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 8, 2019

Abstract

A foundational hypothesis in mcahine learning is that datapoints can be embedded into Euclidean space, and metrics can be generated on the datapoints to solve a wide range of tasks. It is widely believed that these metrics should have the property that two points in the same dense cluster of datapoints should be considered close, even if their Euclidean distance is far. One simple metric with this property is the Nearest neighbor metric. This metric is a mainfold-based metric, and it and its close variants have been studied in the past by multiple researchers.

One key problem on manifold metrics, dating back for four centuries, is computing them exactly. The Nearest Neighbor metric is defined as the infimum cost path over an uncountable number of paths that can go 'anywhere' on a continuous manifold. This makes exactly computing the Nearest Neighbor metric challenging, even for a fixed set of four points in two dimensions. In this paper, we overcome this challenge by equating the Nearest Neighbor metric to a shortest-path distance on a simple geometric graph, in all cases. Remarkably, this equality holds even if the point set is the countable union of compact geometric objects, which are not necessarily convex or even simply connected. We then compute a generalization of this metric, which we call the q-power Nearest Neighbor metric, and prove an analagous equality for point sets that are the union of 4 compact, path-connected geometric objects in arbitrary dimension.**Don: Something about Conformal change of metrics. If you have a manifold and want to put a new Riemannian metric on it, usually the Riemannian metric is a full-on metric tensor, but here we have a simpler case where it's isotropic, same in every direction. In some sense, these transformations are conformal, so this is a conformal change of Riemannian metric. (This is what this general idea we're doing is on. There's a bunch of literature on this, but almost nobody computes it.)**

Our proof uses conservative vector fields, Lipschitz extensions, minimum cost flows, and barycentric subdivisions, all applied to a geometric object we call the q-screw simplex. This work considerably strengthens the work of Cohen et. al., and shows the first non-trivial manifold metric that can be computed exactly with discrete techniques. When the point set is finite, we can use our results to solve a range of classical metric problems for our metric: we can efficiently compute sparse spanners, compute persistent homology, measure the behavior of the metric when the point set is a large number of points drawn from an underlying probability density function, and show links between this metrics and classic goemetric objects like Euclidean MST or single-linkage clustering methods.

1 Introduction

A foundational hypothesis in non-linear dimension reduction and machine learning is that data can be represented as points in Euclidean space, and appropriate metrics on these points can be generated to solve a variety of problems including clustering, classification, regression, surface reconstruction, topological property inference, and more. In machine learning, two data points should intuitively be considered close if they are in the same data cluster, even if their Euclidean distance is far. This property is called the **density-sensitive** property.**Gary: Use data-sensitive instead of density-sensitive**

Density-sensitive metrics are considered fundamental in the study of 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 density-sensitive metrics is an active area of research in machine learning. We consider a simple density-sensitive metric with an underlying manifold structure. This metric is called the Nearest Neighbor Metric, and it 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 problem for any manifold-based metric.

To define the nearest neighbor metric, we first define the notion of a density-based distance. This is a slight variation of the original definition from [].

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 mapping $\gamma : [0,1] \to \mathbb{R}^k$. 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. We then define the **density-based distance** between two points $a, b \in \mathbb{R}^k$ as*

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

Conceptually, the density-based cost of a path is the weighted path length, where each infinitesimal path piece is weighted with cost function c. 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 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 $P \subset \mathbb{R}^k$, there is a real-valued function $\mathbf{r}_P : \mathbb{R}^k \to \mathbb{R}$ defined as $\mathbf{r}_P(z) = 2\min_{x \in p} \|x - z\|$. The **Nearest Neighbor Cost** of a path γ is $\ell_{\mathbf{r}_P}$, which we will shorthand to ℓ_N. The **Nearest Neighbor Metric** between two points is defined as $\mathbf{d}_{\mathbf{r}_P}$, which we short-hand as $\mathbf{d}_N$.*

The factor of 2 in $\mathbf{r}_P(z)$ is a normalizing constant.

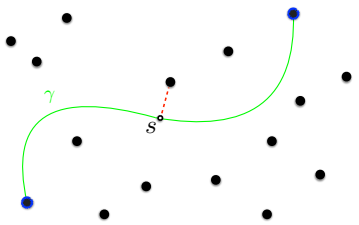

Figure 1: In this figure we have a collection of points. The length or cost of the green curve between the two blue points is the integral along the curve scaled by the distance to the nearest point.

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 collectoins of objects can be extremely compliacted, 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!**

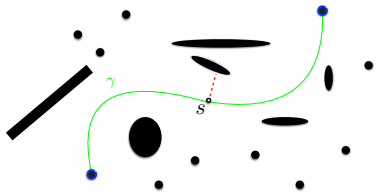

Figure 2: In this figure we have a collection of compact bodies in black. The length or cost of the green curve between the two blue points is the integral along the curve scaled by the distance to the nearest body. A curve may traverse a body at no cost. Theorem 1.4 establishes that the shortest path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

2 Our Results

Our primary results are Theorems 1.4 and 1.5 on the computing the Nearest Neighbor and q-Nearest Neighbor metrics respectively, which rely on fairly intricate and novel mathematical proof techniques. Both of these use a minimum cost flow generated from a conservative vector field to keep track of the q-NN path length, and rely on the geometry of a special simplex called the q-screw simplex. Our proof techniques, theoretically, can prove approximations between metrics as well as exact equalities, and may be uesful for other more general metrics based on data points. This is the first result we know of whose proof combines conservative vector fields, simplex geometry, Lipschitz extensions in geometry, and minimum cost flows on a graph. It is also the first result we know of that eschews calculus of variations for exact manifold metric computation.

These results form the core mathemtaical contribution of our paper.

2.1 Other Contributions

Besides for these contributions, we also tackle a range of problems significant for any metric on a data set. Some of the most important problems on any metric are: how to $(1+\varepsilon)$ approximate them efficiently using sparse data structures, and how the metric behaves in the limit as the point set is a large number of points drawn from a probability distribution. The former is important to compute the metrics in practice. The latter is important since it is highly desirable that this limit converge to a metric that has desirable properties, or else clustering with such a metric may generate non-intelligible results for large datasets. Besides for these, persistent homology of metrics space is also a central tool in topological data analysis, among other fields [], and many papers have devoted themselves to computing such homologies ofr a wide variety of metrics. Finally, it is important to relate metrics like the Nearest Neighbor metric to famous, well-established metrics like l_2, inverse min-cut distance, or the maximum edge length among a path in an MST. We will show that the Nearest Neighbor metric is in fact one of a suite of simple metrics (the q-edge power metrics) which naturally generalize both maximum-edge MST distance and Euclidean distance, the latter of which covers inverse min-cut idstances via the Gomory Hu tree. This helps put our metrics in

a broader and more interpretable context, and ensures that clustering with these q-power metrics is a generalization of more traditional clustering techniques including k-means, k-centers, and level set methods.

In this paper, we present novel results all of these problems. Our spanner and convergence results hold assuming constan dimension, and our persistent homology results and our relation to more famous metrics hold for n points in any dimension. The proofs of these results are mostly simpler than our proofs of Theorem 1.4, and thus should be appreciated mostly for the result statement rather than the complexity or intricacy of their proofs. They all hinge on Theorem 1.4, and are largely included to show that density-sensitive manifold metrics, like the NN metric, can be very flexible when Theorems like Theorem 1.4 hold.

Our spanner theorems are as follows:

Theorem 2.1. *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 2.2. *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.*

Theorem **??** tackles a common setting in machine learning, where points are assumed to be from a well-behaved ditsribution. These assumption is foundational to the field of machine learning. Although the restrictions on the distribution seem fairly limiting (and naively, do not even cover the case of a simple Gaussian), they turn out to be far more flexible than they seem, and are common in the machine learning literature. They can be modified to gain information on most relevant distributions (for example, they cover the case of a Gaussian where the thin tail is removed, which turns out to contain most of the information of a Gaussian). Theorem 2.1 generalizes the results of Cohen et. al. in [**?**], who showed nearly linear size spanners in nearly linear time, but with significantly worse ε dependence. Note that these spanners can be computed even faster than the optimal known Euclidean spanner, indicating that density-sensitive metrics like the Nearest Neighbor metric may have interesting algorithmic possibilities that standard distances like l_2 and l_1 don't have.

Results on the geodesics, etc. will be presented in their appropriate section.

2.2 Definitions and Preliminaries

q-power Nearest Neighbor Metric (q-NN metric):
q-edge power metric:
q-screw simplex:
Min-Cost Flow: Given a graph G with capacities and costs on each edge, the min-cost flow satisfying a set of demands is the minimal cost flow subject to the capacities that satisfies these demands. In this paper, we only deal with uncapacitated min-cost flow.
Unit Flow: A unit flow on a graph from a to b is a flow from a to b that carries a single unit of flow.

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 2.2.

Tim: Cut edge-squared and nearest neighbor here? **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 2.3. *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 geodesic distance: Another metric on the points of P is called the nearest-neighbor geodesic distance, 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 2.4. *The nearest-neighbor geodesic 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 geodesic distance is never greater than the edge-squared distance, as proven in the following lemma.

Lemma 2.5. *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

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

3 Outline

Section 4 contains the proof of Theorem 1.4 and 1.5. This section contains most of the novel mathematical ideas in our work, including how to relate a manifold distance to min-cost flows on a simplex. It also establishes the centrality of the q-screw simplex in our proof techniques.

Section 6 shows how q-screw simplices relate to fractional Laplacians and spectral graph theory, which allows us to prove Theorem **??**. It also contains a new embedding of the q-screw simplex for $q > 2$, proving Theorem **??**. These proofs and results are mostly included for independent interest, and to spur future work relating q-screw simplices to density-sensitive distances.

We show a brief proof in Section **??** that the Nearest Neighbor metric, and the q-edge power metrics in general, generalize common metrics like maximum-edge Euclidean MST metrics and Euclidean distance. This will show how clustering algorithms using q-edge poewr metrics generalize k-means, level-set methods, single linkage clustering, and more.

Persistent homology of the Nearest Neighbor metric is contained in Sectoin 8.1.

Section 8 contains overviews of the proof for both Theorem 2.1 and 2.2, as well as their implications. Full proofs are contained in the Appendix. These results are not particularly tricky to prove, but they nonetheless solve important questions of fast $(1+\varepsilon)$ approximations of the Nearest Neighbor metric.

Section **??** shows the convergence of the Nearest Neighbor metic, and the q-edge power metrics in general, to a beautiful geodesic on an underlying probability distribution previously studied by Hwang et. al. Most of the heavy lifting here was done by past researchers, such as Hwang et. al. and Steele **??**. However, this result lets us interpret Nearest Neighbor metric clustering as an approximation of a very nice clustering on an underlying probability density function.

Section **??** contains conclusions, a summary of work, and open questions for the future.

4 Exactly Computing Nearest Neighbor Metrics for all point sets, and q-NN metrics on small point sets

In this section, we prove our main theorems, Theorem 1.4 and 1.5. We prove that the edge-squared metric exactly equals the nearest neighbor metric on any point set, and that the q-edge power

metric equals the q-NN metric for any set of four points or less. We also conjecture that the q-edge power metric always equals the q-NN metric, and provide a discrete inequality that would imply our larger conjecture. Even though the number of points we handle for the general q-edge power metric is quite small, it still solves a fairly difficult problem: exactly computing the NN or q-NN metric even for four points in two dimensions requires dealing with uncountably number of paths through space.

Our tools for proving both theorems will be use of min-cost flows generated from a conservative vector field. This is the core idea that lets us surmount the difficulties in dealing with uncountably many paths. We hope that our ideas may be more generally applicable to various metrics.

Notice that the Nearest Neighbor cost is upper bounded by the edge-squared metric. This can be done by purely considering Nearest Neighbor paths that are piecewise linear and go straight from data point to data point. Similarly, q-NN metrics are upper bounded by q-edge power metrics for all $q > 1$. Therefore, we only need to prove that the Nearest Neighbor cost is lower bounded by the edge-squared metric, and likewise for q-NN metrics. To do this, we build a graph G' from our point set, which can conceptually be thought of as the edge-squared graph with additional Steiner points. We will show using conservative vector fields and flows that the Nearest Neighbor cost of any path from a to b is bounded below by the shortest path from a to b in G'. If the shortest path in G' were always equal to the shortest path in the edge-squraed graph G, then we'd be done.

However, this is not the case in general. However, it does turn out to be the case always on a 2-screw simplex. Remarkably, we show that proving this equality on the 2 screw simplex is sufficient to prove it for any point set, including point sets with uncountably many points. We show this reduction via the Lipschitz extension theorem and a simple BFS. This proves Theorem 1.4

All our techniques generalize to q-NN metrics (when being related to q-edge power metrics), except the equality between shortest paths on G' and G is less clear for q-screw simplices. We prove that this equality holds when there are four points in three dimension, and conjecture (with computational evidence, but no proof) that this equality holds for any point set. Doing this proves Theorem 1.5, and provides a discrete criterion that would imply Conjecture 4.11 holds.

The gap in Theorem 1.4 and 1.5 is due to the simple geometric structure of the 2-screw simplex, which is known to have a

4.1 Proof Overview

Given points $x_0, x_1, \ldots x_n$, we build a graph G' with vertices v_S for any subset S of $\{0, 1, 2, 3, 4 \ldots n-1\}$. Here, v_S is a vertex representing the circumcenter of the vertex set $\{x_s : s \in S\}$. Graph G' then has some edges, connected in a fashion to be detailed later, and costs on these edges. Graph G' is related to the barycentric subdivision on a graph, used in [].

Tto lower bound the nearest neighbor cost of any path on our original point set, we will lower bound it by the cost of some unit flow (DFEINE) from v_0 to v_n in G'. However, note that the cost of any uncapacitated unit flow from v_0 to v_n is lower bounded by the shortest path (where lengths of edges are their costs) from v_0 to v_n in G'. If the shortest path in G' is equal to the q-edge power cost of going from x_0 to x_n, then we have completed our proof.

To lower bound the cost of the q-NN metric with a unit flow, we aim to build a potential function $\mathcal{C} : \mathbb{R}^n \to \mathbb{R}^{2^n}$ such that $\mathcal{C}(p)$ is a vector whose entries sum to one, supported on the vertices of G'. Here, $\mathcal{C}(p)$ represents a convex combination of the vertices of G'. We construct $\mathcal{C}$ such that $\mathcal{C}(v_S) = e_S$, where e_S represents the unit vector with 1 in the coordinate indexed by S, and 0 elsewhere. Additionally, we would like $\mathcal{C}$ to have the property that the q-NN cost of any

path from $x \in \mathbb{R}^n$ to $x + \Delta(x) \in \mathbb{R}^n$ for small $\Delta(x)$ is bounded below by the min-cost flow on G satisfying demands $\mathcal{C}(x + \Delta(x)) - \mathcal{C}(x)$. If we can construct such a function $\mathcal{C}$, then the q-NN path from v_0 to v_n is thus lower bounded by the min-cost unit flow from v_0 to v_n, which is exactly the shortest path from v_0 to v_n in G'! Therefore, building $\mathcal{C}$ implies that we can lower bound the the q-NN cost of a path from x_0 to x_n with the shortest path in G'.

This bound holds for any initial point set $x_0, x_1, \ldots x_n$. However, in general the shortest path on G' and the shortest path on G are not the same. To overcome this, we show that it suffices to prove Theorem 1.4 on 2-screw simplices, and we show that on 2-screw simplices the shortest path on G' is the shortest path on G. We further show that it suffices to prove Theorem 1.5 on q-screw simplices with 4 points.

Therefore, our proof of Theorem 1.4 is devoted to three things: proving the reduction to the 2-screw simplex, showing the inequality between Nearest Neighbor metrics and shortest paths on G', and proving that the shortest path between points on G' is the same as the shortest path between points on G when G is the q-edge power graph of the q-screw simplex.

Our proof of Theorem 1.5 follows the exact same structure, where the main thing that prevents us from proving Conjecture 4.11 is the fact that we aren't yet able to prove that shortest paths on G' and G are the same for n points in general, and we only do so for sets of 4 points. The rest of the proof still works.

4.2 Reduction to q-screw simplex

In this section, we prove that it suffices to prove Theorem 1.4 and 1.5 on the q-screw simplex. We do so by doing it when $q = 2$, and noting that the proof generalizes naturally for any $q > 2$.

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

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

This will imply that $p_1 = s$. It will suffice to show that for all $p_i \in P$, we have $\mathbf{d}_q(s, p_i) = \mathbf{d}_{qN}(s, p_i)$. The core step is that we will find a Lipschitz map $m : \mathbb{R} \to \mathbb{R}^n$ that preserves $\mathbf{d}_q(s, p)$ for all $p \in P$. We will then show how the Lipschitz extension of m is also Lipschitz as a function between q-NN metrics. If we can do this, then the q-NN cost of any path from s to p_i on our initial point set is lower bounded by the q-NN cost of the image of the path under m, from $m(s)$ to $m(p_i)$ (since the map is Lipschitz as a function between q-NN metrics). If the Theorem holds on $m(p_1), m(p_2), \ldots m(p_{n-1})$, then this q-NN cost is lower bounded by the shortest path from p_0 to p_{n-1}, as desired.

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

Define a mapping $m : P \to \mathbb{R}^n$ so that $m(p_1) = 0$ and otherwise

$$m(p_i) = m(p_{i-1}) + (\mathbf{d}_q(s, p_i) - \mathbf{d}_q(s, p_{i-1}))^{1/q} \, e_i, \tag{1}$$

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

Lemma 4.1. *For all $p_i, p_j \in P$, we have*

(i) $\|m(p_j) - m(p_i)\| = \sqrt{|\mathbf{d}_q(s, p_j) - \mathbf{d}_q(s, p_i)|}$, *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 $i \leq j$.

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

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

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

□

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

Proposition 4.2. *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* (1)*. 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. $\mathbf{d}'(m(s), m(p)) = \mathbf{d}_q(s,p)$ 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.

$$\begin{aligned}
\|m(a) - m(b)\| &= \sqrt{|\mathbf{d}_q(s,a) - \mathbf{d}_q(s,b)|} && \text{[Lemma 4.1(i)]} \\
&\leq \sqrt{\mathbf{d}_q(a,b)} && \text{[triangle inequality]} \\
&\leq \|a - b\| && [\mathbf{d}_q(a,b) \leq \|a-b\|^q \text{ by the definition of } \mathbf{d}]
\end{aligned}$$

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 4.1(ii) implies that removing q_1 gives a shorter sequence. Thus, we may assume $k = 1$ and therefore, by Lemma 4.1(i),

$$\mathbf{d}'(m(s), m(p)) = \|m(s) - m(p)\|^2 = \mathbf{d}_q(s,p).$$ □

4.2.2 The Lipschitz Extension

Proposition 4.2 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 4.3. *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 geodesic.*

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

$$\ell'(f \circ \gamma) = \int_0^1 \mathbf{r}_{f(P)}(f(\gamma(t)))\|(f \circ \gamma)'(t)\|dt \leq \int_0^1 \mathbf{r}_P(\gamma(t))\|\gamma'(t)\|dt = \ell(\gamma), \tag{2}$$

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

$$\begin{aligned}
\mathbf{d}_N(a,b) &= 4 \inf_{\gamma \in \text{path}(a,b)} \ell(\gamma) && \text{[by definition]} \\
&\geq 4 \inf_{\gamma \in \text{path}(a,b)} \ell'(f \circ \gamma) && \text{[by (2)]} \\
&\geq 4 \inf_{\gamma' \in \text{path}(f(a),f(b))} \ell'(\gamma') && [\text{because } f \circ \gamma \in \text{path}(f(a), f(b))] \\
&= \mathbf{d}_N(f(a), f(b)). && \text{[by definition]}
\end{aligned}$$

□

Theorem 4.4. *For any point set $P \subset \mathbb{R}^d$, the edge squared metric $\mathbf{d}$ and the nearest neighbor geodesic $\mathbf{d}_N$ are identical.*

Proof. Fix any pair of points s and p in P. Define the Lipschitz mapping m and its extension f as in (1). Let $\mathbf{d}'$ and $\mathbf{d}'_N$ denote the edge-squared and nearest neighbor geodesics on $f(P)$ in $\mathbb{R}^n$.

$$\begin{aligned}
\mathbf{d}_2(s,p) &= \mathbf{d}'(m(s), m(p)) && \text{[Proposition 4.2(iii)]} \\
&= \mathbf{d}'_N(m(s), m(p)) && [f(P) \text{ are vertices of a box}] \\
&\leq \mathbf{d}_N(s,p) && \text{[Lemma 4.3]}
\end{aligned}$$

We have just shown that $\mathbf{d} \leq \mathbf{d}_N$ and Lemma 2.5 states that $\mathbf{d} \geq \mathbf{d}_N$, so we conclude that $\mathbf{d} = \mathbf{d}_N$ as desired. □

4.3 Construction of G'

Let $x_0, x_1, \dots x_{n-1}$ be our point set. Let G' be a graph on $\mathbb{R}^{2^n}$, with vertices v_S for all $S \subset \{0, 1, \dots n-1\}$. Here, $v_{\{i\}}$ corresponds to x_i. Connect vertices v_S and v_T for $S, T \subset \{0, 1, \dots n-1\}$ iff sets S and T differ by exactly one element. We assign this edge a cost of $|R_S^q - R_T^q|$, where R_S is the circumradius of points $\{x_s : s \in S\}$.

It is not difficult to see that the distance to travel from $v_{\{i\}}$ to $v_{\{i,j\}}$ to $v_{\{j\}}$ is the q-edge poewr distance from x_i to x_j

Lemma 4.5. *There exists a function $\mathcal{C}$ such that for any path piece ϕ running from x and $x+\Delta(x)$, $\ell_{qN}(\phi) \geq MCF_{G'}(B'(x+\Delta(x)) - B'(x))$. Here, $MCF_{G'}(d)$ for $d \in \mathbb{R}^{2^n}$ represents the minimum cost flow on G' satisfying demands d on the vertices of G'.*

4.4 Construction of $\mathcal{C}$

In this section, we prove we Lemma 4.5. We construct a function $\mathcal{C} : \mathbb{R}^n \to \mathbb{R}^{2^n}$ assigning every point in Euclidean space to a vector, representing a convex combination of vertices in G'. We will build $\mathcal{C}$ separately for each Voronoi cell, and show it is piecewise continuous across boundaries of Voronoi cells. However, to simplify our arguments, we further divide up each Voronoi cell into simplices similar to the dissection in a barycentric subdivision. However, rather than barycenters, we use circumcenters, so we are more-precisely dividing up the q-screw simplex with a circumcentric subdivision. **Tim: define circumcentric/barycentric subdivision**

Let p_S be the circumcenter of points $\{p_s | s \in S\}$ when $S \subset \{1, 2, \dots n\}$. A simplicial cell in the circumcentric subdivision is defined by a permutation $a_0, a_1, \dots a_{n-1}$ of $\{0, 1, \dots n-1\}$ as follows: let $A_i = \{a_0, a_1, \dots a_i\}$. Then the vertices $\{p_{A_i} | 0 \leq i < n\}$ are the vertices of the simplicial cell. **Tim: Move this definition upwards somewhere?** Each Voronoi cell is the disjoint **Tim: not exactly disjoint** union of some of these cells, and the cells partition the simplex **Tim: This isn't quite true: its only true for special geometry cells, but we don't particularly care... how can I make this point clear?**. Our general strategy is to create $\mathcal{C}$ on each simplicial cell, and show that is piecewise continuous across the boundaries of simplicial cells.

To simplify our notation, we define $\overline{p}_i$ to be p_{A_i}. **Tim: Make sure all points are p, not x. $\overline{x}_k$ is later used to coordinate-wise split up x.**

By the construction of circumcentric subdivisions, the line $\overline{p}_i\overline{p}_{i+1}$ is perpendicular to $\overline{p}_{i+1}\overline{p}_{i+2}$ for all i. These lines define a natural orthonormal coordinate axis. Thus, for any $\overline{x}$ in the convex hull of $\overline{p}_i$, we can write $\overline{x}$ in coordinates $(\overline{x}_0, \overline{x}_1, \dots \overline{x}_{n-2})$, where the i^{th} coordinate axis is parallel to $\overline{p}_i\overline{p}_{i+1}$.

Next, we introduce how $\overline{\mathcal{C}}$ is defined on the vertices $\overline{p}_i$. Here, $\overline{\mathcal{C}}_i$ is shorthand for the value of $\overline{\mathcal{C}}$ on $\overline{p}_i$.

$$\overline{\mathcal{C}}(\overline{x})_i = \frac{\left(\sum_{s=0}^{i-1} \overline{x}_s{}^2\right)^{q/2} - \left(\sum_{s=0}^{i-2} \overline{x}_s{}^2\right)^{q/2}}{\overline{R}_i^q - \overline{R}_{i-1}^q} - \sum_{i<j<n} \overline{\mathcal{C}}(\overline{x})_j \tag{3}$$

for all $1 \leq i \leq n-1$, and **Tim: make sure you replace all k's with n's, and claim that you're dealing with simplices of full dimension. You may want to say that the proof is counterintuitive since simplices of full dimension are usually considered harder?**

$$\overline{\mathcal{C}}(\overline{x})_0 = 1 - \sum_{0<j<n} \overline{\mathcal{C}}(\overline{x})_j$$

The key feature about $\overline{\mathcal{C}}$ is that

$$\sum_{j=i}^{n} \overline{\mathcal{C}}(\overline{x})_j = \frac{\left(\sum_{s=0}^{i-1} \overline{x}_s{}^2\right)^{q/2} - \left(\sum_{s=0}^{i-2} \overline{x}_s{}^2\right)^{q/2}}{\overline{R}_i^q - \overline{R}_{i-1}^q}$$

for all $i > 0$, and for $i = 0$ the LHS evaluates to 1.

Since we defined this function piecewise, we need to check that this function is piecewise continuous.

Lemma 4.6. *If $\overline{x}$ is on a face of $\overline{p}_0, \ldots \overline{p}_n$, then $\overline{\mathcal{C}}(\overline{x})$ has non-zero coordinates only on that face. Furthermore, the coordinates depend only on SOMETHING.*

Proof. □

Now we are ready to prove our core lemma:

Lemma 4.7. *For x and $x + \Delta(x)$ in the convex hull of $\overline{p}_1, \ldots \overline{p}_k$, the distance:*

$$\|x - \overline{p}_1\|^{q-1} \cdot \|\Delta(x)\| \geq F\left(\overline{\mathcal{C}}(x + \Delta(x)) - \overline{\mathcal{C}}(x)\right)$$

where $F(d)$ is the unique cost of a flow satisfying demand $d \in \mathbb{R}^n$ on $\overline{G}$.

Here, the left hand side represents the q-NN cost of a path piece from x to $x + \Delta(x)$, and the right hand side is the unique cost of the induced flow on graph G', with the restriction that the flow is only nonzero on the vertices $\overline{p}_i\overline{p}_{i+1}$ for any $0 \leq i < n$. This flow is unique since we forced our flow to be non-zero only on the edges $\overline{p}_i\overline{p}_{i+1}$, which form a line graph; and for any set of demands on vertices of a line, there is a unique flow satisfying those demands.

Proof. For any edge $\overline{p}_i\overline{p}_{i+1}$, the cost of a flow (satisfying some set of demands whose sum is 0) on that edge is the absolute value of the sum of the demands on vertices $\overline{p}_{i+1}\overline{p}_{i+2}, \ldots \overline{p}_n$, multiplied by the cost of the edge from $\overline{p}_i$ to $\overline{p}_{i+1}$. This quantity comes out to be:

$$\left(\overline{R}_{i+1}^q - \overline{R}_i^q\right) \sum_{j=i+1}^{n} \overline{\mathcal{C}}(\overline{x})_j \tag{4}$$

$$= \left(\sum_{s=0}^{i-1} \overline{x}_s{}^2\right)^{q/2} - \left(\sum_{s=0}^{i-2} \overline{x}_s{}^2\right)^{q/2}. \tag{5}$$

As $\Delta(x)$ goes to 0, the change in Expression 5 is

$$
\begin{aligned}
&\left(q\overline{x}_0\left(\sum_{s=0}^{i-1}\overline{x}_s^2\right)^{q/2-1} - q\overline{x}_0\left(\sum_{s=0}^{i-2}\overline{x}_s^2\right)^{q/2-1}\right)\Delta(\overline{x})_0 \\
&+\left(q\overline{x}_1\left(\sum_{s=0}^{i-1}\overline{x}_s^2\right)^{q/2-1} - q\overline{x}_1\left(\sum_{s=0}^{i-2}\overline{x}_s^2\right)^{q/2-1}\right)\Delta(\overline{x})_1 \\
&+\dots \\
&+\left(q\overline{x}_{i-2}\left(\sum_{s=0}^{i-1}\overline{x}_s^2\right)^{q/2-1} - q\overline{x}_{i-2}\left(\sum_{s=0}^{i-2}\overline{x}_s^2\right)^{q/2-1}\right)\Delta(\overline{x})_{i-2} \\
&+\left(q\overline{x}_{i-1}\left(\sum_{s=0}^{i-1}\overline{x}_s^2\right)^{q/2-1}\right)\Delta(\overline{x})_{i-1},
\end{aligned}
\tag{6}
$$

Since

$$
q\overline{x}_j\left(\sum_{s=0}^{i-1}\overline{x}_s^2\right)^{q/2-1} - q\overline{x}_j\left(\sum_{s=0}^{i-2}\overline{x}_s^2\right)^{q/2-1}
$$

is always non-negative (only when $q \leq 2$), we get that the absolute value of Expression 6 is bounded above by:

$$
\begin{aligned}
&\left(q\overline{x}_0\left(\sum_{s=0}^{i-1}\overline{x}_s^2\right)^{q/2-1} - q\overline{x}_0\left(\sum_{s=0}^{i-2}\overline{x}_s^2\right)^{q/2-1}\right)|\Delta(\overline{x})_0| \\
&+\left(q\overline{x}_1\left(\sum_{s=0}^{i-1}\overline{x}_s^2\right)^{q/2-1} - q\overline{x}_1\left(\sum_{s=0}^{i-2}\overline{x}_s^2\right)^{q/2-1}\right)|\Delta(\overline{x})_1| \\
&+\dots \\
&+\left(q\overline{x}_{i-2}\left(\sum_{s=0}^{i-1}\overline{x}_s^2\right)^{q/2-1} - q\overline{x}_{i-2}\left(\sum_{s=0}^{i-2}\overline{x}_s^2\right)^{q/2-1}\right)|\Delta(\overline{x})_{i-2}| \\
&+\left(q\overline{x}_{i-1}\left(\sum_{s=0}^{i-1}\overline{x}_s^2\right)^{q/2-1}\right)|\Delta(\overline{x})_{i-1}|,
\end{aligned}
\tag{7}
$$

Expression 7 is an upper bound on the cost of a flow along edge $\overline{v}_i\overline{v}_{i+1}$**Tim: Is this the right indexing for edges?** induced by a path from x to $x + \Delta(x)$. Now we sum this across all i to get an overall cost upper bound, and group by $\Delta(\overline{x})_i$ for fixed i. The sum telescopes beautifully, and

we get:

$$\left(q\overline{x}_0\left(\sum_{s=0}^{n-2}\overline{x}_s^2\right)^{q/2-1}\right)|\Delta(\overline{x})_0| \tag{8}$$

$$+\left(q\overline{x}_1\left(\sum_{s=0}^{n-2}\overline{x}_s^2\right)^{q/2-1}\right)|\Delta(\overline{x})_1|\dots \tag{9}$$

$$+\left(q\overline{x}_{n-1}\left(\sum_{s=0}^{n-2}\overline{x}_s^2\right)^{q/2-1}\right)|\Delta(\overline{x})_{n-2}| \tag{10}$$

This expression, by Cauchy Schwarz, is upper bounded by

$$\sqrt{\sum_{s=0}^{n-2}\Delta(\overline{x})_s^2}\cdot\left(q\sqrt{\sum_{s=0}^{n-2}\overline{x}_s^{q-1}}\right)$$

Which is exactly the q-NN distance.

□

Note that this function is piecewise continuous on the boundary. Therefore, we have shown that the q-NN cost of any path piece is less than the min-cost flow on G satisfying $\mathcal{C}(x+\Delta(x))-\mathcal{C}(x)$ for infinitesimal $\Delta(x)$, as desired.

So far, the only property our flow construction used is that the points $x_0, x_1, \dots x_n$ have Voronoi subdivisions defined by $\overline{p}_{a_0}$, $\overline{p}_{a_0a_1}$, $\dots\overline{p}_{a_0a_1\dots a_k}$, for some $a_0, ldotsa_k \subset \{0,1,\dots n\}$. (DOES THIS WORK FOR ANY GEOMETRY, OR DO I NEED THE INTERIOR CIRCUMCENTER PROPERTY?).

Thus, we have proven a core lemma:

Lemma 4.8. *The q-NN distance between two points in a point set is lower bounded by the shortest path between the two corresponding points in G. Here, G is constructed as in Definition* **??**

We now prove the following two lemmas, completing our proof of Theorem 1.4 and Theorem 1.5 respectively.

Lemma 4.9. *Let G be the edge-squared graph (DEFINE), and let G be defined as in Definition* **??** *for $q = 2$. The shortest path in G is the same as the shortest path in G, when the initial point set generating G and G is a 2-screw simplex.*

Lemma 4.10. *Let $q > 2$. Let G be the q-edge power graph, and G be defined as in Definition* **??** *(MAKE SURE THE DEFINITION IS Q DEPENDENT). The shortest path in G is the same as the shortest path in G, when the initial point set generating G and G is a q-screw simplex with 4 points.*

Combined with Theorem **??**, Lemmas **??** and **??** prove Theorems 1.4 and 1.5 respectively. Moreover, we make the following conjecture, which we have some computational evidence for (See Appendix **??** for details):

Conjecture 4.11. *For $q > 2$, let G and G be defined as in Lemma 4.10. Then the shortest path in G is the same as the shortest path in G.*

If this were true, it would prove that the q-edge power metric and the q-NN metric were equal for all $q > 2$.

5 Proof of Theorem 1.4 and 1.5

5.1 Counterexamples to Theorem 1.4 when $q < 2$

Consider three points A, B, C with distances $AB = 1, BC = 1, AC = 2^{1/q}$. The q-edge power metric from A to C is clearly 2, but the q-NN cost from A to C is less than 2: this can be seen since the Voronoi cell containing B crosses line AC, and thus the q-NN cost of going from A to C in point set $\{A, B, C\}$ is strictly less than the q-NN cost of going from A to C in pointset $\{A, C\}$, the latter of which is 2.

6 Isometries from the q-Screw Simplex

7 Fractional Laplacian

In this section, we prove that the q-screw simplex distances arise as effective resistance of the Fractional Laplacian, for powers $s = -1/2 - 1/q$, when $q > 2$.

Preliminaries: Fractional Laplacian.

We present two definitions: the first definition is based on taking the limit of graph Laplacians raised to the fractional power (which are known in folklore to be graph Laplacians themselves), and the second definition is based on taking fractional powers of the Laplacian differential operator when the latter is written in terms of its Eigenvectors, the Fourier bases.

In this work, we build on Von Neumann and Schoenbergs proof of embeddability of the q-screw simplex. Their work (slightly simplified by the authors of this paper) shows that the q-screw simplex for $q > 1$ can be embedded in infinite dimensional Hilbert space, by the embedding $f : mathbbR \to L_2$ defined as:

$$f(x) = \frac{e^{i\omega x}}{\omega^{1/2+1/q}} \tag{11}$$

Where $f(x)$ is a function in the variable ω.

The proof of their embedding hinges on the following remarkable integral formula:

$$\|x_1 x_2\|_2^{1/q} = \frac{sin^2(\omega(x_1 - x_2))}{\omega^{1+2/q}},$$

the left hand side of which is the norm for Equation 11. This integral formula is a classical integral formula [], and can be proven using Jordans integration theorem from complex analysis [**?**, **?**].

Notice that the step function S_x, which is 1 between $-x$ and x and 0 elsewhere, can be written in Fourier bases as:

$$S(x) = frace^{i\omega x}\omega. \tag{12}$$

Equation 12 is a classic result, dating back to the earliest days of functional analysis. However, this formulation compared with Equation 11 practically invites us to use the Fractional Laplacian, when viewed through the Eigenvector lens in Section **??**.

Therefore, we can write Equation 11 as:

$$f(x) = \Delta^{1/4-1/(2q))} Step = \Delta^{1/4-1/(2q)}(\Delta^{-1/2}\delta_{-x} - \delta_x) = \Delta^{-1/4-1/(2q)}(\delta_{-x} - \delta_x)$$

In the above expression, δ_x represents the Dirac Delta function. DEFINE Δ in this case!!!! Here, the second part of the equation is a standard manipulation in differential equations, as $\Delta^{1/2}$ is conceptually similar to the integral operator. For more on manipulations with this fractional Laplacian operator, see **??**.

And thus $|x-y|^{2/q} = ||f(x)||_2^2$ can be written as:

$$(\delta_{-x} - \delta_x)^T \cdot \Delta^{-1/2-1/q} \cdot (-\delta_{-x}\delta_x) \tag{13}$$

Which is an effective resistance distance. This can be seen since $\Delta^{-1/2-1/q}$ can be written as the limit of fractional graph Laplacians (which are in turn graph Laplacians, by Lemma **??**). Given a finite screw simplex, our distance is thus the limit of the Schur complement of these graph Laplacians onto a finite point set, which is the limit of a sequence of graph Laplacians. It can be seen easily that such graph Laplacians must converge, and the limit of this convergence is a graph Laplacian whose effective resistance distance are the screw simplex distances. This proves Theorem **??**.

8 Spanner Results

8.1 Persistent Homology of the Nearest Neighbor Metric

Tim: This section should be prefaced somewhere – here or in a previous section – with a statement as to why the Nearest Neighbor metric may still be of independent interest. My main claim is that it's useful since it's defined on all points rather than just two.

Tim: My main issue with this section is that I don't have a clean theorem saying how to compute persistent homology, both in ambient and intrinsic setting. Having one or two top-level theorems that say this would be great, rather than having it be written in the exposition. As it stands, I have no clue how Lemma 4.5 is used, or Lemma 4.6, to compute the homologies. In this section, we show how to compute the so-called persistent homology [**?**] of the nearest neighbor distance in two different ways, one ambient and the other intrinsic. The latter relies on Theorem 4.4 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 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 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 8.1. *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 stright 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 distance afterwards. Moreover, one observes that the same equivalence will hold for variants of the nearest neighbor 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, and 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 matric balls by not be so well-behaved. In particular, the nearest neighbor 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 8.2. *Let $P \subset \mathbb{R}^d$ be finite and let $\mathbf{d}_N$ be the nearest neighbor 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 4.4. 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. □

9 Relation to MST distances, and other distances in Geometry