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

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 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 **??** and **??**. 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 **??**

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 **??**, and provides a discrete criterion that would imply Conjecture 1.8 holds.

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

1.1 Reduction to q-screw simplex

In this section, we prove that it suffices to prove Theorem **??** and **??** on the q-screw simplex.

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 γ from s to p_i on our initial point set

is lower bounded by the q-NN cost of $m(\gamma)$ with respect to the point set $\{m(p_0), \ldots, m(p_{n-1})\}$. If the Theorem holds on this point set, the q-NN cost of $m(\gamma)$ is lower bounded by the shortest path from $m(p_0)$ to $m(p_{n-1})$, which is equal to the shortest path from p_0 to p_{n-1}. This completes our reduction.

1.1.1 Single Source Distance Preserving Embedding

We seek to find points $m(p_i) \in \mathbb{R}^n$ such that

$$\mathbf{d}_q(m(s,p_i), m(s,p_{i-1})) = \mathbf{d}_q(s,p_i) - \mathbf{d}_q(s,p_{i-1}) \tag{1}$$

To find m, we perform a breadth-first search to find points on the real line $x_0 < x_1 < \ldots x_{n-1}$ such that $x_i - x_{i-1} = \mathbf{d}_q(s,p_i) - \mathbf{d}_q(s,p_{i-1})$. These points x_i can be found with a simple breadth first search on our points. Note that if we set $m(p_0), \ldots m(p_{n-1})$ as the vertices of the q-screw simplex formed from points $x_0, x_1, \ldots x_n$, then Equation 1 holds.

1.1.2 The Lipschitz Extension

Proposition **??** 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 1.1. *The function f is also 1-Lipschitz as mapping from $\mathbb{R}^d \to \mathbb{R}^n$ with both spaces endowed with the q-NN metric.*

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

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

Tim: propagate the change for q-NN metrics to agree with NN metrics when $q = 1$.
where ℓ' denotes the length with respect to $\mathbf{r}_{f(P)}$. Thus, for all $a, b \in P$,

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

□

This proves that it suffices prove Theorem **??** on all 2-screw simplices and Theorem **??** on all 4 point q-screw simplices.

1.2 From q-NN metrics to flows on a Geometric Graph G'

We give a brief overview of our approach on how to bound q-NN metrics on points in space, with flows on a geometric graph G'. 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 from v_0 to v_n in G'. Note that the min-cost 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're done.

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, as desired. 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, where G is the q-edge power graph of the initial point set. However, these shortest paths are the same on 2-screw simplices (when $q = 2$) and for four point q-screw simplices. By our results from Section **??**, this suffices to prove our relevant theorems.

Thus, the remainder of this section does the following:

1. We construct G' and $\mathcal{C}$ for fixed q from $p_0, \ldots p_{n-1}$, and use $\mathcal{C}$ to show that the q-NN metric on $p_0, \ldots, p_{n-1}$ is lower bounded by the shortest path on G'.

2. We show that shortest ptahs on G and G' are the same in the above-mentioned settings.

1.3 Construction of G'

Let $x_0, x_1, \ldots x_{n-1}$ be our point set. Let G' be a graph on $\mathbb{R}^{2^n}$, with vertices v_S for all $S \subset \{0, 1, \ldots n-1\}$. Here, $v_{\{i\}}$ corresponds to x_i. Connect vertices v_S and v_T for $S, T \subset \{0, 1, \ldots n-1\}$ 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

1.4 Construction of $\mathcal{C}$

In this section, we prove we Lemma 1.2. 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. Maybe put this in the definitions section.**

Lemma 1.2. *There exists a function $\mathcal{C}$ such that for any path piece ϕ running from x and $x+\Delta(x)$, $\ell_{qN}(\phi) \geq MCF_{G'}(\mathcal{C}(x+\Delta(x)) - \mathcal{C}(x))$. 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'.*

Let p_S be the circumcenter of points $\{p_s | s \in S\}$ when $S \subset \{0, 1, \ldots n-1\}$. A simplicial cell in the circumcentric subdivision is defined by a permutation $a_0, a_1, \ldots a_{n-1}$ of $\{0, 1, \ldots n-1\}$ as follows: let $A_i = \{a_0, a_1, \ldots a_i\}$. Then the vertices $\{p_{A_i} | 0 \leq i < n\}$ are the vertices of the simplicial cell. **Tim: Move this definition upwards somewhere?**. We can find a unique subset of these simplicial cells such that they are fully contained inside some Voronoi cell, and this subset happens to paritition the convex hull of p_S for all $S \subset \{0, 1, 2, \ldots n-1\}$. **Tim: Picture**

Our general strategy is to create $\mathcal{C}$ on each simplicial cell in this subset, and ensure that is piecewise continuous across the boundaries of simplicial cells. For this, we build a function $\mathcal{C}$ such that $\mathcal{C}(x)$ takes on non-zero values only at the vertices in G' corresponding to the vertices of the simplicial cell containing x. For now, imagine that such a simplicial cell is defined by $a_0, a_1, \ldots a_{n-1}$.

To simplify our notation, we define $\overline{p}_i$ to be p_{A_i}. Now we show how to construct $\mathcal{C}$. Let $\overline{\mathcal{C}}(x)$ be the restriction of $\mathcal{C}(x)$ to the vertices $\overline{p}_i$. **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 $(\overline{x}_0, \overline{x}_1, \ldots \overline{x}_{n-2})$, 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 $\overline{p}_i$, and $\overline{R}_i$ is the circumcradius of $\overline{p}_0, \ldots \overline{p}_i$.

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

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

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

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 across the boundary.

Lemma 1.3. *If $\overline{x}$ is on a face of $\overline{p}_0, \ldots, \overline{p}_{n-1}$, then $\overline{\mathcal{C}}(\overline{x})$ has non-zero coordinates only on that face. Furthermore, the coordinates depend only on the barycentric coordinates of x with respect to the vertices of that face.*

Proof. **Tim: I can't prove this!!!!** □

Now we are ready to prove our core lemma: **Tim: Make sure you assume the points are full dimensional somewhere in the start.**

Lemma 1.4. *For x and $x + \Delta(x)$ in the convex hull of $\overline{p}_0, \ldots \overline{p}_{n-1}$, 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 $\overline{G}$.

Proving this lemma, combined with Lemma 1.3 will prove Lemma 1.2. 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:

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

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

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

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

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:

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

Therefore, we have that 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 $x_0, x_1, \ldots x_n$ have Voronoi subdivisions defined by $\overline{p}_{a_0}$, $\overline{p}_{a_0 a_1}$, $\ldots \overline{p}_{a_0 a_1 \ldots a_k}$, for some $a_0, ldotsa_k \subset \{0, 1, \ldots n\}$. (DOES THIS WORK FOR ANY GEOMETRY, OR DO I NEED THE INTERIOR CIRCUMCENTER PROPERTY?).

Thus, we have proven a core lemma:

Lemma 1.5. *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 **??** and Theorem **??** respectively.

Lemma 1.6. *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 1.7. *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 **??** and **??** respectively. Moreover, we make the following conjecture, which we have some computational evidence for (See Appendix **??** for details):

Conjecture 1.8. *For $q > 2$, let G and G be defined as in Lemma 1.7. 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 $q > 2$.

2 Proof of Theorem ?? and ??

2.1 Counterexamples to Theorem ?? 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.

3 Isometries from the q-Screw Simplex

4 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 **??**.

5 Spanner Results

5.1 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 **??** 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 5.1. *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 5.2. *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 **??**. 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. □

6 Relation to MST distances, and other distances in Geometry