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 **??** 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 $x_0, x_1, \dots x_n$. 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 $p_0, \dots p_{n-1}$, and use $\mathcal{C}$ to show that the q-NN metric on $p_0, \dots, p_{n-1}$ 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 $x_0, x_1, \dots x_{n-1}$ 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

1.4 Construction of $\mathcal{C}$

In this section, we prove we Lemma **??**. 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