Probabilistic embeddings of the Fréchet distance*

Anne Driemel[1] and Amer Krivošija[2]

1 **Department of Mathematics and Computer Science, TU Eindhoven, The Netherlands**
`a.driemel@tue.nl`

2 **Department of Computer Science, TU Dortmund, Germany**
`amer.krivosija@tu-dortmund.de`

Abstract

The Fréchet distance is a popular distance measure for curves which naturally lends itself to fundamental computational tasks, such as clustering, nearest-neighbor searching, and spherical range searching in the corresponding metric space. However, its inherent complexity poses considerable computational challenges in practice. To address this problem we study distortion of the probabilistic embedding that results from projecting the curves to a randomly chosen line. Such an embedding could be used in combination with, e.g. locality-sensitive hashing. We show that in the worst case and under reasonable assumptions, the discrete Fréchet distance between two polygonal curves in $\mathbb{R}^2$ or $\mathbb{R}^3$ of complexity t degrades by a factor linear in t with constant probability. We show upper and lower bounds on the distortion.

1 Introduction

The Fréchet distance is a distance measure for curves which naturally lends itself to fundamental computational tasks, such as clustering, nearest-neighbor searching, and spherical range searching in the corresponding metric space. However, their inherent complexity poses considerable computational challenges in practice. Indeed, spherical range searching under the Fréchet distance was recently the topic of the yearly ACM SIGSPATIAL GISCUP competition[1], highlighting the relevance and the difficulty of designing efficient data structures for this problem. At the same time, Afshani and Driemel show lower bounds on the space-query-tradeoff in the pointer model [1] that demonstrate that this problem is even harder than simplex-range searching.

The computational complexity of computing a single Fréchet distance between two given curves is a well-studied topic [2, 6–9, 12, 15]. It is believed that it takes time that is quadratic in the length of the curves and this running time can be achieved by applying dynamic programming. In this body of literature, the case of 1-dimensional curves under the continuous Fréchet distance stands out. In particular, no lower bounds are known on computing the continuous Fréchet distance between 1-dimensional curves. It has been observed that the problem has a special structure in this case [10]. Clustering under the Fréchet distance can be done efficiently for 1-dimensional curves [13], but seems to be harder for curves in the plane or higher dimensions. Bringmann and Künnemann use projections to lines to speed up their approximation algorithm for the Fréchet distance [8]. They show that the distance computation can be done in linear time, if the convex hulls of the two curves are disjoint. It is tempting to believe that the curves being restricted to 1-dimensional space makes the

* Driemel has been supported by NWO Veni project “Clustering time series and trajectories (10019853)”. Krivošija has been partly supported by DFG within the Collaborative Research Center SFB 876 “Providing Information by Resource-Constrained Analysis”, project A2.

[1] 6th ACM SIGSPATIAL GISCUP 2017, see also `http://sigspatial2017.sigspatial.org/giscup2017/`

34th European Workshop on Computational Geometry, Berlin, Germany, March 21–23, 2018.
This is an extended abstract of a presentation given at EuroCG’18. It has been made public for the benefit of the community and should be considered a preprint rather than a formally reviewed paper. Thus, this work is expected to appear eventually in more final form at a conference with formal proceedings and/or in a journal.

problem significantly easier. However, in the general case, there are no algorithms known which are faster for 1-dimensional curves than for curves in higher dimensions. In practice, it is very common to separate x and y components of trajectories to simplify computational tasks. It seems that in practice the inherent character of a trajectory is often largely preserved when restricted to one of the coordinates of the ambient space. Mathematically, this amounts to projecting the trajectory to a line.

This motivates our study of probabilistic embeddings of the Fréchet distance into the space of 1-dimensional curves. Concretely, we study distortion of the probabilistic embedding that results from projecting the curves to a randomly chosen line. Such a random projection could be used in combination with probabilistic data structures, e.g. locality-sensitive hashing [14], but also with the multi-level data structures for Fréchet range searching given by Afshani and Driemel [1]. See below for a more in-depth discussion of these data structures.

We show that in the worst case and under certain assumptions, the discrete Fréchet distance between two polygonal curves in $\mathbb{R}^2$ or $\mathbb{R}^3$ of complexity t degrades by a factor linear in t with constant probability. In particular, we show upper and lower bounds on the change in distance for the class of c-packed curves. The notion of the c-packed curves was introduced by Driemel, Har-Peled and Wenk in [12] and has proved useful as a realistic input assumption [3, 6, 11]. A curve is called c-packed for a value $c > 0$ if the length of the intersection of the curve with any ball of any radius r is a most cr. While our study is mostly restricted to the discrete Fréchet distance, we expect that our techniques can be extended to the case of the continuous Fréchet distance.

A closely related distance measure, which is popular in the field of data-mining, is dynamic time warping (DTW). The computational complexity of DTW has also been extensively studied, both empirically and in theory [3, 16]. Some of our lower bounds extend to DTW.

1.1 Related work

The work that is perhaps closest to ours is a recent result by Backurs and Sidiropoulos [4]. They gave an embedding of the Hausdorff distance into constant-dimensional ℓ_∞ space with constant distortion. More precisely, for any $s, d \geq 1$, they obtain an embedding for the Hausdorff distance over point sets of size s in d-dimensional space, into $\ell_\infty^{s^{O(s+d)}}$ with distortion $s^{O(s+d)}$. No such metric embeddings are known for the discrete or continuous Fréchet distance. It has been shown that the doubling dimension of the Fréchet distance is unbounded, even in the case when the metric spaces is restricted to curves of constant complexity [13]. A result of Bartal *et al.* [5] for doubling spaces implies that a metric embedding of the Fréchet distance into an ℓ_p space would have at least super-constant distortion, but it is not known how to find such an embedding.

The complexity of classic data structuring problems for the Fréchet distance is still not very well-understood, despite several papers on the topic. We review what is known for nearest-neighbor searching and range searching. Indyk [17] gave a deterministic and approximate near-neighbor data structure for the discrete Fréchet distance. Given n curves which have at most t vertices, this data structure achieves approximation factor $O(\log t + \log\log n)$ and has query time $O(\text{poly}(t) \log n)$. This data structure requires large space, as it precomputes all queries with curves with $\sqrt{t}$ vertices. For short curves (with $t \in O(\log n)$) Driemel and Silvestri [14] described an approximate near-neighbor structure based on locality-sensitive hashing with approximation factor $O(t)$, query time $O(t \log n)$, using space $O(n \log n + tn)$. LSH is a technique that uses families of hash functions with the property that near points are more likely to be hashed to the same index than far points. Driemel and Silvestri were the first to define locality-sensitive hash functions for the discrete Fréchet distance. No such hash functions are known for the continuous case. It is conceivable that the concept of signatures which was introduced by Driemel, Krivosija and Sohler [13] in the context of clustering of 1-dimensional curves could be used to define an LSH for the continuous case and that this technique could be used in combination with projections to random lines.

Afshani and Driemel recently showed how to leverage semi-algebraic range searching for this problem [1]. Their data structure also supports polygonal curves of low complexity and answers queries exactly. In particular, for the discrete Fréchet distance they describe a data structure which uses space in $O\big(n(\log\log n)^{t_s-1}\big)$ and achieves query time in $O\Big(n^{1-1/d}\cdot\log^{O(t_s)} n\cdot t_q^{O(d)}\Big)$, where t_s denotes the complexity of an input curve and it is assumed that the complexity of the query curves is upper-bounded by a polynomial of $\log n$. For the continuous Fréchet distance they describe a data structure for polygonal curves in the plane which uses space in $O\Big(n(\log\log n)^{O(t_s^2)}\Big)$ and achieves query time in $O\Big(\sqrt{n}\log^{O(t_s^2)} n\Big)$. For the case where the curves lie in dimension higher than 2 and the distance measure is the continuous Fréchet distance, no data structures for range searching or range counting are known.

1.2 Our results

Given two polygonal curves P and Q with t vertices each from $\mathbb{R}^2$ or $\mathbb{R}^3$. Consider sampling a unit vector $\mathbf{u}$ in $\mathbb{R}^2$ (resp. $\mathbb{R}^3$ if the curves lie in $\mathbb{R}^3$) uniformly at random and let P' and Q' be the projections of the two curves to the line supporting $\mathbf{u}$. We show that if the curves P and Q are c-packed for constant c, then, with constant probability, the discrete Fréchet distance between the curves P and Q degrades by at most a linear factor in t.

▶ **Theorem 1.1.** *For any two polygonal curves P and Q and for any $\gamma \in (0,1)$*

$$Pr\left[\frac{d_F(P,Q)}{d_F(P',Q')} \le \frac{12c+16}{\gamma}\cdot t\right] \ge 1-\gamma.$$

We also present a lower bound on the ratio of the two distances. The construction of the lower bound uses c-packed curves with $c < 3$.

▶ **Theorem 1.2.** *There exist polygonal curves P and Q, such that for any $\gamma \in (0, 1/\pi)$*

$$Pr\left[\frac{d_F(P,Q)}{d_F(P',Q')} \ge \frac{5\pi\gamma}{6}\cdot t\right] \ge 1-\gamma.$$

Theorem 1.2 holds for the continuous Fréchet distance and for dynamic time warping distance as well. We also show that there exist polygonal curves P and Q that are not c-packed for sublinear c and their (continuous or discrete) Fréchet distance degrades by a linear factor for any projection line (i.e. with probability 1).

2 Preliminaries

Throughout the paper we use the following notational conventions. Consider two polygonal curves $P = \{p_1, p_2, \dots, p_t\}$ and $Q = \{q_1, q_2, \dots, q_t\}$ in $\mathbb{R}^d$ given by their sequences of vertices. We choose a unit vector $\mathbf{u}$ in $\mathbb{R}^d$ by choosing a point on the $(d-1)$-dimensional unit hypersphere uniformly at random. We denote with L the line through the origin that supports the vector $\mathbf{u}$. Let $P' = \{p'_1, p'_2, \dots, p'_t\}$ and $Q' = \{q'_1, q'_2, \dots, q'_t\}$ be the projections of P and Q to L, defined by $p'_i = \langle p_i, \mathbf{u}\rangle$ and $q'_j = \langle q_j, \mathbf{u}\rangle$, for all $1 \le i \le t$ and $1 \le j \le t$. We denote $\delta_{i,j} = \|q_j - p_i\|$ and $\delta'_{i,j} = \|q'_j - p'_i\|$, for all $1 \le i \le t$ and $1 \le j \le t$, i.e. $\delta_{i,j}$ and

$\delta'_{i,j}$ are the pairwise distances of the vertices for the input curves P and Q and for their respective projections P' and Q'.

We define the discrete Fréchet distance of P and Q as follows: we call the *traversal* T of P and Q the sequence of pairs of indices (i, j) of vertices $(p_i, q_j) \in P \times Q$ such that

i) the traversal T starts with $(1, 1)$ and ends with (t, t), and

ii) the pair (i, j) of T can be followed only by one of $(i + 1, j)$, $(i, j + 1)$ or $(i + 1, j + 1)$.

We notice that every traversal is monotone. If $\mathcal{T}$ is the set of all traversals T of P and Q, then the discrete *Fréchet distance* between P and Q is defined as

$$d_F(P, Q) = \min_{T\in\mathcal{T}} \max_{(i,j)\in T} \|p_i - q_j\|. \tag{1}$$

Furthermore, we define a directed, vertex-weighted graph $G = (V, E)$ on the node set $V = \{(i, j) : 1 \le i, j \le t\}$. A node (i, j) corresponds to a pair of vertices p_i of P and q_j of Q and we assign it the weight $\delta_{i,j}$. The set of edges is defined as $E = \{((i, j),(i', j')) : i' \in \{i, i + 1\}, j' = \{j, j + 1\}, 1 \le i, i', j, j' \le t\}$. The set of paths in the graph G between $(1, 1)$ and (n, n) corresponds to the set of traversals $\mathcal{T}$. We call a path in G which does not start in $(1, 1)$ or end in (t, t) a *partial traversal* of P and Q.

It is useful to picture the nodes of the graph G as a matrix, where rows correspond to the vertices of P and columns correspond to the vertices of Q. For any fixed value $\Delta > 0$, we define the free-space matrix[2] $F_\Delta = (\phi_{i,j})_{1\le i,j\le t}$ with

$$\phi_{i,j} = \begin{cases} 1 & \text{if } \|q_j - p_i\| < \Delta \\ 0 & \text{if } \|q_j - p_i\| \ge \Delta. \end{cases}$$

Overlaying the graph with the free-space matrix for $\Delta > d_F(P, Q)$, we can observe that there exists a path in the graph from $(1, 1)$ to (t, t) that visits only the matrix entries with value 1. Moreover, the existence of such a path in the free-space matrix for some value of Δ implies that $\Delta > d_F(P, Q)$.

We define c-packedness of curves as follows.

▶ **Definition 2.1** (c-packed curve). Given $c > 0$, a curve $P \in \mathbb{R}^d$ is c-packed if for any point $p \in \mathbb{R}^d$ and any radius $r > 0$, the total length of the curve P inside the hypersphere $\texttt{ball}(p, r)$ is at most $c \cdot r$.

We prove the following basic fact about random projections to a line. For a general problem in $\mathbb{R}^d$ the probability bound degrades due to the measure concentration around $\pi/2$.

▶ **Lemma 2.2.** *If the line segment $\overline{pq}$ is projected to the straight line L, supported by the unit vector chosen uniformly at random on the unit hypersphere in $\mathbb{R}^2$ or $\mathbb{R}^3$, the probability that its length will be reduced by a factor greater than φ is at most φ.*

3 Upper bound

The discrete Fréchet distance between curves P and Q is realized by some pair (p_i, q_j) of vertices $p_i \in P$ and $q_j \in Q$, being at the distance $\|p_i - q_j\| = \delta$. We would like to apply Lemma 2.2 to this pair of vertices to show that the distance is preserved up to some constant factor. However, it is possible that the pairwise distances in the projection are such that a

[2] Note that the conventional definition of the free-space matrix for parameter Δ is slightly different, since usually there is an 1-entry iff $\|q_j - p_i\| \le \Delta$. We are using this definition since it better suits our needs.

cheaper traversal is possible that avoids the pair (p_i, q_j) altogether. Therefore, we apply the lemma to a subset of pairs of vertices of P and Q whose distance is large (e.g. larger than $\Delta = \delta/\theta$ for some small value of $\theta \geq 1$) and such that the chosen set forms a hitting set for the set of traversals $\mathcal{T}$. To this end we introduce the notion of the *guarding set*:

▶ **Definition 3.1** (Guarding set). For any two polygonal curves $P = \{p_1, \dots, p_t\}$ and $Q = \{q_1, \dots, q_t\}$ and a given parameter $\theta \geq 1$, a θ-guarding set $B \subseteq V$ for P and Q is a subset of the set of vertices of G that satisfies the following conditions:
a) (distance property) for all $(i, j) \in B$, it holds that $\delta_{i,j} \geq d_F(P, Q)/\theta$, and
b) (guarding property) for any traversal T of P and Q, it is $T \cap B \neq \emptyset$.

Note that the set B "guards" every traversal of P and Q in the sense that any path in G from $(1, 1)$ to (t, t) has non-empty intersection with B. In other words, B is a hitting set for the set of traversals $\mathcal{T}$. We can prove the following lemma using Lemma 2.2 for all elements of B in a union bound.

▶ **Lemma 3.2.** *Given parameter $\theta \geq 1$, if B is a θ-guarding set for the given curves $P = \{p_1, \dots, p_t\}$ and $Q = \{q_1, \dots, q_t\}$ from $\mathbb{R}^2$ or $\mathbb{R}^3$, and if P' and Q' are their projections to the straight line L, whose support unit vector $\mathbf{u}$ is chosen uniformly at random on the unit hypersphere, then for any $\beta > 1$ it holds that*

$$\frac{d_F(P', Q')}{d_F(P, Q)} \geq \frac{1}{\beta \cdot \theta \cdot |B|}$$

with positive constant probability at least $1 - 1/\beta$.

To show existence of a θ-guarding set B for any $\theta \geq 1$ we can construct such a set using breadth-first-search over graph G. Unfortunately, such built set B can have a quadratic number of elements in terms of the input size in the general case.

If the input curves P and Q are c-packed for some constant c, $c \geq 2$, then we construct the 1-guarding set B and modify it using the trimming operation based on Lemma 3.3. The idea is to trim the part of the graph G reachable by a partial traversal from $(1, 1)$ that does not pass through any of the vertices of B. See Figure 1 for an illustration.

▶ **Lemma 3.3.** *Given point p and a c-packed curve $Q = \{q_1, \dots, q_t\}$ from $\mathbb{R}^d$. Then for any value $b > 0$ there exists a value $r \in [b/2, b]$, such that the hypersphere centered at p with radius r intersects or is tangent to at most $2c$ edges of Q.*

$$F_b^{\text{before}} = \begin{pmatrix} \dots & & & & & \\ 0 & 0 & \boxed{0} & \boxed{0} & \boxed{0} & \dots \\ 0 & \boxed{0} & 1 & 1 & \boxed{0} & \dots \\ 0 & 1 & 1 & \boxed{0} & 0 & \dots \\ \boxed{0} & 1 & 1 & \boxed{0} & \boxed{0} & \dots \\ 1 & 1 & 1 & 1 & 1 & \dots \end{pmatrix} \qquad F_{b/2}^{\text{after}} = \begin{pmatrix} \dots & & & & & \\ 0 & 0 & \textcircled{0} & \textcircled{0} & \textcircled{0} & \dots \\ 0 & \textcircled{0} & 1 & 1 & \textcircled{0} & \dots \\ 0 & 1 & 1 & \textcircled{0} & 0 & \dots \\ \boxed{0} & \boxed{0} & \boxed{0} & \boxed{0} & \boxed{0} & \dots \\ 1 & 1 & 1 & 1 & 1 & \dots \end{pmatrix}$$

Figure 1 The elements of a guarding set (marked with boxes) before (left) and after (right) applying the trimming operation to the second row. The removed pairs are marked by circles

We call a pair $(i, j) \in B$ avoidable if there are two traversals of P and Q which guarantee that the pair (i, j) can be removed from the guarding set. Lemma 3.4 describes the algorithm to obtain a 4-guarding set whose size will be at most $(3c + 4) \cdot t$.

We omit discussion of our lower bounds due to space constraints.

▶ **Lemma 3.4.** *Let B be a 1-guarding set.*

(i) After the first phase of the algorithm, which removes all avoidable pairs, the modified set B is a 1-guarding set.

(ii) After the second phase of the algorithm, which applies the trimming operation to each row with $b = d_F(P, Q)$, the modified set B is a 2-guarding set.

(iii) After the third phase of the algorithm, which applies the trimming operation to each column with $b = d_F(P, Q)/2$, the modified set B is a 4-guarding set.

Acknowledgements. We want to thank Kevin Buchin for useful discussions on the topic of this paper.

References

1 P. Afshani and A. Driemel. On the complexity of range searching among curves. In *SODA*, pages 898–917, 2018.

2 P. K. Agarwal, R. Ben Avraham, H. Kaplan, and M. Sharir. Computing the discrete Fréchet distance in subquadratic time. *SIAM Journal of Computing*, 43(2):429–449, 2014.

3 P. K. Agarwal, K. Fox, J. Pan, and R. Ying. Approximating Dynamic Time Warping and Edit Distance for a Pair of Point Sequences. In *SoCG*, volume 51, pages 6:1–6:16, 2016.

4 A. Backurs and A. Sidiropoulos. Constant-distortion embeddings of Hausdorff metrics into constant-dimensional l_p spaces. In *APPROX/RANDOM*, pages 1:1–1:15, 2016.

5 Y. Bartal, L. Gottlieb, and O. Neiman. On the impossibility of dimension reduction for doubling subsets of lp. In *SoCG*, pages 60–66, 2014.

6 K. Bringmann. Why walking the dog takes time: Fréchet distance has no strongly subquadratic algorithms unless SETH fails. In *FOCS*, pages 661–670, 2014.

7 K. Bringmann and M. Künnemann. Quadratic conditional lower bounds for string problems and dynamic time warping. In *FOCS*, pages 79–97, 2015.

8 K. Bringmann and M. Künnemann. Improved approximation for Fréchet distance on c-packed curves matching conditional lower bounds. *IJCGA*, 27(1-2):85–120, 2017.

9 K. Buchin, M. Buchin, W. Meulemans, and W. Mulzer. Four Soviets walk the dog-with an application to Alt's conjecture. In *SODA*, pages 1399–1413, 2014.

10 K. Buchin, J. Chun, M. Löffler, A. Markovic, W. Meulemans, Y. Okamoto, and T. Shiitada. Folding free-space diagrams: Computing the Fréchet distance between 1-dimensional curves (multimedia contribution). In *SoCG*, pages 64:1–64:5, 2017.

11 A. Driemel and S. Har-Peled. Jaywalking your dog – computing the Fréchet distance with shortcuts. *SIAM Journal of Computing*, 42(5):1830–1866, 2013.

12 A. Driemel, S. Har-Peled, and C. Wenk. Approximating the Fréchet distance for realistic curves in near-linear time. *Discrete & Computational Geometry*, 48(1):94–127, 2012.

13 A. Driemel, A. Krivošija, and C. Sohler. Clustering time series under the Fréchet distance. In *SODA*, pages 766–785, 2016.

14 A. Driemel and F. Silvestri. Locally-sensitive hashing of curves. In *SoCG*, pages 37:1–37:16, 2017.

15 T. Eiter and H. Mannila. Computing discrete Fréchet distance. Technical Report CD-TR 94/64, Christian Doppler Laboratory, 1994.

16 O. Gold and M. Sharir. Dynamic time warping and geometric edit distance: Breaking the quadratic barrier. In *ICALP*, pages 25:1–25:14, 2017.

17 P. Indyk. Approximate nearest neighbor algorithms for Fréchet distance via product metrics. In *SoCG*, pages 102–106, 2002.