If $|p_n(x^i + \varepsilon \hat{\nu}_\varepsilon(x^i)) - (x^i + \varepsilon \hat{\nu}_\varepsilon(x^i))| \geq t\varepsilon$ then the set $B(x^i + \varepsilon \hat{\nu}_\varepsilon(x^i), t\varepsilon) \cap \mathcal{X}$ is empty, which by Lemma 2.1 has probability less than $1 - \exp\left(-\frac{1}{3}\omega_d \rho_{min}(n-1)\varepsilon^d t^d\right)$. Union bounding over x^i and using that $n - 1 \geq \frac{1}{2}n$ for $n \geq 2$, we find that

$$\widehat{\nabla}_\nu u(x^i) = \nabla_\nu u(x^i) + \mathcal{O}\left(t + C_\nu \varepsilon + \varepsilon\right)$$

holds for all $x^i \in \partial_\varepsilon \mathcal{X} \subset \partial_{2\varepsilon}\Omega$ with probability at least $1 - n \exp\left(-\frac{1}{6}\omega_d \rho_{min} n \varepsilon^d t^d\right)$. A similar computation can be made for φ, and so we find that

$$|\widehat{\nabla}_\nu \varphi(x^i) - \nabla_\nu \varphi(x^i)|, |\widehat{\nabla}_\nu u(x^i) - \nabla_\nu u(x^i)| \leq C(t + C_\nu \varepsilon + \varepsilon) \tag{6.21}$$

for all $x^i \in \partial_\varepsilon \mathcal{X}$.

2. Let $0 < \lambda \leq \varepsilon^{-1}$. Let φ be the solution of

$$\left.\begin{aligned} -\rho^{-1} \mathrm{div}(\rho^2 \nabla \varphi) &= 1 \quad \text{in } \Omega \\ \gamma \varphi - (1-\gamma) \nabla_\nu \varphi &= 1 \quad \text{on } \partial\Omega. \end{aligned}\right\} \tag{6.22}$$

By assumption, $u, \varphi \in C^4(\bar{\Omega})$, and so by Theorem 6.8, with probability at least $1 - 2n \exp\left(-Cn\varepsilon^{d+2}\lambda^2\right)$ we have

$$|\mathcal{L}_\varepsilon \varphi(x^i) - 1\,|\,, |\mathcal{L}_\varepsilon u(x^i) - f(x^i)| \leq C(\varepsilon^2 + \lambda) \tag{6.23}$$

whenever $\mathrm{dist}(x^i, \partial\Omega) \geq \varepsilon$.

3. Let us now define

$$w(x^i) = u(x^i) - u_\varepsilon(x^i) - K\varphi(x^i),$$

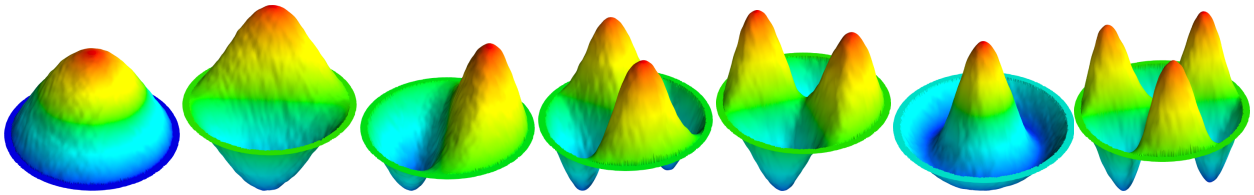


FIGURE 11. First 7 Laplacian Dirichlet eigenfunctions on the disk computed via approximation with graph Laplacian eigenvectors with $n = 10^5$ points.

for K to be determined. Then by (6.23) and (6.21) we have

$$\mathcal{L}_\varepsilon w(x^i) \leq -K + C(\varepsilon^2 + \lambda)$$

for $x^i \in X_\varepsilon$ and

$$\gamma w(x^i) - (1-\gamma)\widehat{\nabla}_\nu w(x^i) \leq -K + \|\gamma u - (1-\gamma)\nabla_\nu u - g\|_{L^\infty(\partial_{2\varepsilon}\Omega)} + C(1-\gamma)(t + C_\nu\varepsilon + \varepsilon)$$

for $x^i \in \partial_\varepsilon \mathcal{X}$. For any choice of K satisfying

$$K > C\left(\|\gamma u - (1-\gamma)\nabla_\nu u - g\|_{L^\infty(\partial_{2\varepsilon}\Omega)} + (1-\gamma)(t + C_\nu\varepsilon + \varepsilon) + \varepsilon^2 + \lambda\right)$$

we can apply Lemma 6.7 to find that $w \leq 0$, and so $u - u_\varepsilon \leq CK\|\varphi\|_{L^\infty(\Omega)}$. The other direction of the proof is similar. □

Remark 6.10. The proof of Theorem 6.9 relies on the maximum principle (Lemma 6.7), which requires $\gamma > 0$. Thus, the result does not apply to the pure Neumann case $\gamma = 0$. This case would require special attention to ensure the compatibility condition

$$\int_\Omega f\, dx = \int_{\partial\Omega} g\, dS$$

holds at both the continuum and discrete level. △

Remark 6.11. Consider the Dirichlet problem in Theorem 6.9 by setting $\gamma = 1$. If we set $\lambda = \varepsilon^2$, then we obtain the rate

$$|u - u_\varepsilon| \leq C(\|u - g\|_{L^\infty(\partial_\varepsilon\Omega)} + \varepsilon^2)$$

with probability at least $1 - 2n\exp\left(-Cn\varepsilon^{d+6}\right)$. If we are able to extend the boundary conditions g to Ω so that $\|u - g\|_{L^\infty(\partial_\varepsilon\Omega)} \leq C\varepsilon^2$, then we obtain a second-order $\mathcal{O}(\varepsilon^2)$ convergence rate in Theorem 6.9. △

Remark 6.12. Finally, we remark that our boundary detection method allows us to consider Dirichlet eigenfunctions of the Laplacian on the point cloud $\mathcal{X}$ by solving the eigenfunction problem

$$\left.\begin{aligned} \mathcal{L}_\varepsilon u(x^i) &= \lambda u(x^i), \quad \text{if } x^i \in X_\varepsilon \\ u(x^i) &= 0, \qquad\quad \text{if } x^i \in \partial_\varepsilon \mathcal{X} \end{aligned}\right\} \tag{6.24}$$

The Dirichlet eigenfunctions of $\mathcal{L}_\varepsilon$ would naturally converge to continuum Dirichlet eigenfunction for the weighted Laplacian $-\rho^{-1}\text{div}(\rho^2\nabla u)$. The proof of this is expected to be more involved than Theorem 6.9, since we cannot use the maximum principle to obtain strong discrete stability results. We expect discrete to continuum convergence results to hold for the eigenvector problem (6.24) using the combined variational and PDE methods from [22, 23, 49]. We show in Figure 11 the first 7 Dirichlet eigenfunctions on the disk computed by solving (6.24) over a graph constructed with $n = 10^5$ random variables independent and uniformly distributed on the disk. △

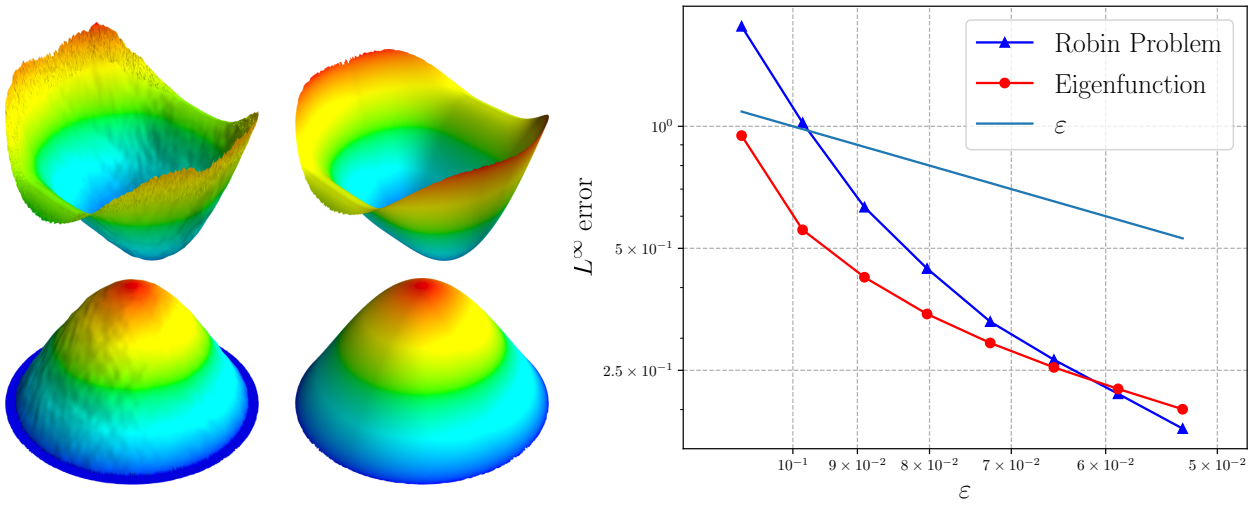


FIGURE 12. On the left, plots of the solution to the Robin problem and principal Dirichlet eigenvector for $n = 10^5$ points on the disk, compared to the exact solutions of each problem. On the right we show an error plot for varying ε averaged over 100 trials.

Remark 6.13. In the case that $f = 0$ and we consider Dirichlet boundary conditions ($\gamma = 1$), we can extend Theorem 6.9 to hold even when $\partial_\varepsilon \mathcal{X}$ is replaced with a thinner boundary $\partial \mathcal{X}_\delta$ for any $\varepsilon^2 \ll \delta \leq \varepsilon$. That is when only the points in a very thin region near the true boundary are identified. In this case we can prove the error rate of $O(\varepsilon^2/\delta)$. The proof is a minor adaptation of [24, Theorem 2.4]. We expect the proof would extend to the case of nonzero f as well, though the incorporation of $\gamma < 1$ seems more difficult. △

6.2.1. *Numerical results.* We ran several numerical experiments to test the rate of convergence in Theorem 6.9 on the disk $\Omega = B(0,1) \subset \mathbb{R}^2$. In this case, $\rho = 1/\pi$. In the first experiment, we set the solution of the Robin problem (6.17) with $\gamma = 1/2$ to be

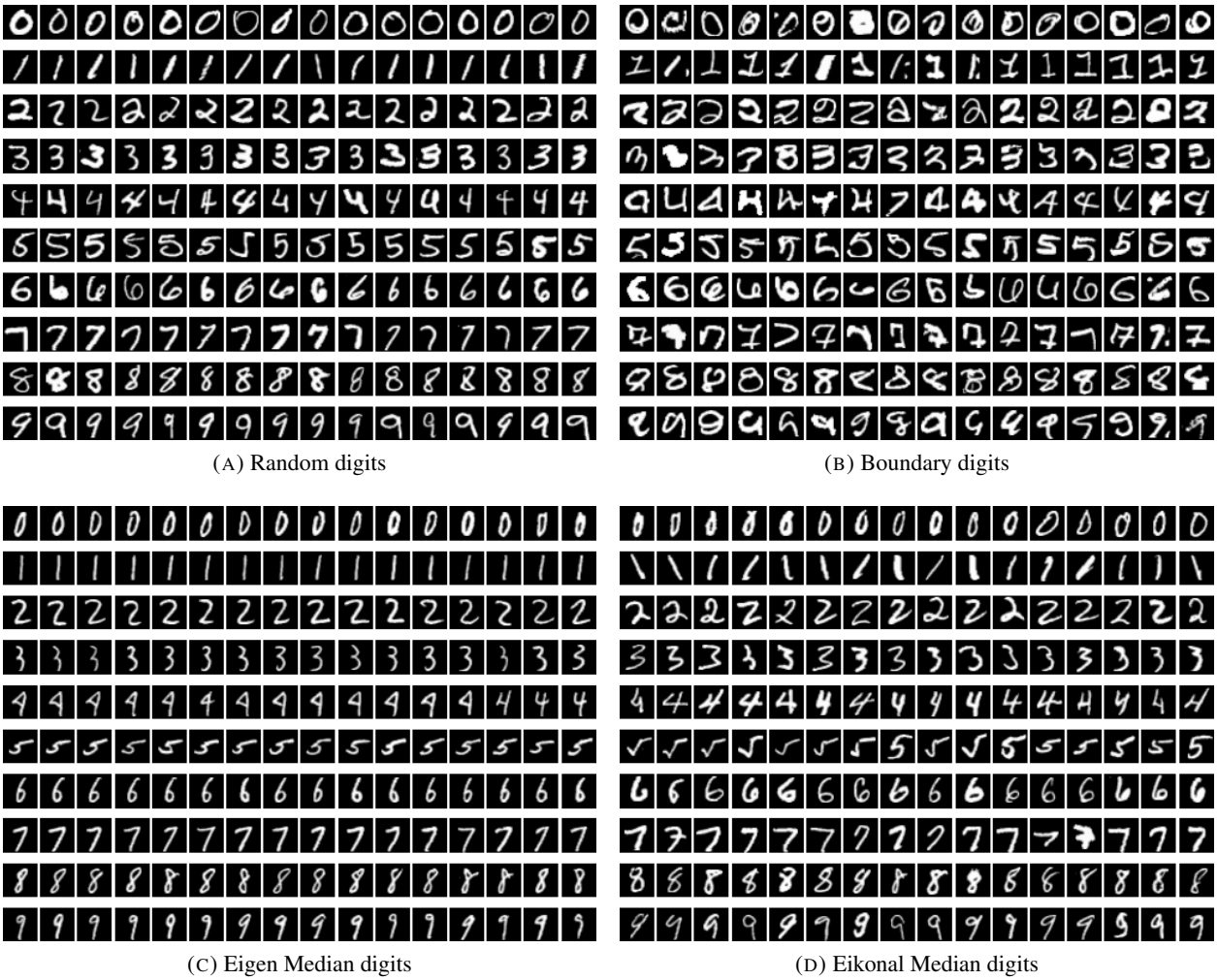
$$u(x) = \sin(2x_1^2) - \cos(2x_1^2)$$

and then set $f = -\frac{1}{\pi}\Delta u$ and $g = \frac{1}{2}(u - \nabla_\nu u)$, and tested how well the solution of the graph Laplace equation (6.16) can reconstruct u. In the second problem, we solved (6.24) for the principal Dirichlet eigenfunction, and compared against the true solution $u(x) = J_0(\lambda|x|)$, where J_0 is the zeroth order Bessel function of the first kind, and λ is the first positive root of J_0. In each case we varied the number n of random variables in the point cloud from $n = 2^{10}$ up to $n = 2^{17} = 131,072$ by powers of 2, and set

$$\varepsilon = \frac{1}{4}\left(\frac{\log n}{n}\right)^{\frac{1}{d+4}},$$

where here, $d = 2$. We approximated the ε boundary using $k = 2\pi n \varepsilon^2$ nearest neighbors. Figure 12 shows plots of the solutions to each graph-based problem, compared to the true solutions of their corresponding PDEs, and a plot of maximum absolute error versus ε, averaged over 100 trials. In both cases we see better convergence rates than the $O(\varepsilon)$ guaranteed by Theorem 6.9. Taking the last three data points on each plot, the empirical convergence rates are $\varepsilon^{1.86}$ for the Robin problem and $\varepsilon^{1.13}$ for the Dirichlet eigenfunction.

6.3. **Experiments with real data.** We now turn to experiments with real data. We use the MNIST [54] and FashionMNIST [79] datasets. MNIST is a standard dataset for handwritten digit recognition, consisting of 70,000 images of handwritten digits 0–9. Each image is a 28×28 grayscale image, which we interpret as a vector in $\mathbb{R}^{784}$. The FashionMNIST dataset is a drop-in replacement for MNIST, with the same number of datapoints and image resolution, except that the 10 classes in FashionMNIST correspond to different items of clothing, with pictures taken from a fashion catalog. In all experiments, we use Euclidean distance between the raw pixel values in $\mathbb{R}^{784}$ to compare images.



(A) Random digits

(B) Boundary digits

(C) Eigen Median digits

(D) Eikonal Median digits

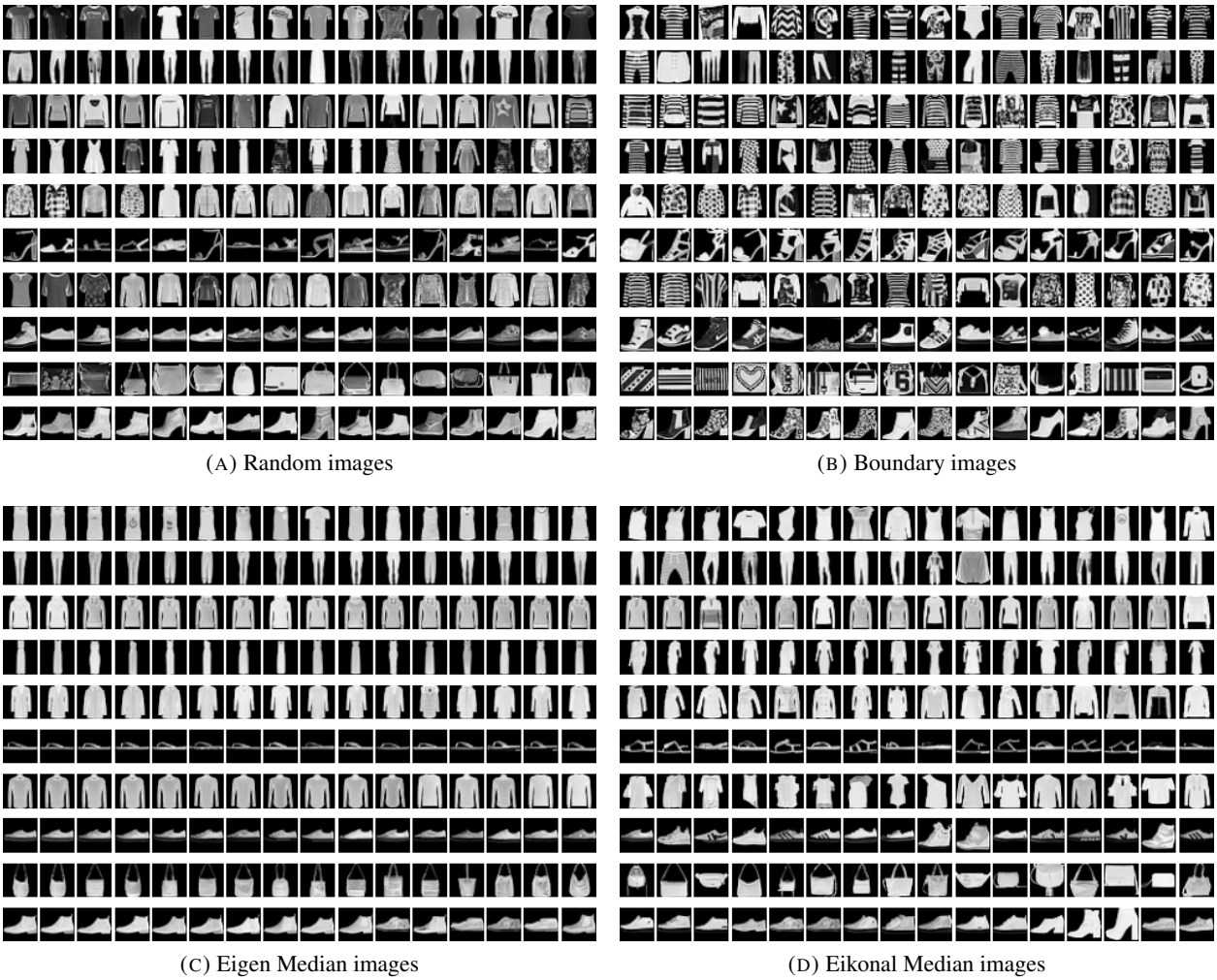
FIGURE 13. MNIST experiments.

We focus our experiments on detecting the boundary images for each class, and then using the discovered boundary to compute a notion of data depth by solving PDEs over the data with Dirichlet boundary conditions. In this way, we also compute a notion of data median, by taking the deepest images in the dataset. To compute the boundary points, we use $k = 10$ Euclidean nearest neighbors and compute $\hat{d}_\varepsilon(x^i)$ for each image x^i by taking ε as the Euclidean distance to the k^{th} nearest neighbor. We then set the images with scores $\hat{d}_\varepsilon(x^i)$ in the lower 10% of all images to be boundary points. This is an implicit way to select the desired width of the boundary by instead specifying how many boundary points are desired. Figures 13 and 14 show that top 10 boundary images in each class compared to randomly selected images.

Once the boundary points are detected, we construct a k nearest neighbor graph over the data points in each class. We use Gaussian weights given by

$$w_{ij} = \exp\left(-\frac{4|x^i - x^j|^2}{\varepsilon_k(x^i)^2}\right),$$

where $\varepsilon_k(x_i)$ is the distance between x^i and its k^{th} nearest neighbor. We used $k = 10$ in all experiments, and the weight matrix was symmetrized by replacing W with $W + W^T$. For a notion of data depth, we compute the principal Dirichlet eigenfunction of the graph Laplacian, i.e., the solution of (6.24) with smallest λ. We



(A) Random images (B) Boundary images

(C) Eigen Median images (D) Eikonal Median images

FIGURE 14. FashionMNIST experiments.

found the symmetric normalization

$$\mathcal{L}u(x) = \sum_{j=1}^{n} w_{ij} \left(\frac{u(x_i)}{\sqrt{d_i}} - \frac{u(x_j)}{\sqrt{d_j}} \right), \quad d_i = \sum_{j=1}^{n} w_{ij}$$

gives slightly more consistent results, and so we report the results with this normalization. The principal Dirichlet eigenfunction has one sign on all of $\mathcal{X}$, and we choose the version that is positive on $\mathcal{X}$. We use $u(x^i)$ as a notion of data depth, and the x^i where $u(x^i)$ is largest can be interpreted as median images for each class. The median images computed this way are shown in Figures 13 (c) and 14 (c). We also computed the median by solving the eikonal equation (6.3), again using our detected boundary images as Dirichlet boundary conditions. The eikonal median images are shown in Figures 13 (d) and 14 (d).

We observe that the eigen-median images are all very similar to each other, compared with the eikonal median images, which have much more variation. There is some work showing that the maximum or minimum points of graph Laplacian eigenvectors correspond to nodes in the graph that are unusually well-connected, in the sense that a random walker will take a long time to escape the region (see, e.g., [4]). These regions then contain groups of highly similar images. In contrast, the eikonal median images are simply those that are furthest from the boundary in the graph geodesic distance, and these images may be scattered around the graph and have far more variability.

We remark that we can also construct a similar notion of data depth by solving the Dirichlet problem (6.16) with $f \equiv 1$, $\gamma = 1$, and $g \equiv 0$. The solution of this Poisson equation has the interpretation that $u(x^i)$ is the mean exit time for random walkers starting at x^i, and exiting at $\partial_\varepsilon \mathcal{X}$. We almost always obtained the same set of median images, up to some minor differences, using the two graph PDEs, so we only show the results using the Dirichlet eigenfunction.

Remark 6.14. It is important to point out that our boundary detection method is designed for data sampled from a distribution with a Lebesgue density on a domain $\Omega \subset \mathbb{R}^d$. That is, our results do not apply to the *manifold assumption*, which is a commonly used modeling assumption in machine learning that assumes the data is sampled from a low dimensional smooth submanifold, possibly with boundary, embedded in $\mathbb{R}^d$. The dimension m of the smooth submanifold is called the *intrinsic dimension* of the data. While the MNIST dataset has extrinsic dimension $d = 784$ (i.e., the number of pixels in each image), it has been estimated that intrinsic dimension of each class of MNIST digits is between $m = 12$ and $m = 14$ [32, 51]. In the manifold setting, it is possible that our approximation of the unit normal vector $\hat{\nu}_\varepsilon$ will point in the direction normal to the data submanifold in regions of higher curvature. This would cause interior points to be incorrectly identified as boundary points. This could be addressed by projecting $\hat{\nu}_\varepsilon$ onto the tangent space to the submanifold, but we leave this for future work. Since we see good results for our method on MNIST and FashionMNIST in Figures 13 (b) and 14 (b), this may indicate that curvature is low for both datasets and does not play a large role in boundary detection. △

REFERENCES

[1] E. AAMARI, C. AARON, AND C. LEVRARD, *Minimax boundary estimation and estimation with boundary*, arXiv preprint arXiv:2108.03135, (2021).

[2] E. AAMARI AND C. LEVRARD, *Nonasymptotic rates for manifold, tangent space and curvature estimation*, The Annals of Statistics, 47 (2019), pp. 177 – 204.

[3] C. AARON AND A. CHOLAQUIDIS, *On boundary detection*, Ann. Inst. Henri Poincaré Probab. Stat., 56 (2020), pp. 2028–2050.

[4] S. S. ADELA DEPAVIA, *Spectral clustering revisited: Information hidden in the Fiedler vector*, Foundations of Data Science, 3 (2021), pp. 225–249.

[5] M. BARDI AND I. CAPUZZO-DOLCETTA, *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations*, Springer Science & Business Media, 2008.

[6] V. BARNETT, *The ordering of multivariate data*, Journal of the Royal Statistical Society: Series A (General), 139 (1976), pp. 318–344.

[7] K. BELLOCK, *Alpha shape toolbox*, 2021. https://github.com/bellockk/alphashape. (accessed 2021/10/22).

[8] J. L. BENTLEY, *Multidimensional divide-and-conquer*, Communications of the ACM, 23 (1980), pp. 214–229.

[9] ———, *Multidimensional divide-and-conquer*, Discrete and Comp. Geom., 4 (1989), p. 101–115.

[10] E. BERNHARDSSON, *Annoy: Approximate nearest neighbors in c++/python*, 2018. https://pypi.org/project/annoy/ (accessed 2020/10/19).

[11] T. BERRY AND T. SAUER, *Density estimation on manifolds with boundary*, Computational Statistics & Data Analysis, 107 (2017), pp. 1–17.

[12] L. BIRBRAIR AND M. P. DENKOWSKI, *Medial axis and singularities*, J. Geom. Anal., 27 (2017), pp. 2339–2380.

[13] A. BOU-RABEE AND P. S. MORFE, *Hamilton-Jacobi scaling limits of pareto peeling in 2d*, arXiv preprint arXiv:2110.06016, (2021).

[14] S. BOUCHERON, G. LUGOSI, AND P. MASSART, *Concentration inequalities: A nonasymptotic theory of independence*, Oxford university press, 2013.

[15] J. CALDER, *The game theoretic p-Laplacian and semi-supervised learning with few labels*, Nonlinearity, 32 (2018), pp. 301–330.

[16] J. CALDER, *Lecture notes on viscosity solutions*, Online Lecture Notes, (2018). http://www-users.math.umn.edu/~jwcalder/viscosity_solutions.pdf.

[17] J. CALDER, *Consistency of Lipschitz learning with infinite unlabeled data and finite labeled data*, SIAM Journal on Mathematics of Data Science, 1 (2019), pp. 780–812.

[18] J. CALDER, *The calculus of variations*, Online Lecture Notes, (2020). http://www-users.math.umn.edu/~jwcalder/CalculusOfVariations.pdf.

[19] ———, *Graph-based clustering and semi-supervised learning*, 2020. https://github.com/jwcalder/GraphLearning. (accessed 2020/10/19).

[20] J. CALDER, S. ESEDOḠLU, AND A. O. HERO, *A Hamilton-Jacobi equation for the continuum limit of non-dominated sorting*, SIAM Journal on Mathematical Analysis, 46 (2014), pp. 603–638.
[21] J. CALDER AND M. ETTEHAD, *Hamilton-Jacobi equations on graphs with applications to semi-supervised learning and data depth*, In preparation, (2021).
[22] J. CALDER AND N. GARCÍA TRILLOS, *Improved spectral convergence rates for graph Laplacians on ε-graphs and k-NN graphs*, arXiv:1910.13476, (2019).
[23] J. CALDER, N. GARCÍA TRILLOS, AND M. LEWICKA, *Lipschitz regularity of graph Laplacians on random data clouds*, arXiv:2007.06679, (2020).
[24] J. CALDER, D. SLEPČEV, AND M. THORPE, *Rates of convergence for Laplacian semi-supervised learning with low labeling rates*, arXiv:2006.02765, (2020).
[25] J. CALDER AND C. K. SMART, *The limit shape of convex hull peeling*, Duke Mathematical Journal, 169 (2020), pp. 2079–2124.
[26] P. CANNARSA AND C. SINESTRARI, *Semiconcave functions, Hamilton-Jacobi equations, and optimal control*, vol. 58, Springer Science & Business Media, 2004.
[27] E. CARRIZOSA, *A characterization of halfspace depth*, Journal of multivariate analysis, 58 (1996), pp. 21–26.
[28] J.-S. CHEN, M. HILLMAN, AND S.-W. CHI, *Meshfree methods: progress made after 20 years*, Journal of Engineering Mechanics, 143 (2017), p. 04017001.
[29] Y.-C. CHEN, C. R. GENOVESE, AND L. WASSERMAN, *Density level sets: asymptotics, inference, and visualization*, J. Amer. Statist. Assoc., 112 (2017), pp. 1684–1696.
[30] CHENYI XIA, W. HSU, M. L. LEE, AND B. C. OOI, *Border: efficient computation of boundary points*, IEEE Transactions on Knowledge and Data Engineering, 18 (2006), pp. 289–303.
[31] V. CHERNOZHUKOV, A. GALICHON, M. HALLIN, AND M. HENRY, *Monge–kantorovich depth, quantiles, ranks and signs*, The Annals of Statistics, 45 (2017), pp. 223–256.
[32] J. A. COSTA AND A. O. HERO, *Determining intrinsic dimension and entropy of high-dimensional shape spaces*, in Statistics and Analysis of Shapes, Springer, 2006, pp. 231–252.
[33] A. CUEVAS, R. FRAIMAN, ET AL., *A plug-in approach to support estimation*, The Annals of Statistics, 25 (1997), pp. 2300–2312.
[34] A. CUEVAS, R. FRAIMAN, AND L. GYÖRFI, *Towards a universally consistent estimator of the Minkowski content*, ESAIM Probab. Stat., 17 (2013), pp. 359–369.
[35] A. CUEVAS, R. FRAIMAN, AND A. RODRÍGUEZ-CASAL, *A nonparametric approach to the estimation of lengths and surface areas*, Ann. Statist., 35 (2007), pp. 1031–1051.
[36] A. CUEVAS AND A. RODRÍGUEZ-CASAL, *On boundary estimation*, Adv. in Appl. Probab., 36 (2004), pp. 340–354.
[37] P. L. DE MICHEAUX, P. MOZHAROVSKYI, AND M. VIMOND, *Depth for curve data and applications*, Journal of the American Statistical Association, (2020), pp. 1–17.
[38] L. DEVROYE AND G. L. WISE, *Detection of abnormal behavior via nonparametric estimation of the support*, SIAM J. Appl. Math., 38 (1980), pp. 480–488.
[39] W. DONG, C. MOSES, AND K. LI, *Efficient k-nearest neighbor graph construction for generic similarity measures*, in Proceedings of the 20th International Conference on World Wide Web, WWW '11, New York, NY, USA, 2011, Association for Computing Machinery, p. 577–586.
[40] H. EDELSBRUNNER, *Alpha shapes—a survey*, Tessellations in the Sciences, (2010).
[41] H. EDELSBRUNNER, D. KIRKPATRICK, AND R. SEIDEL, *On the shape of a set of points in the plane*, IEEE Transactions on Information Theory, 29 (1983), pp. 551–559.
[42] H. EDELSBRUNNER AND E. P. MÜCKE, *Three-dimensional alpha shapes*, ACM Trans. Graph., 13 (1994), p. 43–72.
[43] C. FINLAY AND A. OBERMAN, *Improved accuracy of monotone finite difference schemes on point clouds and regular grids*, SIAM Journal on Scientific Computing, 41 (2019), pp. A3097–A3117.
[44] M. FLORES, J. CALDER, AND G. LERMAN, *Analysis and algorithms for Lp-based semi-supervised learning on graphs*, arXiv:1901.05031, (2019).
[45] N. FLYER AND G. B. WRIGHT, *A radial basis function method for the shallow water equations on a sphere*, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 465 (2009), pp. 1949–1976.
[46] R. L. FOOTE, *Regularity of the distance function*, Proceedings of the American Mathematical Society, 92 (1984), pp. 153–155.
[47] B. D. FROESE, *Meshfree finite difference approximations for functions of the eigenvalues of the Hessian*, Numerische Mathematik, 138 (2018), pp. 75–99.
[48] E. FUSELIER AND G. B. WRIGHT, *Scattered data interpolation on embedded submanifolds with restricted positive definite kernels: Sobolev error estimates*, SIAM Journal on Numerical Analysis, 50 (2012), pp. 1753–1776.
[49] N. GARCÍA TRILLOS, M. GERLACH, M. HEIN, AND D. SLEPČEV, *Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace–Beltrami operator*, Foundations of Computational Mathematics, 20 (2020), pp. 827–887.
[50] N. GARCÍA TRILLOS AND R. W. MURRAY, *A maximum principle argument for the uniform convergence of graph Laplacian regressors*, SIAM Journal on Mathematics of Data Science, 2 (2020), pp. 705–739.

[51] M. HEIN AND J.-Y. AUDIBERT, *Intrinsic dimensionality estimation of submanifolds in rd*, in Proceedings of the 22nd international conference on Machine learning, 2005, pp. 289–296.
[52] R. LACHIÈZE-REY AND S. VEGA, *Boundary density and Voronoi set estimation for irregular sets*, Trans. Amer. Math. Soc., 369 (2017), pp. 4953–4976.
[53] R. LAI, J. LIANG, AND H.-K. ZHAO, *A local mesh method for solving pdes on point clouds*, Inverse Problems & Imaging, 7 (2013), p. 737.
[54] Y. LECUN, L. BOTTOU, Y. BENGIO, AND P. HAFFNER, *Gradient-based learning applied to document recognition*, Proceedings of the IEEE, 86 (1998), pp. 2278–2324.
[55] Z. LI, Z. SHI, AND J. SUN, *Point integral method for solving poisson-type equations on manifolds from point clouds with convergence guarantees*, Communications in Computational Physics, 22 (2017), pp. 228–258.
[56] J. LIANG AND H. ZHAO, *Solving partial differential equations on point clouds*, SIAM Journal on Scientific Computing, 35 (2013), pp. A1461–A1486.
[57] S. LIANG, S. W. JIANG, J. HARLIM, AND H. YANG, *Solving pdes on unknown manifolds with machine learning*, arXiv:2106.06682, (2021).
[58] R. Y. LIU, J. M. PARELIUS, AND K. SINGH, *Multivariate analysis by data depth: descriptive statistics, graphics and inference,(with discussion and a rejoinder by liu and singh)*, The annals of statistics, 27 (1999), pp. 783–858.
[59] P. MCMULLEN, *The maximum numbers of faces of a convex polytope*, Mathematika, 17 (1970), p. 179–184.
[60] M. MOLINA-FRUCTUOSO AND R. MURRAY, *Eikonal depth: an optimal control approach to statistical depths*, In preparation, (2021).
[61] ———, *Tukey depths and Hamilton-Jacobi differential equations*, arXiv:2104.01648, (2021).
[62] A. M. OBERMAN, *Wide stencil finite difference schemes for the elliptic Monge-Ampere equation and functions of the eigenvalues of the Hessian*, Discrete & Continuous Dynamical Systems-B, 10 (2008), p. 221.
[63] C. PIRET, *The orthogonal gradients method: A radial basis functions method for solving partial differential equations on arbitrary surfaces*, Journal of Computational Physics, 231 (2012), pp. 4662–4675.
[64] C. PIRET AND J. DUNN, *Fast rbf ogr for solving pdes on arbitrary surfaces*, in AIP Conference Proceedings, vol. 1776, AIP Publishing LLC, 2016, p. 070005.
[65] W. QIAO AND W. POLONIK, *Nonparametric confidence regions for level sets: statistical properties and geometry*, Electron. J. Stat., 13 (2019), pp. 985–1030.
[66] B.-Z. QIU, F. YUE, AND J.-Y. SHEN, *Brim: An efficient boundary points detecting algorithm*, in Advances in Knowledge Discovery and Data Mining, Z.-H. Zhou, H. Li, and Q. Yang, eds., Berlin, Heidelberg, 2007, Springer Berlin Heidelberg, pp. 761–768.
[67] A. RODRÍGUEZ CASAL, *Set estimation under convexity type assumptions*, Annales de l'I.H.P. Probabilités et statistiques, 43 (2007), pp. 763–774.
[68] J. A. SETHIAN AND A. VLADIMIRSKY, *Fast methods for the eikonal and related Hamilton–Jacobi equations on unstructured meshes*, Proceedings of the National Academy of Sciences, 97 (2000), pp. 5699–5703.
[69] Z. SHI, *Enforce the Dirichlet boundary condition by volume constraint in point integral method*, Commun. Math. Sci., 15 (2017), pp. 1743–1769.
[70] C. G. SMALL, *Multidimensional medians arising from geodesics on graphs*, The Annals of Statistics, (1997), pp. 478–494.
[71] P. SUCHDE AND J. KUHNERT, *A fully lagrangian meshfree framework for pdes on evolving surfaces*, Journal of Computational Physics, 395 (2019), pp. 38–59.
[72] P. SUCHDE AND J. KUHNERT, *A meshfree generalized finite difference method for surface pdes*, Computers & Mathematics with Applications, 78 (2019), pp. 2789–2805.
[73] THE MATHWORKS INC., *alphashape: Matlab documentation*. https://www.mathworks.com/help/matlab/ref/alphashape.html. Accessed: 2021-10-17.
[74] H. TIENG WU AND N. WU, *When locally linear embedding hits boundary*, arXiv:1811.04423, (2019).
[75] N. TRASK AND P. KUBERRY, *Compatible meshfree discretization of surface pdes*, Computational Particle Mechanics, 7 (2020), pp. 271–277.
[76] J. W. TUKEY, *Mathematics and the picturing of data*, in Proceedings of the International Congress of Mathematicians, Vancouver, 1975, vol. 2, 1975, pp. 523–531.
[77] R. VAUGHN, T. BERRY, AND H. ANTIL, *Diffusion maps for embedded manifolds with boundary with applications to pdes*, arXiv preprint arXiv:1912.01391, (2019).
[78] M. WANG, S. LEUNG, AND H. ZHAO, *Modified virtual grid difference for discretizing the Laplace–Beltrami operator on point clouds*, SIAM Journal on Scientific Computing, 40 (2018), pp. A1–A21.
[79] H. XIAO, K. RASUL, AND R. VOLLGRAF, *Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms*, arXiv:1708.07747, (2017).
[80] A. YUAN, J. CALDER, AND B. OSTING, *A continuum limit for the pagerank algorithm*, European Journal of Applied Mathematics, (2020), pp. 1–33.

APPENDIX A. PROOF OF LEMMA 3.1

The following lemma will be useful in proving Lemma 3.1.

Lemma A.1 (Covering with spherical segments)**.** *Let $r \leq 1$ and $0 < a < b \leq r$. For $u \in \mathbb{S}^{d-1}$ and $0 < a < b \leq r$ define the spherical sector by*

$$S^u_{a,b} = \{x \in B(0,r) \,:\, a \leq x \cdot u \leq b\}.$$

Suppose $\Sigma \subset \mathbb{S}^{d-1}$ is a finite set satisfying the following property:

$$\text{for all } u \in \mathbb{S}^{d-1} \text{ there exists } v \in \Sigma \text{ such that } |u - v| \leq \delta. \tag{A.1}$$

Then, for any $u \in \mathbb{S}^{d-1}$ we can find $v \in \Sigma$ such that

$$S^v_{a+\delta b, b-\delta b} \subset S^u_{a,b}.$$

Proof. Let $u \in \mathbb{S}^{d-1}$ and fix a $v \in \Sigma$ satisfying (A.1). Suppose that $x \in S^v_{a+\delta b, b-\delta b}$. Then we have

$$a + \delta b \leq x \cdot v \leq b - \delta b.$$

We have

$$|x \cdot v - x \cdot u| = |x \cdot (v - u)| \leq |x||u - v| \leq \delta |x| \leq \delta b,$$

since $|x| \leq b - \delta \leq b$. Therefore

$$x \cdot u \leq b - \delta b + \delta b = b \;\text{ and }\; x \cdot u \geq a + \delta b - \delta b = a.$$

Therefore $x \in S^u_{a,b}$, which shows that for each $u \in \mathbb{S}^{d-1}$ there exists $v \in \Sigma$ such that

$$S^u_{a,b} \supset S^v_{a+\delta b, b-\delta b}.$$

Hence, the event that $S^u_{a,b}$ is empty for some $u \in \mathbb{S}^{d-1}$ is contained in the event that $S^v_{a+\delta b, b-\delta b}$ is empty for some $v \in \Sigma$—a finite collection of events.

□

Remark A.2 (ε-nets and upper bound on $|\Sigma|$)**.** Recall that an ε-net of $\mathbb{S}^{d-1}$ is the set of points in $\mathbb{S}^{d-1}$ such that the pairwise distance is at least ε. Then we define a maximal ε-net of the sphere to be an ε-net such that no point on $\mathbb{S}^{d-1}$ can be added while preserving the lower bound for the pairwise distance.

Then, observe that any maximal ε-net of the unit sphere satisfies the condition of Lemma A.1. If $\Sigma_\varepsilon = \{x^1, \cdots, x^{N_\varepsilon}\}$ is a maximal ε-net of $\mathbb{S}^{d-1}$, then for each $x \in \mathbb{S}^{d-1}$ there exists $x^i \in \Sigma_\varepsilon$ such that $|x - x^i| \leq \varepsilon$. To see this, suppose $|x^* - x^i| > \varepsilon$ for all $i = 1, \cdots, N_\varepsilon$. Then

$$B(x^*, \varepsilon/2) \cap B(x^i, \varepsilon/2) = \emptyset \text{ for all } x^i \in \Sigma_\varepsilon.$$

Thus $\Sigma_\varepsilon \cap \{x^*\}$ should also be an ε-net, which contradicts the maximality of Σ_ε.

Now, let Σ_δ be any δ-net – i.e. ε-net with $\varepsilon = \delta$. Then $\{B(v^i, \delta/2) \,:\, v^i \in \Sigma_\delta\}$ is a collection of disjoint balls, all contained in $B(0, 1 + \delta/2) \setminus B(0, 1 - \delta/2)$. Thus, base on a simple volumetric argument, we can deduce

$$|\Sigma_\delta| \leq 2d \left(1 + \frac{2}{\delta}\right)^{d-1}, \tag{A.2}$$

△

Proof of Lemma 3.1.

(1) Let $\{v_i\}_{i=1}^M = \Sigma \subset \mathbb{S}^{d-1}$ be a maximal δ-net. By Lemma A.1 and Remark A.2, for any $u \in \mathbb{S}^{d-1}$ we can find $v_k \in \Sigma$ such that

$$S^{v_k}_{a+b\delta, b-b\delta} \subset S^u_{a,b}.$$

This means that if all of $S^{v_i}_{a+b\delta, b-b\delta}$ are nonempty, all of $S^u_{a,b}$ is nonempty for $u \in \mathbb{S}^{d-1}$ hence

$$\hat{d}^1_r(x^0) \geq a.$$

Without loss of generality, assume $x^0 \in \mathbb{R}^d$ is the origin, and let $\alpha = d_\Omega(x^0) \wedge \frac{r}{2}$. Denote by $K^u_{a,b} \subset S^u_{a,b}$ the cone of maximal height sharing the base with $S^u_{a,b}$. Note that $b \leq \alpha$ implies $K^u_{a,b} \subset \overline{B}(x_0, r) \cap \Omega$. On the other hand, we need $a \geq (1-\lambda)\alpha - t$ to deduce the desired lower bound on $\hat{d}^1_r$. Thus choose

$$a = (1-\lambda)\alpha - t,\ b = \alpha.$$

Further, we need the height of $S^{v_i}_{a+b\delta, b-b\delta}$ to scale like t, in order to lower bound the volume. Thus we need

$$b - b\delta - (a + b\delta) = (1-2\delta)b - \alpha = (1-2\delta)\alpha - (1-\lambda)\alpha - t = (\lambda - 2\delta)\alpha + t.$$

As we are interested in $t \lesssim r^2 \ll \alpha \sim \varepsilon$, we need $\lambda - 2\delta \geq 0$, hence

$$\delta \leq \frac{\lambda}{2}.$$

(2) Following the discussion in the previous step, let $\Sigma = \{v^1, \cdots, v^{N_\lambda}\}$ be a maximal $\frac{\lambda}{2}$-net of $\mathbb{S}^{d-1}$, and write

$$S^i = S^{v^i}_{a+b\lambda/2, b-b\lambda/2} \text{ where } a = (1-\lambda)\alpha,\ b = \alpha, \text{ and .}$$

Thus, to show (3.2) holds with probability at least $1 - n^{-\gamma}$, it suffices to show

$$\mathbb{P}(\text{ No point in } S^i) \leq (1 - \rho_{\min}|S^i \cap \Omega|)^n \leq N_\lambda^{-1} n^{-\gamma} \text{ for all } i = 1, \cdots, N_\lambda.$$

(3) We first compute the lower bound for $|S^i \cap \Omega|$. Temporarily write $a' = a + b\lambda/2$, $b' = b - b\lambda/2$. Let $K^i_{a',b'}$ be the cone of height $b' - a' = t$ sharing the base of S^i. Note that $K^i_{a',b'} \subset S^i \cap \Omega$ and its base has radius $\sqrt{r^2 - (a')^2} = r\sqrt{1 - (a'/r)^2}$. As the $|K^i_{a',b'}|$ is independent of i, we may drop the superscript and deduce

$$|S^i \cap \Omega| \geq |K_{a',b'}| = \int_0^t \omega_{d-1} \left(r\sqrt{1-(a'/r)^2}\frac{s}{t} \right)^{d-1} ds = \frac{1}{d}\omega_{d-1} t r^{d-1} (1-(a'/r)^2)^{\frac{d-1}{2}}.$$

As $a' \leq b \leq \alpha \leq r/2$, we have $(1-(a'/r)^2)^{(d-1)/2} \geq 2^{-(d-1)/2}$. Hence, for each $i = 1, \cdots, N_\lambda$

$$\mathbb{P}(\text{ No point in } S^i) \leq (1 - \rho_{\min}|K_{a',b'}|)^n \leq \left(1 - \frac{\rho_{\min}}{d 2^{(d-1)/2}} t r^{d-1}\right)^n.$$

The expression on the right is less than $N_\lambda^{-1} n^{-\gamma}$ if

$$n \log\left(1 - \frac{\rho_{\min}}{d 2^{(d-1)/2}} t r^{d-1}\right) \leq -\gamma \log n - \log N_\lambda,$$

or equivalently

$$t r^{d-1} \geq \frac{d 2^{(d-1)/2}(1 - e^{-\frac{\gamma \log n + \log N_\lambda}{n}})}{\rho_{\min}\omega_{d-1}}.$$

As $1 - e^{-x} \leq x$, it suffices for t, r to satisfy

$$t r^{d-1} \geq \frac{d 2^{(d-1)/2}}{\rho_{\min}\omega_{d-1}} \left(\frac{\gamma \log n + \log N_\lambda}{n}\right).$$

(4) We claim that $\log N_\lambda \leq \gamma(d-1)\log n$. By setting $\delta = \frac{\lambda}{2}$ in (A.2), we know

$$N \leq 2d\left(1 + \frac{4}{\lambda}\right)^{d-1} = 2d\left(\frac{\lambda+4}{\lambda}\right)^{d-1}.$$

By hypothesis $n \geq d \vee \frac{\lambda+4}{\lambda}$ and $\gamma > 2$, we see

$$n^{\gamma(d-1)} \geq n^{d-1} n^{d-1} \geq 2d\left(\frac{\lambda+4}{\lambda}\right)^{d-1} \geq N_\lambda.$$

Thus $\gamma \log n + \log N \leq d\gamma \log n$, and it suffices for t, r to satisfy

$$tr^{d-1} \geq \frac{d^2 2^{(d-1)/2}\gamma}{\rho_{\min}\omega_{d-1}} \left(\frac{\log n}{n}\right).$$

This completes the proof

□

APPENDIX B. PROOF OF PROPOSITION 6.3

Proof. The proof is split into several steps.

1. Let $y \in \partial\Omega$ satisfy $d_\Omega(x_*) = |x_* - y|$. Let $z \in \partial B(x^0, \varepsilon)$ be along the line from x_* to y. Then we have

$$d_\Omega(z) \leq d_\Omega(x_*) - |x_* - z|$$

and so by the property defining x_* we have $x_* = z$; that is $x_* \in \partial B(x^0, \varepsilon)$. Since d_Ω is 1-Lipschitz, we have $d_\Omega(x_*) \geq d_\Omega(x^0) - \varepsilon$. By a similar argument as above, we have $d_\Omega(x_*) \leq d_\Omega(x^0) - \varepsilon$, and so

$$d_\Omega(x_*) = d_\Omega(x^0) - \varepsilon.$$

Now, note that the function

$$g(r) = d_\Omega(x_* + rp)$$

is 1-Lipschitz and satisfies $g(\varepsilon) = d_\Omega(x^0) = g(0) + \varepsilon$. It follows that $g(l) = g(0) + r$ for $0 \leq r \leq \varepsilon$, and so

$$d_\Omega(x_* + rp) = d_\Omega(x_*) + r \quad \text{for } 0 \leq r \leq \varepsilon. \tag{B.1}$$

2. Since $d_\Omega - \frac{1}{R}|x - x_*|^2$ is a concave function, there exists $q \in \mathbb{R}^n$ such that

$$d_\Omega(x) - d_\Omega(x_*) \leq q \cdot (x - x_*) + \frac{1}{R}|x - x_*|^2.$$

for all $x \in \Omega$. By (B.1) we have

$$r = d_\Omega(x_* + rp) - d_\Omega(x_*) \leq rq \cdot p + \frac{r^2}{R}$$

for $0 \leq r \leq \varepsilon$. Therefore

$$q \cdot p \geq 1 - \frac{r}{R}.$$

Sending $r \to 0^+$ we find that $p \cdot q \geq 1$.

3. We now claim that $|q| \leq 1$, which combined with $p \cdot q \geq 1$ from part 2 implies that $p = q$ and completes the proof. To see this, since $B(x^0, \varepsilon) \subset \Omega$, we have $B(x_*, r) \subset \Omega$ for $r > 0$ sufficiently small. Now, the dynamic programming principle gives

$$0 = \min_{x \in B(x_*, r)} \left\{ d_\Omega(x) - d_\Omega(x_*) + |x - x_*| \right\} \leq \min_{x \in B(x_*, r)} \left\{ q \cdot (x - x_*) + |x - x_*| \right\} + \frac{r^2}{R}.$$

Setting $x - x_* = -|x - x_*|q/|q|$ we have

$$0 \leq \min_{x \in B(x_*, r)} \left\{ |x - x_*|(1 - |q|) \right\} + \frac{r^2}{R} = -r(|q| - 1)_+ + \frac{r^2}{R}.$$

Sending $r \to 0^+$ we obtain $|q| \leq 1$, which completes the proof. □

APPENDIX C. CONCENTRATION INEQUALITIES

For reference, we state the Chernoff bounds, Hoeffding inequality, and the Bernstein inequality, which are concentration of measure inequalities used to control the variance of our normal and distance estimators. We refer the reader to [14] for a general reference on concentration inequalties. Proofs of the exact inequalities below can also be found in [18, Chapter 5].

Theorem C.1 (Chernoff bounds)**.** *Let $X_1, X_2 \ldots, X_n$ be a sequence of* i.i.d. *Bernoulli random variables with parameter $p \in [0,1]$ (i.e., $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = 1 - p$). Then for any $\varepsilon > 0$ we have*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq (1+\varepsilon)np\right) \leq \exp\left(-\frac{np\,\varepsilon^2}{2(1+\frac{1}{3}\varepsilon)}\right), \tag{C.1}$$

and for any $0 \leq \varepsilon < 1$ we have

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \leq (1-\varepsilon)np\right) \leq \exp\left(-\frac{1}{2}np\,\varepsilon^2\right), \tag{C.2}$$

Theorem C.2 (Hoeffding inequality)**.** *Let $X_1, X_2 \ldots, X_n$ be a sequence of* i.i.d. *real-valued random variables with finite expectation $\mu = \mathbb{E}[X_i]$, and write $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Assume there exists $b > 0$ such that $|\mathcal{X} - \mu| \leq b$ almost surely. Then for any $t > 0$ we have*

$$\mathbb{P}(S_n - \mu \geq t) \leq \exp\left(-\frac{nt^2}{2b^2}\right). \tag{C.3}$$

Theorem C.3 (Bernstein Inequality)**.** *Let $X_1, X_2 \ldots, X_n$ be a sequence of* i.i.d. *real-valued random variables with finite expectation $\mu = \mathbb{E}[X_i]$ and variance $\sigma^2 = Var(X_i)$, and write $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Assume there exists $b > 0$ such that $|\mathcal{X} - \mu| \leq b$ almost surely. Then for any $t > 0$ we have*

$$\mathbb{P}(S_n - \mu \geq t) \leq \exp\left(-\frac{nt^2}{2(\sigma^2 + \frac{1}{3}bt)}\right). \tag{C.4}$$

APPENDIX D. LIST OF CONSTANTS

We list the explicit constants that appear in *Sections 2 and 3* . Below ω_d is the volume of unit ball in d dimensions, and $\gamma > 2$ is a parameter of choice related to the error rate in the following way: $\mathbb{P}(\text{ Boundary test fails }) = O(n^{-\gamma})$.

$$C_x = 2\omega_{d-1} + \frac{LR\omega_d}{\rho_{min}},$$
$$C_y = \frac{\omega_{d-1}}{2(d+1)},$$
$$C_r = \frac{1}{R}\max\left[\left(\frac{3\gamma\rho_{\max}d^2\omega_d R^2}{{C_x}^2\rho_{\min}^2}\right)^{\frac{1}{d+2}}, \left(\frac{4\gamma C_y d^2 2^{(d-1)/2}}{13\rho_{\min}\omega_{d-1}C_x}\right)^{\frac{1}{d+1}}\right],$$