Exact Computation of a Manifold Metric, via Shortest Paths on a graph

Timothy Chu
CMU
tzchu@andrew.cmu.edu

Gary L. Miller
CMU
glmiller@cs.cmu.edu

Donald Sheehy
University of Connecticut
don.r.sheehy@gmail.com

July 9, 2019

Abstract

Data-sensitive metrics adapt distances locally based the density of data points with the goal of aligning distances and some notion of similarity. In this paper, we give the first exact algorithm for computing a data-sensitive metric called the Nearest Neighbor Metric. In fact, we prove the surprising result that a previously published 3-approximation is an exact algorithm.

The Nearest Neighbor Metric can be viewed as a special case of a density-based distance used in machine learning, or it can be seen as an example of a manifold metric. Previous computational research on such metrics despaired of computing exact distances on account of the apparent difficulty of minimizing over all continuous paths between a pair of points.

We leverage the exact computation of the Nearest Neighbor Metric to compute sparse spanners and persistent homology. We also explore the behavior of the metric built from point sets drawn from an underlying distribution and consider the more general case of inputs that are countable collections of compact sets.

The main results connect several classical theories such as the conformal change of Riemannian metrics and the screw theory of Schoenberg and Von Neumann. We also develop some novel proof techniques based on the combination of screw functions and Lipschitz extensions that may be of independent interest.

1 Introduction

The profound success of nonlinear methods in machine learning such as kernels methods, density-based distances, and neural nets reveals that although data are often represented as points in $\mathbb{R}^n$, the shortest path between two points is *not* a straight line. It is widely believed that a more useful metric on the data points would have the property that two points in a dense cluster will be close in some underlying metric, even if the Euclidean distance is far. That is, distances are scaled inversely according to the density of the data along a path between points. We call such a metric **data-sensitive**.

Data-sensitive metrics arise naturally in machine learning, and are implicitly central in celebrated machine learning methods such as k-NN graph methods, manifold learning, level-set methods, single-linkage clustering, and Euclidean MST-based clustering (See Appendix **??** for details). The construction of appropriate data-sensitive metrics is an active area of research. We consider a simple data-sensitive metric with an underlying manifold structure called the **nearest neighbor metric**. This metric and its close variants have been studied in the past by multiple researchers []. In this paper, we show how to compute the nearest neighbor metric exactly for any dimension, which solves one of the most important and challenging problems for any manifold-based metric.

The nearest neighbor metric is defined with respect to the nearest neighbor function $\mathbf{r}_P$ for the data set P:

$$\mathbf{r}_P(z) = 2 \min_{x \in p} \|x - z\|$$

It will be used as a cost function for a density-based distance defined as follows. The factor of 2 normalizes and simplifies expressions later.

Definition 1.1. *Given a continuous cost function $c : \mathbb{R}^k \to \mathbb{R}$, we define the density-based cost of a path γ relative to c as:*

$$\ell_c(\gamma) = \int_0^1 c(\gamma(t))\|\gamma'(t)\|dt.$$

*Here, the path γ is defined as a continuous map $\gamma : [0, 1] \to \mathbb{R}^k$. Let $\mathrm{path}(a, b)$ denote the set of piecewise-C_1 paths from a to b. We then define the **density-based distance** between two points $a, b \in \mathbb{R}^k$ as*

$$d_c(a, b) \inf_{\gamma \in \mathrm{path}(a,b)} \ell_c(\gamma)$$

This is a slight simplification of the original definition from [**?**] which included other requirements to facilitate approximtion. Conceptually, the density-based cost of a path is the weighted path length, where each infinitesimal path piece is weighted according to c. The cost c is usually some function of an underlying density f (the natural choice would be $c(x) = f(x)^{-\frac{1}{k}}$). Density-based distances have been notable in the machine learning setting for over a decade [**?**, **?**]. To build a density-sensitive metric from density-based distances, we would like a cost function c that is small when close to the data set, and large when far away. The nearest neighbor function $\mathbf{r}_P$ is the most natural candidate, and has been traditionally used as a proximity measure between points and a data set in both the geometry and machine learning settings []. It has been used as such in Nearest Neighbor (and k-NN) classification, k-means/medians/center clustering, finite element methods, and any of the hundreds of methods that use Voronoi diagrams or Delaunay triangulation as intermediate data structures.

Definition 1.2. *Given any finite set $P \subset \mathbb{R}^k$, the* ***Nearest Neighbor Cost*** *of a path γ is $\ell_{\mathbf{r}_P}(\gamma)$, which we will abbreviate as $\ell_N(\gamma)$. The* ***Nearest Neighbor Metric*** *is defined as $\mathbf{d}_{\mathbf{r}_P}$, which we abbreviate as $\mathbf{d}_N$.*

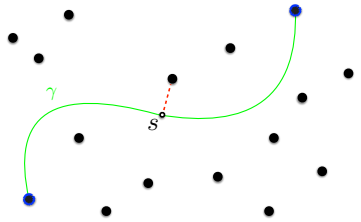

Figure 1: In this figure we have a collection of points. The length or cost of the green curve between the two blue points is the integral along the curve scaled by the distance to the nearest point.

The Nearest Neighbor metric, and density-based distances in general, are examples of manifold geodesics (see [] for details). Manifold geodesics are defined by embedding a point set into a continuous geometric manifold, and computing the infimum length path on the manifold structure between points. Within computer science, dozens of foundational papers in machine learning and surface reconstruction rely on manifold-based metrics to perform clustering, classification, regression, surface reconstruction, persistent homology, and more. Manifold geodesics predate computer science, and are the cornerstone of many fields of physics and mathematics. Exactly computing geodesics is fundamental to countless areas of physics including: the brachistochrone and minimal-drag-bullet problem of Bernoulli and Newton, exactly determining a particle's trajectory in classical physics (Hamilton's Principle of Least Action), computing the path of light through a non-homogeneous medium (Snell's law), finding the evolution of wave functions in quantum mechanics over time (Feynman path integrals), and determining the path of light in the presence of gravitational fields (General Relativity, Schwarzschild metric). In mathematics, manifold geodesics appear in nearly every branch of higher mathematics including differential equations, differential geometry, Lie theory, calculus of variations, algebraic geometry, and topology.

One of the most significant problems on any manifold geodesic is how to compute it exactly. Exact computation of manifold metrics is considered a fundamental problem in mathematics and physics, dating back for four centuries: entire fields of mathematics, including the celebrated calculus of variations, have arisen to tackle this []. Historically, mathematicians placed strong emphasis on exact computation as opposed to constant factor approximations. An algorithmic problem on manifold geodesics, with modern origins, is to $(1+\varepsilon)$ approximate these metrics efficiently on a computer. The core difficulty in the first problem is that geodesics are the minimum cost path out of an uncountable number of paths that can travel 'anywhere' on the manifold structure. This makes exactly computing these metrics challenging, even in the case of the Nearest Neighbor metric for just four fixed points in two dimensions (the authors are unaware of any easy method for this simplified task). The core tool for exactly computing manifold metrics, calculus of variations, is intractable on the nearest neighbor metric due to the metric's heavy dependence on the Voronoi diagram of the point set, which can be quite complicated for even five points in two dimensions (for more on this approach and its limitations, see []). Calculus of variations can show that the optimal nearest neighbor path is piecewise hyperbolic, but this is generally insufficient to exactly compute

the nearest neighbor metric - there are point sets where there are many differentiable, piecewise hyperbolic paths between two data points with different costs.

In this paper, we solve both problems: we exactly compute the Nearest Neighbor metric in all cases, and we $(1+\varepsilon)$ approximate it quickly. Our approach is based on conservative vector fields, contractive embeddings, Lipschitz extensions, and minimum cost flows on a graph. We combine these tools to prove that the nearest neighbor metric is exactly equal to a shortest path distance on a geometric graph, the so-called edge-squared metric, in all cases. This allows us to compute the nearest-neighbor metric exactly for any given point set in polynomial time, and it is the only known (non-trivial) density-based distance that can be computed discretely.

Definition 1.3. *Given points in Euclidean space, the* ***edge-squared graph*** *is the complete graph of Euclidean distances squared. The* ***edge-squared metric*** *is half the shortest path distance between two points on this graph.*

Here, the factor of half in the definition is a normalizing constant.

Theorem 1.4. *The nearest neighbor metric and edge squared metric are equivalent for any compact point set in arbitrary dimension*

The exact equality is realized when the nearest neighbor path is piecewise linear, traveling straight from data point to data point. The edge squared metric has been previously studied by multiple researchers in machine learning and power-efficient wireless networks, but previously has only been linked to the nearest neighbor metric by a fairly weak 3-approximation. Exact equality is considered highly surprising for at least four reasons:

1. The optimal nearest neighbor path for two points not in the dataset is generally piecewise hyperbolic. This holds true even when the dataset is a single point, and was established by [] using tools in Riemannian surfaces and the complex plane. Meanwhile, Theorem **??** implies an optimal nearest neighbor path for two data points is piecewise linear!

2. There are simple and natural variants of the Nearest Neighbor metric, for which no analog of Theorem 1.4 is known nor suspected. These variants are known as the q-Nearest Neighbor metric, for $1 < q < 2$, and we will formally define these metrics later in the introduction. When $q = 2$, these metrics coincide with the Nearest Neighbor metric. This gives us a natural suite of metrics that smoothly converge to the Nearest Neighbor metric, for which no theorem like Theorem 1.4 is known.

3. Even for just three points in a right triangle configuration, there exist an uncountable suite of optimal-cost paths between the two endpoints of the hypotenuse. Each path in this uncountable suite is piecewise hyperbolic, but, surprisingly, they all have the exact same cost as the edge-squared distance. In fact, the union of these paths is the entire right triangle. Thus, lowering the Nearest Neighbor function anywhere inside the triangle and using this to build a new density-based distance will break Theorem 1.4. This establishes that the equality in Theorem 1.4 is fairly tight.

4. This theorem holds for any compact point set, whether its n points in $n-1$ dimensional space or a countable union of compact geometric objects in countably infinite dimension. There is no other restriction on the compact geometric objects, and they need not be convex nor

simply connected. The geometry of these collections of objects can be extremely complicated, and it is generally hard to prove these types of results on such point sets.**Gary: Hard sell this on a union of line segments. Sounds surprising!**

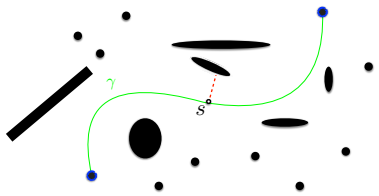

Figure 2: In this figure we have a collection of compact bodies in black. The length or cost of the green curve between the two blue points is the integral along the curve scaled by the distance to the nearest body. A curve may traverse a body at no cost. Theorem 1.4 establishes that the shortest path curve between two points goes straight from convex body to convex body.

We can now tackle a second problem of interest for manifold geodesics, which is efficiently $(1+\varepsilon)$ approximating them. In this paper, we show that the nearest neighbor metric admits $(1+\epsilon)$ spanners computable in nearly-linear time, with linear size, for any point set in constant dimension. Remarkably, these spanners are significantly sparser and faster to compute than the theoretically optimal Euclidean spanners with the same approximation constant, and nearly match the sparsity of the best known Euclidean Steiner spanners. Moreover, if the point set comes from a well-behaved probability distribution in constant dimension (a foundational assumption in machine learning []), we show that the nearest neighbor metric has perfect 1-spanners of nearly linear size. The latter result is impossible for many non-density sensitive metrics, such as the Euclidean metric. Both results rely on Theorem 1.4, and significantly improve the Nearest Neighbor spanners of Cohen et al in [].

Theorem 1.4 and our spanner theorems solve two core problems of interest for the nearest neighbor metric: exactly computing it for any dimension, and approximating it quickly for both general point sets and point sets arising from a well-behaved probability distribution in constant dimension. This is the first work we know of that computes a manifold metric exactly without calculus of variations, and we hope that our tools can be useful for other metric computations and approximations.

Besides for this contribution, we also generalize the Nearest Neighbor Metric to the q-Nearest Neighbor metric (abbreviated q-NN for short), and exactly compute this metric for all point sets with ≤ 4 points for all $q > 2$. We do this by equating it to the q-edge power metrics. Both the q-NN and q-edge power metrics will be defined later.

Theorem 1.5. *For point sets that are the union of up to* 4 *connected compact sets, the q-NN metric is exactly equal to the q-edge power metric when $q > 2$. This equality is false for all $q < 2, q \neq 1$.*

Our equality is robust enough to handle the union of 4 compact sets in any dimension. These unions ch can have very complicated geometry, and their Voronoi diagrams are in general difficult to understand. This is what makes theorems Theorem 1.5 surprising. We further conjecture:

Conjecture 1.6. *For any compact set, the q-NN metric is exactly equal to the q-edge poewr metric when $q > 2$.*

If true, this would give us a quadratic algorithm to compute the q-NN metric for any n point set.

Tim: Move to contributions section? We then use Theorem 1.4 to compute the persistent homology of the Nearest Neighbor metric, a task important in computational geometry. Additionally, we study the behavior of the Nearest Neighbor metric when the points are drawn from a well-behaved distribution, as the number of points goes to infinity. This turns out to converge w.h.p. to an extremely nice, $1 + o(1)$-approximation of a beautiful geodesic defined on the underlying density previously studied by applied probability theorists. This strengthens the work of Hwang, Hero, and Damelin, who showed that the Nearest Neighbor metric converged to a $O(1)$-approximation of this beautiful geodesic. This geodesic is a beautiful and natural generalization of both Euclidean distances and a distance fundamental for clustering using level-set methods. We further show that q-edge power metrics (and thus, it is hoped, the q-Nearest Neighbor metrics) are natural generalizations of maximum-edge-length distances on Euclidean MSTs, which in turn are fundamental for celebrated clustering methods like single-linkage clustering []. This implies that the q-edge power metric, and the Nearest Neighbor metric, can be used to generalize popular methods in clustering.

1.1 Contributions and Past Work

Besides for exactly computing the nearest neighbor metric, we present the following theorems on approximate computatoin:

Theorem 1.7. *For any set of points in $\mathbb{R}^d$ for constant d, there exists a $(1+\varepsilon)$ spanner of the Nearest Neighbor Metric with size $O\left(n\varepsilon^{-d/2}\right)$ computable in time $O\left(n \log n + n\varepsilon^{-d/2} \log \frac{1}{\varepsilon}\right)$. The $\log \frac{1}{\varepsilon}$ term goes away given access to an algorithm computing floor function in $O(1)$ time.*

Theorem 1.8. *Suppose points P in Euclidean space are drawn i.i.d from a Lipschitz probability density bounded above and below by a constant, with support on a smooth, connected, compact manifold with intrinsic dimension d, and smooth boundary of bounded curvature. Then w.h.p. the k-NN graph of P for $k = O(2^d \ln n)$ and edges weighted with Euclidean distance squared, is a 1-spanner of the Nearest Neighbor metric on P.*

These theorems rely on Theorem 1.4 and considerably strengthen the spanner resluts on the Nearest Neighbor metric from [**?**]. Theorem 1.4 will also allow us to compute the persistent homology of $\mathbf{d}_N$.

Theorem 1.7 proves that a $(1+\varepsilon)$-spanner of the edge-squared metric with points in constant dimension is sparser and can be computed more efficiently than the theoretical optimal Euclidean spanner []. Previously, sparse spanners of the edge-squared metric were shown to exist in two dimensions via Yao graphs and Gabriel graphs [**?**], but these did not generalize well to constant dimension: Yao graphs are not very efficient to compute, and Gabriel graphs can have quadratically many edges even in 3 dimension.

Theorem 1.8 proves that a 1-spanner of the edge-squared metric can be found assuming points are samples from a probability density, by using a k-NN graph for appropriate k. Our result is tight when d is constant. This is not possible for Euclidean distance, as a 1-spanner is almost surely the complete graph. Without the probability density assumption, there are point sets in $\mathbb{R}^4$ where 1-spanners of the edge-squared metric require $\Omega(n^2)$ edges.

Finally, we show how the nearest neighbor metric generalizes Euclidean distance and maximum-edge Euclidean MST distance [**?**].

1.2 Definitions and Preliminaries

Edge-squared metric: For $x \in \mathbb{R}^d$, let $\|x\|$ denote the Euclidean norm. For a set of points $P \subset \mathbb{R}^d$:

Definition 1.9. *The edge-squared metric for $a, b \in P$ is*

$$\mathbf{d}_2(a,b) = \min_{(p_0,\ldots,p_k)} \sum_{i=1}^{k} \|p_i - p_{i-1}\|^2,$$

where the minimum is over sequences of points $p_0, \ldots, p_k \in P$ with $p_0 = a$ and $p_k = b$.

Nearest Neighbor Metric: Another metric on the points of P is called the Nearest Neighbor Metric, and is denoted $\mathbf{d}_N$. This distance was first defined and studied in [**?**]. Before we can define it, we need a couple other definitions.

Given any finite set $P \subset \mathbb{R}^k$, there is a real-valued function $\mathbf{r}_P : \mathbb{R}^k \to \mathbb{R}$ defined as $\mathbf{r}_P(z) = \min_{x \in P} \|x - z\|$. A path is a continuous mapping $\gamma : [0,1] \to \mathbb{R}^d$. Let $\text{path}(a,b)$ denote the set of piecewise-C_1 paths from a to b. We will compute the lengths of paths relative to the distance function $\mathbf{r}_P$ as follows.

$$\ell(\gamma) := \int_0^1 \mathbf{r}_P(\gamma(t))\|\gamma'(t)\|dt.$$

By considering the velocity of γ, this definition is independent of the parameterization of the path.

Definition 1.10. *The Nearest Neighbor Metric distance is defined as:*

$$\mathbf{d}_N(a,b) := 4 \inf_{\gamma \in \text{path}(a,b)} \ell(\gamma).$$

The factor of 4 normalizes the metrics.

In particular, when P has only two points a and b, $\mathbf{d}_2(a,b) = \mathbf{d}_N(a,b)$. This reduces to a high school calculus exercise as the minimum path γ will be a straight line between the points and the nearest neighbor geodesic is

$$\mathbf{d}_N(a,b) = 4\int_0^1 \mathbf{r}_P(\gamma(t))\|\gamma'(t)\|dt = 8\int_0^{\frac{1}{2}} t\|a-b\|^2 dt = \|a-b\|^2 = \mathbf{d}_2(a,b).$$

This observation about pairs of points makes it easy to see that the Nearest Neighbor Metric distance is never greater than the edge-squared distance, as proven in the following lemma.

Lemma 1.11. *For all $s, p \in P$, we have $\mathbf{d}_N(s,p) \leq \mathbf{d}_2(s,p)$.*

Proof. Fix any points $s, p \in P$. Let $q_0, \ldots, q_k \in P$ be such that $q_0 = s$, $q_k = p$ and

$$\mathbf{d}_2(s,p) = \sum_{i=1}^{k} \|q_i - q_{i-1}\|^2.$$

Let $\psi_i(t) = tq_i + (1-t)q_{i-1}$ be the straight line segment from q_{i-1} to q_i. Observe that $\ell(\psi_i) = \|q_i - q_{i-1}\|^2/4$, by the same argument as in the two point case. Then, let ψ be the concatenation of the ψ_i and it follows that

$$\mathbf{d}_2(s,p) = 4\ell(\psi) \geq 4 \inf_{\gamma \in \text{path}(s,p)} \ell(\gamma) = \mathbf{d}_N(s,p). \qquad \square$$

Spanners: For real value $t \geq 1$, a t-spanner of a weighted graph G is a subgraph S such that $d_G(x,y) \leq d_S(x,y) \leq t \cdot d_G(x,y)$ where d_G and d_S represent the shortest path distance functions between vertex pairs in G and S. Spanners of Euclidean distances, and general graph distances, have been studied extensively, and their importance as a data structure is well established. [**?**, **?**, **?**, **?**].

k-nearest neighbor graphs: The k-nearest neighbor graph (k-NN graph) for a set of objects V is a graph with vertex set V and an edge from $v \in V$ to its k most similar objects in V, under a given distance measure. In this paper, the underlying distance measure is Euclidean, and the edge weights are Euclidean distance squared. k-NN graph constructions are a key data structure in machine learning [**?**, **?**], clustering [**?**], and manifold learning [**?**].

Gabriel Graphs: The Gabriel graph is a graph where two vertices p and q are joined by an edge if and only if the disk with diameter pq has no other points of S in the interior. The Gabriel graph is a subgraph of the Delaunay triangulation [**?**], and a 1-spanner of the edge-squared metric [**?**]. Gabriel graphs will be used in the proof of Theorem 1.8.

2 Outline

Section **??** contains the proof of Theorem 1.4, equating the edge-squared metric and Nearest Neighbor Metric distance in all cases. It should be noted that our proof is robust enough to handle not just finite point sets, but also countably infinite collections of disjoint path-connected, compact sets. Remarkably, there is no restriction on the convexity or simply-connectedness of these sets.

We then compute the persistent homology of the Nearest Neighbor Metric distance. Section 4.1 outlines a proof of Theorem 1.7, and compares our spanner to new lower bounds on the sparsity of $(1+\varepsilon)$-spanners of the Euclidean metric. We outline a proof of Theorem 1.8 in Section 4 and discuss its implications.

Section 6 introduces the p-power metrics. We show that Euclidean spanners and Euclidean MSTs are special cases of p-power spanners. We show how clustering algorithms including k-means, level-set methods, and single linkage clustering, are special cases of clustering with p-power metrics.

Conclusions and open questions are in Section 5. Full proofs for Theorems 1.8, 1.7 are contained in the Appendix.

3 Exactly Computing the Nearest Neighbor Metric

In this section, we prove Theorem 1.4. By Lemma 1.11, it suffices to show that for a and b on Euclidean point set P, we have $\mathbf{d}_N(a,b) \geq \mathbf{d}_2(a,b)$ for $a, b \in P$. This allows us exact computation of the Nearest Neighbor Metric. We note that remarkably, our proof of equality between metrics is robust enough to handle the case of a finite disjoint union of path-connected compact sets in finite dimension, which in general can have very complicated geometry.

Let $P \subset \mathbb{R}^d$ be a set of n points. Pick any *source* point $s \in P$. Order the points of P as $p_1, \ldots, p_n$ so that

$$\mathbf{d}_2(s,p_1) \leq \cdots \leq \mathbf{d}_2(s,p_n).$$

This will imply that $p_1 = s$. It will suffice to show that for all $p_i \in P$, we have $\mathbf{d}_2(s,p_i) = \mathbf{d}_N(s,p_i)$. There are three main steps:

1. We first show that when P is a subset of the vertices of an axis-aligned box, $\mathbf{d} = \mathbf{d}_N$. In this case, shortest paths for $\mathbf{d}$ are single edges and shortest paths for $\mathbf{d}_N$ are straight lines.
2. We then show how to lift the points from $\mathbb{R}^d$ to $\mathbb{R}^n$ by a Lipschitz map m that places all the points on the vertices of a box and preserves $\mathbf{d}_2(s, p)$ for all $p \in P$.
3. Finally, we show how the Lipschitz extension of m is also Lipschitz as a function between Nearest Neighbor Metric distances. We combine these pieces to show that $\mathbf{d} \leq \mathbf{d}_N$. As $\mathbf{d} \geq \mathbf{d}_N$ (Lemma 1.11), this will conclude the proof that $\mathbf{d} = \mathbf{d}_N$.

3.0.1 Boxes

Let Q be the vertices of a box in $\mathbb{R}^n$. That is, there exist some positive real numbers $\alpha_1, \ldots, \alpha_n$ such that each $q \in Q$ can be written as $q = \sum_{i \in I} \alpha_i e_i$, for some $I \subseteq [n]$.

Let the source s be the origin. Let $\mathbf{r}_Q : \mathbb{R}^n \to \mathbb{R}$ be the distance function to the set Q. Setting $r_i(x) := \min\{x_i, \alpha_i - x_i\}$ (a lower bound on the difference in the ith coordinate to a vertex of the box), it follows that

$$\mathbf{r}_Q(x) \geq \sqrt{\sum_{i=1}^{n} r_i(x)^2}. \tag{1}$$

Let $\gamma : [0, 1] \to \mathbb{R}^n$ be a curve in $\mathbb{R}^n$. Define $\gamma_i(t)$ to be the projection of γ onto its ith coordinate. Thus,

$$r_i(\gamma(t)) = \min\{\gamma_i(t), \alpha_i - \gamma_i(t)\} \tag{2}$$

and

$$\|\gamma'(t)\| = \sqrt{\sum_{i=1}^{n} \gamma_i'(t)^2}. \tag{3}$$

We can bound the length of γ as follows. For simplicity of exposition we only present the case of a path from the origin to the far corner, $p = \sum_{i=1}^{n} \alpha_i e_i$.

$$\begin{aligned}
\ell(\gamma) &= \int_0^1 \mathbf{r}_Q(\gamma(t))\|\gamma'(t)\|dt && \text{[by definition]} \\
&\geq \int_0^1 \left(\sqrt{\sum_{i=1}^{n} r_i(\gamma(t))^2} \sqrt{\sum_{i=1}^{n} \gamma_i'(t)^2} \right) dt && \text{[by (1) and (3)]} \\
&\geq \sum_{i=1}^{n} \int_0^1 r_i(\gamma(t))\gamma_i'(t)dt && \text{[Cauchy-Schwarz]} \\
&\geq \sum_{i=1}^{n} \left(\int_0^{\ell_i} \gamma_i(t)\gamma_i'(t)dt + \int_{\ell_i'}^1 (\alpha_i - \gamma_i(t))\gamma_i'(t)dt \right) \\
&\quad \text{[by (2) where } \gamma_i(\ell_i) = \alpha_i/2 \text{ for the first time and } \gamma_i(\ell_i') = \alpha_i/2 \text{ for the last time.]} \\
&= \sum_{i=1}^{n} 2 \int_0^{\ell_i} \gamma_i(t)\gamma_i'(t)dt && \text{[by symmetry]} \\
&\geq \sum_{i=1}^{n} \frac{\alpha_i^2}{4} && \text{[basic calculus]}
\end{aligned}$$

It follows that if γ is any curve that starts at s and ends at $p = \sum_{i=1}^{n} \alpha_i e_i$, then $\mathbf{d}_N(s,p) = \mathbf{d}_2(s,p)$.

3.0.2 Lifting the points to $\mathbb{R}^n$

Define a mapping $m : P \to \mathbb{R}^n$. We do this by adding the points $p_1, \ldots, p_n$, as defined above, one point at a time. For each new point we will introduce a new dimension. We start by setting $m(p_1) = 0$ and by induction:

$$m(p_i) = m(p_{i-1}) + \sqrt{\mathbf{d}_2(s,p_i) - \mathbf{d}_2(s,p_{i-1})}e_i, \tag{4}$$

where the vectors e_i are the standard basis vectors in $\mathbb{R}^n$.

Lemma 3.1. *For all $p_i, p_j \in P$, we have*

(i) $\|m(p_j) - m(p_i)\| = \sqrt{|\mathbf{d}_2(s,p_j) - \mathbf{d}_2(s,p_i)|}$, *and*

(ii) $\|m(s) - m(p_j)\|^2 \leq \|m(p_i)\|^2 + \|m(p_i) - m(p_j)\|^2$.

Proof. Proof of (i). Without loss of generality, let $i \leq j$.

$$\begin{aligned}
\|m(p_j) - m(p_i)\| &= \left\| \sum_{k=i+1}^{j} \sqrt{\mathbf{d}_2(s,p_k) - \mathbf{d}_2(s,p_{k-1})}e_k \right\| && \text{[from the definition of } m\text{]} \\
&= \sqrt{\sum_{k=i+1}^{j} (\mathbf{d}_2(s,p_k) - \mathbf{d}_2(s,p_{k-1}))} && \text{[expand the norm]} \\
&= \sqrt{\mathbf{d}_2(s,p_j) - \mathbf{d}_2(s,p_i)}. && \text{[telescope the sum]}
\end{aligned}$$

Proof of (ii). As $m(s) = 0$, it suffice to observe that

$$\begin{aligned}
\|m(p_j)\|^2 &= \mathbf{d}_2(s,p_j) && \text{[by } (i)\text{]} \\
&\leq \mathbf{d}_2(s,p_i) + |\mathbf{d}_2(s,p_j) - \mathbf{d}_2(s,p_i)| && \text{[basic arithmetic]} \\
&= \|m_(p_i)\|^2 + \|m(p_i) - m(p_j)\|^2 && \text{[by } (i)\text{]}
\end{aligned}$$

□

We can now show that m has all of the desired properties.

Proposition 3.2. *Let $P \subset \mathbb{R}^d$ be a set of n points, let $s \in P$ be a designated source point, and let $m : P \to \mathbb{R}^n$ be the map defined as in (4). Let $\mathbf{d}'$ denote the edge squared metric for the point set $m(P)$ in $\mathbb{R}^n$. Then,*

(i) m is 1-Lipschitz as a map between Euclidean metrics,

(ii) m maps the points of P to the vertices of a box, and

(iii) m preserves the edge squared distance to s, i.e. $\mathbf{d}'(m(s), m(p)) = \mathbf{d}_2(s,p)$ for all $p \in P$.

Proof. Proof of (i). To prove the Lipschitz condition, fix any $a, b \in P$ and bound the distance as follows.

$$\begin{aligned}
\|m(a) - m(b)\| &= \sqrt{|\mathbf{d}_2(s,a) - \mathbf{d}_2(s,b)|} && \text{[Lemma 3.1(i)]}\\
&\leq \sqrt{\mathbf{d}_2(a,b)} && \text{[triangle inequality]}\\
&\leq \|a - b\| && [\mathbf{d}_2(a,b) \leq \|a-b\|^2 \text{ by the definition of } \mathbf{d}]
\end{aligned}$$

Proof of (ii). That m maps P to the vertices of a box is immediate from the definition. The box has side lengths $\|m_i - m_{i-1}\|$ for all $i > 1$ and $p_i = \sum_{k=1}^{i} \|m_k - m_{k-1}\| e_k$.
Proof of (iii). We can now show that the edge squared distance to s is preserved. Let $q_0, \ldots, q_k$ be the shortest sequence of points of $m(P)$ that realizes the edge-squared distance from $m(s)$ to $m(p)$, i.e., $q_0 = m(s)$, $q_k = m(p)$, and

$$\mathbf{d}'(m(s), m(p)) = \sum_{i=1}^{k} \|m(q_i) - m(q_{i-1})\|^2.$$

If $k > 1$, then Lemma 3.1(ii) implies that removing q_1 gives a shorter sequence. Thus, we may assume $k = 1$ and therefore, by Lemma 3.1(i),

$$\mathbf{d}'(m(s), m(p)) = \|m(s) - m(p)\|^2 = \mathbf{d}_2(s,p). \qquad \square$$

3.0.3 The Lipschitz Extension

Proposition 3.2 and the Kirszbraun theorem on Lipschitz extensions imply that we can extend m to a 1-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}^n$ such that $f(p) = m(p)$ for all $p \in P$ [**?**, **?**, **?**].

Lemma 3.3. *The function f is also 1-Lipschitz as mapping from $\mathbb{R}^d \to \mathbb{R}^n$ with both spaces endowed with the Nearest Neighbor Metric distance.*

Proof. We are interested in two distance functions $\mathbf{r}_P : \mathbb{R}^d \to \mathbb{R}$ and $\mathbf{r}_{f(P)} : \mathbb{R}^n \to \mathbb{R}$. Recall that each is the distance to the nearest point in P or $f(P)$ respectively.

$$\begin{aligned}
\mathbf{r}_{f(P)}(f(x)) &= \min_{q \in f(P)} \|q - f(x)\| && \text{[by definition]}\\
&= \min_{p \in P} \|f(p) - f(x)\| && [q = f(p) \text{ for some } p]\\
&\leq \min_{p \in P} \|p - x\| && [f \text{ is 1-Lipschitz}]\\
&= \mathbf{r}_P(x). && \text{[by definition]}
\end{aligned}$$

For any curve $\gamma : [0,1] \to \mathbb{R}^d$ and for all $t \in [0,1]$, we have $\|(f \circ \gamma)'(t)\| \leq \|\gamma'(t)\|$. It then follows that

$$\ell'(f \circ \gamma) = \int_0^1 \mathbf{r}_{f(P)}(f(\gamma(t)))\|(f \circ \gamma)'(t)\| dt \leq \int_0^1 \mathbf{r}_P(\gamma(t))\|\gamma'(t)\| dt = \ell(\gamma), \tag{5}$$

where ℓ' denotes the length with respect to $\mathbf{r}_{f(P)}$. Thus, for all $a, b \in P$,

$$\begin{aligned}
\mathbf{d}_N(a,b) &= 4 \inf_{\gamma \in \text{path}(a,b)} \ell(\gamma) && \text{[by definition]}\\
&\geq 4 \inf_{\gamma \in \text{path}(a,b)} \ell'(f \circ \gamma) && \text{[by (5)]}\\
&\geq 4 \inf_{\gamma' \in \text{path}(f(a),f(b))} \ell'(\gamma') && [\text{because } f \circ \gamma \in \text{path}(f(a), f(b))]\\
&= \mathbf{d}_N(f(a), f(b)). && \text{[by definition]}
\end{aligned}$$

□

We now restate Theorem 1.4 for convenience, and prove it.

Theorem 3.4. *For any point set $P \subset \mathbb{R}^d$, the edge squared metric $\mathbf{d}$ and the Nearest Neighbor Metric distance $\mathbf{d}_N$ are identical.*

Proof. Fix any pair of points s and p in P. Define the Lipschitz mapping m and its extension f as in (4). Let $\mathbf{d}'$ and $\mathbf{d}'_N$ denote the edge-squared and Nearest Neighbor Metric distances on $f(P)$ in $\mathbb{R}^n$.

$$\begin{aligned}
\mathbf{d}_2(s,p) &= \mathbf{d}'(m(s),m(p)) && \text{[Proposition 3.2(iii)]}\\
&= \mathbf{d}'_N(m(s),m(p)) && \text{[}f(P)\text{ are vertices of a box]}\\
&\leq \mathbf{d}_N(s,p) && \text{[Lemma 3.3]}
\end{aligned}$$

We have just shown that $\mathbf{d} \leq \mathbf{d}_N$ and Lemma 1.11 states that $\mathbf{d} \geq \mathbf{d}_N$, so we conclude that $\mathbf{d} = \mathbf{d}_N$ as desired. □

3.1 Persistent Homology of the Nearest-neighbor Geodesic Distance

In this section, we show how to compute the so-called persistent homology [**?**] of the Nearest Neighbor Metric distance in two different ways, one ambient and the other intrinsic. The latter relies on Theorem 1.4 and would be quite surprising without it.

Persistent homology is a popular tool in computational geometry and topology to ascribe quantitative topological invariants to spaces that are stable with respect to perturbation of the input. In particular, it's possible to compare the so-called persistence diagram of a function defined on a sample to that of the complete space [**?**]. These two aspects of persistence theory—the intrinsic nature of topological invariants and the ability to rigorously compare the discrete and the continuous—are both also present in our theory of Nearest Neighbor Metric distances. Indeed, the primary motivation for studying these metrics was to use them as inputs to persistence computations for problems such as persistence-based clustering [**?**] or metric graph reconstruction [**?**].

The input for persistence computation is a *filtration*—a nested sequence of spaces, usually parameterized by a real number $\alpha \geq 0$. The output is a set of points in the plane called a *persistence diagram* that encodes the birth and death of topological features like connected components, holes, and voids.

The Ambient Persistent Homology Perhaps the most popular filtration to consider on a Euclidean space is the sublevel set filtration of the distance to a sample P. This filtration is $(F_\alpha)_{\alpha\geq 0}$, where

$$F_\alpha := \{x \in \mathbb{R}^d \mid \mathbf{r}_P(x) \leq \alpha\},$$

for all $\alpha \geq 0$. If one wanted to consider instead the Nearest Neighbor Metric distance $\mathbf{d}_N$, one gets instead a filtration $(G_\alpha)_{\alpha\geq 0}$, where

$$G_\alpha := \{x \in \mathbb{R}^d \mid \min_{p\in P} \mathbf{d}_N(x,p) \leq \alpha\},$$

for all $\alpha \geq 0$.

Both the filtrations (F_α) and (G_α) are unions of metric balls. In the former, they are Euclidean. In the latter, they are the metric balls of $\mathbf{d}_N$. These balls can look very different, for example, for $\mathbf{d}_N$, the metric balls are likely not even convex. However, these filtrations are very closely related.

Lemma 3.5. *For all $\alpha \geq 0$, $F_\alpha = G_{2\alpha^2}$.*

Proof. The key to this exercise is to observe that the nearest point $p \in P$ to a point x is also the point that minimizes $\mathbf{d}_N(x, p)$. To prove this, we will show that for any $p \in P$ and any path $\gamma \in \text{path}(x, p)$, we have $\ell(\gamma) \geq \frac{1}{2}\mathbf{r}_P(x)^2$. Consider any such x, p, and γ. The euclidean length of γ must be at least $\mathbf{r}_P(x)$, so we will assume that $\|\gamma'\| = \mathbf{r}_P(x)$ and will prove the lower bound on the subpath starting at x of length exactly $\mathbf{r}_P(x)$. This will imply a lower bound on the whole path. Because $\mathbf{r}_P$ is 1-Lipschitz, we have $\mathbf{r}_P(\gamma(t)) \geq (1-t)\mathbf{r}_P(x)$ for all $t \in [0, 1]$. It follows that

$$\ell(\gamma) = \int_0^1 \mathbf{r}_P(\gamma(t))\|\gamma'(t)\|dt \geq \mathbf{r}_P(x)^2 \int_0^1 (1-t)dt = \frac{1}{2}\mathbf{r}_P(x)^2$$

The bound above applies to any path from x to a point $p \in P$, and so,

$$\mathbf{d}_N(x, p) = 4 \inf_{\gamma \in \text{path}(x,p)\ell(\gamma)} \geq 2\mathbf{r}_P(x).$$

If p is the nearest neighbor of x in P, then $\mathbf{d}_N(x, p) = 2\mathbf{r}_P(x)$, by taking the path to be a straight line. It follows that $\min_{p \in P} \mathbf{d}_N(x, p) = 2\mathbf{r}_P(x)$. □

The preceding lemma shows that the two filtrations are equal up to a monotone change in parameters. By standard results in persistent homology, this means that their persistence diagrams are also equal up to the same change in parameters. This means that one could use standard techniques such as α-complexes [**?**] to compute the persistence diagram of the Euclidean distance and convert it to the Nearest Neighbor Metric distance afterwards. Moreover, one observes that the same equivalence will hold for variants of the Nearest Neighbor Metric distance that take other powers of the distance.

Intrinsic Persistent Homology Recently, several researchers have considered intrinsic nerve complexes on metric data, especially data coming from metric graphs [**?**, **?**]. These complexes are defined in terms of the intersections of metric balls in the input. The vertex set is the input point set. The edges at scale α are pairs of points whose α-radius balls intersect. In the intrinsic Čech complex, triangles are defined for three way intersections, tetrahedra for four-way intersections, etc.

In Euclidean settings, little attention was given to the difference between the intrinsic and the ambient persistence, because a classic result, the Nerve Theorem [**?**], and its persistent version [**?**] guaranteed there is no difference. The Nerve theorem, however, requires the common intersections to be contractible, a property easily satisfied by convex sets such as Euclidean balls. However, in many other topological metric spaces, the metric balls might not be so well-behaved. In particular, the Nearest Neighbor Metric distance has metric balls which may take on very strange shapes, depending on the density of the sample. This is similarly true for graph metrics. So, in these cases, there is a difference between the information in the ambient and the intrinsic persistent homology.

Theorem 3.6. *Let $P \subset \mathbb{R}^d$ be finite and let $\mathbf{d}_N$ be the Nearest Neighbor Metric distance with respect to P. The edges of the intrinsic Čech filtration with respect to $\mathbf{d}_N$ can be computed exactly in polynomial time.*

Proof. The statement is equivalent to the claim that $\mathbf{d}_N$ can be computed exactly between pairs of points of P, a corollary of Theorem 3.4. Two radius α balls will intersect if and only of the distance between their centers is at most 2α. The bound on the distance necessarily implies a path and the common intersection will be the midpoint of the path. □

4 Exact-spanners of Edge-Squared Metrics in the Probability Density Setting

Theorem 1.8 states that for $k = O(2^d \log n)$, the k-NN graph of n points drawn i.i.d from a nicely behaved probability distribution is a 1-spanner of the edge-squared metric. Theorem 1.8 is impossible for general point sets: Lemma **??** gives an example where a 1-spanner of the edge-squared metric in 4 dimensions requires $\Omega(n^2)$ edges. This result is also impossible for Euclidean distances, whose 1-spanner is the complete graph almost surely. Our theorem implies any off-the-shelf k-nearest neighbor graph generator can compute edge-squared metric. In this section, we outline a proof and defer the analytical details to Appendix 9.

First, let us assume that the support of our probability density D has the same dimension as our ambient space. This simplifies our calculations without changing the problem much. Then, we note that as our number of sample points get large, the density inside a k-NN ball around any point x (the ball with radius k^{th}-NN distance, center at x) looks like the uniform distribution on that ball, possibly intersected with a halfspace. The bounding plane of our halfspace represents the boundary of our density D.

For simplicity in the outline, let's suppose that D is convex. If we condition on the radius of the k-NN ball, then the $k - 1^{st}$ nearest neighbors of x are distributed roughly according to the above distribution, described by the ball intersected with a halfspace. For any other point p in D, we project p onto the k-NN ball to point p', and show that the ball $p'x$ contains a k^{th} nearest neighbor w.h.p, when $k = O(2^d \log n)$. This implies ball with diameter px contains a k^{th} nearest neighbor of x, and thus px is not necessary in any 1-spanner of the edge-squared metric. Then we take union bound over all x. A rigorous proof of Theorem 1.8 requires careful analysis, and is contained in Section 9. Our proof can be tweaked to show:

Theorem 4.1. *Given a Lipschitz distribution bounded above and below with support on convex set $C \subset \mathbb{R}^d$, the k-NN graph is Gabriel w.h.p. for $k = O(2^d \log n)$.*

4.1 Fast, Sparse Spanner for the Edge-Squared Metric

Now we outline a proof for Theorem 1.7, which shows that one can construct a $(1 + \varepsilon)$ Nearest Neighbor Metric spanner of size $O(n\varepsilon^{-d/2})$ in time $O\left(n \log n + n\varepsilon^{-d/2} \log\left(\frac{1}{\varepsilon}\right)\right)$, for points in constant dimensional space. The full proof is in Appendix 8. We critically rely on Theorem 1.4 for this work, which shows edge-squared metric.

Note that this spanner is sparser and faster in terms of epsilon dependency than the theoretical optimal spanner for Euclidean distances []. We rely extensively on well-separated pair decompositions (WSPDs), and this outline assumes familiarity with that notation. For a comprehensive set of definitions and notations on well separated pairs, refer to any of [**?**, **?**, **?**, **?**]. Our proof consists of three parts.

1. Showing that connecting a $(1+O(\delta^2))$-approximate shortest edge in a $1/\delta$ well separated pair for all the pairs in the decomposition gives a $1+O(\delta^2)$ edge-squared spanner. The processing for this step takes $O(n \log n + \delta^{-d} n)$ time.

2. Previous work contains an algorithm computing $1 + O(\delta^2)$-approximate shortest edge in a $1/\delta$ well separated pair for all the pairs in a WSPD, and takes $O(1)$ time per pair. The pre-processing for this step will be bounded by $O(\delta^{-d} n \log\left(\frac{1}{\delta}\right))$ time. The $\log\left(\frac{1}{\delta}\right)$ factor goes away given a fast floor function. This procedure was first introduced in [**?**].

3. Putting these two together, and setting $\epsilon = \delta^2$ gives us a $1+\epsilon$ spanner with $O(\epsilon^{-d/2} n)$ edges in $O(n \log n + \epsilon^{-d/2} n)$ time.

Full details of this proof are contained in Appendix 8

4.2 Lower Bounds for Sparsity of Euclidean Spanners

Theorem 4.2. *For constant d and any fixed ε, there exists a set of points such that any $(1+\varepsilon)$ Euclidean spanner in $\mathbb{R}^d$ needs $\Omega\left(n\varepsilon^{-\lfloor d/2 \rfloor + 1}\right)$ edges.*

5 Conclusions and Open Questions

We examined a graph-based distance on Euclidean point sets, showed it equaled a special density-based distance, and built sparser and faster spanners on this metric than is known for Euclidean distances. Such sparse data structures may be surprising given that the metric can have high doubling dimension. Many problems remain open.

Is there a generalization of Theorem 1.4 to p-power metrics? This would require defining a new version of the Nearest Neighbor Metric distance. Separately, are the proof techniques for Theorem 1.4 of use for computing or approximating other density-based distances? Can non-spanner data structures for clustering with the edge-squared metric be computed efficiently? Such data structures include core-sets and distance oracles [**?**, **?**, **?**].

Can we efficiently compute $o(\log n)$-spanners of the p-power metric in high dimension with a nearly linear number of edges? The existence of such spanners has been studied for Euclidean metrics in [**?**], where the stretch obtained is $\sqrt{\log n}$. Good constructions for $(1+\varepsilon)$-spanners of the normalized ∞-power metric are known: many (but not all) approximate Euclidean MST constructions are $(1+\varepsilon)$-spanners of this metric [**?**, **?**]. Can high-dimensional approximate Euclidean MST algorithms [**?**, **?**, **?**] be adapted to create efficient p-power spanners? Any spanner for high-dimensional edge-squared metrics must give the same quality spanner for negative type distances [**?**, **?**], which include l_2 and l_1.

Does computing k-NN graphs with approximate nearest neighbor methods give 1-spanners of the edge-squared metric with high probability? Approximate nearest neighbors have been studied extensively [**?**, **?**, **?**], including locality-sensitive hashing for high dimensional point sets [**?**] and more [**?**]. Recent work by Andoni et al. [**?**] showed how to compute approximate nearest neighbors for any non-Euclidean norm. Perhaps there is a rigorous theory about density-sensitive metrics generated from any such norm? Similar to how the edge-squared metric is generated from the Euclidean distance.

It remains an open question how well clustering or classification with edge-squared metrics and Nearest Neighbor Metric distances performs on real-world data. Experiments have been done by

Bijral, Ratliff, and Srebro in [**?**]. Theorem 1.8 implies that future experiments can be done using any k-nearest-neighbor graph.

6 Relating the Nearest Neighbor Metric to Euclidean MSTs, Euclidean Spanners, and More

The Nearest Neighbor metric, as seen in Theorem 1.4, is equal to the edge-squared metric. This allows us to connect this manifold distance to a graph distance, which we will in turn show is a generalization of maximum-edge distance on minimum spanning trees.

The edge-squared metric on a Euclidean point set, as we recall, is defined by taking the Euclidean distances squared and finding the shortest paths. We could have taken any such power p of the Euclidean distances. We will soon see that taking $p = 1$ gives us the Euclidean distance, and finding spanners of the graph as $\lim p \to \infty$ is the Euclidean MST problem. Let the p-power metric be defined on a Euclidean point set by taking Euclidean distances to the power of p, and performing all-pairs shortest path on the resulting distance graph.

Theorem 6.1. *For all $q > p$, any 1-spanner of the p-power metric is a 1-spanner of the q-power metric on the same point set*

Proof. A 1-spanner of the q-power metric can be made by taking edges uv where

$$\min_{p_0=u,\ldots p_k=v,k\neq 1} \sum_k ||p_i - p_{i-1}||^q > ||u-v||^q. \tag{6}$$

If $\sum_{i=1}^k ||p_i - p_{i-1}||^q > ||u-v||^q$ for any points $p_1, \ldots p_k$, then $\sum_{i=1}^k ||p_i - p_{i-1}||^p > ||u-v||^p$ for any $q > p$. Thus, for all such edges uv satisfying Equation 6:

$$\min_{p_0=u,\ldots p_k=v,k\neq 1} \sum_k ||p_i - p_{i-1}||^p > ||u-v||^p.$$

Such edges uv must be included in any 1-spanner of the p-power metric. □

Corollary 6.1.1. *Let P be a set of points in Euclidean space drawn i.i.d. from a Lipschitz probability density bounded above and below, with support on a smooth, compact manifold with intrinsic dimension d, bounded curvature, and smooth boundary of bounded curvature. Then the k-NN graph on P when $k = O(2^d \log n)$ is a 1-spanner of the p-power metric for every $p \geq 2$, w.h.p.*

This follows from combining Theorem 1.8 and Theorem 6.1.

6.1 Relation to the Euclidean MST problem

Definition 6.2. *Let the **normalized p-power metric** between two points in $\mathbb{R}^d$ be the p-power metric between the two points, raised to the $\frac{1}{p}$ power. Define the normalized ∞-power metric as the limit of the normalized p-power metric as $p \to \infty$.*

Lemma 6.3. *The Euclidean MST is a 1-spanner for the normalized ∞-power metric.*

This lemma follows from basic properties of the MST. The normalized p-power metrics give us a suite of metrics such that $p = 1$ is the Euclidean distance and $p = \infty$ gives us the distance of the longest edge on the unique MST-path. Setting $p = 2$ gives the edge-squared metric, which sits between the Euclidean and max-edge-on-MST-path distance. Theorem 6.1 establishes that minimal 1-spanners of the (normalized) p-power metric are contained in each other, as p varies from 1 to ∞. The minimal spanner for a general point set when $p = 1$ is the complete graph, and the Euclidean MST is the minimal spanner for $p = \infty$. Thus:

Theorem 6.4. *For points in $\mathbb{R}^d$, every 1-spanner of the p-power metric on that set of points contains every Euclidean MST.*

Corollary 6.4.1. *Every 1-spanner for the edge-squared metric and/or Nearest Neighbor Geodesic contains every Euclidean MST.*

6.2 Generalizing Single Linkage Clustering, Level Sets, and k-Centers clustering

If our point set is drawn from a well-behaved probability density, then the normalized edge-power metrics converge to a nice geodesic distance detailed in [**?**]. When $p = 1$, clustering with this metric is the same as Euclidean metric clustering (k-means, k-medians, k-centers), and when $p = \infty$, clustering with this metric is the same as the widely used level-set method [**?**, **?**, **?**, **?**]. Thus, clustering with normalized edge-power metrics generalizes these two very popular methods, and interpolates between their advantages. Definitions of the level-set method and a full discussion are contained in Appendix 7

7 Edge-Power Metrics relate to Single Linkage Clustering, Level Sets, and k-Centers clustering

Many popular clustering algorithms, including k-centers, k-means, and k-medians clustering, use Euclidean distance as a measure of distance between points in $\mathbb{R}^d$. These methods are useful when clusters are spherical and well-separated. However, it is believed by practitioners that density-sensitive distances more accurately capture intrinsic distances between data [**?**].

The celebrated single-linkage clustering algorithm [**?**, **?**], which is clustering based on an MST, is a widely used tool in machine learning, and gets around many of the problems of the Euclidean distance clustering. In single-linkage clustering, two points are considered similar if the maximum length edge on the path between them in the MST is small. This turns out to be equivalent to computing the normalized ∞-power metric between the two points. Therefore, single linkage clustering can be seen as clustering using the normalized ∞-power metric. Generally, normalized p-power metrics can be seen as an intermediary between Euclidean distances (1-power metrics) and Euclidean MST-based clustering.

Clustering with p-power metric relates to another popular clustering method in machine learning, known as level-set clustering. Loosely speaking, level set clustering involves finding an estimate for the probability density that points are drawn from, finding a cut threshold t, and then taking as clusters all regions with probability density $> t$. Level set clustering has appeared in many incarnations [**?**, **?**, **?**], including the celebrated and widely used DBScan method [**?**] and its considerable

number of variations [**?**]. It is known that level-set clustering is related to single-linkage clustering, as the latter is an approximation of the former [**?**, **?**]. Level-set methods have the advantage that they can find arbitrarily shaped clusters [**?**], but can cause two points that are very close in Euclidean distance to be considered far apart.

Clustering with the p-power metric incorporates the advantages of both Euclidean distance clustering and level set clustering, as it is both density-sensitive and takes into account overall Euclidean distance between two points. Here, p can be toggled to change the sensitivity of the metric to the underlying density. As the number of samples drawn from our probability density grows large, it has been proven that the behavior of normalized p-power metrics converges to a natural geodesic distance on the underlying probability density [**?**]. Clustering with this geodesic distance for $p = 1$ is exactly Euclidean clustering, and for $p = \infty$ is exactly the level set method. Thus, clustering with p-power metric converges to a clustering method that smoothly interpolates between Euclidean-distance clustering and level set clustering.

8 Proving Faster and Sparser-than-Euclidean Approximate Spanners

In this appendix, we finish the proof of Theorem 1.7 based on the outline given in Section 4.1.

8.1 $1 + O(\delta^2)$ spanners can be generated from a $1/\delta$ WSPD

Definition 8.1. *Let e be a **critical** edge in a shortest path metric on any graph if the (possibly-not-unique) shortest path between the endpoints of e is the edge e.*

Lemma 8.2. *The set of critical edges on any graph forms a 1-spanner of the shortest path metric.*

The above lemma is known in the literature.

To check that any graph H is a $(1 + O(\delta^2)$ spanner of any graph G, it suffices to prove that all critical edges in the edge-squared metric have a stretch no larger than $1 + O(\delta^2)$. Let G be the edge-squared graph arising from points $P \subset \mathbb{R}^d$. Build a well-separated pair decomposition on P, with pairs given as $\{A_1, B_1\}, \{A_2, B_2\}, \ldots \{A_m, B_m\}$. Create a spanner H as follows: for each pair $\{A_i, B_i\}$, connect an edge $\{a, b\}, a \in A_i, b \in B_i$ such that the Euclidean distance between a and b is a $(1 + c\delta^2)$ approximation of the shortest distance between point sets A_i and B_i, for some constant c independent of i. This can be accomplished in $O(1)$ time assuming a preprocessing step of $O(\delta^{-d} \log\left(\frac{1}{\delta}\right)$ time, as noted in Callahan's paper on constructing a Euclidean MST [**?**]. Do this for all $1 \leq i \leq m$.

For each critical edge (s, t), consider the well-separated pair $\{A, B\}$ that (s, t) is part of. Let $s \in A$ and $t \in B$. Let (a, b) be a $(1 + c\delta^2)$-approximate shortest edge between A and B ($a \in A, b \in B$). Scale $||a - b||_2$ to be 1. A and B have Euclidean radius at most δ, by the definition of a well separated pair. By induction on Euclidean distance, H is an edge-squared 2-spanner of the edge-squared metric for all points in A and B and all points in B (assuming sufficiently small δ).

Lemma 8.3.

$$dist_H(s, t) \leq dist_H(s, a) + dist_H(a, b) + dist_H(b, t) \leq 1 + O(\delta^2)$$

Proof. We know $dist_H(a,b) = 1$ by our scaling, and

$$dist_H(s,a) \leq 2 \cdot (dist_G(s,a)) \leq 2 \cdot ||s-a||^2 \leq 8\delta^2$$

The first inequality follows by the inductive hypothesis that H is a 2-spanner of G in A. The third inequality follows since both s and a are contained in a ball of radius δ.

The same bound applies for $dist_H(b,t)$. □

Lemma 8.4.

$$(1+c\delta^2)(dist_G(s,t)) \geq dist_G(a,b) = 1$$
$$\Rightarrow dist_G(s,t) \geq \frac{1}{1+c\delta^2}$$

Lemma 8.4 follows from the fact that (a,b) is a $(1+c\delta^2)$ approximate shortest distance between A and B.

Therefore

$$stretch_H(s,t) \leq \frac{dist_H(s,t)}{dist_G(s,t)} \leq (1+16\delta^2)(1+c\delta^2) = 1 + O(\delta^2) \tag{7}$$

Thus we have proven that H is a $1+16\delta^2$ spanner. Now set $\epsilon = \delta^2$, which completes proof of Theorem 1.7.

9 Spanners in the Probability Density Setting: Full Proof

We prove Theorem 1.8 in full. Through this section, we assume that D is a probability density function with support on smooth connected compact manifold with intrinsic dimension d embedded in ambient space $\mathbb{R}^s$, with smooth boundary of bounded curvature. This probability density function is further assumed to be bounded above and below, and to be Lipschitz. For simplicity, we assume that $s = d$, and we can prove all our results when $s > d$ by taking coordinate charts from the manifold into Euclidean space. We will show at the end of the section that if the distribution is supported on a convex set of full dimension in the ambient space, then the k-NN graph is Gabriel for the same k. It is not difficult to see that Gabriel graphs are 1-spanners of the edge-squared metric [**?**].

Lemma 9.1. *Let M be a compact object in $\mathbb{R}^d$, whose boundary is a smooth manifold of dimension $d-1$ with bounded curvature. Let $\mathbb{B}$ be any ball with sufficiently small radius r_B with center in M, that intersects the boundary of D at some point x. Let H be the halfspace tangent to M at x containing the center of the ball.*

For any point $Q \in M$, let Q' be the point in B closest to Q. If $d(Q',H)/r_B > c$ for arbitrary constant c, then $d(Q,H) \geq c'$ for some constant c'.

This is a basic fact about the smoothness and bounded curvature of the boundary.

Lemma 9.2. *Pick n points from D. W.h.p, any two points in $Support(D)$ with Euclidean distance $\geq \Omega(1)$ have Nearest Neighbor Metric distance of $o(1)$.*

This is implicit in [**?**].

Lemma 9.3. *For any ball $\mathbb{B}$ with center O and any point Q' on the boundary of B, let $B_{Q'O}$ be the ball with diameter $Q'O$. Let H be any halfspace containing O. If $d(Q', H)/r_B \leq c$ for some constant c possibly depending on the dimension d, then $\text{Vol}(\mathbb{B}_{Q'O} \cap H) \geq \frac{1-c'}{2^d} \text{Vol}(\mathbb{B} \cap H)$ for some constant c', where c' goes to 0 as c goes to 0.*

Proof. First, let us consider the case where $d(Q'H) = 0$, that is, Q' is contained in halfspace H'. In this case, dilating $B_{Q'O} \cap H$ by a factor of 2 about point Q' gives a superset of $B \cap H$, as $B_{Q'O}$ maps to B and H maps to a halfspace strictly containing H. In this case, $\text{Vol}(\mathbb{B}_{Q'O} \cap H) \geq \frac{1}{2^d} \text{Vol}(\mathbb{B} \cap H)$ as desired. The case when $d(Q', H)/r_B$ is bounded follows in a straightforward manner. □

This leads us to our following theorem:

Theorem 9.4. *For any n point set P picked i.i.d from D, consider any point O. Let $\mathbb{B}$ be the k-NN ball of O. Let $Q \in Support(D)$ be any point outside $\mathbb{B}$, and let the closest point to Q in $\mathbb{B}$ be Q'. For a point x inside B on the boundary of D (assuming such a point exists), let H be the tangent halfplane containing the center of $\mathbb{B}$.*

Then: either $d(Q', H)/r_B \leq c'$ for some constant c' or there exists a constant c where $|QO| > c$. Here, c and c' are independent of the number of points chosen, and c' can be set arbitrarily small.

In the latter case, w.h.p. QO is not in the edge-squared 1-spanner. In the former case, setting c' to be a very small constant ϵ lets us say:

$$\text{Vol}(\mathbb{B}_{Q'O} \cap H) \geq \frac{1-\epsilon}{2^d} \text{Vol}(\mathbb{B} \cap H), \tag{8}$$

or equivalently:

$$\mathop{\mathbb{P}}_{x \sim D}[x \in \mathbb{B}_{QO} | x \in \mathbb{B}] \tag{9}$$

$$\geq \mathop{\mathbb{P}}_{x \sim D}\left[x \in \mathbb{B}_{Q'O} | x \in \mathbb{B}\right] \tag{10}$$

$$\geq \frac{1-\varepsilon-o(1)}{2^d} \tag{11}$$

Expression 10 > Expression 11 follows from Equation 8, and the fact that the radius of the k-NN ball goes to 0 as n gets large, and thus the probability density of sampling x from D conditioned on x being in $\mathbb{B}$ approaches the uniform density in $\mathbb{B} \cap Support(D)$. Also, $B \cap H$ approaches $B \cap Support(D)$ as the radius of B goes to 0.

Expression 9 > Expression 10 since $\mathbb{B}_{QO} \supset B_{Q'O}$. (Here, the k-NN ball B w.r.t. point O is defined as the ball centered at O with radius equal to the distance of the k^{th} nearest neighbor to O).

Note that the $k-1$ nearest neighbors of O, conditioned only on the radius of B, are distributed equivalently to $k-1$ i.i.d samples of D conditioned on containment in $\mathbb{B}$. It follows that for any point Q outside B and in the support of D, where $|QO| < c$:

$$\mathop{\mathbb{P}}_{P \sim D^k}[QO \text{ is not Gabriel w.r.t. } P | Q \notin B] \geq 1 - \left(1 - \frac{1-\varepsilon-o(1)}{2^d}\right)^k \tag{12}$$

Thus, setting $\epsilon = 0.1$ and $k > O(\log n/2^d)$, and factoring in the case where $|QO| > c$, then w.h.p.:

$$\mathbb{P}_{P \sim D^k}[QO \text{ is not critical w.r.t. } P | Q \notin B]$$

Here, we recall that an edge AB is Gabriel with respect to a point set P if and only if $\mathbb{B}_{AB}$ does not contain any points in P. Note that every non-Gabriel edge is non-critical, where a critical edge is an edge that must be in the 1-spanner (as in Definition 8.1). Thus taking the union bound over $Q, O \in P$ gives us that no edge outside the k-NN graph is critical w.h.p, and thus the k-NN graph contains all critical edges and is a 1-spanner w.h.p.

This proves Theorem 1.8 when the support of D has the same intrinsic dimension as the ambient space. If the support of D has dimension $d < d'$ (where d' is the ambient dimension of the space), simply take coordinate charts from D onto $\mathbb{R}^d$ and the previous arguments will still carry through . We should note that if no point x inside B on the boundary of D exists, then we can ignore H and all the steps of the proof still follow.