Products of Euclidean Metrics and Applications to Proximity Questions among Curves

Ioannis Z. Emiris[1]

Department of Informatics & Telecommunications
National & Kapodistrian University of Athens, Greece, and
ATHENA Research & Innovation Center, Greece
emiris@di.uoa.gr

Ioannis Psarros[2]

Department of Informatics & Telecommunications
National & Kapodistrian University of Athens, Greece
ipsarros@di.uoa.gr

Abstract

The problem of Approximate Nearest Neighbor (ANN) search is fundamental in computer science and has benefited from significant progress in the past couple of decades. However, most work has been devoted to pointsets whereas complex shapes have not been sufficiently treated. Here, we focus on distance functions between discretized curves in Euclidean space: they appear in a wide range of applications, from road segments and molecular backbones to time-series in general dimension. For ℓ_p-products of Euclidean metrics, for any $p \geq 1$, we design simple and efficient data structures for ANN, based on randomized projections, which are of independent interest. They serve to solve proximity problems under a notion of distance between discretized curves, which generalizes both discrete Fréchet and Dynamic Time Warping distances. These are the most popular and practical approaches to comparing such curves. We offer the first data structures and query algorithms for ANN with arbitrarily good approximation factor, at the expense of increasing space usage and preprocessing time over existing methods. Query time complexity is comparable or significantly improved by our algorithms; our approach is especially efficient when the length of the curves is bounded.

1 Introduction

The problem of Approximate Nearest Neighbor (ANN) search is fundamental in computer science: one has to preprocess a dataset so as to answer proximity queries efficiently, for a given query object. ANN has been enjoying a lot of attention and significant progress has been achieved in the past couple of decades. However, most work has been devoted to vector spaces, and complex objects have not been sufficiently treated. Here, we focus on

distance functions for polygonal curves which lie in the Euclidean space. Polygonal curves are essentially point sequences of varying length and have a wide range of applications ranging from road segments in low dimensions to time-series in arbitrary dimension and protein backbone structures. In general, the problem we aim to solve is as follows.

▶ **Definition 1** (ANN). Input are n polygonal curves $V_1, \ldots, V_n$, where each V_i is a sequence $v_{i1}, \ldots, v_{im_i}$ with each $v_{ij} \in \mathbb{R}^d$, and each $m_i \leq m$ for some pre-specified m. Given distance function $d(\cdot, \cdot)$, $\epsilon > 0$, preprocess $V_1, \ldots, V_n$ into a data structure s.t. for any query polygonal curve Q, the data structure reports V_j for which the following holds

$$\forall i : \ d(Q, V_j) \leq (1 + \epsilon) \cdot d(Q, V_i).$$

There are various ways to define dissimilarity or distance between two curves. Two popular dissimilarity measures are the Discrete Fréchet Distance (DFD) and the Dynamic Time Warping (DTW) distance which are both widely studied and applied to classification and retrieval problems for various types of data. DFD satisfies the triangular inequality, unlike DTW.

It is common, in distance functions of curves, to involve the notion of a traversal for two curves. Intuitively, a traversal corresponds to a time plan for traversing the two curves simultaneously, starting from the first point of each curve and finishing at the last point of each curve. With time advancing, the traversal advances in at least one of the two curves. DFD is the minimum over traversals, maximum distance of points while traversing. DTW is the minimum over traversals, sum of distances while traversing.

We denote by ℓ_p^d the normed space $(\mathbb{R}^d, \| \cdot \|_p)$, where for any $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$. We also use the notation $\tilde{O}(f(d, m, n))$, which hides polylogarithmic factors in $f(d, m, n)$ or polynomial factors in $1/\epsilon$.

1.1 Previous work

The ANN problem has been mainly addressed for datasets consisting of points. Efficient deterministic solutions exist when the dimension is constant, e.g. [5], while for high-dimensional data the state-of-the-art solutions are mainly based on the notion of Locality Sensitive Hashing, e.g. [8, 4], or on random projections, e.g. [1, 2]. Another line of work focuses on subsets of general metrics which satisfy some sort of low intrinsic dimension assumption, e.g. [9]. This is only a small fraction of the body of work on pointsets; however, very little is known about distances between curves.

Let us start with point sequences, which are closely related to curves. For metrics $M_1, \ldots, M_k$, we define the ℓ_p-product of $M_1, \ldots, M_k$ as the metric with domain $M_1 \times \cdots \times M_k$ and distance function

$$d((x_1, \ldots, x_k), (y_1, \ldots, y_k)) = \left(\sum_{i=1}^k d_{M_i}^p(x_i, y_i) \right)^{1/p}.$$

If there exists an ANN data structure with approximation factor c for each $M_1, \ldots, M_k$, then one can build a data structure for the ℓ_p-product with approximation factor $O(c \log \log n)$ [3, 10].

Let us now focus on curves: the two existing approaches both solve the approximate near neighbor problem, instead of the optimization ANN. It is known that a data structure for the approximate near neighbor problem can be used as a building block for solving the ANN problem. This procedure has provable guarantees on metrics [8], but it is not clear whether it can be extended to non-metric distances such as the DTW.

The first result for DFD by Indyk [10], defined by any metric $(X, d(\cdot, \cdot))$, achieved approximation factor $O((\log m + \log \log n)^{t-1})$, where m is the maximum length of a curve, and $t > 1$ is a trade-off parameter. The solution is based on an efficient data structure for ℓ_∞-products of arbitrary metrics, and achieves space and preprocessing time in $O(m^2|X|)^{tm^{1/t}} \cdot n^{2t}$, and query time in $(m \log n)^{O(t)}$. Table 1 states these bounds for appropriate $t = 1 + o(1)$, hence a constant approximation factor. It is not clear whether the approach may achieve a $1 + \epsilon$ approximation factor by employing more space.

Quite recently, a new data structure was devised for the DFD of curves defined by the Euclidean metric [7]. The approximation factor is $O(d^{3/2})$. The space required is $O(2^{4md}n \log n + mn)$ and each query costs $O(2^{4md}m \log n)$. They also provide a trade-off between space/query time, and the approximation factor. At the other extreme of this trade-off, they achieve space in $O(n \log n + mn)$, query time in $O(m \log n)$ and approximation factor $O(m)$. Our methods can achieve any user-desired approximation factor at the expense of a reasonable increase in the space and time complexities.

Furthermore, they show in [7] that the result establishing an $O(m)$ approximation extends to DTW, whereas the other extreme of the trade-off has remained open.

Table 1 summarizes space and query time complexities, and approximation factors of the main methods for searching among discrete curves under the two main dissimilarity measures.

1.2 Our contribution

Our first contribution is a simple data structure for the ANN problem in ℓ_p-products of finite subsets of ℓ_2^d, for any p. The key ingredient is a random projection from points in ℓ_2 to points in ℓ_p. Although this has proven a relevant approach for ANN of pointsets, it is quite unusual to employ randomized embeddings from ℓ_2 to ℓ_p, $p > 2$, because such norms are considered "harder" than ℓ_2 in the context of proximity searching. After the random projection, the algorithm "vectorizes" all point sequences. The original problem is then translated to the ANN problem for points in $\ell_p^{d'}$, for $d' \approx d \cdot m$ to be specified later, and can be solved by simple bucketing methods in space $\tilde{O}\left(d'n \cdot (1/\epsilon)^{d'}\right)$ and query time $\tilde{O}(d' \log n)$, which is very efficient when $d \cdot m$ is low.

Then, we present a notion of distance between two polygonal curves, which generalizes both DFD and DTW (for a formal definition see Definition 13). The ℓ_p-*distance* of two curves minimizes, over all traversals, the ℓ_p norm of the vector of all Euclidean distances between paired points. Hence, DFD corresponds to ℓ_∞-distance of polygonal curves, and DTW corresponds to ℓ_1-distance of polygonal curves.

Our main contribution is an ANN structure for the ℓ_p-distance of curves, when $1 \le p < \infty$. This easily extends to ℓ_∞-distance of curves by solving for the ℓ_p-distance, where p is sufficiently large. Our target are methods with approximation factor $1+\epsilon$. Such approximation factors are obtained for the first time, at the expense of larger space or time complexity. Moreover, a further advantage is that our methods solve ANN directly instead of requiring to reduce it to near neighbor search. While a reduction to the near neighbor problem has provable guarantees on metrics [8], we are not aware of an analogous result for non-metric distances such as the DTW.

Specifically, when $p > 2$, there exists a data structure with space and preprocessing time in

$$\tilde{O}\left(n \cdot \left(\frac{d}{p\epsilon} + 2\right)^{O(dm \cdot \alpha_{p,\epsilon})}\right),$$

■ **Table 1** Summary of previous results compared to this paper's: X denotes the domain set of the input metric. The first method is deterministic while the rest are randomized. All previous results are tuned to optimize the approximation factor. The parameters ρ_u, ρ_q satisfy $(1+\epsilon)\sqrt{\rho_q} + \epsilon\sqrt{\rho_u} \geq \sqrt{1+2\epsilon}$.

	Space	Query	Approx.	Comments
DFD	$O(m^2|X|)^{m^{1-o(1)}} \times O(n^{2-o(1)})$	$(m\log n)^{O(1)}$	$O(1)$	any metric [10]
	$\tilde{O}(2^{4md}n)$	$\tilde{O}(2^{4md}\log n)$	$O\left(d^{3/2}\right)$	ℓ_2^d, [7]
	$\tilde{O}(n) \times \left(\frac{d}{\log m} + 2\right)^{O(m \cdot d \log(1/\epsilon))}$	$\tilde{O}(d \cdot 2^{2m}\log n)$	$1 + \epsilon$	ℓ_2^d, Thm 16
DTW	$\tilde{O}(mn)$	$O(m\log n)$	$O(m)$	ℓ_2^d, rand.[7]
	$\tilde{O}(n) \times 2^{O(m \cdot d \log(1/\epsilon))}$	$\tilde{O}(d \cdot 2^{2m}\log n)$	$1 + \epsilon$	ℓ_2^d, Thm 17
	$\tilde{O}\left(2^{2m}n^{1+\rho_u}\right)$	$\tilde{O}\left(2^{2m}n^{\rho_q}\right)$	$1 + \epsilon$	ℓ_2^d, Thm 18

where $\alpha_{p,\epsilon}$ depends only on p, ϵ, and query time in $\tilde{O}(2^{2m}\log n)$. When specialized to DFD and compared to [7], our space and preprocessing time complexity is higher by the exponent $\log(1/\epsilon)$ but our query time is linear instead of being exponential in d.

When $p \in [1, 2]$, there exists a data structure with space and preprocessing time in

$$\tilde{O}\left(n \cdot 2^{O(dm \cdot \alpha_{p,\epsilon})}\right),$$

where $\alpha_{p,\epsilon}$ depends only on p, ϵ, and query time in $\tilde{O}\left(2^{2m}\log n\right)$. This leads to the first approach that achieves $1 + \epsilon$ approximation for DTW at the expense of space, preprocessing and query time complexities being exponential in m. Hence our method is best suited when the curve size is small.

Our results for DTW and DFD are summarized in Table 1 and juxtaposed to existing approaches in [7, 10].

The rest of the paper is structured as follows. In Section 2, we present a data structure for ANN in ℓ_p-products of ℓ_2, which is of independent interest. In Section 3, we employ this result to address the ℓ_p-distance of curves. We conclude with future work.

2 ℓ_p-products of ℓ_2

In this section, we present a simple data structure for ANN in ℓ_p-products of finite subsets of ℓ_2. Recall that the ℓ_p-product of $X_1, \ldots, X_m$, which are finite subsets of ℓ_2, is a metric space with ground set $X_1 \times X_2 \times \cdots \times X_m$ and distance function:

$$d((x_1, \ldots, x_m), (y_1, \ldots, y_m)) = \|\|x_1 - y_1\|_2, \ldots, \|x_m - y_m\|_2\|_p = \left(\sum_{i=1}^m \|x_i - y_i\|_2^p\right)^{1/p}.$$

For ANN, the algorithm first randomly embeds points from ℓ_2 to ℓ_p. Then, it is easy to translate the original problem to ANN in ℓ_p for large vectors corresponding to point sequences.

2.1   Concentration inequalities

In this subsection, we prove concentration inequalities for central absolute moments of the normal distribution. Most of these results are probably folklore and the reasoning is quite similar to the one followed by proofs of the Johnson-Lindenstrauss lemma, e.g. [11].

The 2-stability property of standard normal variables, along with standard facts about their absolute moments imply the following claim.

▶ **Claim 2.** *Let $v \in \mathbb{R}^d$ and let G be $k \times d$ matrix with i.i.d random variables following $N(0,1)$. Then,*

$$\mathbb{E}\left[\|Gv\|_p^p\right] = c_p \cdot k \cdot \|v\|_2^p,$$

where $c_p = \frac{2^{p/2} \cdot \Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}$ is a constant depending only on $p > 1$.

Proof. Let $g = (X_1, \ldots, X_d)$ be a vector of random variables which follow $N(0,1)$ and any vector $v \in \mathbb{R}^d$. The 2-stability property of gaussian random variables implies that $\langle g, v \rangle \sim N(0, \|v\|_2^2)$. Recall the following standard fact for central absolute moments of $Z \sim N(0, \sigma^2)$:

$$\mathbb{E}[|Z|^p] = \sigma^p \cdot \frac{2^{p/2} \cdot \Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}.$$

Hence,

$$\mathbb{E}\left[\|Gv\|_p^p\right] = \mathbb{E}\left[\sum_{i=1}^{k} |\langle g_i, v \rangle|^p\right] = k \cdot \|v\|_2^p \cdot \frac{2^{p/2} \cdot \Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}. \qquad \blacktriangleleft$$

In the following lemma, we give a simple upper bound on the moment generating function of $|X|^p$, where $X \sim N(0,1)$.

▶ **Lemma 3.** *Let $X \sim N(0, \sigma^2)$, $p \geq 1$, and $t > 0$, then $\mathbb{E}[exp(-t|X|^p)] \leq exp(-t\mathbb{E}[|X|^p] + t^2\mathbb{E}[|X|^{2p}])$.*

Proof. We use the easily verified fact that for any $x \leq 1$, $exp(x) \leq 1 + x + x^2$ and the standard inequality $1 + x \leq e^x$, for all $x \in \mathbb{R}$.

$$\mathbb{E}\left[e^{-t|X|^p}\right] \leq 1 - t \cdot \mathbb{E}\left[|X|^p\right] + t^2 \cdot \mathbb{E}\left[|X|^{2p}\right] \leq e^{-t\mathbb{E}[|X|^p] + t^2\mathbb{E}[|X|^{2p}]}. \qquad \blacktriangleleft$$

▶ **Claim 4.** *Let $X \sim N(0,1)$. Then, there exists constant $C > 0$ s.t. for any $p \geq 1$, $\mathbb{E}[|X|^{2p}] \leq C \cdot \mathbb{E}[|X|^p]^2$.*

Proof. In the following, we denote by $f(p) \approx g(p)$ the fact that there exist constants $0 < c < C$ s.t. for any $p > 1$, $f(p) \leq C \cdot g(p)$ and $f(p) \geq c \cdot g(p)$. In addition, $f(p) \gtrsim g(p)$ means that $\exists C > 0$ s.t. $\forall p > 1$, $C \cdot f(p) \geq g(p)$. In other words, $g(p) \approx f(p) \iff g(p) = \Theta(f(p))$ and $f(p) \gtrsim g(p) \iff f(p) = \Omega(g(p))$. In the following we make use of the Stirling approximation and standard facts about moments of normal variables.

$$\mathbb{E}\left[|X|^{2p}\right] = \frac{2^p \cdot \Gamma\left(\frac{2p+1}{2}\right)}{\sqrt{\pi}} \approx (2p-1)!! = \frac{(2p)!}{2^p \cdot p!} \approx \left(\frac{(2p)^{2p}}{e}\right) \cdot \sqrt{p} \cdot \frac{1}{2^p \cdot \left(\frac{p}{e}\right)^p} \approx \frac{2^p p^p \sqrt{p}}{e^p} \approx 2^p \cdot p!$$

$$\mathbb{E}\left[|X|^p\right]^2 \approx ((p-1)!!)^2 \gtrsim \left(2^{p/2+1/2} \cdot \left(\frac{(p/2+1/2)}{e}\right)^{(p/2+1/2)}\right)^2 \approx 2^p \cdot \frac{p^{p+1}}{e^{p+1}} \gtrsim 2^p \cdot p! \qquad \blacktriangleleft$$

The following lemma is the main ingredient of our embedding, since it provides us with a lower tail inequality for one projected vector.

▶ **Lemma 5.** *Let G be a $k \times d$ matrix with i.i.d. random variables following $N(0,1)$ and consider vector $v \in \mathbb{R}^d$, s.t. $\|v\|_2 = 1$. For appropriate constant $c' > 1$, for $p \geq 1$ and $\delta \in (0,1)$,*

$$Pr[\|Gv\|_p^p \leq (1-\delta) \cdot \mathbb{E}[\|Gv\|_p^p]] \leq e^{-c' \cdot k \cdot \delta^2}.$$

Proof. For $X \sim N(0,1)$ and any $t > 0$,

$$\Pr\left[\|Gv\|_p^p \leq (1-\delta) \cdot \mathbb{E}\left[\|Gv\|_p^p\right]\right] \leq \mathbb{E}\left[e^{-t|X|^p}\right]^k \cdot e^{(t(1-\delta)k \cdot \mathbb{E}[|X|^p])} \leq$$

$$\leq e^{k(-t \cdot \mathbb{E}[|X|^p] + t^2 \cdot C \cdot \mathbb{E}[|X|^p]^2 + t \cdot (1-\delta) \cdot \mathbb{E}[|X|^p])}.$$

The last inequality derives from Claim 4. Now, we set $t = \frac{\delta}{2\sqrt{C} \cdot \mathbb{E}[|X|^p]}$, and we assume wlog $C > 4$. Hence,

$$\Pr[\|Gv\|_p^p \leq (1-\delta) \cdot \mathbb{E}[\|Gv\|_p^p]] \leq e^{-c' \cdot k \cdot \delta^2},$$

for some constant $c' > 1$. ◀

Finally, we make use of the following one-sided Johnson-Lindenstrauss lemma (see e.g. [11]).

▶ **Theorem 6.** *Let G be a $k \times d$ matrix with i.i.d. random variables following $N(0,1)$ and consider vector $v \in \mathbb{R}^d$. Then, for constant $C > 0$,*

$$\Pr\left[\|Gv\|_2 \geq (1+\epsilon)\|v\|_2\sqrt{k}\right] \leq e^{-C \cdot k \cdot \epsilon^2}.$$

Standard properties of ℓ_p norms imply a loose upper tail inequality.

▶ **Corollary 7.** *Let G be a $k \times d$ matrix with i.i.d. random variables following $N(0,1)$ and consider vector $v \in \mathbb{R}^d$. Let $p \geq 2$. Then, for constant $C > 0$,*

$$\Pr\left[\|Gv\|_p \geq (1+\epsilon)\|v\|_2\sqrt{k}\right] \leq e^{-C \cdot k \cdot \epsilon^2}.$$

Proof. Since $p \geq 2$, we have that $\forall x \in \mathbb{R}^d \ \|x\|_p \leq \|x\|_2$. Hence, by Theorem 6,

$$\Pr[\|Gv\|_p \geq (1+\epsilon)\|v\|_2\sqrt{k}] \leq \Pr[\|Gv\|_2 \geq (1+\epsilon)\|v\|_2\sqrt{k}] \leq e^{-C \cdot k \cdot \epsilon^2}. \qquad ◀$$

However, an improved upper tail inequality can be derived when $p \in [1,2]$.

▶ **Lemma 8.** *Let G be a $k \times d$ matrix with i.i.d. random variables following $N(0,1)$ and consider vector $v \in \mathbb{R}^d$. Let $p \in [1,2]$. Then, for constant $C > 0$,*

$$\Pr\left[\|Gv\|_p \geq (3 \cdot c_p \cdot k)^{1/p}\|v\|_2\right] \leq e^{-C \cdot k}.$$

Proof. Let $X \sim N(0,1)$.

$$\mathbb{E}\left[e^{|X|^p/3}\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{|x|^p/3 - x^2/2} \mathrm{d}x \leq \frac{\sqrt{2}}{\sqrt{\pi}} \int_0^{+\infty} e^{x^2/3 - x^2/2} \mathrm{d}x = \sqrt{3}.$$

Now, assume wlog $\|v\|_2 = 1$,

$$\Pr\left[\|Gv\|_p^p \geq 3 \cdot \mathbb{E}\left[\|Gv\|_p^p\right]\right] \leq \mathbb{E}\left[e^{|X|^p/3}\right]^k \cdot e^{-k \cdot \mathbb{E}[|X|^p]} \leq e^{-k(c_p - 2/3)} \leq e^{-k/10},$$

where $c_p = \frac{2^{p/2} \cdot \Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}$. ◀

2.2   Embedding ℓ_2 into ℓ_p

In this subsection, we present our main results concerning ANN for ℓ_p-products of ℓ_2. First, we show that a simple random projection maps points from ℓ_2^d to ℓ_p^k, where $k = \tilde{O}(d)$, without arbitrarily contracting norms. The probability of failure decays exponentially with k. For our purposes, there is no need for an almost isometry between norms. Hence, our efforts focus on proving lower tail inequalities which imply that, with good probability, no far neighbor corresponds to an approximate nearest neighbor in the projected space.

We now prove bounds concerning the contraction of distances of the embedded points. Our proof builds upon the inequalities developed in Subsection 2.1.

▶ **Theorem 9.** *Let G be a $k \times d$ matrix with i.i.d. random variables following $N(0,1)$. Then,*
- *if $2 < p < \infty$ then,*

$$\Pr\left[\exists v \in \mathbb{R}^d : \ \|Gv\|_p \leq \frac{(c_p \cdot k)^{1/p}}{1+\epsilon} \cdot \|v\|_2\right] \leq O\left(\frac{k^{\frac{1}{2}-\frac{1}{p}}}{p\epsilon} + 2\right)^d \cdot \mathrm{e}^{-c' \cdot k \cdot (p\epsilon/(2+p\epsilon))^2},$$

- *if $p \in [1,2]$ then,*

$$\Pr\left[\exists v \in \mathbb{R}^d : \ \|Gv\|_p \leq \frac{(c_p \cdot k)^{1/p}}{1+\epsilon} \cdot \|v\|_2\right] \leq O\left(\frac{1}{\epsilon}\right)^d \cdot \mathrm{e}^{-c' \cdot k \cdot (p\epsilon/(2+p\epsilon))^2},$$

where $c' > 1$ is a constant, $\epsilon \in (0, 1/2)$.

Proof. By Lemma 5:

$$\Pr\left[\|Gv\|_p^p \leq \frac{c_p \cdot k}{(1+\epsilon)^p} \cdot \|v\|_2^p\right] \leq \Pr\left[\|Gv\|_p^p \leq \frac{c_p \cdot k}{1+p\epsilon/2} \cdot \|v\|_2^p\right] \leq \mathrm{e}^{-c' \cdot k \cdot (p\epsilon/(2+p\epsilon))^2}.$$

In order to bound the probability of contraction among all distances, we argue that it suffices to use the strong bound on distance contraction, which is derived in Lemma 5, and the weak bound on distance expansion from Corollary 7 or Lemma 8, for a δ-dense set $N \subset \mathbb{S}^{d-1}$ for δ to be specified later. First, a simple volumetric argument [8] shows that there exists $N \subset \mathbb{S}^{d-1}$ s.t. $\forall x \in \mathbb{S}^{d-1} \ \exists y \in N \ \|x - y\|_2 \leq \delta$, and $|N| = O\left(1/\delta\right)^d$.

We first consider the case $p > 2$. From now on, we assume that for any $u \in N$, $\|Gu\|_p \geq (c_p \cdot k)^{1/p}/(1+\epsilon)$ and $\|Gu\|_p \leq 2\sqrt{k}$ which is achieved with probability

$$\geq 1 - O\left(\frac{1}{\delta}\right)^d \cdot \mathrm{e}^{-c' \cdot k \cdot (p\epsilon/(2+p\epsilon))^2}.$$

Now let x be an arbitrary vector in $\mathbb{R}^d$ s.t. $\|x\|_2 = 1$. Then, there exists $u \in N$ s.t. $\|x - u\|_2 \leq \delta$. Also, by the triangular inequality we obtain the following,

$$\|Gx\|_p \leq \|Gu\|_p + \|G(x-u)\|_p = \|Gu\|_p + \|x-u\|_2 \left\|G\frac{(x-u)}{\|x-u\|_2}\right\|_p \implies$$

$$\implies \|Gx\|_p \leq \|Gu\|_p + \delta \left\|G\frac{(x-u)}{\|x-u\|_2}\right\|_p. \tag{1}$$

Let $M = \max_{x \in \mathbb{S}^{d-1}} \|Gx\|_p$. The existence of M is implied by the fact that $\mathbb{S}^{d-1}$ is compact and $x \mapsto \|x\|_p$, $x \mapsto Gx$ are continuous functions. Then, by plugging M into (1),

$$M \leq \|Gu\|_p + \delta M \implies M \leq \frac{\|Gu\|_p}{1-\delta} \leq \frac{2\sqrt{k}}{1-\delta},$$

where the last inequality is implied by Corollary 7. Again, by the triangular inequality,

$$\|Gx\|_p \geq \|G_u\|_p - \|G(x-u)\|_p \geq \frac{(c_p \cdot k)^{1/p}}{1+\epsilon} - \frac{2\delta\sqrt{k}}{1-\delta} \geq \frac{1-\epsilon/2}{1+\epsilon} \cdot (c_p \cdot k)^{1/p},$$

for $\delta \leq \frac{\epsilon \cdot (c_p \cdot k)^{1/p}}{2\sqrt{k} + \epsilon \cdot (c_p \cdot k)^{1/p}}$.

Notice now that

$$\frac{1}{\delta} = O\left(\frac{k^{1/2-1/p}}{p\epsilon}\right) + 1.$$

In the case $p \in [1,2]$, we are able to use a better bound on the distance expansion; namely Lemma 8. We now assume that for any $u \in N$, $\|Gu\|_p \geq (c_p \cdot k)^{1/p}/(1+\epsilon)$ and $\|Gu\|_p \leq (3 \cdot c_p \cdot k)^{1/p}$ which is achieved with probability

$$\geq 1 - O\left(\frac{1}{\delta}\right)^d \cdot e^{-c' \cdot k \cdot (p\epsilon/(2+p\epsilon))^2}.$$

Once again, we use inequality (1) to obtain:

$$M \leq \frac{\|Gu\|_p}{1-\delta} \leq \frac{(3 \cdot c_p \cdot k)^{1/p}}{1-\delta} \implies$$

$$\implies \|Gx\|_p \geq \|Gu\|_p - \|Gx - Gu\|p \geq (c_p \cdot k)^{1/p}\left(\frac{1}{1+\epsilon} - \frac{3^{1/p} \cdot \delta}{1-\delta}\right) \implies$$

$$\implies \|Gx\|_p \geq (c_p \cdot k)^{1/p} \cdot \frac{1-\epsilon/2}{1+\epsilon},$$

for $\delta \leq \epsilon/(6(1+\epsilon)+\epsilon) = \Omega(\epsilon)$. ◀

Theorem 9 implies that the ANN problem for ℓ_p products of ℓ_2 translates to the ANN problem for ℓ_p products of ℓ_p. The latter easily translates to the ANN problem in $\ell_p^{d'}$. One can then solve the approximate near neighbor problem in $\ell_p^{d'}$, by approximating $\ell_p^{d'}$ balls of radius 1 with a regular grid with side length $\epsilon/(d')^{1/p}$. Each approximate ball is essentially a set of $O(1/\epsilon)^{d'}$ cells [8]. Building not-so-many approximate near neighbor data structures for various radii leads to an efficient solution for the ANN problem [8].

▶ **Theorem 10.** *There exists a data structure which solves the ANN problem for point sequences in ℓ_p-products of ℓ_2, and satisfies the following bounds on performance:*
- *If $p \in [1,2]$, then space usage and preprocessing time is in*

$$\tilde{O}(dmn) \times \left(\frac{1}{\epsilon}\right)^{O(m \cdot d \cdot \alpha_{p,\epsilon})},$$

  query time is in $\tilde{O}(dm\log n)$, and $\alpha_{p,\epsilon} = \log(1/\epsilon) \cdot (2+p\epsilon)^2 \cdot (p\epsilon)^{-2}$.
- *If $2 < p < \infty$, then space usage and preprocessing time is in*

$$\tilde{O}(dmn) \times \left(\frac{d}{p\epsilon} + 2\right)^{O(m \cdot d \cdot \alpha_{p,\epsilon})},$$

  query time is in $\tilde{O}(dm\log n)$, and $\alpha_{p,\epsilon} = \log(1/\epsilon) \cdot (2+p\epsilon)^2 \cdot (p\epsilon)^{-2}$.

We assume $\epsilon \in (0, 1/2]$. The probability of success is $\Omega(\epsilon)$ and can be amplified to $1 - \delta$, by building $\Omega(\log(1/\delta)/\epsilon)$ independent copies of the data-structure.

Proof. Let $\delta_{p,\epsilon} = p\epsilon/(2 + p\epsilon)$. We first consider the case $p > 2$. We employ Theorem 9 and we map point sequences to point sequences in ℓ_p^k, for

$$k = \Theta\left(\frac{d \log \frac{d}{p\epsilon}}{\delta_{p,\epsilon}^2}\right).$$

Hence, Theorem 9 implies that,

$$\Pr\left[\exists v \in \mathbb{R}^d : \|Gv\|_p \leq \frac{(c_p \cdot k)^{1/p}}{1 + \epsilon} \cdot \|v\|_2\right] \leq \epsilon/10.$$

Then, by concatenating vectors, we map point sequences to points in ℓ_p^{km}.

Now, fix query point sequence $Q = q_1, \ldots, q_m \in \left(\mathbb{R}^d\right)^m$ and its nearest neighbor $U_* = u_1, \ldots, u_m \in \left(\mathbb{R}^d\right)^m$. By a union bound, the probability of failure for the embedding is at most

$$\epsilon/2 + \Pr\left[\sum_{i=1}^m \|Gu_i - Gq_i\|_p^p \leq (1 + \epsilon)^p \cdot c_p \cdot k \sum_{i=1}^m \|u_i - q_i\|_2^p\right],$$

since we already know that

$$\Pr\left[\exists v \in \mathbb{R}^d : \|Gv\|_p \leq \frac{(c_p \cdot k)^{1/p}}{1 + \epsilon} \cdot \|v\|_2\right] \leq \epsilon/2.$$

Hence, we now bound the second probability. Notice that

$$\mathbb{E}\left[\sum_{i=1}^m \|Gu_i - Gq_i\|_p^p\right] = \sum_{i=1}^m \mathbb{E}\left[\|G(u_i - q_i)\|_p^p\right] = c_p \cdot k \sum_{i=1}^m \|u_i - q_i\|_2^p.$$

By Markov's inequality, we obtain,

$$\Pr\left[\sum_{i=1}^m \|Gu_i - Gq_i\|_p^p \leq (1 + \epsilon)^p \cdot c_p \cdot k \sum_{i=1}^m \|u_i - q_i\|_2^p\right] \leq (1 + \epsilon)^{-p}.$$

Hence, the total probability of failure is $\frac{1 + \epsilon/10}{(1 + \epsilon)^p}$. In the projected space, we build AVDs[8]. The total space usage, and the preprocessing time is

$$\tilde{O}(dmn) \times O(1/\epsilon)^{km} = \tilde{O}(dmn) \times \left(\frac{d}{p\epsilon} + 2\right)^{O(m \cdot d \cdot \log(1/\epsilon)/\delta_{p,\epsilon}^2)}.$$

The query time is $\tilde{O}(dm \log n)$. The probability of success can be amplified by repetition. By building $\Theta\left(\frac{\log(1/\delta)}{\epsilon}\right)$ data structures as above, the probability of failure becomes δ.

The same reasoning is valid in the case $p \in [1, 2]$, but it suffices to set

$$k = \Theta\left(\frac{d \log \frac{1}{\epsilon}}{\delta_{p,\epsilon}^2}\right). \tag*{◀}$$

When $p \in [1, 2]$, we can also utilize "high-dimensional" solutions for ℓ_p and obtain data structures with complexities polynomial in $d \cdot m$. Combining Theorem 9 with the data structure of [4], we obtain the following result.

▶ **Theorem 11.** *There exists a data structure which solves the ANN problem for point sequences in ℓ_p-products of ℓ_2, $p \in [1, 2]$, and satisfies the following bounds on performance: space usage and preprocessing time is in $\tilde{O}(n^{1+\rho_u})$, and the query time is in $\tilde{O}(n^{\rho_q})$, where ρ_q, ρ_u satisfy:*

$$(1 + \epsilon)^p \sqrt{\rho_q} + ((1 + \epsilon)^p - 1)\sqrt{\rho_u} \geq \sqrt{2(1 + \epsilon)^p - 1}$$

We assume $\epsilon \in (0, 1/2]$. The probability of success is $\Omega(\epsilon)$ and can be amplified to $1 - \delta$, by building $\Omega(\log(1/\delta)/\epsilon)$ independent copies of the data-structure.

Proof. We proceed as in the proof of Theorem 10. We employ Theorem 9 and by Markov's inequality, we obtain,

$$Pr\left[\sum_{i=1}^{m} \|Gv_i - Gu_i\|_p^p \leq (1 + \epsilon)^p \cdot c_p \cdot k \sum_{i=1}^{m} \|v_i - u_i\|_2^p\right] \leq (1 + \epsilon)^{-p}.$$

Then, by concatenating vectors, we map point sequences to points in ℓ_p^{km}, where $k = \tilde{O}(d)$. For the mapped points in ℓ_p^{km}, we build the LSH-based data structure from [4] which succeeds with high probability $1 - o(1)$. By independence, both the random projection and the LSH-based structure succeed with probability $\Omega(\epsilon) \times (1 - o(1)) = \Omega(\epsilon)$. ◀

3 Polygonal curves

In this section, we show that one can solve the ANN problem for a certain class of distance functions defined on polygonal curves. Since this class is related to ℓ_p-products of ℓ_2, we invoke results of Section 2, and we show an efficient data structure for the case of short curves, i.e. when m is relatively small compared to the other complexity parameters.

First, we need to introduce a formal definition of the traversal of two curves.

▶ **Definition 12.** Given polygonal curves $V = v_1, \ldots, v_{m_1}$, $U = u_1, \ldots, u_{m_2}$, a traversal $T = (i_1, j_1), \ldots, (i_t, j_t)$ is a sequence of pairs of indices referring to a pairing of vertices from the two curves such that:
1. $i_1, j_1 = 1$, $i_t = m_1$, $j_t = m_2$.
2. $\forall(i_k, j_k) \in T : i_{k+1} - i_k \in \{0, 1\}$ and $j_{k+1} - j_k \in \{0, 1\}$.
3. $\forall(i_k, j_k) \in T : (i_{k+1} - i_k) + (j_{k+1} - j_k) \geq 1$.
Now, we define a class of distance functions for polygonal curves. In this definition, it is implied that we use the Euclidean distance to measure distance between any two points. However, the definition could be easily generalized to arbitrary metrics.

▶ **Definition 13** (ℓ_p-distance of polygonal curves). Given polygonal curves $V = v_1, \ldots, v_{m_1}$, $U = u_1, \ldots, u_{m_2}$, we define the ℓ_p-distance between V and U as the following function:

$$d_p(V, U) = \min_{T \in \mathcal{T}} \left(\sum_{(i_k, j_k) \in T} \|v_{i_k} - u_{j_k}\|_2^p\right)^{1/p},$$

where $\mathcal{T}$ denotes the set of all possible traversals for V and U.

The above class of distances for curves includes some widely known distance functions. For instance, $d_\infty(V, U)$ coincides with the DFD of V and U (defined for the Euclidean distance). Moreover $d_1(V, U)$ coincides with DTW for curves V, U.

▶ **Theorem 14.** *Suppose that there exists a randomized data structure for the ANN problem in ℓ_p products of ℓ_2, with space in $S(n)$, preprocessing time $T(n)$ and query time $Q(n)$, with probability of failure less than $2^{-2m-1}m^{-1}$. Then, there exists a data structure for the ANN problem for the ℓ_p-distance of polygonal curves, $1 \leq p < \infty$, with space in $m \cdot 2^{2m} \cdot S(n)$, preprocessing time $m \cdot 2^{2m} \cdot T(n)$ and query time $m \cdot 2^{2m} \cdot Q(n)$, where m denotes the maximum length of a polygonal curve, and the probability of failure is less than $1/2$.*

Proof. We denote by X the input dataset. Given polygonal curves $V = v_1, \ldots, v_{m_1}$, $Q = q_1, \ldots, q_{m_2}$, and traversal T, one can define $V_T = v_1, \ldots, v_l$, $Q_T = q_1, \ldots, q_l$, sequences of l points (allowing consecutive duplicates) s.t. $\forall k$, $v_{i_k} = V_T[k]$ and $q_{j_k} = Q_T[k]$, if and only if $(i_k, j_k) \in T$.

One traversal of V, Q is uniquely defined by its length $l \in \{\max(m_1, m_2), \ldots, m_1 + m_2\}$, the set of indices $A = \{k \in \{1, \ldots, l\} \mid i_{k+1} - i_k = 0 \text{ and } j_{k+1} - j_k = 1\}$ for which only Q is progressing and the set of indices $B = \{k \in \{1, \ldots, l\} \mid i_{k+1} - i_k = 1 \text{ and } j_{k+1} - j_k = 1\}$ for which both Q and V are progressing. We can now define $V_{l,A,B}$, $Q_{l,A,B}$ to be the corresponding sequences of l points. In other words if l, A, B corresponds to traversal T, $V_{l,A,B} = V_T$, $Q_{l,A,B} = Q_T$. Observe that it is possible that curve V is not compatible with some triple l, A, B.

We build one ANN data structure, for ℓ_p products of ℓ_2, for each possible l, A, B. Each data structure contains at most $|X|$ point sequences which correspond to curves that are compatible to l, A, B. We denote by $m = \max(m_1, m_2)$. The total number of data structures is upper bounded by

$$\sum_{l=m}^{2m} \sum_{t=0}^{m} \binom{l}{t} \cdot \binom{l-t}{m-t} \leq \sum_{l=m}^{2m} \sum_{t=0}^{m} \binom{l}{t} \cdot \binom{l}{m-t} = \sum_{l=m}^{2m} \binom{2l}{m} \leq m \cdot \binom{2m}{m} \leq m \cdot 2^{2m}.$$

For any query curve Q, we create all possible combinations of l, A, B and we perform one query per ANN data structure. We report the best answer. The probability that the building of one of the $\leq m \cdot 2^{2m}$ data structures is not successful is less than $1/2$ due to a union bound. ◀

We now investigate applications of the above results, to the ANN problem for some popular distance functions for curves.

Discrete Fréchet Distance

DFD is naturally included in the distance class of Definition 13 for $p = \infty$. However, Theorem 14 is valid only when p is bounded. To overcome this issue, p is set to a suitable large value.

▶ **Claim 15.** *Let $V = v_1, \ldots, v_{m_1} \in \mathbb{R}^d$ and $U = u_1, \ldots, u_{m_2} \in \mathbb{R}^d$ be two polygonal curves. Then for any traversal T of V and U:*

$$(1 + \epsilon)^{-1} \cdot \left(\sum_{(i_k, j_k) \in T} \|v_{i_k} - u_{j_k}\|^p \right)^{1/p} \leq \max_{(i_k, j_k) \in T} \|v_{i_k} - u_{j_k}\| \leq \left(\sum_{(i_k, j_k) \in T} \|v_{i_k} - u_{j_k}\|^p \right)^{1/p},$$

for $p \geq \log(|T|) / \log(1 + \epsilon)$.

Proof. For any $x \in \mathbb{R}^{|T|}$, it is known that $\|x\|_\infty \leq \|x\|_p \leq (|T|)^{1/p} \|x\|_\infty$. ◀

▶ **Theorem 16.** *There exists a data structure for the ANN problem for the DFD of curves, with space and preprocessing time*

$$\tilde{O}(dm^2n) \times \left(\frac{d}{\log m} + 2\right)^{O(m \cdot d \cdot \log(1/\epsilon))},$$

and query time $\tilde{O}(d \cdot 2^{2m} \log n)$, where m denotes the maximum length of a polygonal curve, and $\epsilon \in (0, 1/2]$. The data structure succeeds with probability $1/2$, which can be amplified by repetition.

Proof. We combine Theorem 14 with Theorem 10 for $p \geq \log m / \log(1 + \epsilon) \geq \epsilon^{-1} \log m$. Notice that in order to plug the data structure of Theorem 10 into Theorem 14 we need to amplify the probability of success to $1 - 2^{-2m-1} \cdot m^{-1}$. Hence, the data structure for the ANN problem for ℓ_p-products of ℓ_p needs space and preprocessing time

$$\tilde{O}(dm^2n) \times \left(\frac{d}{p\epsilon} + 2\right)^{O(m \cdot d \cdot \alpha_{p,\epsilon})},$$

and each query time costs $O(dm^2)$, where $\alpha_{p,\epsilon} = \log(1/\epsilon) \cdot (2 + p\epsilon)^2 \cdot (p\epsilon)^{-2}$. Now, substituting p and invoking Theorem 14 completes our proof. ◀

Dynamic Time Warping

DTW corresponds to the ℓ_1-distance of polygonal curves as defined in Definition 13. Now, we combine Theorem 14 with each of the Theorems 10 and 11.

▶ **Theorem 17.** *There exists a data structure for the ANN problem for DTW of curves, with space and preprocessing time*

$$\tilde{O}(dm^2n) \times \left(\frac{1}{\epsilon}\right)^{O(m \cdot d \cdot \epsilon^{-2})},$$

and query time $\tilde{O}(d \cdot 2^{2m} \log n)$, where m denotes the maximum length of a polygonal curve, and $\epsilon \in (0, 1/2]$. The data structure succeeds with probability $1/2$, which can be amplified by repetition.

Proof. We first amplify the probability of success for the data structure of Theorem 10 to $1 - 2^{-2m-1} \cdot m^{-1}$. Hence, the data structure for the ANN problem for ℓ_1-products of ℓ_1 needs space and preprocessing time

$$\tilde{O}(dm^2n) \times 2^{O(m \cdot d \cdot \alpha_{p,\epsilon})},$$

and each query time costs $O(dm^2)$, where $\alpha_{p,\epsilon} = \log(1/\epsilon) \cdot (2 + \epsilon)^2 \cdot (\epsilon)^{-2}$. We plug this data structure into Theorem 14. ◀

▶ **Theorem 18.** *There exists a data structure for the ANN problem for DTW of curves, with space and preprocessing time $\tilde{O}(2^{2m}n^{1+\rho_u})$, and the query time is in $\tilde{O}(2^{2m}n^{\rho_q})$, where ρ_q, ρ_u satisfy:*

$$(1 + \epsilon)\sqrt{\rho_q} + \epsilon\sqrt{\rho_u} \geq \sqrt{1 + 2\epsilon}.$$

We assume $\epsilon \in (0, 1/2]$. The data structure succeeds with probability $1/2$, which can be amplified by repetition.

Proof. First amplify the probability of success for the data structure of Theorem 11 to $1 - 2^{-2m-1} \cdot m^{-1}$, by building independently $\tilde{O}(m)$ such data structures. We plug the resulting data structure into Theorem 14. ◀

4   Conclusion

Thanks to the simplicity of the approach, it should be easy to implement it and should have practical interest. We plan to apply it to real scenarios with data from road segments or time series.

The key ingredient of our approach is a randomized embedding from ℓ_2 to ℓ_p which is the first step to the ANN solution for ℓ_p-products of ℓ_2. The embedding is essentially a gaussian projection and it exploits the 2-stability property of normal variables, along with standard properties of their tails. We expect that a similar result can be achieved for ℓ_p-products of ℓ_q, where $q \in [1, 2)$. One related result for ANN [6], provides with dimension reduction for ℓ_q, $q \in [1, 2)$.

─── **References** ───

1   E. Anagnostopoulos, I. Z. Emiris, and I. Psarros. Low-quality dimension reduction and high-dimensional approximate nearest neighbor. In *Proc. 31st Intern. Symp. on Computational Geometry (SoCG)*, pages 436–450, 2015. `doi:10.4230/LIPIcs.SOCG.2015.436`.

2   E. Anagnostopoulos, I. Z. Emiris, and I. Psarros. Randomized embeddings with slack, and high-dimensional approximate nearest neighbor. In *ACM Transactions on Algorithm*, 2018, To appear.

3   A. Andoni. *NN search: the old, the new, and the impossible.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2009. URL: `http://hdl.handle.net/1721.1/55090`.

4   A. Andoni, T. Laarhoven, I. Razenshteyn, and E. Waingarten. Optimal hashing-based time-space trade-offs for approximate near neighbors. In *Proc. Symp. Discrete Algorithms (SODA)*, 2017. Also as arxiv.org/abs/1608.03580.

5   S. Arya, T. Malamatos, and D. M. Mount. Space-time tradeoffs for approximate nearest neighbor searching. *J. ACM*, 57(1):1:1–1:54, 2009. `doi:10.1145/1613676.1613677`.

6   Y. Bartal and L. -A. Gottlieb. Dimension reduction techniques for ₚ (1<p<2), with applications. In *32nd International Symposium on Computational Geometry, SoCG 2016, June 14-18, 2016, Boston, MA, USA*, pages 16:1–16:15, 2016. `doi:10.4230/LIPIcs.SoCG.2016.16`.

7   A. Driemel and F. Silvestri. Locality-sensitive hashing of curves. In *Proc. 33rd Intern. Symposium on Computational Geometry*, pages 37:1–37:16, 2017. `doi:10.4230/LIPIcs.SoCG.2017.37`.

8   S. Har-Peled, P. Indyk, and R. Motwani. Approximate nearest neighbor: Towards removing the curse of dimensionality. *Theory of Computing*, 8(1):321–350, 2012. `doi:10.4086/toc.2012.v008a014`.

9   S. Har-Peled and M. Mendel. Fast construction of nets in low dimensional metrics, and their applications. In *Proc. 21st Annual Symp. Computational Geometry*, SCG'05, pages 150–158, 2005. `doi:10.1145/1064092.1064117`.

10   P. Indyk. Approximate nearest neighbor algorithms for frechet distance via product metrics. In *Proc. 18th Annual Symp. on Computational Geometry*, SCG '02, pages 102–106, New York, NY, USA, 2002. ACM. `doi:10.1145/513400.513414`.

11   J. Matoušek. On variants of the Johnson-Lindenstrauss lemma. *Random Struct. Algorithms*, 33(2):142–156, 2008. `doi:10.1002/rsa.v33:2`.