# Exact Computation of a Manifold Metric with Desirable Machine Learning Properties, via Shortest Paths in a Geometric 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 7, 2019

**Abstract**

In the machine learning setting, distances between two datapoints in a Euclidean point set are considered short if they are in the same data cluster - even if their Euclidean distance is long. A simple metric with this property is the Nearest neighbor metric, defined by finding a shortest path on a manifold constructed from the point set. This metric 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, like any manifold-based metric, is the infimum cost path over an uncountable number of paths that can go 'anywhere' on a continuous manifold. This makes computing the Nearest Neighbor metric exactly challenging, even for a fixed set of four points in two dimensions. In this paper, we overcome this difficulty by equating the Nearest Neighbor metric to a shortest-path distance on a simple geometric graph, in all cases. This result considerably strengthens the work of Cohen et al. We then show how to exactly compute a generalization of the Nearest Neighbor metric, called the $q$-Nearest Neighbor metric, for small point sets. Our tools include conservative vector fields, Lipschitz extensions, minimum cost flows, barycentric subdivisions, and matrix inversion, all applied to a geometric object we call the $q$-screw simplex.

The key geometric object in our proof, the $q$-screw simplex, was first discovered by John Von Neumann and Issai Schoenberg. This simplex is defined by taking $n$ points on a line, applying the $1/q$ power to all pairwise distances, and isometrically embedding the resulting point set into Euclidean space. In this paper, we prove that the $q$-screw simplex has a deep connection to fractional Laplacians, a differential operator that

appears in a wide variety of physics and mathematical settings. We use this connection to prove that the $q$-screw simplex embeds isometrically into Effective resistance distance, and that its volume and circumcenter can be approximated efficiently using a Laplacian determinant estimator and Laplacian system solver respectively. We also show that any finite $l_1$ metric raised to the $1/q$ power is isometrically embeddable into $l_1$, mirroring a famous theorem by Schoenberg on $l_2$.

Finally, we compute sparse spanners for the Nearest Neighbor metric and show links to existing problems ranging from persistent homology, geodesic approximation, density-based clustering, and Euclidean MST.

# 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.

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, whose close variants have been studied by multiple researchers. This metric is called the Nearest Neighbor Metric. 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 on 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 $\gamma$ relative to $c$ as:*

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

*Here, the path $\gamma$ 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*

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

*PICTURE*

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 $\mathbf{r}_P(z) = \min_{x \in p} \|x - z\|$. The **Nearest Neighbor Cost** of a path $\gamma$ is $2 \cdot \ell_{\mathbf{r}_P}$, which we will shorthand to $\ell_N$.*

*The **Nearest Neighbor Metric** between two points is defined as $\mathbf{d}_{\mathbf{r}_P}$, which we shorthand as $\mathbf{d}_N$.*

The factor of 2 in the definition of the Nearest Neighbor Cost is a normalizing constant.

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. Manifold geodesics have a long history of study in computer science, and are fundamental to some of the oldest and most successful methods of classification and clustering. Study of manifold geodesics predates computer science, and are the cornerstone of many fields of physics and mathematics. Exactly computing these 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. 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.

One of the most significant problems on any manifold geodesic is how to compute it exactly. This problem's study dates back for four centuries, and has spawned off entire fields of mathematics including the celebrated calculus of variations. Historically, mathematicians place strong emphasis on exact computation as opposed to constant factor approximations. Another problem of algorithmic importance is to $(1+\epsilon)$ approximate these metrics efficiently. The core difficulty in the first problem is that a manifold metric is the minimum cost path over uncountably many paths that can go 'anywhere' through a manifold. 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 primary 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 smooth, piecewise hyperbolic paths between two data points with different costs.

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 geodesics, calculus of variations, since the Nearest Calculus of variations *can* show that the optimal nearest neighbor path between two points must be piecewise hyperbolic. However, there can be many piecewise hyperbolic paths between two data points, and in general these paths can have very 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. This makes the Nearest Neighbor metric is the *only* known example of a density-based distance that can be computed exactly. Our approach is based on conservative vector fields, 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.

**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 the shortest path distance between two points on this graph.*

**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 is piecewise linear when the start and end points are in the dataset!

2. There are simple and natural variants of the Nearest Neighbor metric, for which no analog of Theorem 1.6 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. gives us a natural suite of metrics that smoothly converge to the Nearest Neighbor metric, for which no theorem like Theorem 1.6 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 function to build a density-based distance will necessarily break Theorem 1.6 on these points. This establishes that Theorem 1.6 is fairly tight, and won't work for any cost function meaningfully less than the Nearest Neighbor function on the right triangle point set.

4. This theorem holds for any compact point set, whether its $n$ points in $n-1$ dimensional space or a finite union of compact geometric blobs in countably infinite dimension. The geometry of both can be quite complicated, and it is generally hard to prove these types of results on arbitrary point sets (TIM:CHECK AND MAKE SURE THE PROOF WORKS)

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 1.6, and significantly improve the Nearest Neighbor spanners of Cohen et al in [].

Theorem 1.6 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 small point sets for all $q > 2$. We do this by equating it to the $q$-edge power metrics, which we will define later. We then use Theorem 1.6 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.

Our final set of theorems regards the $q$-screw simplex, the core geometric object in our proof of Theorem 1.6 and its generalizations. The $q$-screw simplex was first discovered by John Von Neumann and Issai Schoenberg. It is defined by taking $n$ points anywhere on a line, taking the $1/q$ power of the distances, and isometrically embedding the resulting distances into Euclidean space. The fact that such an embedding exists was the core contribution of Schoenberg and Von Neumann in **??**. The central role of $q$-screw simplices in our proofs motivates us to develop new theorems on the geometry of these objects. Surprisingly, we find that these simplices are useful for proving generalizations of Von Neumann's work, and are deeply related to spectral graph theory.

Isometric embedding is a topic of wide interest in the field of metric geometry, and has been studied for many decades. Von Neumann and Issai Schoenberg proved in their seminal work that any $q$-screw simplex is isometrically embeddable in $l_2$. We extend their work to prove a stronger result: the $q$-screw simplex isometrically embeds into the space of Effective Resistance metrics. Simple metrics like the square in $l_2$ are not isometrically embeddable into this class of metrics, and thus, most Euclidean metrics are not expected to isometrically embed into Effective Resistance distance. Isometric embedding into effective resistance metrics has been a popular question in spectral graph theory [], and this is the first result we know of where a geometric distance defined without an obvious underlying electrical network embeds isometrically into Effective Resistances. We prove this isometry by showing a deep link between $q$-screw simplices and a differential operator known as the fractional Laplacian, which has wide applications in fields including fractional quantum physics [], cell membrane biology [], financial mathematics [], Brownian motion [], differential equations [], semi-groups [], Fourier analysis [], and more []. This further allows us to show that fundamental geometric quantities like circumcenters (essential for Voronoi diagram construction) and volumes on the $q$-screw simplex can be determined using fundamental primitives on graph Laplacians, in this case Laplacian system solving and Laplacian determinant estimation respectively. The fractional Laplacian can be interpreted as a natural example of a geometric resistive graph, first introduced by Alman et. al. in []. We further conjecture that taking the $q^{th}$ root of any tree metric is isometrically embeddable into effective resistance, which would imply that the Gomory Hu tree (and thus the inverse min-cut distance) embeds isometrically into effective resistances.

We also provide the first known closed form finite-dimensional embedding of the $q$-screw

simplex into Euclidean space, when $q > 2$. The work of Von Neumann et. Al. proved the simplex's existence for $q > 1$ by embedding it into infinite dimensional Hilbert space using theorems from complex analysis, functional analysis, infinite dimensional Hilbert space theory, and Fourier analysis. Our embedding uses only elementary techniques of eigenvector computation on finite matrices. We hope that this embedding makes the work of Von Neumann and Schoenberg more accessible. We use our new embedding to state and prove a generalization of Von Neumann and Schoenbergs theorem (on the embeddability of the $q$-screw simplex):

**Theorem 1.5.** *For points $p_1, \ldots p_n \in \mathbb{R}^n$ and any $q > 1$, the metric $D(p_i, p_j) = |p_i - p_j|_1^{1/q}$ is isometrically embeddable into $l_1$.*

This mirrors Schoenberg's famous theorem that any finite $l_2$ metric, raised to the $1/q$ power for $q > 1$, is isometrically embeddable in $l_2$.

## 1.1 Contributions

Our paper has three main theorems.

**Theorem 1.6.** *Given a point set $P \in \mathbb{R}^d$, the edge-squared metric on $P$ and the nearest-neighbor geodesic on $P$ are always equivalent.*

**Theorem 1.7.** *For any set of points in $\mathbb{R}^d$ for constant $d$, there exists a $(1+\varepsilon)$ spanner of the edge-squared 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 a fast floor function.*

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

Theorem 1.6 considerably strengthens a result from in [**?**], which showed $\mathbf{d}_2$ is a 3-approximation of $\mathbf{d}_N$. Our theorem finds $\mathbf{d}_N$ exactly, and lets us compute the persistent homology of $\mathbf{d}_N$. $\mathbf{d}_N$ is defined on all points in space, and is thus a metric extension [**?**] of the edge-squared metric and of negative type distances [**?**] to the entire space.

Theorem 1.7 proves that a $(1+\varepsilon)$-spanner of the edge-squared metric with points in constant dimension is sparser and can be computed more quickly than the Euclidean spanners of Callahan and Kosaraju [**?**]. The latter spanners have $O(n\varepsilon^{-d})$ edges and are computable in $O(n \log n + n\varepsilon^{-d})$ time. To the authors' knowledge, these are the sparsest quickly-constructable Euclidean spanners in terms of $\varepsilon$ dependence. Later works on spanners have focused on bounding diameter, degree, or total edge weight [**?**, **?**]. We give a size lower bound for $(1+\varepsilon)$-Euclidean spanners, which is close to the sparsity of our $(1+\varepsilon)$ spanner of the edge-squared metric. Previously, sparse spanners of the edge-squared metric were shown to exist in two dimensions via Yao graphs and Gabriel graphs [**?**].

Theorem 1.8 proves that a 1-spanner of the edge-squared metric can be found assuming points are samples from a probability density, by using a $k$-$NN$ graph for appropriate $k$. Our result is tight when $d$ is constant. This is not possible for Euclidean distance, as a 1-spanner is almost surely the complete graph. Without the probability density assumption, there are point sets in $\mathbb{R}^4$ where 1-spanners of the edge-squared metric require $\Omega(n^2)$ edges. Finally, we show that spanners of $p$-power metrics, which are edge-squared metrics but with powers of $p$ instead of 2, generalize Euclidean spanners and Euclidean MSTs. $p$-power metrics were considered in [**?**].

# 2 Outline

Section **??** contains the proof of Theorem 1.6, equating the edge-squared metric and nearest-neighbor geodesic distance in all cases. We then compute the persistent homology of the nearest-neighbor geodesic distance. Section **??** outlines a proof of Theorem 1.7, and compares our spanner to new lower bounds on the sparsity of $(1+\varepsilon)$-spanners of the Euclidean metric. We outline a proof of Theorem 1.8 in Section **??** and discuss its implications.

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

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