SIAM J. MATH. DATA SCI.
Vol. 4, No. 1, pp. 72–99

Balancing Geometry and Density: Path Distances on High-Dimensional Data*

Anna Little†, Daniel McKenzie‡, and James M. Murphy§

Abstract. New geometric and computational analyses of power-weighted shortest path distances (PWSPDs) are presented. By illuminating the way these metrics balance geometry and density in the underlying data, we clarify their key parameters and illustrate how they provide multiple perspectives for data analysis. Comparisons are made with related data-driven metrics, which illustrate the broader role of density in kernel-based unsupervised and semisupervised machine learning. Computationally, we relate PWSPDs on complete weighted graphs to their analogues on weighted nearest neighbor graphs, providing high probability guarantees on their equivalence that are near-optimal. Connections with percolation theory are developed to establish estimates on the bias and variance of PWSPDs in the finite sample setting. The theoretical results are bolstered by illustrative experiments, demonstrating the versatility of PWSPDs for a wide range of data settings. Throughout the paper, our results generally require only that the underlying data is sampled from a compact low-dimensional manifold, and depend most crucially on the intrinsic dimension of this manifold, rather than its ambient dimension.

Key words. path distances, manifold learning, clustering, machine learning, high-dimensional statistics, kernel methods

AMS subject classifications. 60, 62, 68

DOI. 10.1137/20M1386657

1. Introduction.

The analysis of high-dimensional data is a challenge in modern statistical and machine learning. In order to defeat the *curse of dimensionality* [37, 33, 10], distance metrics that efficiently and accurately capture intrinsically low-dimensional latent structure in high-dimensional data are required. Indeed, this need to capture low-dimensional linear and nonlinear structure in data has led to the development of a range of data-dependent distances and related dimension reduction methods, which have been widely employed in applications [43, 55, 8, 26, 21, 56]. Understanding how these metrics trade off fundamental properties in the data (e.g., local versus global structure, geometry versus density) when making pointwise comparisons is an important challenge in their use and may be understood as a form of model selection in unsupervised and semisupervised machine learning problems.

*Received by the editors December 17, 2020; accepted for publication (in revised form) September 21, 2021; published electronically January 24, 2022.
https://doi.org/10.1137/20M1386657
Funding: The first author acknowledges partial support from the US National Science Foundation under grants DMS-1912906 and DMS-2131292. The second author acknowledges partial support from the US National Science Foundation under grant DMS-1720237 and the Office of Naval Research under grant N000141712162. The third author acknowledges partial support from the US National Science Foundation under grants DMS-1912737 and DMS-1924513.

†Department of Mathematics, University of Utah, Salt Lake City, UT 84112 USA (little@math.utah.edu).
‡Department of Mathematics, University of California, Los Angeles, CA 90095 USA (mckenzie@math.ucla.edu).
§Department of Mathematics, Tufts University, Medford, MA 02155 USA (jm.murphy@tufts.edu).

1.1. Power-weighted shortest path distances. In this paper we analyze *power-weighted shortest path distances (PWSPDs)* and develop their applications to problems in machine learning. These metrics compute the shortest path between two points in the data, accounting for the underlying density of the points along the path. Paths through low-density regions are penalized, so that the optimal path must balance being "short" (in the sense of the classical geodesic distance) with passing through high-density regions. We consider a finite data set $\mathcal{X} = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$, which we usually assume to be intrinsically low-dimensional, in the sense that there exists a compact d-dimensional Riemannian *data manifold* $\mathcal{M} \subset \mathbb{R}^D$ with $d \leq D$ and a probability density function $f(x)$ supported on $\mathcal{M}$ such that $\{x_i\}_{i=1}^n \overset{i.i.d.}{\sim} f(x)$, where i.i.d. is independent and identically distributed.

Definition 1.1. *For $p \in [1, \infty)$ and for $x, y \in \mathcal{X}$, the* (discrete) PWSPD *from x to y is*

$$\ell_p(x,y) = \min_{\pi = \{x_{i_j}\}_{j=1}^T} \left(\sum_{j=1}^{T-1} \|x_{i_j} - x_{i_{j+1}}\|^p \right)^{\frac{1}{p}}, \tag{1.1}$$

where π is a path of points in $\mathcal{X}$ with $x_{i_1} = x$ and $x_{i_T} = y$ and $\|\cdot\|$ is the Euclidean norm.

Early uses of density-based distances for interpolation [52] led to the formulation of PWSPD in the context of unsupervised and semisupervised learning and applications [30, 58, 17, 51, 18, 13, 46, 45, 41, 62, 16]. It will occasionally be useful to think of $\ell_p^p(\cdot,\cdot)$ as the path distance in the complete graph on $\mathcal{X}$ with edge weights $\|x_i - x_j\|^p$, which we shall denote $\mathcal{G}_{\mathcal{X}}^p$. When $p = 1$, $\ell_1(x,y) = \|x - y\|$, i.e., the Euclidean distance. As p increases, the largest elements in the set of path edge lengths $\{\|x_{i_j} - x_{i_{j+1}}\|\}_{j=1}^{T-1}$ begin to dominate the optimization (1.1), so that paths through higher-density regions (with shorter edge lengths) are promoted. When $p \to \infty$, ℓ_p converges (up to rescaling by the number of edges achieving maximal length) to the longest-leg path distance $\ell_\infty(x,y) = \min_{\pi = \{x_{i_j}\}_{j=1}^T} \max_{j=1,\ldots,T-1} \|x_{i_j} - x_{i_{j+1}}\|$ [41] and is thus driven by the density function f. Outside these extremes, ℓ_p balances taking a "short" path and taking one through regions of high density. Note that ℓ_p can be defined for $p < 1$, but it does not satisfy the triangle inequality and is thus not a metric (ℓ_p^p, however, *is* a metric for all $p > 0$). This case was studied in [2], where it is shown to have counterintuitive properties that should preclude its use in machine learning and data analysis.

While (1.1) is defined for finite data, it admits a corresponding continuum formulation.

Definition 1.2. *Let $(\mathcal{M}, g)$ be a compact, d-dimensional Riemannian manifold and f a continuous density function on $\mathcal{M}$ that is lower bounded away from zero (i.e., $f_{\min} := \min_{x \in \mathcal{M}} f(x) > 0$ on $\mathcal{M}$). For $p \in [1, \infty)$ and $x, y \in \mathcal{M}$, the* (continuum) PWSPD *from x to y is*

$$\mathcal{L}_p(x,y) = \left(\inf_\gamma \int_0^1 \frac{1}{f(\gamma(t))^{\frac{p-1}{d}}} \sqrt{g\left(\gamma'(t), \gamma'(t)\right)} dt \right)^{\frac{1}{p}}, \tag{1.2}$$

where $\gamma : [0,1] \to \mathcal{M}$ is a $\mathcal{C}^1$ path with $\gamma(0) = x, \gamma(1) = y$.

Note $\mathcal{L}_1$ is simply the geodesic distance on $\mathcal{M}$. However, for $p > 1$ and a nonuniform density, the optimal path γ is generally not the geodesic distance on $\mathcal{M}$: $\mathcal{L}_p$ favors paths which travel along high-density regions, and detours off the classical $\mathcal{L}_1$ geodesics are thus

acceptable. The parameter p controls how large a detour is optimal; for large p, optimal paths may become highly nonlocal and different from classical geodesic paths.

It is known [38, 32] that when f is continuous and positive, for $p > 1$ and all $x, y \in \mathcal{M}$,

$$\lim_{n\to\infty} n^{\frac{p-1}{pd}} \ell_p(x,y) = C_{p,d}\mathcal{L}_p(x,y) \tag{1.3}$$

for an absolute constant $C_{p,d}$ depending only on p and d, i.e., that the discrete PWSPD computed on an i.i.d. sample from f (appropriately rescaled) is a consistent estimator for the continuum PWSPD. In particular, (1.3) is established by [32] for C^1, isometrically embedded manifolds and by [38] for smooth, compact manifolds without boundary and for ℓ_p defined using geodesic distance. We thus define the normalized (discrete) path metric

$$\tilde{\ell}_p(x,y) := n^{\frac{p-1}{pd}} \ell_p(x,y)\,. \tag{1.4}$$

The $n^{\frac{p-1}{pd}}$ normalization factor accounts for the fact that for $p > 1$, ℓ_p converges uniformly to 0 as $n \to \infty$ [45]. Note that the $1/p$ exponent in (1.1) and (1.3) is necessary to obtain a metric that is homogeneous. Moreover, as $p \to \infty$, $\mathcal{L}_p$ is constant on regions of constant density, but $\mathcal{L}_p^p$ is not. Indeed, consider a uniform distribution on $[0,1]^d$, which has density $f = \mathbb{1}_{[0,1]^d}$. Then for all $x, y \in [0,1]^d$ and for all p, $\mathcal{L}_p^p(x,y) = \|x - y\|$. On the other hand, for all $x, y \in [0,1]^d$, $\mathcal{L}_p(x,y) = \|x - y\|^{1/p} \to 1$ as $p \to \infty$, i.e., all points are equidistant in the limit $p \to \infty$. Thus the $1/p$ exponent in (1.1) and (1.3) is necessary to obtain an entirely density-based metric for large p.

In practice, it is more efficient to compute PWSPDs in a sparse graph instead of a complete graph. It is thus natural to define PWSPDs *with respect to a subgraph* $\mathcal{H}$ *of* $\mathcal{G}_{\mathcal{X}}^p$.

Definition 1.3. *Let* $\mathcal{H}$ *be any subgraph of* $\mathcal{G}_{\mathcal{X}}^p$. *For* $x, y \in X$, *let* $\mathcal{P}_{\mathcal{H}}(x,y)$ *be the set of paths connecting* x *and* y *in* $\mathcal{H}$. *For* $p \in [1,\infty)$ *and for* $x, y \in \mathcal{X}$, *the* (discrete) PWSPD with respect to $\mathcal{H}$ *from* x *to* y *is*

$$\ell_p^{\mathcal{H}}(x,y) = \min_{\pi=\{x_{i_j}\}_{j=1}^T \in \mathcal{P}_{\mathcal{H}}(x,y)} \left(\sum_{j=1}^{T-1} \|x_{i_j} - x_{i_{j+1}}\|^p \right)^{\frac{1}{p}}.$$

Clearly $\ell_p^{\mathcal{G}_{\mathcal{X}}^p}(\cdot,\cdot) = \ell_p(\cdot,\cdot)$. In order to compute all-pairs PWSPDs in a complete graph with n nodes (i.e., $\ell_p(x_i, x_j)$ for all $x_i, x_j \in \mathcal{X}$), a direct application of Dijkstra's algorithm has complexity $O(n^3)$. Let $\mathcal{G}_{\mathcal{X}}^{p,k}$ denote the *kNN graph*, constructed from $\mathcal{G}_{\mathcal{X}}^p$ by retaining only edges $\{x, y\}$ if x is among the k nearest neighbors of y in $\mathcal{X}$ (we say "x is a kNN of y" for short) or vice versa. In some cases the PWSPDs with respect to $\mathcal{G}_{\mathcal{X}}^{p,k}$ are known to coincide with those computed in $\mathcal{G}_{\mathcal{X}}^p$ [32, 20]. If so, we say the kNN graph is a 1-*spanner* of $\mathcal{G}_{\mathcal{X}}^p$. This provides a significant computational advantage, since kNN graphs are much sparser, and reduces the complexity of computing all-pairs PWSPD to $O(kn^2)$ [39].

1.2. Summary of contributions. This article develops new analyses, computational insights, and applications of PWSPDs, which may be summarized in three major contributions. First, we establish that when $\frac{p}{d}$ is not too large, PWSPDs locally are density-rescaled Euclidean distances. We give precise error bounds that improve over known bounds [38] and are

tight enough to prove the local equivalence of Gaussian kernels constructed with PWSPD and density-rescaled Euclidean distances. We also develop related theory which clarifies the role of density in machine learning kernels more broadly. A range of machine learning kernels that normalize in order to mitigate or leverage differences in underlying density are considered and compared to PWSPD. Relatedly, we analyze how PWSPDs become increasingly influenced by the underlying density as $p \to \infty$. We also illustrate the role of density and benefits of PWSPDs on illustrative data sets.

Second, we improve and extend known bounds on k [32, 45, 20] guaranteeing that the kNN graph is a 1-spanner of $\mathcal{G}_{\mathcal{X}}^{p}$. Specifically, we show that for any $1 < p < \infty$, the kNN graph is a 1-spanner of $\mathcal{G}_{\mathcal{X}}^{p}$ with probability exceeding $1 - 1/n$ if $k \geq C_{p,d,f,\mathcal{M}} \cdot \log(n)$, for an explicit constant $C_{p,d,f,\mathcal{M}}$ that depends on the density power p, intrinsic dimension d, underlying density f, and the geometry of the manifold $\mathcal{M}$, but is crucially independent of n. These results are proved both in the case that the manifold is isometrically embedded and in the case that the edge lengths are in terms of intrinsic geodesic distance on the manifold. Our results provide an essential computational tool for the practical use of PWSPDs, and their key dependencies are verified numerically with extensive large-scale experiments.

Third, we bound the convergence rate of PWSPD to its continuum limit using a percolation theory framework, thereby quantifying the [38, 32] asymptotic convergence result (1.4). Specifically, we develop bias and variance estimates by relating results on Euclidean first passage percolation (FPP) to the PWSPD setting. Surprisingly, these results suggest that the variance of PWSPD is essentially independent of p and depends on the intrinsic dimension d in complex ways. Numerical experiments verify our theoretical analyses and suggest several conjectures related to Euclidean FPP that are of independent interest.

1.3. Notation. We shall use the notation in Table 1 consistently, though certain specialized notation will be introduced as required. We assume throughout that the data $\mathcal{X}$ is drawn from a compact Riemannian data manifold $(\mathcal{M}, g)$, with additional assumptions imposed on $\mathcal{M}$ as needed; we do not rigorously consider the more general case that $\mathcal{X}$ is drawn from a distribution supported *near* $\mathcal{M}$. If $\mathcal{M} \subset \mathbb{R}^D$, we assume that it is isometrically embedded in $\mathbb{R}^D$, i.e., g is the unique metric induced by restricting the Euclidean metric on $\mathbb{R}^D$ to $\mathcal{M}$, unless otherwise stated. If an event holds with probability $1 - c/n$, where $n = |\mathcal{X}|$ and c is independent of n, we say it holds *with high probability (w.h.p.)*.

2. Local analysis: Density and kernels.

Density-driven methods are commonly used for unsupervised and semisupervised learning [19, 27, 21, 49, 13, 7, 50]. Despite this popularity, the role of density is not completely clear in this context. Indeed, some methods seek to leverage variations in density while others mitigate it. In this section, we explore the role that density plays in popular machine learning kernels, including those used in self-tuning spectral clustering and diffusion maps. We compare with the effect of density in ℓ_p-based kernels and illustrate the primary advantages and disadvantages on toy data sets.

2.1. Role of density in graph Laplacian kernels. A large family of algorithms [8, 9, 54, 47, 59] view data points as the nodes of a graph and define the corresponding edge weights via a kernel function. In general, by kernel we mean a function $\mathcal{K} : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ that captures a notion of *similarity* between elements of $\mathbb{R}^D$. More precisely, we suppose that $\mathcal{K}$ is of the

Table 1
Notation used throughout the paper.

Notation	Definition
$\mathcal{X}$	$\mathcal{X} = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$, a finite data set
D	ambient dimension of data set $\mathcal{X}$
d	intrinsic dimension of data set $\mathcal{X}$
$\|v\|_p$	$\|v\|_p = (\sum_{i=1}^D \|v_i\|^p)^{\frac{1}{p}}$, the Euclidean p-norm of $v \in \mathbb{R}^D$
$\|v\|$	$\|v\|_2$, the Euclidean 2-norm
$\|c\|$	the absolute value of $c \in \mathbb{R}$
$\mathcal{G}_{\mathcal{X}}^p$	complete graph on $\mathcal{X}$ with edge weight $\|x_i - x_j\|^p$ between $x_i, x_j \in \mathcal{X}$
$\{x, y\}$	edge between nodes x, y in a graph
$(\mathcal{M}, g)$	a Riemannian manifold with associated metric g
κ	measure of curvature on $\mathcal{M}$; see Definition 2.1
κ_0	measure of regularity on $\mathcal{M}$; see Definition 3.7
ζ	reach of a manifold $\mathcal{M}$; see Definition 3.8
$f(x)$	probability density function from which $\mathcal{X}$ is drawn
$f_{\min}$, $f_{\max}$	minimum and maximum values of density f defined on compact manifold $\mathcal{M}$
$\{\pi_i\}_{i=1}^T$, $\gamma(t)$	discrete, continuous path
$\ell_p(x, y)$	discrete PWSPD; see (1.1)
$\bar{\ell}_p(x, y)$	rescaled version of $\ell_p(x, y)$; see (1.4)
$\ell_p^{\mathcal{H}}(x, y)$	discrete PWSPD defined on the subgraph $\mathcal{H} \subset \mathcal{G}_{\mathcal{X}}^p$; see Definition 1.3
$\mathcal{L}_p(x, y)$	continuum PWSPD; see (1.2)
$\mathscr{D}(x, y)$	geodesic distance on manifold $\mathcal{M}$
$\mathscr{D}_{f,\mathrm{Euc}}(x, y)$	density-based *stretch* of Euclidean distance with respect to f
$\mathfrak{L}$	Lipschitz constant of the density f, satisfying $\|f(x) - f(y)\| \leq \mathfrak{L}\mathscr{D}(x, y)$
W, Deg, L	weight, degree, and Laplacian matrices associated to a graph
$\delta(\cdot, \cdot)$	arbitrary metric
$B_\delta(x, \epsilon)$	$\{y \mid \delta(x, y) \leq \epsilon\}$, ball of radius $\epsilon > 0$ centered at x with respect to δ
$B(x, \epsilon)$	Euclidean ball of radius $\epsilon > 0$ centered at x, dimension determined by context
$\mathcal{D}_{\alpha,p}(x, y)$	p-elongated set of radius α based at points x, y; see Definition 3.4
k	number of nearest neighbors, sometimes dependent on n (i.e., $k = k(n)$)
μ, χ	percolation time, fluctuation constants
λ	intensity parameter in a Poisson point process
$\bar{A}$	complement of the set A
$\mathbb{E}[\xi], \mathrm{Var}[\xi]$	expectation, variance of a random variable ξ
$\mathrm{diam}(A)$	$\sup_{x,y \in A} \|x - y\|$, the Euclidean diameter of a set A
$\mathrm{vol}(A)$	volume of a set A, with dimension depending on context
$\bar{A}$	complement of a set A
∂A	boundary of a set A
$a \lesssim b$	$a \leq Cb$ for a constant C independent of the dependencies of a, b
$a \propto b$	quantity a is proportional to quantity b, i.e., $a \lesssim b$ and $b \lesssim a$

form $\mathcal{K}(x_i, x_j) = h(\delta(x_i, x_j))$ for some metric δ on $\mathbb{R}^D$ and smooth, positive, rapidly decaying (hence integrable) function $h : \mathbb{R} \to \mathbb{R}$. Our technical results will pertain exclusively to the Gaussian kernel $\mathcal{K}(x_i, x_j) = \exp(-\delta(x_i, x_j)^2/\epsilon^2)$ for some metric δ and scaling parameter $\epsilon > 0$, albeit more general kernels have been considered in the literature [4, 23, 11]. Given $\mathcal{X} \subset \mathbb{R}^D$, one first defines a weight matrix $W \in \mathbb{R}^{n \times n}$ by $W_{ij} = \mathcal{K}(x_i, x_j)$ for some kernel $\mathcal{K}$, and diagonal degree matrix $\mathrm{Deg} \in \mathbb{R}^{n \times n}$ by $\mathrm{Deg}_{ii} = \sum_{j=1}^n W_{ij}$. A *graph Laplacian* L is then defined using W, Deg. Then, the K lowest frequency eigenvectors of L, denoted $\phi_1, \ldots, \phi_K$,

define a K-dimensional spectral embedding of the data by $x_i \mapsto (\phi_1(x_i), \phi_2(x_i), \ldots, \phi_K(x_i))$, where $\phi_j(x_i) = (\phi_j)_i$. Commonly, a standard clustering algorithm such as K-means is then applied to the spectral embedding. This procedure is known as *spectral clustering* (SC). In unnormalized SC, $L = \mathrm{Deg} - W$, while in normalized SC either the random walk Laplacian $L_{\mathrm{RW}} = \mathrm{Deg}^{-1}L$ or the symmetric normalized Laplacian $L_{\mathrm{SYM}} = \mathrm{Deg}^{-1/2}L\mathrm{Deg}^{-1/2}$ is used.

Many modifications of this general framework have been considered. Although SC is better able to handle irregularly shaped clusters than many traditional algorithms [5, 53], it is often unstable in the presence of low degree points and sensitive to the choice of scaling parameter ϵ when using the Gaussian kernel [59]. These shortcomings motivated [61] to apply SC with the *self-tuning kernel* $W_{ij} = \exp(-\frac{\|x_i - x_j\|^2}{\sigma_{i,k}\sigma_{j,k}})$, where $\sigma_{i,k}$ is Euclidean distance of x_i to its kth NN. To clarify how the data density influences this kernel, consider how $\sigma_{i,k}$ relates to the kNN density estimator at x_i:

$$f_n(x_i) := \frac{k}{n\mathrm{vol}(B(0,1))\sigma_{i,k}^d}. \tag{2.1}$$

It is known [42] that if $k = k(n)$ is such that $k(n) \to \infty$ while $k(n)/n \to 0$, then $f_n(x_i)$ is a consistent estimator of $f(x_i)$, as long as f is continuous and positive. Furthermore, if f is uniformly continuous and $k(n)/\log n \to \infty$ while $k(n)/n \to 0$, then $\sup_i |f_n(x_i) - f(x_i)| \to 0$ with probability 1 [25]. Although these results assume the density f is supported in $\mathbb{R}^d$, the density estimator (2.1) is consistent in the general case when f is supported on a d-dimensional Riemannian manifold $\mathcal{M} \subseteq \mathbb{R}^D$ for $\log n \ll k(n) \ll n$ [28]. For such $k(n)$, $\sigma_{i,k} \to \epsilon_{n,d} f(x_i)^{-\frac{1}{d}}$ for some constant $\epsilon_{n,d}$ depending on n, d. Thus, for n large the kernel for self-tuning SC is approximately

$$W_{ij} \approx \exp\left(-f(x_i)^{\frac{1}{d}} f(x_j)^{\frac{1}{d}} \frac{\|x_i - x_j\|^2}{\epsilon_{n,d}^2}\right). \tag{2.2}$$

Relative to a standard SC kernel, (2.2) weakens connections in high-density regions and strengthens connections in low-density regions.

Diffusion maps [22, 21] is a more general framework which reduces to SC for certain parameter choices. More specifically, [21] considered the family of kernels

$$W_{ij} = \frac{\exp\left(-\|x_i - x_j\|^2/\epsilon^2\right)}{\deg(x_i)^a \deg(x_j)^a} \quad , \quad \deg(x_i) = \sum_{j=1}^{n} \exp\left(-\|x_i - x_j\|^2/\epsilon^2\right) \tag{2.3}$$

parametrized by $a \in [0, 1]$, which determines the degree of density normalization. Since $\deg(x_i) \propto f(x_i) + O(\epsilon^2)$, $\deg(x_i)$ is a kernel density estimator of the density $f(x_i)$ [12] and, up to higher-order terms,

$$W_{ij} \propto \frac{\exp\left(-\|x_i - x_j\|^2/\epsilon^2\right)}{f(x_i)^a f(x_j)^a}. \tag{2.4}$$

Note that f has an effect on the kernel similar to the self-tuning kernel (2.2): connections in high-density regions are weakened, and connections in low-density regions are strengthened.

Let $L_{\mathrm{RW}}^{a,\epsilon}$ denote the discrete random walk Laplacian using the weights W_{ij} given in (2.3). The discrete operator $-L_{\mathrm{RW}}^{a,\epsilon}/\epsilon^2$ converges to the continuum Kolmogorov operator $\mathscr{L}\psi = \Delta\psi + (2-2a)\nabla\psi \cdot \frac{\nabla f}{f}$ as $n \to \infty, \epsilon \to 0^+$ for Laplacian operator Δ and gradient ∇, both taken with respect to the Riemannian metric inherited from the ambient space [8, 21, 12]. When $a = 0$, we recover standard SC; there is no density renormalization in the kernel but the limiting operator is density dependent. When $a = 1$, $-L_{\mathrm{RW}}^{1,\epsilon}/\epsilon^2 \to \Delta$; in this case the discrete operator is density dependent but the limiting operator is purely geometric, since the density term is eliminated. We note that Laplacians and diffusion maps with various metrics and norms have been considered in a range of settings [60, 15, 57, 40].

2.2. Local characterization of PWSPD-based kernels. While the kernels discussed in section 2.1 compensate for discrepancies in density, PWSPD-based kernels strengthen connections through high-density regions and weaken connections through low-density regions. To illustrate more clearly the role of density in PWSPD-based kernels, we first show that locally the continuum PWSPD $\mathcal{L}_p^p$ is well-approximated by the density-based *stretch* of Euclidean distance $\mathscr{D}_{f,\mathrm{Euc}}(x,y) = \frac{\|x-y\|}{(f(x)f(y))^{\frac{p-1}{2d}}}$, as long as f does not vary too rapidly and $\mathcal{M}$ does not curve too quickly. This is quantified in Lemma 2.2, which is then used to prove Theorem 2.3, which bounds the local deviation of $\mathcal{L}_p$ from $\mathscr{D}_{f,\mathrm{Euc}}^{1/p}$. Finally, Corollary 2.4 establishes that Gaussian kernels constructed with $\mathcal{L}_p$ and $\mathscr{D}_{f,\mathrm{Euc}}^{1/p}$ are locally similar. Throughout this section we assume $\mathcal{M} \in S(d,\kappa,\epsilon_0)$ as defined below.

Definition 2.1. *An isometrically embedded Riemannian manifold $\mathcal{M} \subset \mathbb{R}^D$ is an element of $S(d,\kappa,\epsilon_0)$ if it is compact with dimension $d \le D$, vol$(\mathcal{M}) = 1$, and $\mathscr{D}(x,y) \le \|x-y\|(1+\kappa\|x-y\|^2)$ for all $x,y \in \mathcal{M}$ such that $\mathscr{D}(x,y) \le \epsilon_0$, where $\mathscr{D}(\cdot,\cdot)$ is geodesic distance on $\mathcal{M}$.*

The condition $\mathscr{D}(x,y) \le \|x-y\|(1+\kappa\|x-y\|^2)$ for all $x,y \in \mathcal{M}$ such that $\mathscr{D}(x,y) \le \epsilon_0$ is equivalent to an upper bound on the second fundamental form: $\|II_x\| \le \kappa$ for all $x \in \mathcal{M}$ [4, 44]. Note that this is also equivalent to a positive lower bound on the *reach* [29] of $\mathcal{M}$ (e.g., Proposition 6.1 in [48] and Proposition A.1 in [1]); see Definition 3.8.

Let $B_{\mathcal{L}_p^p}(x,\epsilon)$ and $B_{\mathscr{D}}(x,\epsilon)$ denote, respectively, the (closed) $\mathcal{L}_p^p$ and geodesic balls centered at x of radius ϵ. Let $f_{\max} = \max_y\{f(y) : y \in \mathcal{M}\}$, $f_{\min} = \min_y\{f(y) : y \in \mathcal{M}\}$ be the global density maximum and minimum. Define the following local quantities:

$$f_{\min}(x,\epsilon) = \min_y \left\{ f(y) : y \in B_{\mathscr{D}}\left(x, \epsilon\left(1+\kappa\epsilon^2\right)\right)\right\},$$
$$f_{\max}(x,\epsilon) = \max_y \left\{ f(y) : y \in B_{\mathcal{L}_p^p}\left(x, \epsilon(1+\kappa\epsilon^2)/f_{\min}(x,\epsilon)^{\frac{p-1}{d}}\right)\right\}.$$

Let $\rho_{x,\epsilon} = f_{\max}(x,\epsilon)/f_{\min}(x,\epsilon)$, which characterizes the local discrepancy in density in a ball of radius $O(\epsilon)$ around the point x.

The following lemma establishes that $\mathcal{L}_p^p$ and $\mathscr{D}_{f,\mathrm{Euc}}$ are locally equivalent and that discrepancies depend on $(\rho_{x,\epsilon})^{\frac{p-1}{d}}$ and the curvature constant κ. We note similar estimates appear in [2] for the special case $p = 0$. The proof appears in Appendix A.

Lemma 2.2. *Let $\mathcal{M} \in S(d, \kappa, \epsilon_0)$. Then for all $y \in \mathcal{M}$ with $\mathscr{D}(x, y) \leq \epsilon_0$ and $\|x - y\| \leq \epsilon$,*

$$\frac{1}{(\rho_{x,\epsilon})^{\frac{p-1}{d}}} \mathscr{D}_{f,Euc}(x, y) \leq \mathcal{L}_p^p(x, y) \leq (\rho_{x,\epsilon})^{\frac{p-1}{d}} (1 + \kappa\epsilon^2) \mathscr{D}_{f,Euc}(x, y)\,. \tag{2.5}$$

Note that corresponding bounds in terms of geodesic distance follow easily from the definition of $\mathcal{L}_p$: $f_{\max}(x,\epsilon)^{-\frac{p-1}{d}} \mathscr{D}(x, y) \leq \mathcal{L}_p^p(x, y) \leq f_{\min}(x,\epsilon)^{-\frac{p-1}{d}} \mathscr{D}(x, y)$. Lemma 2.2 thus establishes that the metrics $\mathcal{L}_p^p$ and $\mathscr{D}_{f,\text{Euc}}$ are locally equivalent when (i) $\rho_{x,\epsilon}$ is close to 1, (ii) $\frac{p-1}{d}$ is not too large, and (iii) κ is not too large. However, when $\frac{p-1}{d} \gg 1$, $\mathcal{L}_p^p$ balls may become highly nonlocal in terms of geodesics.

The following theorem establishes the local equivalence of $\mathcal{L}_p$ and $\mathscr{D}_{f,\text{Euc}}^{1/p}$ (and thus kernels constructed using these metrics). Assuming the density does not vary too quickly, Lemma 2.2 can be used to show that locally the difference between $\mathscr{D}_{f,\text{Euc}}^{1/p}$ and $\mathcal{L}_p$ is small. Variations in density are controlled by requiring that f is $\mathfrak{L}$-Lipschitz with respect to geodesic distance, i.e., $|f(x) - f(y)| \leq \mathfrak{L}\mathscr{D}(x, y)$. This Lipschitz assumption allows us to establish a higher-order equivalence compared to existing results (e.g., Corollary 9 in [38]), which we leverage to obtain the local kernel equivalence stated in Corollary 2.4. The following analysis also establishes explicit dependencies of the equivalence on $d, p, \mathfrak{L}, \kappa$.

Theorem 2.3. *Assume $\mathcal{M} \in S(d, \kappa, \epsilon_0)$ and that f is a bounded $\mathfrak{L}$-Lipschitz density function on $\mathcal{M}$ with $f_{min} > 0$. Let $\epsilon > 0$ and let*

$$\rho = \max_{x \in \mathcal{M}} \rho_{x,\epsilon},\ C_1 = \frac{\mathfrak{L}\left(\rho^{\frac{p-1}{d}} + 1\right)(p-1)}{f_{min}^{1+\frac{p-1}{pd}}\, pd},\ C_2 = \frac{\kappa}{f_{min}^{\frac{p-1}{pd}}\, p}.$$

Then for all $x, y \in \mathcal{M}$ such that $\mathscr{D}(x, y) \leq \epsilon_0$ and $\|x - y\| \leq \epsilon$,

$$\left|\mathcal{L}_p(x, y) - \mathscr{D}_{f,Euc}^{1/p}(x, y)\right| \leq C_1 \epsilon^{1+\frac{1}{p}} + C_2 \epsilon^{2+\frac{1}{p}} + O\left(\epsilon^{3+\frac{1}{p}}\right).$$

Proof. We first show that $\rho_{x,\epsilon}$ is close to 1. Let $y_1 \in B_{\mathcal{L}_p^p}(x, \epsilon(1 + \kappa\epsilon^2)/f_{\min}(x,\epsilon)^{\frac{p-1}{d}})$ satisfy $f(y_1) = f_{\max}(x,\epsilon)$ and $y_2 \in B_{\mathscr{D}}(x, \epsilon(1 + \kappa\epsilon^2))$ satisfy $f(y_2) = f_{\min}(x,\epsilon)$ (since these sets are compact, these points must exist). Then by the Lipschitz condition,

$$|\rho_{x,\epsilon} - 1| = \frac{|f(y_1) - f(y_2)|}{f(y_2)} \leq \frac{\mathfrak{L}\mathscr{D}(y_1, y_2)}{f(y_2)} \leq \frac{\mathfrak{L}\mathscr{D}(x, y_1) + \mathfrak{L}\mathscr{D}(x, y_2)}{f(y_2)}\,.$$

Let $\gamma_2(t)$ be a path achieving $\mathcal{L}_p^p(x, y_1)$. Note that

$$\frac{\mathscr{D}(x, y_1)}{f_{\max}(x,\epsilon)^{\frac{p-1}{d}}} \leq \int_0^1 \frac{1}{f(\gamma_2(t))^{\frac{p-1}{d}}} |\gamma_2'(t)|\, dt = \mathcal{L}_p^p(x, y_1) \leq \frac{\epsilon(1 + \kappa\epsilon^2)}{f_{\min}(x,\epsilon)^{\frac{p-1}{d}}}$$

so that $\mathscr{D}(x, y_1) \leq \rho_{x,\epsilon}^{\frac{p-1}{d}} \epsilon(1 + \kappa\epsilon^2)$, $\mathscr{D}(x, y_2) \leq \epsilon(1 + \kappa\epsilon^2)$. We thus obtain

$$\rho_{x,\epsilon} \leq 1 + \mathfrak{L}\left(\frac{\rho_{x,\epsilon}^{\frac{p-1}{d}} + 1}{f_{\min}(x,\epsilon)}\right)\epsilon(1 + \kappa\epsilon^2)\,. \tag{2.6}$$

Letting $C_{x,\epsilon} = \mathfrak{L}(\rho_{x,\epsilon}^{\frac{p-1}{d}} + 1)/f_{\min}(x,\epsilon)$, Taylor expanding around $\epsilon = 0$ and (2.6) give $\rho_{x,\epsilon}^{\frac{p-1}{pd}} \leq (1 + C_{x,\epsilon}\epsilon(1+\kappa\epsilon^2))^{\frac{p-1}{pd}} = 1 + C_{x,\epsilon}\frac{(p-1)}{pd}\epsilon + O(\epsilon^3)$. Applying Lemma 2.2 yields $(\rho_{x,\epsilon})^{-\frac{p-1}{pd}}\mathscr{D}_{f,\text{Euc}}^{1/p}(x,y) \leq \mathcal{L}_p(x,y) \leq (\rho_{x,\epsilon})^{\frac{p-1}{pd}}(1+\kappa\epsilon^2)^{\frac{1}{p}}\mathscr{D}_{f,\text{Euc}}^{1/p}(x,y)$, which gives

$$\frac{\mathscr{D}_{f,\text{Euc}}^{1/p}(x,y)}{\left(1 + C_{x,\epsilon}\frac{(p-1)}{pd}\epsilon + O(\epsilon^3)\right)} \leq \mathcal{L}_p(x,y) \leq \left(1 + C_{x,\epsilon}\frac{(p-1)}{pd}\epsilon + \frac{\kappa}{p}\epsilon^2 + O(\epsilon^3)\right)\mathscr{D}_{f,\text{Euc}}^{1/p}(x,y).$$

Rewriting the above yields

$$\left(1 - C_{x,\epsilon}\frac{(p-1)}{pd}\epsilon - \frac{\kappa}{p}\epsilon^2 + O(\epsilon^3)\right)\mathcal{L}_p(x,y) \leq \mathscr{D}_{f,\text{Euc}}^{1/p}(x,y) \leq \mathcal{L}_p(x,y)\left(1 + C_{x,\epsilon}\frac{(p-1)}{pd}\epsilon + O(\epsilon^3)\right).$$

We thus obtain

$$\begin{aligned}
&\left|\mathcal{L}_p(x,y) - \mathscr{D}_{f,\text{Euc}}^{1/p}(x,y)\right| \\
&\quad\leq \left(C_{x,\epsilon}\frac{(p-1)}{pd}\epsilon + \frac{\kappa}{p}\epsilon^2 + O(\epsilon^3)\right)\mathcal{L}_p(x,y) \\
&\quad\leq \left(C_{x,\epsilon}\frac{(p-1)}{pd}\epsilon + \frac{\kappa}{p}\epsilon^2 + O(\epsilon^3)\right)\frac{\epsilon^{\frac{1}{p}}(1+\kappa\epsilon^2)^{\frac{1}{p}}}{f_{\min}(x,\epsilon)^{\frac{p-1}{pd}}} \\
&\quad= \left(\frac{C_{x,\epsilon}}{f_{\min}(x,\epsilon)^{\frac{p-1}{pd}}}\frac{(p-1)}{pd}\epsilon^{1+\frac{1}{p}} + \frac{\kappa}{pf_{\min}(x,\epsilon)^{\frac{p-1}{pd}}}\epsilon^{2+\frac{1}{p}} + O\left(\epsilon^{3+\frac{1}{p}}\right)\right).
\end{aligned}$$

■

Note the coefficient C_1 increases exponentially in p; thus the equivalence between $\mathcal{L}_p$ and $\mathscr{D}_{f,\text{Euc}}^{1/p}$ is weaker for large p. We also emphasize that in a Euclidean ball of radius ϵ, the metric $\mathcal{L}_p$ scales like $\epsilon^{\frac{1}{p}}$; Theorem 2.3 thus guarantees that the relative error of approximating $\mathcal{L}_p$ with $\mathscr{D}_{f,\text{Euc}}^{1/p}$ is $O(\epsilon)$.

When $\mathcal{L}_p$ is locally well-approximated by $\mathscr{D}_{f,\text{Euc}}^{1/p}$, the kernels constructed from these two metrics are also locally similar. The following corollary leverages the error term in Theorem 2.3 to make this precise for Gaussian kernels. It is a direct consequence of Theorem 2.3 and Taylor expanding the Gaussian kernel, and its proof is given in the supplementary material file PWSPD_Supplement_Final.pdf [local/web 1.06MB]. Let $h_a(x) = \exp(-x^{2a})$ so that $h_1(\frac{\delta(\cdot,\cdot)}{\epsilon})$ is the Gaussian kernel with metric $\delta(\cdot,\cdot)$ and scaling parameter $\epsilon > 0$. Note $h_1(\frac{\mathcal{L}_p}{\epsilon^{1/p}}) = h_{\frac{1}{p}}(\frac{\mathcal{L}_p^p}{\epsilon})$.

Corollary 2.4. *Under the assumptions and notation of Theorem 2.3, for* $\tilde{C}_i = C_i/f_{min}^{\frac{p-1}{pd}}$,

$$\frac{\left|h_{\frac{1}{p}}\left(\mathcal{L}_p^p(x,y)/\epsilon\right) - h_{\frac{1}{p}}\left(\mathscr{D}_{f,Euc}(x,y)/\epsilon\right)\right|}{h_{\frac{1}{p}}\left(\mathcal{L}_p^p(x,y)/\epsilon\right)} \leq \tilde{C}_1\epsilon + \left(\tilde{C}_2 + \frac{1}{2}\tilde{C}_1^2\right)\epsilon^2 + O(\epsilon^3).$$

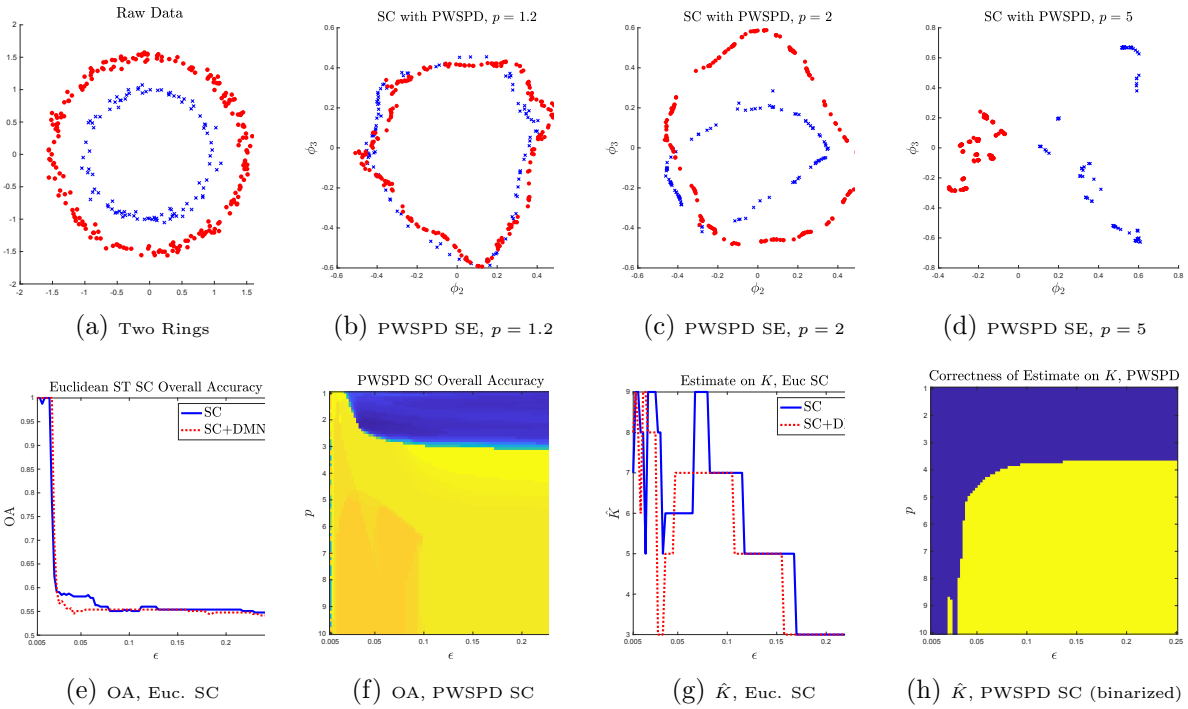
When $p-1$ is not too large relative to d, a kernel constructed with $\mathcal{L}_p$ is locally well-approximated by a kernel constructed with $\mathscr{D}_{f,\text{Euc}}^{1/p}$. Thus, in a Euclidean ball of radius ϵ, we may think of the Gaussian $\mathcal{L}_p$ kernel as

$$h_1\left(\frac{\mathcal{L}_p(x_i,x_j)}{\epsilon^{1/p}}\right) \approx h_{\frac{1}{p}}\left(\frac{\|x_i-x_j\|}{\epsilon(f(x_i)f(x_j))^{\frac{p-1}{2d}}}\right).$$

Density plays a different role in this kernel compared with those of section 2.1. This kernel strengthens connections in high-density regions and weakens them in low-density regions.

We note that the $\frac{1}{p}$-power in Definition 1.2 has a large impact, in that $\mathcal{L}_p$-based and $\mathcal{L}_p^p$-based kernels have very different properties. More specifically, $h_1(\mathcal{L}_p^p/\epsilon)$ is a local kernel as defined in [12], so it is sufficient to analyze the kernel locally. However, $h_1(\mathcal{L}_p/\epsilon^{1/p})$ is a nonlocal kernel, so that nontrivial connections between distant points are possible. The analysis in this section thus establishes the global equivalence of $h_1(\mathcal{L}_p^p/\epsilon)$ and $h_1(\mathscr{D}_{f,\mathrm{Euc}}/\epsilon)$ (when p is not too large relative to d) but only the local equivalence of $h_{\frac{1}{p}}(\mathcal{L}_p^p/\epsilon)$ and $h_{\frac{1}{p}}(\mathscr{D}_{f,\mathrm{Euc}}/\epsilon)$.

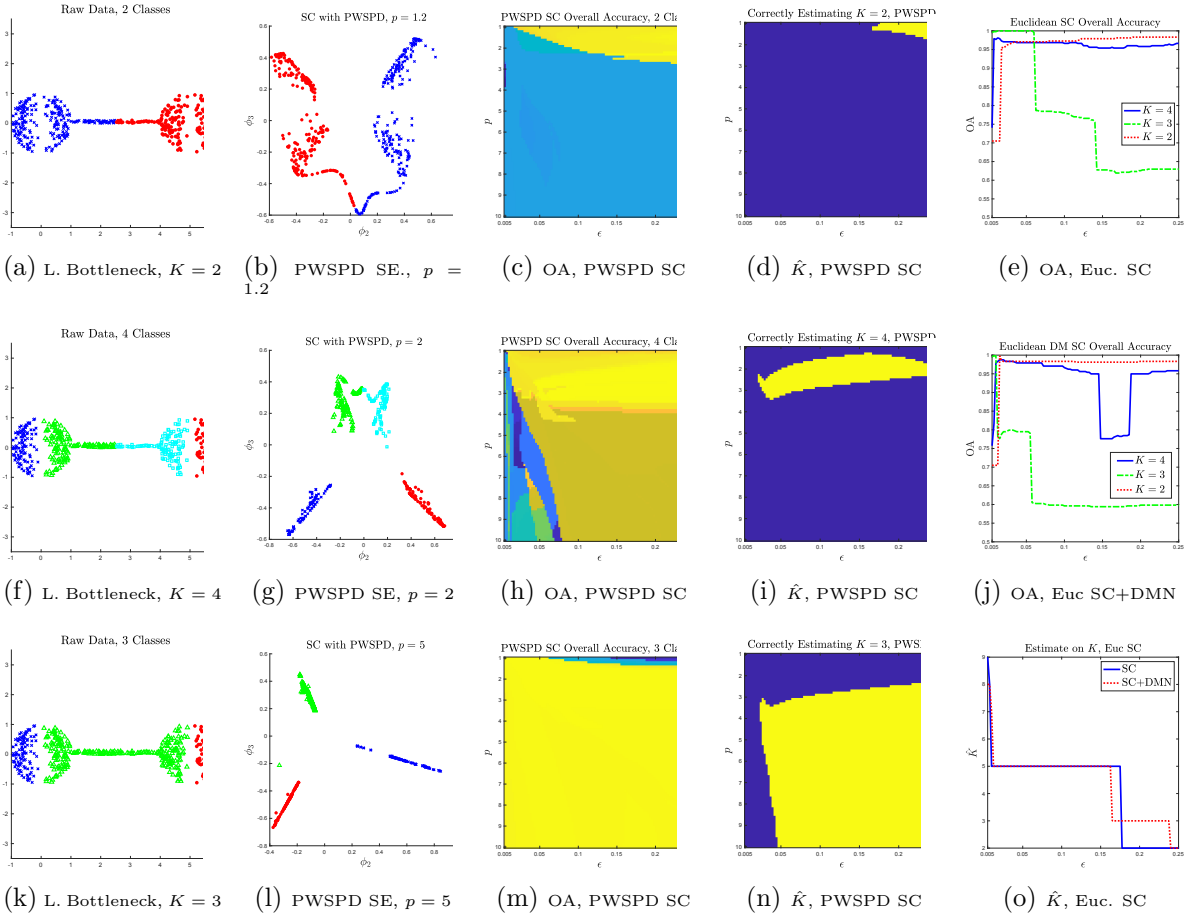
2.3. The role of p: Examples. This subsection illustrates the useful properties of PWSPDs and the role of p on three synthetic data sets in $\mathbb{R}^2$: (1) *two rings* data, consisting of two nonconvex clusters that are well-separated by a low-density region; (2) *long bottleneck* data, consisting of two isotropic clusters each with a density gap connected by a long, thin bottleneck; (3) *short bottleneck* data, where two elongated clusters are connected by a short bottleneck. The data sets are shown in Figures 1, 2, and 3, respectively. We also show the



(a) Two Rings (b) PWSPD SE, $p = 1.2$ (c) PWSPD SE, $p = 2$ (d) PWSPD SE, $p = 5$

(e) OA, Euc. SC (f) OA, PWSPD SC (g) $\hat{K}$, Euc. SC (h) $\hat{K}$, PWSPD SC (binarized)

Figure 1. *Two rings data set. Because the underlying cluster structure is density driven, the PWSPD SE separates the clusters for large p (see (d)). While taking ϵ small in Euclidean SC can allow for good clustering accuracy (see (e)), the range is narrow and does not permit accurate estimation of K via the eigengap (see (g)). On the other hand, PWSPD consistently clusters well and correctly captures $K = 2$ for a wide range of (ϵ, p) pairs (see (f), (h)). Generally, PWSPD allows for fully unsupervised clustering as long as p is sufficiently large and ϵ not too small.*



(a) L. Bottleneck, $K = 2$ (b) PWSPD SE., $p = 1.2$ (c) OA, PWSPD SC (d) $\hat{K}$, PWSPD SC (e) OA, Euc. SC

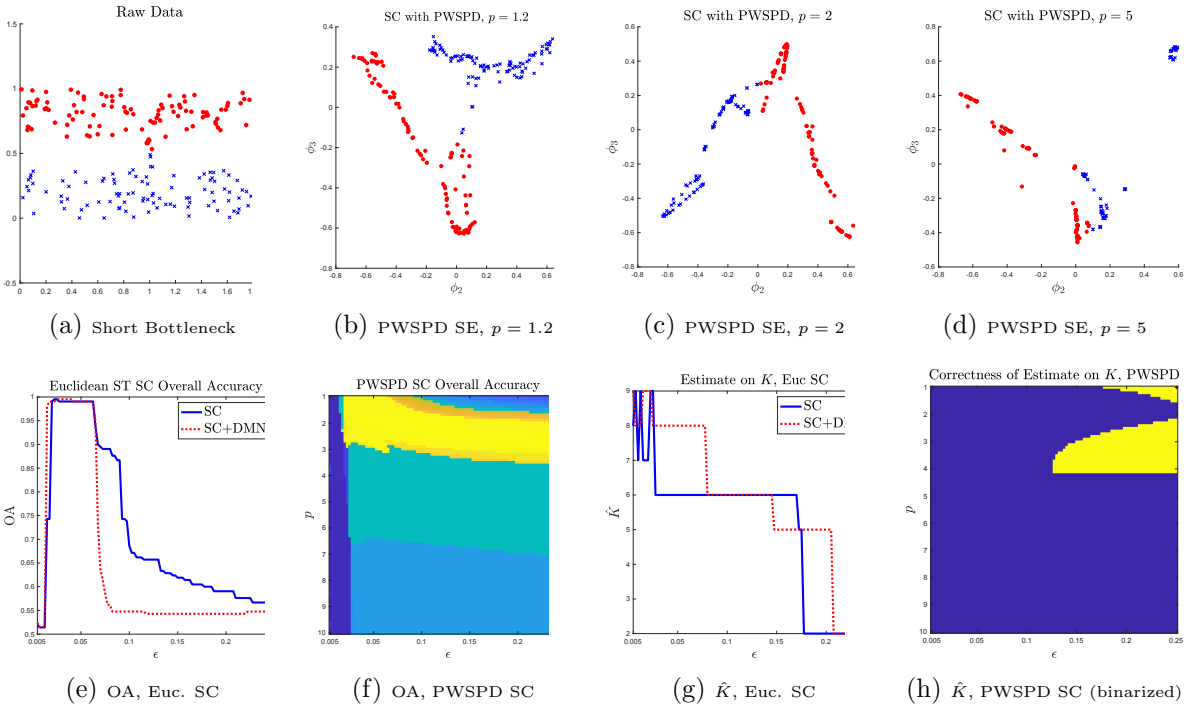
(f) L. Bottleneck, $K = 4$ (g) PWSPD SE, $p = 2$ (h) OA, PWSPD SC (i) $\hat{K}$, PWSPD SC (j) OA, Euc SC+DMN

(k) L. Bottleneck, $K = 3$ (l) PWSPD SE, $p = 5$ (m) OA, PWSPD SC (n) $\hat{K}$, PWSPD SC (o) $\hat{K}$, Euc. SC

Figure 2. *Long bottleneck data set. Different latent cluster structures exist in this data, driven by geometry ((a), $K = 2$), density ((k), $K = 3$), and a combination of geometry and density ((f), $K = 4$). When varying p, the PWSPD SE separates by geometry (see (b)) for p near 1, before separating by density for $p \gg 1$ (see (l)). Given the correct choice of ϵ and a priori knowledge of K, any of the three natural clusterings can be learned by Euclidean SC (see (e), (j)). However, in the Euclidean SC case, correct estimation of K fails to coincide with parameters that give good clustering results (see (o)). On the other hand, PWSPD SC is able to correctly estimate each of $K = 2, 3, 4$ for some choice of (ϵ, p) parameters in the same region that such parameters yield high clustering accuracy ((c), (d) for $K = 2$; (m), (n) for $K = 3$; (h), (i) for $K = 4$).*

PWSPD spectral embedding (denoted PWSPD SE) for various p, computed from a symmetric normalized Laplacian constructed with PWSPD. The scaling parameter ϵ for each data set is chosen as the 15th percentile of pairwise PWSPD distances.

Different aspects of the data are emphasized in the low-dimensional PWSPD embedding as p varies. Indeed, in Figure 1, we see the PWSPD embedding separates the rings for large p but not for small p. In Figure 2, we see separation across the bottleneck for p small, while for p large there is separation with respect to the density gradients that appear in the two bells of the dumbbell. Interestingly, separation with respect to both density and geometry is observed for $p = 2$ (see Figure 2(g)). In Figure 3, the clusters both are elongated and lack robust



(a) Short Bottleneck (b) PWSPD SE, $p = 1.2$ (c) PWSPD SE, $p = 2$ (d) PWSPD SE, $p = 5$

(e) OA, Euc. SC (f) OA, PWSPD SC (g) $\hat{K}$, Euc. SC (h) $\hat{K}$, PWSPD SC (binarized)

Figure 3. *Short bottleneck data set. Because the underlying cluster structure is not driven entirely by geometry or density, the PWSPD SE separates the clusters for moderate p (see (c)). We note PWSPD is able to correctly learn K and cluster accurately for ϵ somewhat large and p between 2 and 3 (see (f), (h)), while Euclidean SC cannot simultaneously learn K and cluster accurately (see (e), (g)).*

density separation, but the PWSPD embedding well separates the two clusters for moderate p. In general, p close to 1 emphasizes the geometry of the data, large p emphasizes the density structure of the data, and moderate p defines a metric balancing these two considerations.

2.3.1. Comparison with Euclidean spectral clustering. To evaluate how p impacts the clusterability of the PWSPD spectral embedding, we consider experiments in which we run SC under various graph constructions. We run K-means for a range of parameters on the spectral embedding $x_i \mapsto (\phi_2(x_i), \ldots, \phi_K(x_i))$, where ϕ_k is the kth lowest frequency eigenvector of the Laplacian. We construct the symmetric normalized Laplacian using PWSPD (denoted PWSPD SC) and also using Euclidean distances (denoted SC) and the Laplacian with diffusion maps normalization $a = 1$ (denoted SC+DMN). We vary ϵ in the SC and SC+DMN methods and both ϵ and p in the PWSPD SC method. Results for self-tuning SC, in which the kNN used to compute the local scaling parameter varies, are in the supplementary material file PWSPD_Supplement_Final.pdf [local/web 1.06MB]. To allow for comparisons across figures, ϵ is varied across the percentiles of the pairwise distances in the underlying data, up to the 25th percentile. We measure two outputs of the clustering experiments:

(i) The *overall accuracy (OA)*, namely the proportion of data points correctly clustered when K is known a priori. For $K = 2$, similar results were observed when thresholding ϕ_2 at 0 instead of running K-means; see the supplementary material file PWSPD_Supplement_Final.pdf [local/web 1.06MB].

(ii) The *eigengap estimate* of the number of latent clusters: $\hat{K} = \arg\max_{k \geq 2} \lambda_{k+1} - \lambda_k$, where $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the eigenvalues of the corresponding graph Laplacian. We note that experiments estimating K by considering the ratio of consecutive eigenvalues were also performed, with similar results. In the case of PWSPD SC, we plot heatmaps of where K is correctly estimated, with yellow corresponding to success ($\hat{K} = K$) and blue corresponding to failure ($\hat{K} \neq K$).

The results in terms of OA and $\hat{K}$ as a function of ϵ and p are in Figures 1–3. We see that when density separates the data clearly, as in the two rings data, PWSPD SC with large p gives accurate clustering results, while small p may fail. In this data set, ϵ very small allows for the data to be correctly clustered with SC and SC+DMN when K is known a priori. However, the regime of ϵ is so small that the eigenvalues become unhelpful for estimating the number of latent clusters. Unlike Euclidean SC, PWSPD SC correctly estimates $\hat{K} = 2$ for a range of parameters and achieves near-perfect clustering results for those parameters as well. Indeed, as shown by Figures 1(f) and 1(h), PWSPD SC with p large is able to do fully unsupervised clustering on the two rings data.

In the case of the long bottleneck data set, there are three reasonable latent clusterings, depending on whether geometry, density, or both matter (see Figures 2(a), 2(k), 2(f)). PWSPD is able to balance between the geometry and density-driven cluster structure in the data. Indeed, all of the cluster configurations shown in Figures 2(a), 2(k), and 2(f) are learnable without supervision for some choice of parameters (ϵ, p). To capture the density cluster structure ($K = 3$), p should be taken large, as suggested in Figures 2(m) and 2(n). To capture the geometry cluster structure ($K = 2$), p should be taken small and ϵ large, as suggested by Figures 2(c) and 2(d). Interestingly, both cluster and geometry ($K = 4$) can be captured by choosing p moderate, as in Figures 2(h) and 2(i). For Euclidean SC, varying ϵ is insufficient to capture the rich structure of this data.

In the case of the short bottleneck, taking ϵ large allows for the Euclidean methods to correctly estimate the number of clusters. But, in this ϵ regime, the methods do not cluster accurately. On the other hand, taking p between 2 and 3 and ϵ large allows PWSPD to correctly estimate K and also cluster accurately.

Overall, this suggests that varying p in PWSPD SC has a different impact than varying the scaling parameter ϵ and can allow for richer cluster structures to be learned when compared to SC with Euclidean distances. In addition, PWSPDs generally allow for the underlying cluster structures to be learned in a *fully unsupervised manner*, while Euclidean methods may struggle to simultaneously cluster well and estimate K accurately.

3. Spanners for PWSPD.

Let $\mathcal{H} \subset \mathcal{G}_{\mathcal{X}}^p$ denote a subgraph and recall the definition of $\ell_p^{\mathcal{H}}(\cdot, \cdot)$ given in Definition 1.3.

Definition 3.1. *For $t \geq 1$, $\mathcal{H} \subset \mathcal{G}_{\mathcal{X}}^p$ is a* t-spanner *if $\ell_p^{\mathcal{H}}(x, y) \leq t\ell_p(x, y)$ for all $x, y \in \mathcal{X}$.*

Clearly $\ell_p(x, y) \leq \ell_p^{\mathcal{H}}(x, y)$ always, as any path in $\mathcal{H}$ is a path in $\mathcal{G}_{\mathcal{X}}^p$. Hence if $\mathcal{H}$ is a 1-*spanner* we have equality: $\ell_p^{\mathcal{H}}(x, y) = \ell_p(x, y)$. Define the kNN graph, $\mathcal{G}_{\mathcal{X}}^{p,k}$, by retaining only edges $\{x, y\}$ if x is a kNN of y or vice versa. For appropriate k, p, and $\mathcal{M}$ it is known that $\mathcal{G}_{\mathcal{X}}^{p,k}$ is a 1-spanner of $\mathcal{G}_{\mathcal{X}}^p$ w.h.p. Specifically, [32] shows this when $\mathcal{M}$ is an open connected set with C^1 boundary, $1 < p < \infty$ and $k = O(c_{p,d} \log(n))$ for a constant $c_{p,d}$ depending on p, d.

One can deduce $c_{p,d} \geq 2^{d+1}3^d d^{d/2}$, while the dependence on p is more obscure. A different approach is used in [20] to show this for arbitrary smooth, closed, isometrically embedded $\mathcal{M}$, $2 \leq p < \infty$, and $k = O(2^d \log(n))$, where O hides constants depending on the geometry of $\mathcal{M}$. In both cases f must be continuous and bounded away from zero.

Under these assumptions, we prove $\mathcal{G}_{\mathcal{X}}^{p,k}$ is a 1-spanner w.h.p. for any smooth, closed, isometrically embedded $\mathcal{M}$ with mild restrictions on its curvature. Our results hold generally for $1 < p < \infty$ and enjoy improved dependence of k on d and explicit dependence of k on p and the geometry of $\mathcal{M}$ compared to [32, 20]. We also consider an *intrinsic* version of PWSPD,

$$\ell_{\mathcal{M},p}(x,y) = \left(\min_{\pi=\{x_{i_j}\}_{j=1}^{T}} \sum_{j=1}^{T-1} \mathscr{D}(x_{i_j}, x_{i_{j+1}})^p \right)^{1/p},$$

where $\mathscr{D}(\cdot,\cdot)$ is assumed known, which is not typically the case in data science. However, this situation can occur when $\mathcal{X}$ is presented as a subset of $\mathbb{R}^D$, but one wishes to analyze $\mathcal{X}$ with an exotic metric (i.e., not $\|\cdot\|$). For example, if each $x_i \in \mathcal{X}$ is an image, a Wasserstein metric may be more appropriate than $\|\cdot\|$. As this case closely mirrors the statement and proof of Theorem 3.9 we leave it to the supplementary material file PWSPD_Supplement_Final.pdf [local/web 1.06MB]. Before proceeding we introduce some further terminology.

Definition 3.2. *The edge $\{x,y\}$ is* critical *if it is in the shortest path from x to y in $\mathcal{G}_{\mathcal{X}}^p$.*

Lemma 3.3 ([20]). *$\mathcal{H} \subset \mathcal{G}_{\mathcal{X}}^p$ is a 1-spanner if it contains every critical edge of $\mathcal{G}_{\mathcal{X}}^p$.*

3.1. Nearest neighbors and PWSPD spanners. A key proof ingredient is the following definition, which generalizes the role of spheres in the proof of Theorem 1.3 in [20].

Definition 3.4. *For any $x, y \in \mathbb{R}^d$ and $\alpha \in (0,1]$, the* p-elongated set *associated to x, y is*

$$\mathcal{D}_{\alpha,p}(x,y) = \left\{ z \in \mathbb{R}^d : \ \|x-z\|^p + \|y-z\|^p \leq \alpha \|x-y\|^p \right\}.$$

Visualizations of $\mathcal{D}_{1,p}(x,y) \subset \mathbb{R}^2$ are shown in Figure 4. $\mathcal{D}_{1,p}(x,y)$ is the set of points z such that the two-hop path, $x \to z \to y$, is ℓ_p-shorter than the one-hop path, $x \to y$. Hence, we have the following.

Lemma 3.5. *If there exists $z \in \mathcal{D}_{1,p}(x,y) \cap \mathcal{X}$, then the edge $\{x,y\}$ is not critical.*

We defer the proof of the following technical lemma to Appendix B.

Lemma 3.6. *Let $r := \|x-y\|$, $x_M = \frac{x+y}{2}$, and $r^\star := r\sqrt{\frac{\alpha^{2/p}}{4^{1/p}} - \frac{1}{4}}$ for $\alpha > 2^{1-p}$. Then,*

$$B(x_M, r^\star) \subset \mathcal{D}_{\alpha,p}(x,y) \subset B(x,r).$$

For $\alpha = 1$, [32] makes a similar claim but crucially does not quantify the dependence of the radius of this ball on p. Before proceeding, we introduce two regularity assumptions.

Definition 3.7. *$\mathcal{M} \subset \mathbb{R}^D$ is in $V(d, \kappa_0, \epsilon_0)$ for $\kappa_0 \geq 1$ and $\epsilon_0 > 0$ if it is connected and for all $x \in \mathcal{M}$, $\epsilon \in (0, \epsilon_0)$ we have $\kappa_0^{-1}\epsilon^d \leq vol(\mathcal{M} \cap B(x,\epsilon))/vol(B(0,1)) \leq \kappa_0 \epsilon^d$.*

Definition 3.8. *A compact manifold $\mathcal{M} \subset \mathbb{R}^D$ has* reach *$\zeta > 0$ if every $x \in \mathbb{R}^D$ satisfying $dist(x,\mathcal{M}) := \min_{y \in \mathcal{M}} \|x-y\| < \zeta$ has a unique projection onto $\mathcal{M}$.*

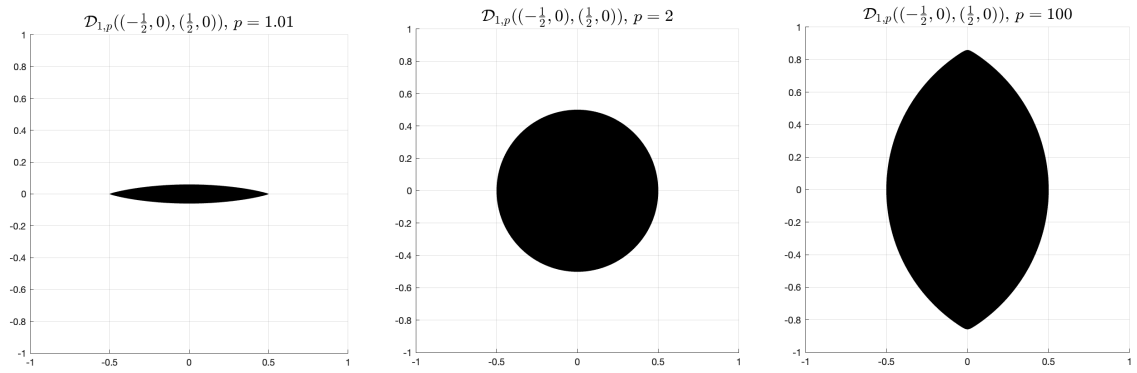


Figure 4. *Plots of* $\mathcal{D}_{1,p}((-\frac{1}{2},0),(\frac{1}{2},0))$ *for* $p = 1.01, 2, 100$. *We see that for smaller* p, *the set becomes quite small, converging to a line segment as* $p \to 1^+$. *For* $p = 2$, *the* p*-elongated set is a circle. As* p *increases,* $\mathcal{D}_{1,p}((-\frac{1}{2},0),(\frac{1}{2},0))$ *converges to a set resembling a vertically oriented American football.*

Theorem 3.9. *Let* $\mathcal{M} \in V(d, \kappa_0, \epsilon_0)$ *be a compact manifold with reach* $\zeta > 0$. *Let* $\mathcal{X} = \{x_i\}_{i=1}^n$ *be drawn i.i.d. from* $\mathcal{M}$ *according to a probability distribution with continuous density* f *satisfying* $0 < f_{\min} \le f(x) \le f_{\max}$ *for all* $x \in \mathcal{M}$. *For* $p > 1$ *and* n *sufficiently large,* $\mathcal{G}_{\mathcal{X}}^{p,k}$ *is a* 1*-spanner of* $\mathcal{G}_{\mathcal{X}}^{p}$ *with probability at least* $1 - 1/n$ *if*

$$k \ge 4\kappa_0^2 \left[\frac{f_{\max}}{f_{\min}}\right] \left[\frac{4}{4^{1-1/p} - 1}\right]^{d/2} \log(n). \tag{3.1}$$

Proof. In light of Lemma 3.3 we prove that, with probability at least $1-1/n$, $\mathcal{G}_{\mathcal{X}}^{p,k}$ contains every critical edge of $\mathcal{G}_{\mathcal{X}}^{p}$. Equivalently, we show every edge of $\mathcal{G}_{\mathcal{X}}^{p}$ not contained in $\mathcal{G}_{\mathcal{X}}^{p,k}$ is not critical.

For any $c, \epsilon > 0$, $\mathbb{P}[\max_{x,y\in\mathcal{X}} \ell_p(x,y) \le \epsilon] \ge 1 - c/n$ for n sufficiently large [45]. So, let n be sufficiently large so that $\mathbb{P}[\ell_p(x,y) \le \min\{\epsilon_0, \frac{\zeta}{d}\sqrt{\frac{1}{4^{1/p}} - \frac{1}{4}}\}$ for all $x, y \in \mathcal{X}] \ge (1 - \frac{1}{2n})$. Pick any $x, y \in \mathcal{X}$ which are not kNNs and let $r := \|x - y\|$. If $r > \min\{\epsilon_0, \frac{\zeta}{d}\sqrt{\frac{1}{4^{1/p}} - \frac{1}{4}}\}$, then $\ell_p(x,y) < \|x - y\|$ and thus the edge $\{x, y\}$ is not critical. So, suppose without loss of generality in what follows that $r \le \min\{\epsilon_0, \frac{\zeta}{d}\sqrt{\frac{1}{4^{1/p}} - \frac{1}{4}}\}$.

Define $r_1^\star := r\sqrt{\frac{1}{4^{1/p}} - \frac{1}{4}}$ and $r_2^\star := r(\sqrt{\frac{1}{4^{1/p}} - \frac{1}{4}} - \frac{r}{4\zeta})$; note that $r_2^\star > 0$ by the assumption $r \le \frac{\zeta}{d}\sqrt{\frac{1}{4^{1/p}} - \frac{1}{4}}$. Let $x_M := \frac{x+y}{2}$ and let $\tilde{x}_M := \arg\min_{z\in\mathcal{M}} \|x_M - z\|$ be the projection of x_M onto $\mathcal{M}$, which is unique because $r < \zeta$. By Lemma 3.6, $B(x_M, r_1^\star) \subset \mathcal{D}_{1,p}(x,y) \subset B(x,r)$. By Lemma B.1, $B(\tilde{x}_M, r_2^\star) \subset B(x_M, r_1^\star)$. Let $x_{i_1}, \dots, x_{i_k}$ denote the kNNs of x, ordered randomly. Because y is not a kNN of x, $\|x - x_{i_j}\| \le \|x - y\| = r$ for $j = 1, \dots, k$. Thus, $x_{i_j} \in B(x,r)$ and so by Lemma B.2 we bound for fixed j

$$\mathbb{P}\left[x_{i_j} \in \mathcal{D}_{1,p}(x,y) \mid x_{i_j} \in B(x,r)\right] \ge \mathbb{P}\left[x_{i_j} \in B(\tilde{x}_M, r_2^\star) \mid x_{i_j} \in B(x,r)\right] \tag{3.2}$$

$$\ge \frac{3}{4}\kappa_0^{-2}\frac{f_{\min}}{f_{\max}}\left(\frac{1}{4^{1/p}} - \frac{1}{4}\right)^{d/2} =: \varepsilon_{\mathcal{M},p,f}. \tag{3.3}$$

Because the x_{i_j} are all independently drawn,

$$\mathbb{P}\left[\nexists j \text{ with } x_{i_j} \in \mathcal{D}_{1,p}(x,y)\right] = \prod_{j=1}^{k} \mathbb{P}\left[x_{i_j} \notin \mathcal{D}_{1,p}(x,y) \mid x_{i_j} \in B(x,r)\right] \leq (1-\varepsilon_{\mathcal{M},p,f})^k.$$

A routine calculations reveals that for $k \geq \frac{3\log n}{-\log(1-\varepsilon_{\mathcal{M},p,f})}$,

$$\mathbb{P}\left[\exists j \text{ with } x_{i_j} \in \mathcal{D}_{1,p}(x,y)\right] = 1 - \mathbb{P}\left[\nexists j \text{ with } x_{i_j} \in \mathcal{D}_{1,p}(x,y)\right] \geq 1 - \frac{1}{n^3}. \tag{3.4}$$

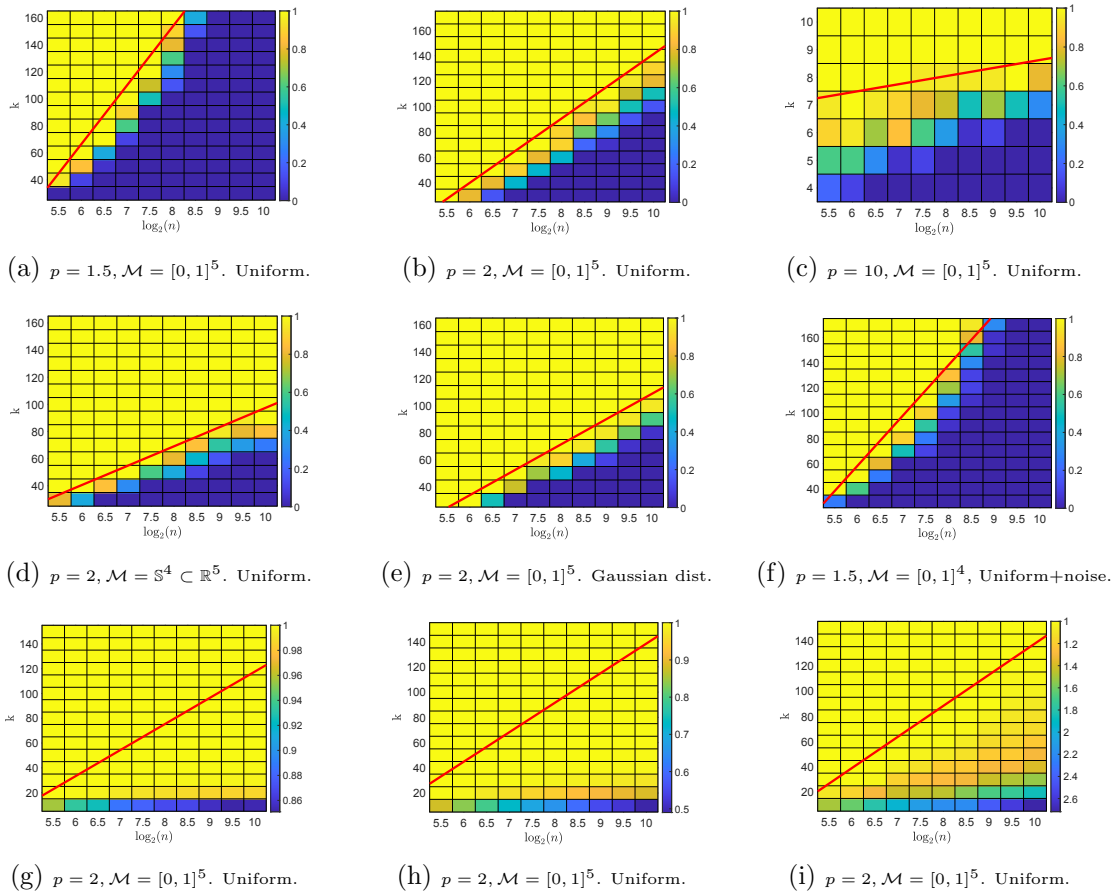
By Lemma 3.5 we conclude the edge $\{x,y\}$ is not critical with probability exceeding $1-\frac{1}{n^3}$. There are fewer than $n(n-1)/2$ such non-kNN pairs $x,y \in \mathcal{X}$. These edges $\{x,y\}$ are precisely those contained in $\mathcal{G}_{\mathcal{X}}^{p}$ but not in $\mathcal{G}_{\mathcal{X}}^{p,k}$. By the union bound and (3.4) we conclude that none of these are critical with probability greater than $1-\frac{n(n-1)}{2}\frac{1}{n^3} \geq 1-\frac{1}{2n}$. This was conditioned on $\ell_p(x,y) \leq \min\{\epsilon_0, \frac{\zeta}{d}\sqrt{\frac{1}{4^{1/p}}-\frac{1}{4}}\}$ for all $x,y \in \mathcal{X}$, which holds with probability exceeding $1-\frac{1}{2n}$. Thus, all critical edges are contained in $\mathcal{G}_{p,k}^{\mathcal{X}}$ with probability exceeding $1-(\frac{1}{2n}+\frac{1}{2n}) = 1-\frac{1}{n}$. Unpacking $\varepsilon_{\mathcal{M},p,f}$ yields the claimed lower bound on k. ∎

In (3.1), the explicit dependence of k on κ_0, p, and d is shown. The $4\kappa_0^2$ factor corresponds to the geometry of $\mathcal{M}$. The numerical constant 4, which is not tight, stems from accounting for the reach of $\mathcal{M}$. If $\mathcal{M}$ is convex (i.e., $\zeta = \infty$), then it can be replaced with 3. The second factor in (3.1) is controlled by the probability distribution, while the third corresponds to p and d. For $p=2$ and ignoring geometric and density factors we attain $k = O(2^d \log(n))$ as in [20]. For large p we get $k \approx O((\frac{4}{3})^{d/2}\log(n))$, thus improving the dependence of k on d given in [32, 20]. Finally, using Corollary 4.4 of [45] we can sharpen the qualitative requirement that n be "sufficiently large" to the quantitative lower bound $n \geq C\max\{[\frac{d}{\zeta}]^{\frac{pd}{p-1}}[\frac{4}{4^{1-1/p}-1}]^{\frac{pd}{2(p-1)}}, [\frac{1}{\epsilon_0}]^{\frac{pd}{p-1}}\}$ for a constant C depending on the geometry of $\mathcal{M}$. So, when $\mathcal{M}$ is intrinsically high-dimensional or has small reach, or when p is close to 1, n may need to be quite large for k as in (3.1) to yield a 1-spanner.

3.2. Numerical experiments. We verify the claimed dependence of k on n, p, and d ensures that $\mathcal{G}_{\mathcal{X}}^{p,k}$ is a 1-spanner of $\mathcal{G}_{\mathcal{X}}^{p}$ numerically. We generate Figures 5(a)–5(f) as follows:

(1) Fix $p, d, \mathcal{M}$, and f, then generate a sequence of (n,k) pairs.
(2) For each (n,k), do:
  (i) Generate $\mathcal{X} = \{x_i\}_{i=1}^n$ by sampling i.i.d. from f on $\mathcal{M}$.
  (ii) For all pairs $\{x_i, x_j\}$ compute $\ell_p(x_i,x_j)$ and $\ell_p^{\mathcal{G}_{\mathcal{X}}^{p,k}}(x_i,x_j)$.
  (iii) If $\max_{1\leq i<j\leq n} |\ell_p(x_i,x_j) - \ell_p^{\mathcal{G}_{\mathcal{X}}^{p,k}}(x_i,x_j)| > 10^{-10}$ record "failure"; else, record "success."
(3) Repeat step (2) 20 times and compute the proportion of successes.

As can be seen from Figure 5, there is a sharp transition between an "all failures" and an "all successes" regime. The transition line is roughly linear when viewed using semi-log-x axes, i.e., $k \propto \log(n)$. Moreover the slope of the line-of-best-fit to this transition line decreases with increasing p (compare Figures 5(a)–5(c)) and depends on intrinsic, not extrinsic dimension (compare Figures 5(b) and 5(d)), as predicted by Theorem 3.9. Intriguingly, there is little



(a) $p = 1.5, \mathcal{M} = [0,1]^5$. Uniform. (b) $p = 2, \mathcal{M} = [0,1]^5$. Uniform. (c) $p = 10, \mathcal{M} = [0,1]^5$. Uniform.

(d) $p = 2, \mathcal{M} = \mathbb{S}^4 \subset \mathbb{R}^5$. Uniform. (e) $p = 2, \mathcal{M} = [0,1]^5$. Gaussian dist. (f) $p = 1.5, \mathcal{M} = [0,1]^4$, Uniform+noise.

(g) $p = 2, \mathcal{M} = [0,1]^5$. Uniform. (h) $p = 2, \mathcal{M} = [0,1]^5$. Uniform. (i) $p = 2, \mathcal{M} = [0,1]^5$. Uniform.

Figure 5. *Figures* (a)–(f) *show the proportion of randomly generated data sets for which* $\mathcal{G}_{\mathcal{X}}^{p,k}$ *is a* 1*-spanner of* $\mathcal{G}_{\mathcal{X}}^{p}$*. The red line is the line of best fit through the cells representing the first value of* k*, for each value of* n*, for which all trials were successful, i.e., it is the line ensuring* $\mathcal{G}_{\mathcal{X}}^{p,k}$ *is a* 1*-spanner. The slopes of* (a)–(f) *are, respectively,* 43.43*,* 25.33*,* 0.29*,* 14.18*,* 18.79*, and* 40.0*. Figures* (g) *and* (h) *show the minimal* ω *(averaged across simulations) for which* $\mathcal{G}_{\mathcal{X}}^{p,k}$ *is a* $(1.1, \omega)$*-spanner and a* $(1.01, \omega)$*-spanner, respectively; the red lines trace out the requirements for* $\mathcal{G}_{\mathcal{X}}^{p,k}$ *to be a* $(1.1, 1)$*-spanner and a* $(1.01, 1)$*-spanner respectively. Figure* (i) *shows the minimal* $t \geq 1$ *such that* $\mathcal{G}_{\mathcal{X}}^{p,k}$ *is a* $(t, 1)$*-spanner of* $\mathcal{G}_{\mathcal{X}}^{p}$*, and the red line traces out the* $(1, 1)$*-spanner requirement. Experiments with a smaller range of* k *but the same data setting as* (g) *and* (h) *are in the supplementary material file* PWSPD_Supplement_Final.pdf [local/web 1.06MB].

difference between Figure 5(b) (uniform distribution) and Figure 5(e) (Gaussian distribution), suggesting that perhaps the assumption $f_{\min} > 0$ in Theorem 3.9 is unnecessary. Finally, we observe that the constant of proportionality (i.e., C such that $k = C \log n$) predicted by Theorem 3.9 appears pessimistic. For Figures 5(a)–5(c), Theorem 3.9 predicts $C = 484.03, 128$, and 21.76, respectively (taking $\kappa_0 = 1$ due to the flat domain), while empirically the slope of the line-of-best-fit is 43.43, 25.33, and 0.29, respectively.

In Figure 5(f), we consider an intrinsically four-dimensional set corrupted with Gaussian noise (standard deviation 0.1) in the fifth dimension. Interestingly, the scaling with k is more efficient than as shown in Figure 5(a) for the intrinsically five-dimensional data. This

suggests that measures which concentrate near low-dimensional sets benefit from that low-dimensionality, even if they are not supported exactly on it.

We also consider relaxing the success condition (2)(iii). We define $\mathcal{H}$ to be a (t,ω)-*spanner* if $\ell_p^{\mathcal{H}}(x,y) \le t\ell_p(x,y)$ for $\omega \in (0,1]$ proportion of the edges, so that Theorem 3.9 pertains to (1,1)-spanners. Figures 5(g) and 5(h) show the minimal ω (averaged across simulations) for which $\mathcal{G}_{\mathcal{X}}^{p,k}$ is a $(1.1,\omega)$-spanner and a $(1.01,\omega)$-spanner, respectively; the red lines trace out the requirements for $\mathcal{G}_{\mathcal{X}}^{p,k}$ to be a $(1.1,1)$-spanner and a $(1.01,1)$-spanner, respectively. Comparing with Figure 5(b), we see that the required scaling for $\mathcal{G}_{\mathcal{X}}^{p,k}$ to be a $(1+\epsilon,1)$-spanner is similar to the required scaling to be a $(1,1)$-spanner, at least for $\epsilon > 0$ small. However, the required scaling for $(1+\epsilon,\omega)$-spanners ($\omega < 1$) is quite different and much less restrictive, even for ω very close to 1; for example, the requirement for $\mathcal{G}_{\mathcal{X}}^{p,k}$ to be a $(1.01,0.95)$-spanner appears sublinear in the $\log_2(n)$ versus k plot (see Figure 5(h)). If this notion of approximation is acceptable, our empirical results suggest one can enjoy much greater sparsity. Finally, in Figure 5(i) we compute the minimal $t \ge 1$ such that $\mathcal{G}_{\mathcal{X}}^{p,k}$ is a $(t,1)$-spanner of $\mathcal{G}_{\mathcal{X}}^{p}$; again the overall transition patterns for $(t,1)$-spanners are similar to the $(1,1)$-spanner case in Figure 5(b) when t is close to 1. Overall we see that greater sparsity is permissible in these relaxed cases; rigorous analysis is a topic of ongoing research.

4. Global analysis: Statistics on PWSPD and percolation. We recall that after a suitable normalization, ℓ_p is a consistent estimator for $\mathcal{L}_p$. Indeed, [38, 32] prove that for any $d \ge 1$, $p > 1$, there exists a constant $C_{p,d}$ independent of n such that $\lim_{n\to\infty} \tilde{\ell}_p(x,y) = C_{p,d}\mathcal{L}_p(x,y)$. The important question then arises: how quickly does $\widetilde{\ell}_p$ converge? How large does n need to be to guarantee the error incurred by approximating $\mathcal{L}_p$ with $\widetilde{\ell}_p$ is small? To answer this question we turn to results from Euclidean FPP [35, 36, 6, 24]. For any discrete set $\mathcal{X}$, we let $\ell_p(x,y,\mathcal{X})$ denote the PWSPD computed in the set $\mathcal{X} \cup \{x\} \cup \{y\}$.

4.1. Overview of Euclidean first passage percolation. Euclidean FPP analyzes $\ell_p^p(0, z, H_1)$, where H_1 is a homogeneous, unit intensity Poisson point process (PPP) on $\mathbb{R}^d$.

Definition 4.1. *A* (homogeneous) PPP *on $\mathbb{R}^d$ is a point process such that for any bounded subset $A \subset \mathbb{R}^d$, n_A (the number of points in A) is a random variable with distribution $\mathbb{P}[n_A = m] = \frac{1}{m!}(\lambda|A|)^m e^{-\lambda|A|}$; λ is the* intensity *of the PPP.*

It is known that

$$\lim_{\|z\|\to\infty} \frac{\ell_p^p(0,z,H_1)}{\|z\|} = \mu\,, \tag{4.1}$$

where $\mu = \mu_{p,d}$ is a constant depending only on p, d known as the *time constant*. The convergence of $\ell_p^p(0,z,H_1)$ is studied by decomposing the error into random and deterministic fluctuations, i.e.,

$$\ell_p^p(0,z,H_1) - \mu\|z\| = \underbrace{\ell_p^p(0,z,H_1) - \mathbb{E}[\ell_p^p(0,z,H_1)]}_{\text{random}} + \underbrace{\mathbb{E}[\ell_p^p(0,z,H_1)] - \mu\|z\|}_{\text{deterministic}}\,.$$

In terms of mean squared error (MSE), one has the standard bias-variance decomposition: $\mathbb{E}[(\ell_p^p(0,z,H_1) - \mu\|z\|)^2] = (\mathbb{E}[\ell_p^p(0,z,H_1)] - \mu\|z\|)^2 + \mathrm{Var}[\ell_p^p(0,z,H_1)]$. The following proposition is well known in the Euclidean FPP literature.

Proposition 4.2. *Let $d \geq 2$ and $p > 1$. Then $\mathbb{E}[(\ell_p^p(0, z, H_1) - \mu\|z\|)^2] \leq C\|z\| \log^2(\|z\|)$ for a constant C depending only on p, d.*

Proof. By Theorem 2.1 in [36], $\mathrm{Var}[\ell_p^p(0, z, H_1)] \leq C\|z\|$. By Theorem 2.1 in [3], $(\mathbb{E}[\mu\|z\|] - \mu\|z\|)^2 \leq C\|z\| \log^2(\|z\|)$. ∎

Although $\mathrm{Var}[\ell_p^p(0, z, H_1)] \leq C\|z\|$ is the best bound which has been proved, the fluctuation rate is known to in fact depend on the dimension, i.e., $\mathrm{Var}[\ell_p^p(0, z, H_1)] \sim \|z\|^{2\chi}$ for some exponent $\chi = \chi(d) \leq \frac{1}{2}$. Strong evidence is provided in [24] that the bias can be bounded by the variance, so the exponent χ very likely controls the total convergence rate.

The following tail bound is also known [36].

Proposition 4.3. *Let $d \geq 2, p > 1, \beta_1 = \min\{1, d/p\}$, and $\beta_2 = 1/(4p + 3)$. For any $\epsilon \in (0, \beta_2)$, there exist constants C_0 and C_1 (depending on ϵ) such that for $\|z\| > 0$ and $\|z\|^{\frac{1}{2}+\epsilon} \leq t \leq \|z\|^{\frac{1}{2}+\beta_2-\epsilon}$, $\mathbb{P}[|\ell_p^p(0, z, H_1) - \mu\|z\|| \geq t] \leq C_1 \exp(-C_0(t/\sqrt{\|z\|})^{\beta_1})$.*

4.2. Convergence rates for PWSPD. We wish to utilize the results in section 4.1 to obtain convergence rates for PWSPD. However, we are interested in PWSPD computed on a compact set with boundary M and the convergence rate of ℓ_p rather than ℓ_p^p. To simplify the analysis, we restrict our attention to the following idealized model.

Assumption 1. Let $M \subseteq \mathbb{R}^d$ be a convex, compact, d-dimensional set of unit volume containing the origin. Assume we sample n points independently and uniformly from M, i.e., $f = 1_M$, to obtain the discrete set $\mathcal{X}_n$. Let M_τ denote the points in M which are at least distance τ from the boundary of M, i.e., $M_\tau := \{x \in M : \min_{y \in \partial M} \|x - y\| > \tau\}$.

We establish three things: (i) Euclidean FPP results apply away from ∂M; (ii) the time constant μ equals the constant $C_{p,d}$ in (1.3); (iii) ℓ_p has the same convergence rate as ℓ_p^p.

To establish (i), we let H_n denote a homogeneous PPP with rate $\lambda = n$ and let $\ell_p(0, y, H_n)$ denote the length of the shortest path connecting 0 and y in H_n. We also let $\mathcal{X}_N = H_n \cap M$ and $\ell_p(0, y, \mathcal{X}_N)$ denote the PWSPD in $\mathcal{X}_N$; note $\mathbb{E}[|\mathcal{X}_N|] = n$. To apply percolation results to our setting, the statistical equivalence of $\ell_p(0, y, \mathcal{X}_n)$, $\ell_p(0, y, \mathcal{X}_N)$, and $\ell_p(0, y, H_n)$ must be established. For n large, the equivalence of $\ell_p(0, y, \mathcal{X}_n)$ and $\ell_p(0, y, \mathcal{X}_N)$ is standard and we omit any analysis. The equivalence of $\ell_p(0, y, \mathcal{X}_N)$ and $\ell_p(0, y, H_n)$ is less clear. In particular, how far away from ∂M do $0, y$ need to be to ensure these metrics are the same? The following proposition is a direct consequence of Theorem 2.4 from [36] and essentially guarantees the equivalence of the metrics as long as 0 and y are at least distance $O(n^{-\frac{1}{4d}})$ from ∂M.

Proposition 4.4. *Let $d \geq 2$, $p > 1$, $\beta_1 = \min\{1, \frac{d}{p}\}$, $\beta_2 = 1/(4p + 3)$, and $\epsilon \in (0, \frac{\beta_2}{2})$, and $\tau = n^{-\frac{1}{4d}+\frac{\epsilon}{d}} diam(M)^{\frac{3}{4}+\epsilon}$. Then for constants C_0, C_1 (depending on ϵ), for all $0, y \in M_\tau$, the geodesics connecting $0, y$ in $\mathcal{X}_N$ and H_n are equal with probability at least $1 - C_1 \exp(-C_0(n^{\frac{1}{d}}\|y\|)^{\frac{3}{4}\epsilon\beta_1})$, so that $\ell_p(0, y, \mathcal{X}_N) = \ell_p(0, y, H_n)$.*

Next we establish the equivalence of $\mu_{p,d}$ (percolation time constant) and $C_{p,d}$ (PWSPD discrete-to-continuum normalization constant).

Proposition 4.5. *Let $\mu_{p,d}$ be as in (4.1) and $C_{p,d}$ as in (1.3). Then $\mu_{p,d}^{1/p} = C_{p,d}$.*

Proof. Suppose Assumption 1 holds and choose $y \in M$ with $\|y\| = 1$ and let M be such that $0, y$ are not on the boundary. By Proposition 4.4, $\lim_{n\to\infty} \ell_p(0, y, \mathcal{X}_n) = \lim_{n\to\infty} \ell_p(0,$

$y, H_n)$. Let H_1 be the unit intensity PPP obtained from H_n by rescaling each axis by $n^{1/d}$, so that $\ell_p(0, y, H_n) = n^{-\frac{1}{d}}\ell_p(0, n^{\frac{1}{d}}y, H_1)$. For notational convenience, let $z = n^{\frac{1}{d}}y$. Then

$$\begin{aligned}\lim_{n\to\infty} \widetilde{\ell}_p(0, y, \mathcal{X}_n) &= \lim_{n\to\infty} \widetilde{\ell}_p(0, y, H_n)\\ &= \lim_{n\to\infty} n^{\frac{p-1}{pd}}\ell_p(0, y, H_n)\\ &= \lim_{n\to\infty} n^{\frac{p-1}{pd}} n^{-\frac{1}{d}}\ell_p\left(0, n^{\frac{1}{d}}y, H_1\right)\\ &= \lim_{\|z\|\to\infty} \|z\|^{\frac{p-1}{p}}\|z\|^{-1}\ell_p(0, z, H_1)\\ &= \lim_{\|z\|\to\infty} \frac{\ell_p(0, z, H_1)}{\|z\|^{1/p}}.\end{aligned}$$

Thus, $C_{p,d} = C_{p,d}\mathcal{L}_p(0, y) = \lim_{n\to\infty} \widetilde{\ell}_p(0, y, \mathcal{X}_n) = \lim_{\|z\|\to\infty} \frac{\ell_p(0,z,H_1)}{\|z\|^{1/p}} = \mu_{p,d}^{1/p}$. ∎

Finally, we bound our real quantity of interest: the convergence rate of $\widetilde{\ell}_p$ to $C_{p,d}\mathcal{L}_p$.

Theorem 4.6. *Assume Assumption 1, $d \geq 2$, $\beta_2 = 1/(4p+3)$, $\tau = n^{-\frac{(1-\beta_2)}{4d}} diam(M)^{\frac{3+\beta_2}{4}}$, $p > 1$, and $0, y \in M_\tau$. Then for n large enough, $\mathbb{E}[(\tilde{\ell}_p(0, y, \mathcal{X}_n) - C_{p,d}\mathcal{L}_p(0, y))^2] \lesssim n^{-\frac{1}{d}}\log^2(n)$.*

Proof. To simplify notation throughout the proof we denote $\mathcal{L}_p(0, y)$ simply by $\mathcal{L}_p$. By Proposition 4.5 and for n large enough,

$$\mathbb{E}\left[\left(\tilde{\ell}_p(0, y, \mathcal{X}_n) - C_{p,d}\mathcal{L}_p\right)^2\right] \lesssim \mathbb{E}\left[\left(\tilde{\ell}_p(0, y, \mathcal{X}_N) - \mu^{1/p}\mathcal{L}_p\right)^2\right] =: (I)\,,$$

where $\mathcal{X}_N = H_n \cap M$ and H_n is a homogeneous PPP with rate n. Let A be the event that the geodesics from 0 to y in $\mathcal{X}_N$ and H_n are equal. Since we assume $\tau = n^{-\frac{(1-\beta_2)}{4d}}\text{diam}(M)^{\frac{3+\beta_2}{4}}$, we may apply Proposition 4.4 with $\epsilon = \beta_2/4$ to conclude $\mathbb{P}[A] \geq 1 - C_1\exp(-C_0\|y\|^{\frac{\nu}{d}}n^\nu)$ for $\nu = \frac{3\beta_2}{16}\min\{1, \frac{d}{p}\}$. Conditioning on A, and observing $\tilde{\ell}_p(0, y, \mathcal{X}_N) = n^{\frac{p-1}{pd}}\ell_p(0, y, \mathcal{X}_N) \leq n^{\frac{p-1}{pd}}\|y\|$, we obtain

$$\begin{aligned}(I) &= \mathbb{E}\left[\left(\tilde{\ell}_p(0, y, \mathcal{X}_N) - \mu^{1/p}\mathcal{L}_p\right)^2 \mid A\right]\mathbb{P}[A] + \mathbb{E}\left[\left(\tilde{\ell}_p(0, y, \mathcal{X}_N) - \mu^{1/p}\mathcal{L}_p\right)^2 \mid \bar{A}\right]\mathbb{P}[\bar{A}]\\ &\leq \mathbb{E}\left[\left(\tilde{\ell}_p(0, y, H_n) - \mu^{1/p}\mathcal{L}_p\right)^2 \mid A\right] + \left(n^{\frac{2(p-1)}{pd}}\|y\|^2 + \mu^{2/p}\mathcal{L}_p^2\right)C_1\exp\left(-C_0\|y\|^{\frac{\nu}{d}}n^\nu\right)\\ &\leq \mathbb{E}\left[\left(\tilde{\ell}_p(0, y, H_n) - \mu^{1/p}\mathcal{L}_p\right)^2\right] + q_1,\end{aligned}$$

where q_1 decays exponentially in n (for the last line note that conditioning on A means conditioning on the geodesics being local, which can only decrease the expected error).

A Lipschitz analysis applied to the function $g(x) = x^{1/p}$ yields

$$\left(\widetilde{\ell}_p(0, y, H_n) - \mu^{1/p}\mathcal{L}_p\right)^2 \leq p^{-2}\widetilde{\ell}_p(0, y, H_n)^{2(1-p)/p}\cdot\left(\widetilde{\ell}_p^p(0, y, H_n) - \mu\mathcal{L}_p^p\right)^2\,.$$

By Proposition 4.3,

$$\widetilde{\ell}_p^p(0,y,H_n) \geq \mu\mathcal{L}_p^p - \|y\|^{\frac{1}{2}+\epsilon}/n^{\frac{1}{d}\left(\frac{1}{2}-\epsilon\right)} \tag{4.2}$$

with probability at least $1-C_1\exp(-C_0\|y\|^{\epsilon\beta_1}n^{\frac{\epsilon\beta_1}{d}})$ for any $\epsilon \in (0,\beta_2)$, where $\beta_1 = \min\{1, d/p\}$. Fix $\epsilon \in (0,\beta_2)$ and let B be the event that (4.2) is satisfied. On B,

$$\begin{aligned}
\widetilde{\ell}_p(0,y,H_n)^{\frac{2(1-p)}{p}} &\leq (\mu^{1/p}\mathcal{L}_p)^{\frac{2(1-p)}{p}}\left(1 - \frac{\|y\|^{\frac{1}{2}+\epsilon}}{\mu\mathcal{L}_p^p n^{\frac{1}{d}\left(\frac{1}{2}-\epsilon\right)}}\right)^{\frac{2(1-p)}{p^2}} \\
&\leq (\mu^{1/p}\mathcal{L}_p)^{\frac{2(1-p)}{p}}\left(1 + \frac{2(p-1)\|y\|^{\frac{1}{2}+\epsilon}}{p^2\mu\mathcal{L}_p^p n^{\frac{1}{d}\left(\frac{1}{2}-\epsilon\right)}} + \text{higher order terms}\right) \\
&\leq 2(\mu^{1/p}\mathcal{L}_p)^{\frac{2(1-p)}{p}}
\end{aligned}$$

for n large enough. Note also that

$$\mathbb{E}\left[\left(\widetilde{\ell}_p(0,y,H_n) - \mu^{1/p}\mathcal{L}_p\right)^2 \mid \bar{B}\right]\mathbb{P}[\bar{B}] \leq \left(n^{\frac{2(p-1)}{pd}}\|y\|^2 + \mu^{2/p}\mathcal{L}_p^2\right)\exp\left(-C_0\|y\|^{\epsilon\beta_1}n^{\frac{\epsilon\beta_1}{d}}\right) := q_2$$

and q_2 decreases exponentially in n. We thus obtain

$$\begin{aligned}
\mathbb{E}\left[\left(\widetilde{\ell}_p(0,y,H_n) - \mu^{1/p}\mathcal{L}_p\right)^2\right] &\leq \mathbb{E}\left[\left(\widetilde{\ell}_p(0,y,H_n) - \mu^{1/p}\mathcal{L}_p\right)^2 \mid B\right]\mathbb{P}[B] + q_2 \\
&\leq \frac{2}{p^2}(\mu^{1/p}\mathcal{L}_p)^{\frac{2(1-p)}{p}}\mathbb{E}\left[\left(\widetilde{\ell}_p^p(0,y,H_n) - \mu\mathcal{L}_p^p\right)^2 \mid B\right] + q_2 \\
&= C\mathbb{E}\left[\left(\widetilde{\ell}_p^p(0,y,H_n) - \mu\mathcal{L}_p^p\right)^2\right] + q_2,
\end{aligned}$$

where C is a constant depending on $p, d, \|y\|$, and the last line follows since once again the expected error is lower conditioned on B than unconditionally. We have thus established

$$\mathbb{E}\left[\left(\tilde{\ell}_p(0,y,\mathcal{X}_n) - C_{p,d}\mathcal{L}_p\right)^2\right] \lesssim \mathbb{E}\left[\left(\widetilde{\ell}_p^p(0,y,H_n) - \mu\mathcal{L}_p^p\right)^2\right] + q_1 + q_2$$

for q_1, q_2 exponentially small in n. Finally let H_1 be the unit intensity homogeneous PPP obtained from H_n by multiplying each axis by $n^{1/d}$. By Proposition 4.2,

$$\begin{aligned}
&\mathbb{E}\left[\left(\ell_p^p\left(0, n^{\frac{1}{d}}y, H_1\right) - \mu n^{\frac{1}{d}}\|y\|\right)^2\right] \lesssim n^{\frac{1}{d}}\|y\|\log^2\left(n^{\frac{1}{d}}\|y\|\right) \\
\Rightarrow\ &\mathbb{E}\left[\left(n^{\frac{p}{d}}\ell_p^p(0,y,H_n) - n^{\frac{1}{d}}\mu\mathcal{L}_p^p\right)^2\right] \lesssim n^{\frac{1}{d}}\|y\|\log^2\left(n^{\frac{1}{d}}\|y\|\right) \\
\Rightarrow\ &\mathbb{E}\left[\left(n^{\frac{p-1}{d}}\ell_p^p(0,y,H_n) - \mu\mathcal{L}_p^p\right)^2\right] \lesssim n^{-\frac{1}{d}}\|y\|\log^2\left(n^{\frac{1}{d}}\|y\|\right) \\
\Rightarrow\ &\mathbb{E}\left[\left(\widetilde{\ell}_p^p(0,y,H_n) - \mu\mathcal{L}_p^p\right)^2\right] \lesssim n^{-\frac{1}{d}}\log^2(n)\,.
\end{aligned}$$

For n large, the above dominates q_1, q_2, so that for a constant C depending on $p, d, \|y\|$,

$$\mathbb{E}\left[\left(\tilde{\ell}_p(0,y,\mathcal{X}_n) - C_{p,d}\mathcal{L}_p\right)^2\right] \leq Cn^{-\frac{1}{d}}\log^2(n).$$

■

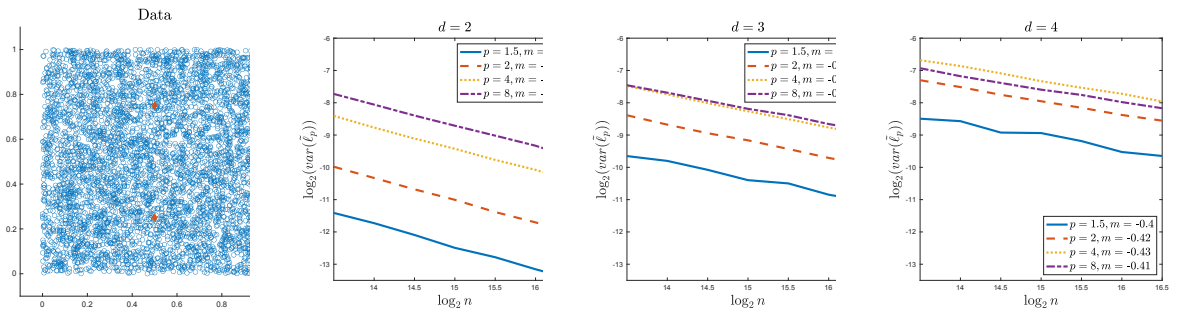
4.3. Estimating the fluctuation exponent. As an application, we utilize the 1-spanner results of section 3 to empirically estimate the fluctuation rate $\chi(d)$. Since there is evidence that the variance dominates the bias, this important parameter likely determines the convergence rate of $\widetilde{\ell}_p$ to $\mathcal{L}_p$. Once again utilizing the change of variable $z = n^{\frac{1}{d}}y$, we note

$$\mathrm{Var}\left[\ell_p^p(0,z,H_1)\right] \lesssim \|z\|^{2\chi} \iff \mathrm{Var}\left[\widetilde{\ell}_p(0,y,\mathcal{X}_n)\right] \lesssim n^{\frac{2(\chi-1)}{d}},$$

and we estimate the right-hand side from simulations. Specifically, we sample n points uniformly from the unit cube $[0,1]^d$ and compute $\widetilde{\ell}_p(x,y,\mathcal{X}_n)$ for $x = (0.25, 0.5, \ldots, 0.5)$, $y = (0.75, 0.5, \ldots, 0.5)$ in a kNN graph on $\mathcal{X}_n$, with $k = \lceil 1+3(\frac{4}{4^{1-1/p}-1})^{d/2}\log(n)\rceil$ as suggested by Theorem 3.9 (note that $f_{min} = f_{max}, \zeta = \infty, \kappa_0 = 1$ in this example). We vary n from $n_{\min} = 11586$ to $n_{\max} = 92682$, and for each n we estimate $\mathrm{Var}[\widetilde{\ell}_p(x,y,\mathcal{X}_n)]$ from N_{sim} simulations. Figure 6 shows the resulting log-log variance plots for $d = 2, 3, 4$ and various p, as well as the slopes m from a linear regression. The observed slopes are related to χ by $\chi = md/2+1$, and one thus obtains the estimates for χ reported in Table 2. See the supplementary material file PWSPD_Supplement_Final.pdf [local/web 1.06MB] for confidence interval estimates.

These simulations confirm that χ is indeed independent of p. It is conjectured in the percolation literature that $\chi(d) \to 0^+$ as d increases, with $\chi(2) = \frac{1}{3}$, $\chi(3) \approx \frac{1}{4}$, which is consistent with our results. For $d = 2$, the empirical convergence rate is thus $n^{-\frac{2}{3}}$ (not $n^{-\frac{1}{2}}$ as given in Theorem 4.6), and for large d one expects an MSE of order $n^{-\frac{2}{d}}$ instead of $n^{-\frac{1}{d}}$. However, estimating χ empirically becomes increasingly difficult as d increases, since one has less sparsity in the kNN graph, and because χ is obtained from m by $\chi = md/2 + 1$, so errors incurred in estimating the regression slopes are amplified by a factor of d. Table 2 also reports the factor $n_{\max}/k$, which can be interpreted as the expected computational speed-up obtained by running the simulation in a kNN graph instead of a complete graph. We were unable to obtain empirical speed-up factors since computational resources prevented running the simulations in a complete graph.

An important open problem is establishing that $\widetilde{\ell}_p$ computed from a nonuniform density enjoys the same convergence rate (with respect to n) as the uniform case. Although this



(a) Data for $d = 2$ (b) $d = 2$ (c) $d = 3$ (d) $d = 4$

Figure 6. *Variance plots for $\tilde{\ell}_p$. For each n, the variance was estimated from a maximum of $N_{sim} = 24000$ simulations, with a smaller N_{sim} when p was small and/or the dimension was large. Specifically, when $d = 2$, $N_{sim} = 14000$ was used for $p = 1.5$; when $d = 3$, $N_{sim} = 5000, 12000$ was used for $p = 1.5, 2$; when $d = 4$, $N_{sim} = 2000, 6000, 19000$ was used for $p = 1.5, 2, 4$.*

Table 2

The slopes of $\log(n)$ *versus* $Var[\tilde{\ell}_p]$ *are shown for uniform data for different density weightings* (p) *and different dimensions* (d).

d	p	$\hat{\chi}$	$n_{\max}/k$	d	p	$\hat{\chi}$	$n_{\max}/k$	d	p	$\hat{\chi}$	$n_{\max}/k$
2	1.5	0.30	394	3	1.5	0.28	152	4	1.5	0.19	58
2	2	0.31	667	3	2	0.23	336	4	2	0.16	169
2	4	0.33	1204	3	4	0.24	820	4	4	0.14	558
2	8	0.34	1545	3	8	0.29	1204	4	8	0.19	927

seems intuitively true and preliminary simulation results support this equivalence, to the best of our knowledge it has not been proven, as the current proof techniques rely on "straight line" geodesics.

5. Conclusion and future work.

This article establishes local equivalence of PWSPD to a density-based stretch of Euclidean distance. We derive a near-optimal condition on k for the kNN graph to be a 1-spanner for PWSPD, quantifying and improving the dependence on p and d. Moreover, we leverage the theory of Euclidean FPP to establish statistical convergence rates for PWSPD to its continuum limit, and apply our spanner results to empirically support conjectures on the optimal dimension-dependent rates of convergence.

Many directions remain for future work. Our statistical convergence rates for PWSPD in section 4 are limited to uniform distributions. Preliminary numerical experiments indicate that these rates also hold for PWSPDs defined with varying density, but rigorous convergence rates for nonhomogeneous PPPs are lacking in the literature.

The analysis of section 2 proved the local equivalence of PWSPDs with density-stretched Euclidean distances. These results and the convergence results of section 4 are the first steps in a program of developing a discrete-to-continuum limit analysis for PWSPDs and PWSPD-based operators. A major problem is to develop conditions so that the discrete graph Laplacian (defined with $\tilde{\ell}_p$) converges to a continuum second-order differential operator as $n \to \infty$. A related direction is the analysis of how data clusterability with PWSPDs depends on p for various random data models and in specific applications.

The numerical results of section 3.2 confirm that $k \propto \log(n)$ is required for the kNN graph to be a 1-spanner, as predicted by theory. Relaxing the notion of t-spanners to (t, ω)-spanners, as suggested in section 3.2, is a topic of future research.

Finally, the results of this article require data to be generated from a distribution supported exactly on a low-dimensional manifold $\mathcal{M}$. An arguably more realistic setting is the noisy one in which the data is distributed only approximately on $\mathcal{M}$. Two potential models are of interest: (i) replacing $\mathcal{M}$ with $B(\mathcal{M}, \tau) = \{x \in \mathbb{R}^D \mid \text{dist}(x, \mathcal{M}) \leq \tau\}$ (tube model) and (ii) considering a density that *concentrates* on $\mathcal{M}$, rather than being supported on it (concentration model). PWSPDs may exhibit very different properties under these two noise models, for example, under bounded uniform noise and Gaussian noise, especially for large p. For the concentration model one expects noisy PWSPDs to converge to manifold PWSPDs for p large, since the optimal PWSPD paths are density driven. Preliminary empirical results (Figure 5(f)) suggest that when the measure concentrates sufficiently near a low-dimensional set $\mathcal{M}$, the number of nearest neighbors needed for a 1-spanner benefits from the intrinsic

low-dimensional structure. For the tube model, although noisy PWSPDs will not converge to manifold PWSPDs, they will still scale according to the intrinsic manifold dimension for τ small. For both models, incorporating a denoising procedure such as local averaging [31] or diffusion [34] before computing PWSPDs is expected to be advantageous. Future research will investigate robust denoising procedures for PWSPD and which type of noise distributions are most adversarial to PWSPD.

Appendix A. Proofs for section 2.

Proof of Lemma 2.2. Let $\gamma_1(t)$ be a path which achieves $\mathscr{D}(x,y)$. Since $\mathscr{D}(x,y) \le \epsilon(1+\kappa\epsilon^2)$, $f(\gamma_1(t)) \ge f_{\min}(x,\epsilon)$ for all t. Then,

$$\mathcal{L}_p^p(x,y) \le \int_0^1 \frac{1}{f(\gamma_1(t))^{\frac{p-1}{d}}} |\gamma_1'(t)|\, dt \le \frac{\mathscr{D}(x,y)}{f_{\min}(x,\epsilon)^{\frac{p-1}{d}}} \le \frac{\epsilon(1+\kappa\epsilon^2)}{f_{\min}(x,\epsilon)^{\frac{p-1}{d}}}.$$

Note $y \in B_{\mathcal{L}_p^p}(x, \epsilon(1+\kappa\epsilon^2)/f_{\min}(x,\epsilon)^{\frac{p-1}{d}})$ implies $f(y) \le f_{\max}(x,\epsilon)$, and thus $\frac{f_{\max}(x,\epsilon)^{\frac{p-1}{d}}}{(f(x)f(y))^{\frac{p-1}{2d}}} \ge 1$, so that $\mathcal{L}_p^p(x,y) \le \frac{\mathscr{D}(x,y)}{f_{\min}(x,\epsilon)^{\frac{p-1}{d}}} \frac{f_{\max}(x,\epsilon)^{\frac{p-1}{d}}}{(f(x)f(y))^{\frac{p-1}{2d}}}$. This yields

$$\mathcal{L}_p^p(x,y) \le (\rho_{x,\epsilon})^{\frac{p-1}{d}} \frac{\|x-y\|(1+\kappa\|x-y\|^2)}{(f(x)f(y))^{\frac{p-1}{2d}}} \le (\rho_{x,\epsilon})^{\frac{p-1}{d}}(1+\kappa\epsilon^2)\mathscr{D}_{f,\mathrm{Euc}}(x,y),$$

which proves the upper bound. Now let $\gamma_0(t)$ be a path achieving $\mathcal{L}_p^p(x,y)$; note that since $\mathcal{L}_p^p(x,y) \le \frac{\mathscr{D}(x,y)}{f_{\min}(x,\epsilon)^{\frac{p-1}{d}}}$, the path γ_0 is contained in $B_{\mathcal{L}_p^p}(x, \epsilon(1+\kappa\epsilon^2)/f_{\min}(x,\epsilon)^{\frac{p-1}{d}})$. Thus

$$\mathcal{L}_p^p(x,y) = \int_0^1 \frac{1}{f(\gamma_0(t))^{\frac{p-1}{d}}} |\gamma_0'(t)|\, dt \ge \frac{\mathscr{D}(x,y)}{f_{\max}(x,\epsilon)^{\frac{p-1}{d}}} \ge \frac{\mathscr{D}(x,y)}{f_{\max}(x,\epsilon)^{\frac{p-1}{d}}} \cdot \frac{f_{\min}(x,\epsilon)^{\frac{p-1}{d}}}{(f(x)f(y))^{\frac{p-1}{2d}}}$$

so that

$$\mathcal{L}_p^p(x,y) \ge \frac{\mathscr{D}(x,y)}{(\rho_{x,\epsilon})^{\frac{p-1}{d}}(f(x)f(y))^{\frac{p-1}{2d}}} \ge \frac{\|x-y\|}{(\rho_{x,\epsilon})^{\frac{p-1}{d}}(f(x)f(y))^{\frac{p-1}{2d}}} = \frac{1}{(\rho_{x,\epsilon})^{\frac{p-1}{d}}}\mathscr{D}_{f,\mathrm{Euc}}(x,y).$$

■

Appendix B. Proofs for section 3.

Proof of Lemma 3.6. Let $s := \|x-y\|$ and choose a coordinate system $x^{(1)}, \dots, x^{(n)}$ such that $y = (-s/2, 0, \dots 0)$, $x = (s/2, 0, \dots, 0)$, and $x_M = \mathbf{0}$. $\mathcal{D}_{\alpha,p}(x,y)$ is now the interior of

$$\left(\left(x^{(1)} + \frac{s}{2}\right)^2 + (x^{(2)})^2 + \cdots + (x^{(n)})^2\right)^{p/2} + \left(\left(x^{(1)} - \frac{s}{2}\right)^2 + (x^{(2)})^2 + \cdots + (x^{(n)})^2\right)^{p/2} = \alpha s^p.$$

In spherical coordinates the boundary of this region may be expressed as

$$\left(r^2 + sr\cos\theta_1 + s^2/4\right)^{p/2} + \left(r^2 - sr\cos\theta_1 + s^2/4\right)^{p/2} = \alpha s^p, \tag{B.1}$$

where $(x^{(1)})^2 + \cdots + (x^{(n)})^2 = r^2$ and $x_1 = r\cos\theta_1$. Define $r = H(\theta_1)$ as the unique positive solution of (B.1). Implicitly differentiating in θ_1 yields

$$\frac{p}{2}\left(r^2 + sr\cos(\theta_1) + \frac{s^2}{4}\right)^{\frac{p}{2}-1}\left(2r\frac{dr}{d\theta_1} - sr\sin(\theta_1) + s\cos(\theta_1)\frac{dr}{d\theta_1}\right)$$
$$+ \frac{p}{2}\left(r^2 - sr\cos(\theta_1) + \frac{s^2}{4}\right)^{\frac{p}{2}-1}\left(2r\frac{dr}{d\theta_1} + sr\sin(\theta_1) - s\cos(\theta_1)\frac{dr}{d\theta_1}\right) = 0\,.$$

Solving for $\frac{dr}{d\theta_1}$ and setting the result to 0 yields

$$\left[\left(r^2 + sr\cos(\theta_1) + \frac{s^2}{4}\right)^{\frac{p-2}{2}} - \left(r^2 - sr\cos(\theta_1) + \frac{s^2}{4}\right)^{\frac{p-2}{2}}\right]\sin(\theta_1) := (\mathrm{I})\cdot(\mathrm{II}) = 0.$$

Thus we obtain two solutions to $\frac{dr}{d\theta_1} = 0$:

$$(\mathrm{I}) = 0 \Rightarrow \cos(\theta_1) = 0 \Rightarrow \theta_1 = \frac{\pi}{2} \qquad (\text{min.}) \quad (\mathrm{II}) = 0 \Rightarrow \sin(\theta_1) = 0 \Rightarrow \theta_1 = 0 \qquad (\text{max.}).$$

Thus the minimal radius occurs when $\theta_1 = \frac{\pi}{2}$. Substituting $\theta_1 = \frac{\pi}{2}$ into (B.1) yields

$$r = s\sqrt{\alpha^{2/p}/4^{1/p} - 1/4} = \|x - y\|\sqrt{\alpha^{2/p}/4^{1/p} - 1/4}.$$

Hence $B(x_M, r) \subset \mathcal{D}_{\alpha,p}(x_i, x_j)$, as desired. To see $\mathcal{D}_{\alpha,p}(x, y) \subset B(x, r)$ observe that if $z \notin B(x, r)$, then

$$\|x - z\| > r = \|x - y\| \Rightarrow \|x - z\|^p > \alpha\|x - y\|^p \quad \text{for all } \alpha \in (0, 1] \text{ and } p \geq 1, \tag{B.2}$$

hence z cannot be in $\mathcal{D}_{\alpha,p}(x, y)$. ∎

Lemma B.1. *With assumptions and notation as in Theorem 3.9, $B(\tilde{x}_M, r_2^\star) \subset B(x_M, r_1^\star)$.*

Proof. By [14, Lemma 1], $\|x_M - \tilde{x}_M\| \leq \zeta - \sqrt{\zeta^2 - r^2/4} < r^2/(4\zeta)$. Now, suppose $y \in B(\tilde{x}_M, r_2^\star)$. Then

$$\|x_M - y\| \leq \|x_M - \tilde{x}_M\| + \|\tilde{x}_M - y\| \leq r^2/(4\zeta) + r\left(\sqrt{1/4^{1/p} - 1/4} - r/(4\zeta)\right) = r_1^\star,$$

so that $y \in B(x_M, r_1^\star)$, as desired. ∎

Lemma B.2. *With notation and assumptions as in Theorem 3.9,*

$$\mathbb{P}\left[x_{i_j} \in B(\tilde{x}_M, r_2^\star) \mid x_{i_j} \in B(x, r)\right] \geq \frac{3}{4}\kappa_0^{-2}\frac{f_{\min}}{f_{\max}}\left(\frac{1}{4^{1/p}} - \frac{1}{4}\right)^{d/2}. \tag{B.3}$$

Proof. By the definition of conditional probability and $B(\tilde{x}_M, r_2^\star) \subset B(x, r)$,

$$\mathbb{P}\left[x_{i_j} \in B(\tilde{x}_M, r_2^\star) \mid x_{i_j} \in B(x, r)\right] = \int_{B(\tilde{x}_M, r_2^\star)\cap\mathcal{M}} f \Big/ \int_{B(x,r)\cap\mathcal{M}} f. \tag{B.4}$$

By Definition 3.7, $\int_{B(\tilde{x}_M, r_2^\star)\cap\mathcal{M}} f \geq f_{\min}\mathrm{vol}\left(B(\tilde{x}_M, r_2^\star)\cap\mathcal{M}\right) \geq f_{\min}\kappa_0^{-1}(r_2^\star)^d\mathrm{vol}\left(B(0,1)\right)$ and $\int_{B(x,r)\cap\mathcal{M}} f \leq f_{\max}\mathrm{vol}\left(B(x,r)\cap\mathcal{M}\right) \leq f_{\max}\kappa_0 r^d\mathrm{vol}\left(B(0,1)\right)$. Returning to (B.4),

$$\mathbb{P}\left[x_{i_j} \in B(\tilde{x}_M, r_2^\star) \mid x_{i_j} \in B(x,r)\right] \geq \kappa_0^{-2}\frac{f_{\min}}{f_{\max}}\left(\frac{r_2^\star}{r}\right)^d. \tag{B.5}$$

The result follows by noting

$$\left(\frac{r_2^\star}{r}\right)^d = \left(\sqrt{\frac{1}{4^{1/p}}-\frac{1}{4}}-\frac{r}{4\zeta}\right)^d \geq \left(\sqrt{\frac{1}{4^{1/p}}-\frac{1}{4}}-\frac{1}{4d}\sqrt{\frac{1}{4^{1/p}}-\frac{1}{4}}\right)^d \geq \frac{3}{4}\left(\frac{1}{4^{1/p}}-\frac{1}{4}\right)^{d/2}.$$

Acknowledgments. DM thanks Matthias Wink for several useful discussions on Riemannian geometry. We thank the two reviewers and the associate editor for many helpful comments that greatly improved the manuscript.

REFERENCES

[1] E. Aamari, J. Kim, F. Chazal, B. Michel, A. Rinaldo, and L. Wasserman, *Estimating the reach of a manifold*, Electron. J. Stat., 13 (12019), pp. 1359–1399.
[2] M. Alamgir and U. Von Luxburg, *Shortest path distance in random k-nearest neighbor graphs*, in Proceedings of ICML, 2012, pp. 1251–1258.
[3] K. S. Alexander, *A note on some rates of convergence in first-passage percolation*, Ann. Appl. Probab., pp. 81–90, 1993.
[4] H. Antil, T. Berry, and J. Harlim, *Fractional diffusion maps*, Appl. Comput. Harmon. Anal., 54 (2021), pp. 145–175.
[5] E. Arias-Castro, *Clustering based on pairwise distances when the data is of mixed dimensions*, IEEE Trans. Inform. Theory, 57 (2011), pp. 1692–1706.
[6] A. Auffinger, M. Damron, and J. Hanson, 50 *Years of First-Passage Percolation*, Univ. Lecture Ser. 68, AMS, Providence, RI, 2017.
[7] M. Azizyan, A. Singh, and L. Wasserman, *Density-sensitive semisupervised inference*, Ann. Statist., 41 (2013), pp. 751–771.
[8] M. Belkin and P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural Comput., 15 (2003), pp. 1373–1396.
[9] M. Belkin and P. Niyogi, *Convergence of Laplacian eigenmaps*, in Proceedings of NIPS, pp. 129–136, 2007.
[10] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, Princeton, NJ, 2015.
[11] T. Berry and J. Harlim, *Variable bandwidth diffusion kernels*, Appl. Comput. Harmon. Anal., 40 (2016), pp. 68–96.
[12] T. Berry and T. Sauer, *Local kernels and the geometric structure of data*, Appl. Comput. Harmon. Anal., 40 (2016), pp. 439–469.
[13] A. S. Bijral, N. Ratliff, and N. Srebro, *Semi-supervised learning with density based distances*, in Proceedings of UAI, 2011, pp. 43–50.
[14] J.-D. Boissonnat, A. Lieutier, and M. Wintraecken, *The reach, metric distortion, geodesic convexity and the variation of tangent spaces*, J. Appl. Comput. Topol., 3 (2019), pp. 29–58.
[15] L. Boninsegna, G. Gobbo, F. Noé, and C. Clementi, *Investigating molecular kinetics by variationally optimized diffusion maps*, J. Chemical Theory Comput., 11 (2015), pp. 5947–5960.
[16] E. Borghini, X. Fernández, P. Groisman, and G. Mindlin, *Intrinsic Persistent Homology via Density-Based Metric Learning*, preprint, arXiv:2012.07621, 2020.

[17] O. Bousquet, O. Chapelle, and M. Hein, *Measure based regularization*, in Proceedings of NIPS, 2003, pp. 1221–1228.
[18] H. Chang and D.-Y. Yeung, *Robust path-based spectral clustering*, Pattern Recognition, 41 (2008), pp. 191–203.
[19] Y. Cheng, *Mean shift, mode seeking, and clustering*, IEEE Trans. Pattern Anal. Machine Intelligence, 17 (1995), pp. 790–799.
[20] T. Chu, G. L. Miller, and D. R. Sheehy, *Exact computation of a manifold metric, via Lipschitz embeddings and shortest paths on a graph*, in Proceedings of SODA, 2017, pp. 411–425.
[21] R. R. Coifman and S. Lafon, *Diffusion maps*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 5–30.
[22] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, *Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps*, Proc. Natl. Acad. Sci. USA, 102 (2005), pp. 7426–7431.
[23] S. B. Damelin, F. J. Hickernell, D. L. Ragozin, and X. Zeng, *On energy, discrepancy and group invariant measures on measurable subsets of euclidean space*, J. Fourier Anal. Appl., 16 (2010), pp. 813–839.
[24] M. Damron and X. Wang, *Entropy reduction in Euclidean first-passage percolation*, Electron. J. Probab., 21 (2016).
[25] L. P. Devroye and T. J. Wagner, *The strong uniform consistency of nearest neighbor density estimates*, Ann. Statist., 5 (1977), pp. 536–540.
[26] D. L. Donoho and C. Grimes, *Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data*, Proc. Natl. Acad. Sci., 100 (2003), pp. 5591–5596.
[27] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, *A density-based algorithm for discovering clusters in large spatial databases with noise*, in Proceedings of KDD, 1996, pp. 226–231.
[28] A. M. Farahmand, C. Szepesvári, and J.-Y. Audibert, *Manifold-adaptive dimension estimation*, in Proceedings of ICML, 2007, pp. 265–272.
[29] H. Federer, *Curvature measures*, Trans. Amer. Math. Soc., 93 (1959), pp. 418–491.
[30] B. Fischer, T. Zöller, and J. M. Buhmann, *Path based pairwise data clustering with application to texture segmentation*, in International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition, Springer, New York, 2001, pp. 235–250.
[31] N. García Trillos, D. Sanz-Alonso, and R. Yang, *Local regularization of noisy point clouds: Improved global geometric estimates and data analysis*, J. Mach. Learn. Res., 20 (2019), pp. 1–37.
[32] P. Groisman, M. Jonckheere, and F. Sapienza, *Nonhomogeneous Euclidean First-Passage Percolation and Distance Learning*, preprint, arXiv:1810.09398, 2018.
[33] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*, Springer, 2006.
[34] M. Hein and M. Maier, *Manifold denoising*, in Proceedings of NIPS, Vol. 19, pp. 561–568, 2006.
[35] C. D. Howard and C. M. Newman, *Euclidean models of first-passage percolation*, Probab. Theory Related Fields, 108 (1997), pp. 153–170.
[36] C. D. Howard and C. M. Newman, *Geodesics and spanning trees for Euclidean first-passage percolation*, Ann. Probab., 29 (2001), pp. 577–623.
[37] G. Hughes, *On the mean accuracy of statistical pattern recognizers*, IEEE Trans. Inform. Theory, 14 (1968), pp. 55–63.
[38] S. J. Hwang, S. B. Damelin, and A. Hero, *Shortest path through random points*, Ann. Appl. Probab., 26 (2016), pp. 2791–2823.
[39] D. B. Johnson, *Efficient algorithms for shortest paths in sparse networks*, J. ACM, 24 (1977), pp. 1–13.
[40] J. Kileel, A. Moscovich, N. Zelesko, and A. Singer, *Manifold Learning with Arbitrary Norms*, preprint, arXiv:2012.14172, 2020.
[41] A. Little, M. Maggioni, and J. M. Murphy, *Path-based spectral clustering: Guarantees, robustness to outliers, and fast algorithms*, J. Mach. Learn. Res., 21 (2020), pp. 1–66.
[42] D. O. Loftsgaarden and C. P. Quesenberry, *A nonparametric estimate of a multivariate density function*, Ann. Math. Statist., 36 (1965), pp. 1049–1051.
[43] P. C. Mahalanobis, *On the Generalized Distance in Statistics*, National Institute of Science of India, 1936.

[44] J. Malik, C. Shen, H.-T. Wu, and N. Wu, *Connecting dots: From local covariance to empirical intrinsic geometry and locally linear embedding*, Pure Appl. Anal., 1 (2019), pp. 515–542.
[45] D. Mckenzie and S. Damelin, *Power weighted shortest paths for clustering Euclidean data*, Found. of Data Science, 1 (2019), pp. 307–327.
[46] A. Moscovich, A. Jaffe, and B. Nadler, *Minimax-optimal semi-supervised regression on unknown manifolds*, in Proceedings of AISTATS, 2017, pp. 933–942.
[47] A. Y. Ng, M. I. Jordan, and Y. Weiss, *On spectral clustering: Analysis and an algorithm*, in Proceedings of NIPS, pp. 849–856, 2002.
[48] P. Niyogi, S. Smale, and S. Weinberger, *Finding the homology of submanifolds with high confidence from random samples*, Discrete Comput. Geom., 39 (2008), pp. 419–441.
[49] A. Rinaldo and L. Wasserman, *Generalized density clustering*, Ann. Statist., 38 (2010), pp. 2678–2722.
[50] A. Rodriguez and A. Laio, *Clustering by fast search and find of density peaks*, Science, 344 (2014), pp. 1492–1496.
[51] Sajama and A. Orlitsky, *Estimating and computing density based distance metrics*, in Proceedings of ICML, 2005, pp. 760–767.
[52] L. K. Saul and M. I. Jordan, *A variational principle for model-based interpolation*, in Proceedings of NIPS, 1997, pp. 267–273.
[53] G. Schiebinger, M. J. Wainwright, and B. Yu, *The geometry of kernelized spectral clustering*, Ann. Statist., 43 (2015), pp. 819–846.
[54] J. Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE Trans. Pattern Anal. Machine Intelligence, 22 (2000), pp. 888–905.
[55] J. B. Tenenbaum, V. De Silva, and J. C. Langford, *A global geometric framework for nonlinear dimensionality reduction*, Science, 290 (2000), pp. 2319–2323.
[56] L. van der Maaten and G. Hinton, *Visualizing data using t-SNE*, J. Machine Learning Research, 9 (2008), pp. 2579–2605.
[57] D. Van Dijk, R. Sharma, J. Nainys, K. Yim, P. Kathail, A. J. Carr, C. Burdziak, K. R Moon, C. L. Chaffer, D. Pattabiraman, B. Bierie, L. Mazutis, G. Wolf, S. Krishnaswamy, and D. Pe'er, *Recovering gene interactions from single-cell data using data diffusion*, Cell, 174 (2018), pp. 716–729.
[58] P. Vincent and Y. Bengio, *Density-sensitive metrics and kernels*, in Snowbird Learning Workshop, 2003.
[59] U. Von Luxburg, *A tutorial on spectral clustering*, Stat. Comput., 17 (2007), pp. 395–416.
[60] R. Xu, S. Damelin, B. Nadler, and D. C. Wunsch II, *Clustering of high-dimensional gene expression data with feature filtering methods and diffusion maps*, Artificial Intelligence Medicine, 48 (2010), pp. 91–98.
[61] L. Zelnik-Manor and P. Perona, *Self-tuning spectral clustering*, in Proceedings of NIPS, pp. 1601–1608, 2005.
[62] S. Zhang and J. M. Murphy, *Hyperspectral image clustering with spatially-regularized ultrametrics*, Remote Sensing, 13 (2021), 955.