On the hardness of computing an average curve

On the hardness of computing an average curve

Kevin Buchin
Department of Mathematics and Computing Science, TU Eindhoven, The Netherlands
k.a.buchin@tue.nl

Anne Driemel
University of Bonn, Hausdorff Center for Mathematics, Bonn, Germany
driemel@cs.uni-bonn.de

Martijn Struijs
Department of Mathematics and Computing Science, TU Eindhoven, The Netherlands
m.a.c.struijs@tue.nl

Abstract

We study the complexity of clustering curves under k-median and k-center objectives in the metric space of the Fréchet distance and related distance measures. The k-center problem has recently been shown to be NP-hard, even in the case where $k = 1$, i.e. the minimum enclosing ball under the Fréchet distance. We extend these results by showing that also the k-median problem is NP-hard for $k = 1$. Furthermore, we show that the 1-median problem is W[1]-hard with the number of curves as parameter. We show this under the discrete and continuous Fréchet and Dynamic Time Warping (DTW) distance. Our result generalizes an earlier result by Bulteau et al. from 2018 for a variant of DTW that uses squared distances. Moreover, closing some gaps in the literature, we show positive results for a variant where the center curve may have complexity at most ℓ under the discrete Fréchet distance. In particular, for fixed k, ℓ and ε, we give $(1+\varepsilon)$-approximation algorithms for the (k, ℓ)-median and (k, ℓ)-center objectives and a polynomial-time exact algorithm for the (k, ℓ)-center objective.

2012 ACM Subject Classification Theory of computation → Computational geometry; Theory of computation → Problems, reductions and completeness

Keywords and phrases Curves, Clustering, Algorithms, Hardness, Approximation

1 Introduction

Clustering is an important tool in data analysis, used to split data into groups of similar objects. Their dissimilarity is often based on distance between points in Euclidean space. However, the dissimilarity of (polygonal) curves is more accurately measured by specialised measures: Dynamic Time Warping (DTW) [18], continuous and discrete Fréchet distance [1, 8].

We focus on *centroid-based clustering*, where each cluster has a centre curve and the quality of the clustering is based on the similarity between the centre and the elements inside the cluster. In particular, given a distance measure δ, we consider the following problems:

▷ Problem 1 (k-**median for curves with distance** δ). Given a set $\mathcal{G} = \{g_1, \ldots, g_m\}$ of polygonal curves, find a set $\mathcal{C} = \{c_1, \ldots c_k\}$ of polygonal curves that minimizes $\sum_{g \in \mathcal{G}} \min_{i=1}^{k} \delta(c_i, g)$.

▷ Problem 2 (k-**center for curves with distance** δ). Given a set $\mathcal{G} = \{g_1, \ldots, g_m\}$ of polygonal curves, find a set $\mathcal{C} = \{c_1, \ldots c_k\}$ of polygonal curves that minimizes $\max_{g \in \mathcal{G}} \min_{i=1}^{k} \delta(c_i, g)$.

The k-median problem for general metric spaces is well studied, often in the context of the closely related facility location problem [14, 13, 15]. Usually, they consider what we will call the *discrete k-median problem*, where the centers must be selected from the set of input points. In this paper, we are interested in the *unconstrained k-median problem*, where the centers can be any element of the metric space. Often, we will simply write k-median problem

to denote the unconstrained version. For approximation in metric spaces, the two versions are closely related: any set of curves that realises an α-approximation for the discrete k-median problem realises an 2α-approximation for the unconstrained k-median problem. There is an elegant $O(kn)$ time 2-approximation algorithm for the k-center problem in metric spaces [10]. This approximation factor is tight for clustering curves under the discrete and continuous Fréchet distance [4]. Finding approximate solutions for k-median is more challenging: the best known polynomial-time approximation algorithm for discrete k-median in general metric space achieves a factor of $3+\varepsilon$ for any $\varepsilon > 0$ [2] and it is NP-hard to achieve an approximation factor of $1 + 2/e$ [13]. Since we still want efficient algorithms to do curve clustering, we look at a variant of these problems: we only look for centre curves with at most a fixed complexity, denoted by ℓ. More formally, the (k, ℓ)-*center problem* is to find a set of curves $\mathcal{C} = \{c_1, \ldots c_k\}$, each of complexity at most ℓ, that minimizes $\max_{g \in \mathcal{G}} \min_{i=1}^{k} \delta(c_i, g)$. The (k, ℓ)-*median problem* is defined analogously. Finding short centre curves is also useful for applications, as it can prevent overfitting the centre to details of individual input curves. Although the general case for this variant is still NP-hard, we can find efficient algorithms when k and ℓ are fixed. The (k, ℓ)-center and (k, ℓ)-median problems were introduced by Driemel et al. [7], who obtained an $\widetilde{O}(mn)$-time $(1 + \varepsilon)$-approximation algorithm for the (k, ℓ)-center and (k, ℓ)-median problem under the Fréchet distance for curves in 1D, assuming k, ℓ, ε are constant. In [4], Buchin et al. gave polynomial-time constant-factor approximation algorithms for the (k, ℓ)-center problem under the discrete and continuous Fréchet distance for curves in arbitrary dimension.

1.1 The average curve problem.

We call the 1-median problem the *average curve* problem. Denote the set of all warping paths (or alignments, see also [18]) between curves x and y by $\mathcal{W}_{x,y}$. For any integers $p, q \geq 1$, we define the Dynamic Time Warping Distance as

$$\mathrm{DTW}_p^q(x, y) := \left(\min_{W \in \mathcal{W}_{x,y}} \sum_{(i,j) \in W} |x_i - y_j|^p \right)^{q/p}.$$

Similarly, define the discrete Fréchet distance as

$$\mathrm{d}_{dF}(x, y) := \min_{W \in \mathcal{W}_{x,y}} \max_{(i,j) \in W} |x_i - y_j|.$$

For the continuous version of the Fréchet distance, we refer to [1]. The average curve problem for the (p, q)-DTW distance is to answer for a given set of curves $\mathcal{G}$ and a real number r, whether there exists a center curve c such that $\sum_{g \in \mathcal{G}} \mathrm{DTW}_p^q(c, g) \leq r$.

While clustering on points for general k in the plane or higher dimension is often NP-hard [17], some point clustering problems can be solved efficiently when $k = 1$ in low dimension. In contrast, clustering curves tends to be hard even when $k = 1$ and the curves lie in 1D. For instance, the 1-center problem in the plane can be solved in linear time [16], and there are practical algorithms for higher dimensional Euclidean space [9]. Buchin et. al. [4] show that the 1-center problem for the discrete and continuous Fréchet distance in 1D is NP-hard and that for the discrete Fréchet distance, it is NP-hard to approximate with any ratio strictly smaller than 2.

The average curve problem for the $(2, 2)$-DTW distance has resisted efficient algorithms so far, which motivated several heuristic approaches [11, 12, 18]. Brill et al. showed that dynamic programming yields an exponential-time exact algorithm [3]. A formal proof of

Table 1 Overview of results. In these tables, n denotes the length of the input curves, m denotes the number of input curves and d denotes the ambient dimension of the curves.

(a) Results on exact computation.

Problem	Result	Restrictions	Reference
1-median, DTW_p^q	$O(n^{2m+1}2^m m)$ $O(mn^3)$	$d=1$ Binary	Brill et al. [3]
	NP-hard W[1]-hard in m	$p=q=2$	Bulteau et al. [5]
	NP-hard W[1]-hard in m	$p,q\in\mathbb{N}$	Theorem 7
1-median, Fréchet	NP-hard W[1]-hard in m		Theorem 4
1-center, discrete Fréchet	NP-hard		Buchin et al. [4]
(k,ℓ)-center, discrete Fréchet	$O((mn)^{2k\ell}k\ell m\log(mn))$	$d\leq 2$	Theorem 12

(b) Approximation algorithms. (In stating the running times we assume k and ℓ are constants independent of n and m.)

Problem	Result	Approximation factor	Restrictions	Reference
(k,ℓ)-median, continuous Fréchet	$\widetilde{O}(nm)$	$(1+\varepsilon)$	$d=1$	Driemel et al. [7]
(k,ℓ)-median, discrete Fréchet	$\widetilde{O}(nm)$ $\widetilde{O}(nm^2)$ $\widetilde{O}(nm^{dk\ell+1})$	$O(1)$ $(1+\varepsilon)$ $(1+\varepsilon)$	 $k=1$ $k>1$	Driemel et al. [7] Theorem 11 Theorem 11
(k,ℓ)-center, discrete Fréchet	$\widetilde{O}(nm)$ $\widetilde{O}(nm)$	3 $(1+\varepsilon)$		Buchin et al. [4] Theorem 10

NP-hardness has only recently been given by Bulteau et. al. [5], who additionally show the $(2,2)$-DTW problem is W[1]-hard when parametrised in the number of input curves m and there exists no $f(m)\cdot n^{o(m)}$-time algorithm unless the ETH fails.

1.2 Our results

We show that the average curve problem for discrete and continuous Fréchet distance in 1D is NP-complete and W[1]-hard when parametrised in the number of curves m. In addition, we prove hardness of the average curve problem for the (p,q)-DTW distance for any $p,q\in\mathbb{N}$, thereby generalizing the result by Bulteau et. al. [5]. Our proof uses a different reduction than theirs. Moreover, closing some gaps in the literature, we give a $(1+\varepsilon)$-approximation algorithm that runs in $\widetilde{O}(mn)$ time and a polynomial-time exact algorithm to solve the (k,ℓ)-center problem for the discrete Fréchet distance, when k,ℓ and ε are fixed. Additionally, we show that the $(1+\varepsilon)$-approximation algorithm can be adapted to the (k,ℓ)-median case, where we get an improved running time for $k=1$. Table 1 gives an overview of our results.

2 Hardness of the average curve problem for discrete and continuous Fréchet

In this section, we will show that the 1-median problem (or average curve problem) is NP-hard for the discrete and continuous Fréchet distance. The average curve problem for the discrete Fréchet distance is as follows: given a set of curves $\mathcal{G}$ and an integer r, determine whether there exists a center curve c such that $\sum_{g\in\mathcal{G}} \mathrm{d}_{dF}(c,g) \leq r$. We will show that this problem is NP-hard. To find a reasonable algorithm, we can look at a parametrised version of the problem. A natural parameter is the number of input curves, which we will denote by m. However, we will show that this parametrised problem is W[1]-hard, which rules out any $f(m)\cdot n^{O(1)}$-time algorithm, unless FPT = W[1]. To achieve these reductions, we create a reduction from a variant of the shortest common supersequence (SCS) problem.

2.1 The FCCS problem

To show the hardness of the average curve problem for the Fréchet and DTW distance, we reduce from a variant of the *Shortest Common Supersequence* (SCS) problem, which we will call the *Fixed Character Common Supersequence* (FCCS) problem. If s is a string and x is a character, $\#_x(s)$ denotes the number of occurrences of x in s.

▷ Problem 3 (**Shortest Common Supersequence (SCS)**). Given a set S of m strings with length at most n over the alphabet Σ and an integer t, does there exists a string s^* of length t that is a supersequence of each string $s \in S$?

▷ Problem 4 (**Fixed Character Common Supersequence (FCCS)**). Given a set S of m strings with length at most n over the alphabet $\Sigma = \{A, B\}$ and $i, j \in \mathbb{N}$, does there exists a string s^* with $\#_A(s^*) = i$ and $\#_B(s^*) = j$ that is a supersequence of each string $s \in S$?

The SCS problem with a binary alphabet is known to be NP-hard [20] and $W[1]$-hard [19]. The same holds for our variant:

▶ **Lemma 1.** *The FCCS problem is NP-hard. The FCCS problem with m as parameter is W[1]-hard. There exists no $f(m)\cdot n^{o(m)}$ time algorithm for FCCS unless ETH fails.*

Proof. We reduce from SCS with the binary alphabet $\{A, B\}$ to FCCS. Given an instance (S, t) of SCS, construct $S' = \{s + AB^{2t}A + c(s) \mid s \in S\}$, where $c(s)$ denotes the string constructed by replacing all A characters in s by B and vice versa. We reduce to the instance $(S', t+2, 3t)$ of FCCS and claim that (S, t) is a true instance of SCS if and only if $(S', t+2, 3t)$ is a true instance of FCCS.

If (S, t) is a true instance of SCS, then there exists a string q of length t that is a supersequence of each string in S. Therefore, the string $q' = q+AB^{2t}+c(q)$ is a supersequence of all strings in S'. Since $\#_A(q') = 2+\#_A(q+c(q)) = 2+t$ and $\#_B(q') = 2t+\#_B(q+c(q)) = 3t$, $(S', t+2, 3t)$ is a true instance of FCCS.

If $(S', t+2, 3t)$ is a true instance of FCCS, there is string q' with $\#_A(q') = t+2$ and $\#_B(q') = 3t$ that is a supersequence of each string $s' \in S'$. Consider a pair of strings $s'_1 = s_1 + AB^{2t}A + c(s_1)$ and $s'_2 = s_2 + AB^{2t}A + c(s_2)$ from S'. If there is no matching such that the first character of the AB^{2t} substring in s'_1 is matched to the same character of q' as the first character of that substring in s'_1, then q' is a supersequence of $AB^{2t}AB^{2t}A$ and so $\#_B(q') > 3t$, a contradiction. By symmetry, the same holds for the last character of the substring $AB^{2t}A$ and therefore $q = q_1 + q_2 + q_3$, where q_1 is a supersequence of S, q_2 is a supersequence of $AB^{2t}A$ and q_3 is a supersequence of $\{c(s) \mid s \in S\}$. Note

that $c(q_3)$ is a supersequence of S. Also, $\#_A(q_1 + c(q_3)) = \#_A(q) - \#_A(q_2) \leq t$ and $\#_B(q_1 + c(q_3)) = \#_B(q) - \#_B(q_2) \leq t$. So, $|q_1| + |c(q_3)| \leq 2t$, which means that $|q_1| \leq t$ or $|c(q_3)| \leq t$ and thus (S, t) is a true instance of SCS.

Note that this reduction is both a polynomial-time reduction and a parametrised reduction in the parameter m. Since the SCS problem over the binary alphabet $\{A, B\}$ is NP-hard [20] and W[1]-hard when parametrised with the number of strings m [19], the first two parts of the claim follow. The final part of the claim follows from the fact that this reduction together with the reduction from [19] give a parametrised reduction from Clique that is linear in the parameter. ◀

2.2 Complexity of the average curve problem under the discrete and continuous Fréchet distance

We will show the hardness of finding the average curve under the discrete and continuous Fréchet distance is via the following reduction from FCCS. Given an instance (S, i, j) of FCCS, we construct a set of curves using the following vertices in $\mathbb{R}$: $g_a = -1$, $g_b = 1$, $g_A = -3$, and $g_B = 3$. For each string $s \in S$, we map each character to a subcurve in $\mathbb{R}$:

$$A \to (g_a g_b)^{i+j} g_A (g_a g_b)^{i+j} \qquad B \to (g_b g_a)^{i+j} g_B (g_b g_a)^{i+j}.$$

The curve $\gamma(s)$ is constructed by concatenating the subcurves resulting from this mapping, $G = \{\gamma(s) \mid s \in S\}$ denotes the set of these curves. Additionally, we use the curves

$$A_i = g_b (g_A g_b)^i \qquad B_j = g_a (g_B g_a)^j.$$

We will call subcurves containing only g_A or g_B vertices *letter gadgets* and subcurves containing only g_a or g_b vertices *buffer gadgets*. Let $R_{i,j}$ contain curves A_i and B_j, both with multiplicity $\alpha = |S|(|S| - 1) + 1$. We reduce to the instance $(G \cup R_{i,j}, r)$ of the average curve problem, where $r = |S| + 2\alpha$. We use the same construction for the discrete and continuous case.

For an example of this construction, take $S = \{ABB, BBA, ABA\}$, $i = 2$, $j = 2$. Then $ABBA$ is a supersequence of S with the correct number of characters. Note that the curve with vertices $0 g_A 0 g_B 0 g_B 0 g_A 0$ has a (discrete) Fréchet distance of at most 1 to the curves in $G \cup R_{i,j}$, see Figure 1, so the sum of those distances is at most $|S| + 2\alpha = r$.

▶ **Lemma 2.** *If (S, i, j) is a true instance of FCCS, then $(G \cup R_{i,j}, r)$ is a true instance of the average curve problem for discrete and continuous Fréchet.*

Proof. We will show the proof for the discrete Fréchet distance. Since the discrete Fréchet distance is an upper bound of the continuous version, this proves the continuous case as well.

Since (S, i, j) is a true instance of FCCS, there exists a common supersequence s^* of S with $\#_A(s^*) = i$ and $\#_B(s^*) = j$. Construct the curve c of complexity $2|s^*| + 1$, given by

$$c_l = \begin{cases} 0 & \text{if } l \text{ is odd} \\ -2 & \text{if } l \text{ is even and } s^*_{l/2} = A \\ 2 & \text{if } l \text{ is even and } s^*_{l/2} = B \end{cases},$$

for each $l \in \{1, \ldots, 2|s^*| + 1\}$. Let $s \in S$, then note that the sequence of letter gadgets in $\gamma(s)$ is a subsequence of the letter gadgets in c, because s is a subsequence of s^*. So, all letter gadgets in $\gamma(s)$ can be matched with a letter gadget in c, the remaining letter gadgets in c with a buffer gadget in $\gamma(s)$ and all remaining buffer gadgets with another buffer gadget,

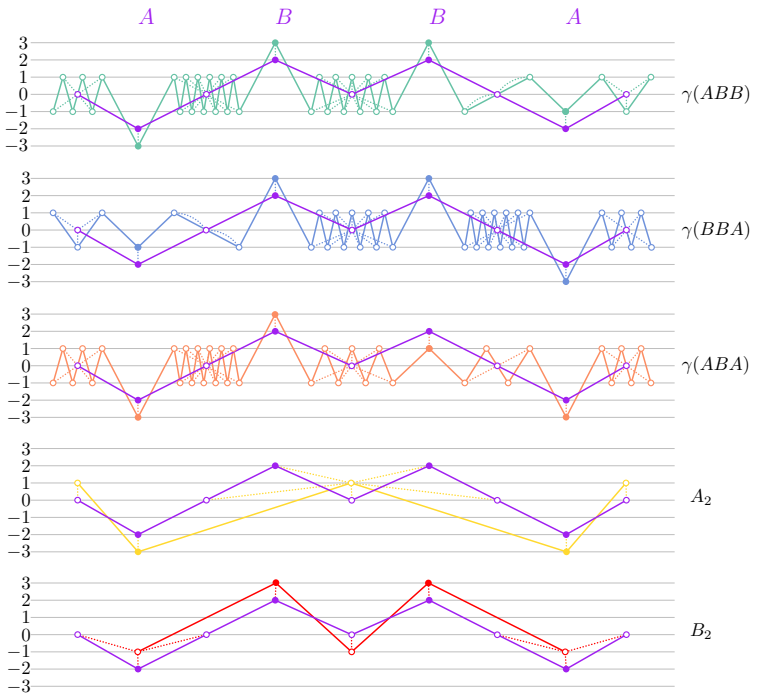

Figure 1 Five curves from $G \cup R_{i,j}$ in the reduction for the Fréchet average curve problem and a center curve constructed from $ABBA$ (purple) as in Lemma 2. Matchings are indicated by dotted lines. Note that each of these matchings achieves a (discrete) Fréchet distance of 1.

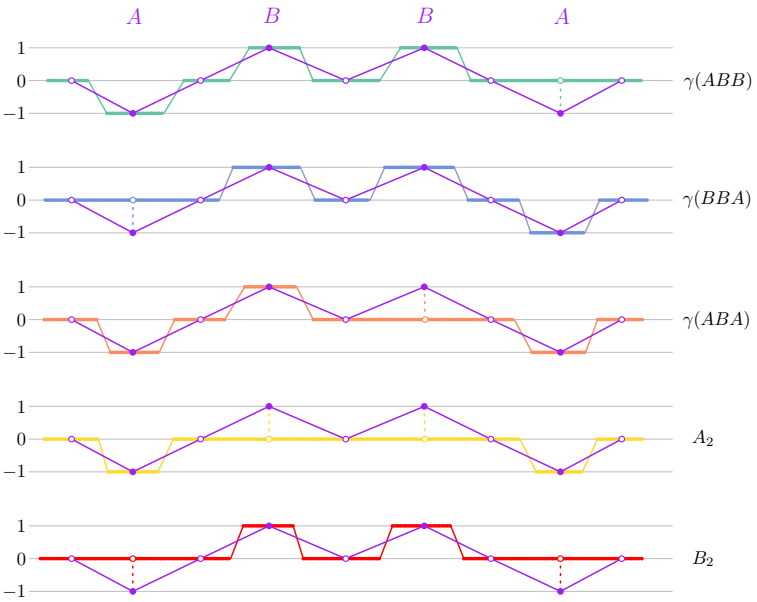

Figure 2 Five curves from $G \cup R_{i,j}$ in the reduction for the DTW average curve problem and a center curve constructed from the string $ABBA$ (purple) as in Lemma 5. Fat horizontal lines indicate β consecutive vertices. Vertices that match at distance 0 touch, vertices that match at distance 1 are indicated by dotted lines. The center has 1 mismatch with the first 3 curves and 2 with the final two, so the total cost here is $3 \cdot (1^p)^{q/p} + 2\alpha \cdot (2 \cdot 1^p)^{q/p} = 3 + 2\alpha \cdot 2^q$

such that $\mathrm{d}_{dF}(c, \gamma(s)) \leq 1$. For the matching with A_i, note that c has exactly i g_A vertices, so these can be matched with the i g_A vertices in A_i. All other vertices in c have distance 1 to the remaining buffer gadgets in A_i, so $\mathrm{d}_{dF}(c, A_i) \leq 1$. Analogously, $\mathrm{d}_{dF}(c, B_j) \leq 1$. So, we get $\sum_{g \in G \cup R_{i,j}} \mathrm{d}_{dF}(c, g) = \sum_{s \in S} \mathrm{d}_{dF}(c, \gamma(s)) + \alpha(\mathrm{d}_{dF}(c, A_i) + \mathrm{d}_{dF}(c, B_j)) \leq |S| + 2\alpha = r$, and $(G \cup R_{i,j}, r)$ is a true instance of average curve for discrete Fréchet. ◀

▶ **Lemma 3.** *If $(G \cup R_{i,j}, r)$ is a true instance of the average curve problem for discrete and continuous Fréchet, then (S, i, j) is a true instance of FCCS.*

Proof. We will show the proof for the continuous Fréchet distance. Since the continuous Fréchet distance is a lower bound of the discrete version, this proves the discrete case as well.

Since $(G \cup R_{i,j}, r)$ is a true instance of the average curve problem for continuous Fréchet, there exists a curve c^* such that $\sum_{g \in G \cup R_{i,j}} \mathrm{d}_F(c^*, g) \leq r = |S| + 2\alpha$. We start by deriving bounds for distance between c^* and the individual curves in $G \cup R_{i,j}$. Without loss of generality, we can assume that c^* has same matching among all copies of A_i and B_j. So, the previous bound gives $\mathrm{d}_F(c^*, A_i) + \mathrm{d}_F(c^*, B_j) \leq 2 + \varepsilon$, where $\varepsilon = |S|/\alpha$. Note that $\mathrm{d}_F(c^*, A_i) \geq 1$: if not, then for all vertices v of c^* we have $|v - g_A| < 1$ or $|v - g_b|$, so $|p - g_B| > 1$ for each point p on c^*. We can assume without loss of generality that each string in S contains at least one B character (if there is a string s with only A characters, any supersequence with the correct number of A-characters is a supersequence of s when $|s| \leq i$ and none when $|s| > i$), which means $\mathrm{d}_F(c^*, \gamma(s)) > 1$ for any $s \in S$. As $\mathrm{d}_F(c^*, A_i) + \mathrm{d}_F(c^*, B_j) \geq \mathrm{d}_F(A_i, B_j) = 2$, we get $r \geq \sum_{g \in G \cup R_{i,j}} \mathrm{d}_F(c^*, g) > |S| + 2\alpha = r$, a contradiction. So it indeed holds that $\mathrm{d}_F(c^*, A_i) \geq 1$ and analogously, $\mathrm{d}_F(c^*, A_i) \geq 1$ as well.

Using the previous upper bound of $2 + \varepsilon$ for their sum, it follows that $\mathrm{d}_F(c^*, A_i) \leq 1 + \varepsilon$ and $\mathrm{d}_F(c^*, B_j) \leq 1 + \varepsilon$. This means that for each point p on c^*, $|p| \leq 2 + \varepsilon$ (otherwise, p has distance at least $1 + \varepsilon$ to all points on A_i or B_j), so $\mathrm{d}_F(c^*, \gamma(s)) \geq 1 - \varepsilon$ for all $s \in S$, since we can assume s contains at least one A and B character. Therefore, $\mathrm{d}_F(\gamma(s), c^*) \leq r - \alpha(\mathrm{d}_F(A_i, c^*) + \mathrm{d}_F(B_j, c^*)) - \sum_{s' \in S \setminus \{s\}} \mathrm{d}_F(\gamma(s'), c^*) \leq |S| - \sum_{s' \in S \setminus \{s\}} \mathrm{d}_F(\gamma(s'), c^*) \leq |S| - (|S| - 1)(1 - \varepsilon) = 1 + (|S| - 1)\varepsilon < 2$ for all $s \in S$. In summary, we have shown $\mathrm{d}_F(c^*, g) < 2$ for all $g \in G \cup R_{i,j}$.

We partition c^* into subcurves based on the location in $\mathbb{R}$ of each point on this curve. We call each maximal contiguous subcurve with $p < -1$ for each p on the subcurve an A-part, a maximal contiguous subcurve with $p > 1$ a B-part and all maximal contiguous subcurves with $|p| < 1$ a buffer part. Note that since $\mathrm{d}_F(c^*, A_i) < 2$, there are exactly i A-parts and since $\mathrm{d}_F(c^*, B_j) < 2$, there are exactly j B-parts. Construct the string s' from the sequence of A/B-parts in c^* by mapping A-parts to A characters and B-parts to B characters. So, by construction, $\#_A(s') = i$ and $\#_B(s') = j$. Let $s \in S$. Since $\mathrm{d}_F(\gamma(s), c^*) < 2$, each letter gadget in $\gamma(s)$ is matched to (part of) a corresponding A or B-part and each A or B-part is matched to at most one letter gadget.Furthermore, matchings are monotone, so the sequence of letter gadgets in $\gamma(s)$ is a subsequence of the sequence of A/B-parts when we identify A-parts with A-gadgets and B-parts with B-gadgets. So, s is a subsequence of s', which means s' is a supersequence of S with $\#_A(s') = i$ and $\#_B(s') = j$. ◀

Since the reduction above runs in polynomial time and FCCS is NP-hard (1), the 1-median problem for discrete Fréchet is NP-hard. Note that the number of curves in the reduced 1-median instance is $k + 2k(k-1) + 2$, where k is the number of input sequences of the FCCS instance. So, this reduction is also a parametrised reduction from FCCS with the number of sequences as parameter to the 1-median problem for discrete Fréchet with the number of curves as a parameter. As FCCS with the number of sequences as parameter is W[1]-hard,

1-median for discrete Fréchet with the number of curves as a parameter is W[1]-hard. We summarize these results in the following theorem:

▶ **Theorem 4.** *The average curve problem for discrete and continuous Fréchet distance is NP-hard. When parametrised in the number of input curves m, this problem is W[1]-hard.*

3 Hardness of DTW average problems

In this section, we will show that some variants of the average curve problem for the DTW distance are computationally hard.

3.1 Hardness of the (p, q)-DTW average curve problem

We will show that the average curve problem under the (p, q)-DTW distance is NP-hard for all $p, q \in \mathbb{N}$. This generalises the result of [5], who use different methods to achieve the same hardness results for the $(2, 2)$-DTW average curve problem only. We again reduce from FCCS instance (S, i, j). Given a string $s \in S$ over the binary alphabet $\{A, B\}$, we map each character to a subcurve in $\mathbb{R}$:

$$A \to g_0^\beta g_a^\beta g_0^\beta \qquad B \to g_0^\beta g_b^\beta g_0^\beta,$$

where $g_0 = 0, g_a = -1, g_b = 1$ as before and β is a large constant that will be determined later. The curve $\gamma(s)$ is constructed by concatenating these subcurves and $G = \{\gamma(s) \mid s \in S\}$. We additionally use the curves

$$A_i = g_0^\beta (g_a^\beta g_0^\beta)^i \qquad B_j = g_0^\beta (g_b^\beta g_0^\beta)^j.$$

Call any subcurve consisting of g_a or g_b vertices a letter gadget and any subcurve consisting of g_0 a buffer gadget. Let $R_{i,j}$ contain curves A_i and B_j, both with multiplicity α. We reduce to the instance $(G \cup R_{i,j}, r)$ of (p, q)-DTW average curve, where $r = \sum_{s \in S} (i + j - |s|)^{q/p} + \alpha(i^{q/p} + j^{q/p})$, $\beta = \lceil r/\varepsilon^q \rceil + 1$, $\alpha = |S|$ and $\varepsilon = 1 - (1 - \min_{x \in \{i,j\}} \frac{(x+1)^{q/p} - x^{q/p}}{4(i+j)^{q/p}})^{1/q}$. See Figure 2 for an example of this construction with $S = \{ABB, BBA, ABA\}$ and $i = j = 2$.

▶ **Lemma 5.** *If (S, i, j) is a true instance of FCCS, then $(G \cup R_{i,j}, r)$ is a true instance of (p, q)-DTW average curve.*

Proof. If (S, i, j) is a true instance of FCCS, then there exists a string s^* that is a supersequence of S, with $\#_A(s^*) = i$ and $\#_B(s^*) = j$. Construct the curve c of length $2(i + j) + 1$:

$$c_l = \begin{cases} 0 & \text{if } l \text{ is odd} \\ g_a & \text{if } l \text{ is even and } s^*_{l/2} = A \\ g_b & \text{if } l \text{ is even and } s^*_{l/2} = B \end{cases},$$

for each $l \in \{1, \ldots, 2(i + j) + 1\}$. Analogously to Lemma 2, we can match the letter gadgets from $\gamma(s)$ to g_A or g_B in c as s^* is a supersequence of s, the letter gadgets of A_i, B_j to g_A, g_B in c as the number of curves match, and g_0 vertices to buffer gadgets. This gives a matching such that $\sum_{g \in G \cup R_{i,j}} \mathrm{DTW}_p^q(c, g) \leq r$. ◀

To help prove the next lemma, we introduce some terminology. Take a vertex p on some curve c^*. If $|p - g_a| < \varepsilon$, we call p an *A-signal vertex*. If $|p - g_b| < \varepsilon$ we call p an *B-signal*

vertex. If p is not a signal vertex, then we call p a *buffer vertex*. Note that ε is chosen small enough such that no vertex is both an A- and B-signal vertex. We will show that the sequence of signal vertices in the curve satisfying $(G \cup R_{i,j}, r)$ is a supersequence satisfying (S, i, j).

▶ **Lemma 6.** *If $(G \cup R_{i,j}, r)$ is a true instance of (p,q)-DTW average curve, then (S, i, j) is a true instance of FCCS.*

Proof. If $(G \cup R_{i,j}, r)$ is a true instance of (p,q)-DTW average curve, then there exists a curve c^* such that $\sum_{g \in G \cup R_{i,j}} \mathrm{DTW}_p^q(c^*, g) \leq r$. Take a curve $g \in G \cup R_{i,j}$. First note that there is at least one signal vertex in c^* matched to each letter gadget in g: otherwise, matching all β vertices in the gadget costs at least $\varepsilon^q \cdot \beta = \varepsilon^q \cdot (r/\varepsilon^q + 1) > r$, which contradicts the choice of c^*. Similarly, each signal vertex is matched to at most one letter gadget in g, since otherwise it would have to match a g_0^β subcurve in between the letter gadgets, which would have a cost of at least $(1-\varepsilon)^q \cdot \beta > \varepsilon^q \cdot \beta > r$. This means that the sequence of letter gadgets in $\gamma(s)$ is a subsequence of the sequence of signal vertices in c^*. So, if we construct s' from the sequence of signal vertices in c^* by mapping A-signal vertices to A characters and B-signal vertices to B characters, we have that s' is a supersequence of S. What remains to be proven is that $\#_A(s') = i$ and $\#_B(s') = j$, i.e. there are exactly i A-signal vertices and j B-signal vertices.

First, note that the sequence of A letter gadgets in A_i is a subsequence of the sequence of signal vertices in c^* (using the same argument as above), so there are at least i A-signal vertices. Analogously, there are at least j B-signal vertices. Now if we can show that there are at most $i + j$ signal vertices, then we are done.

Observe that there is at least one buffer vertex within a distance ε to g_0 in between signal vertices that are matched to letter gadget in A_i or B_j, as such a vertex must cover a g_0^β subcurve between the letter gadgets. We call signal vertices that are matched to the same letter gadget in either A_i or B_j a group. (Note that by definition, a signal vertex cannot be matched to letter gadgets in both A_i and B_j) This means that there are at least i groups of A-signal vertices and at least j groups of B-signal vertices.

When matching c^* and $\gamma(s)$ for some $s \in S$, we can only match at most $|s|$ groups of signal vertices to a g_a or g_b vertex in a letter gadget in $\gamma(s)$. So, for the at least $i + j - |s|$ remaining groups of signal vertices, we can either match them to a g_0 vertex in $\gamma(s)$, or to a corresponding g_a or g_b vertex. In the latter case, the signal vertex is matched to the same g_a^β or g_b^β subcurve in $\gamma(s)$ as another signal vertex in a different group. This means that the buffer vertex that separates the two signal vertices is matched to a g_a or g_b vertex in the letter gadget. So in all cases, we match two vertices at distance at leasts $1 - \varepsilon$. Since we do this for at least $i + j - |s|$ vertices, $\mathrm{DTW}_p(c^*, \gamma(s)) \geq (1-\varepsilon)(i + j - |s|)^{1/p}$.

Now, we have

$$\begin{aligned}
\alpha(\mathrm{DTW}_p^q(c^*, A_i) + \mathrm{DTW}_p^q(c^*, B_j)) &\leq r - \sum_{s \in S} \mathrm{DTW}_p^q(c^*, \gamma(s)) \\
&\leq r - \sum_{s \in S} (1-\varepsilon)^q (i + j - |s|)^{q/p} \\
&= \alpha(i^{q/p} + j^{q/p}) + \sum_{s \in S} (1 - (1-\varepsilon)^q)(i + j - |s|)^{q/p} \\
&\leq \alpha(i^{q/p} + j^{q/p}) + (1 - (1-\varepsilon)^q)|S|(i+j)^{q/p},
\end{aligned}$$

so that $\mathrm{DTW}_p^q(c^*, A_i) + \mathrm{DTW}_p^q(c^*, B_j) \leq i^{q/p} + j^{q/p} + (1 - (1-\varepsilon)^q)(i+j)^{q/p} < i^{q/p} + j^{q/p} + \frac{1}{2}\min_{x \in \{i,j\}}(x+1)^{q/p} - x^{q/p}$. This means that there are at most $i + j$ signal vertices: suppose there are at least $i + 1$ A-signal vertices, then $\mathrm{DTW}_p^q(c^*, A_i) + \mathrm{DTW}_p^q(c^*, B_j) \geq$

$(1-\varepsilon)^q((i+1)^{q/p}+j^{q/p}) \geq i^{q/p}+j^{q/p}+((i+1)^{q/p}-i^{q/p})/2$, a contradiction. Analogously, at least $j+1$ B-signal vertices lead to a contradiction. ◀

Since the reduction runs in polynomial time (note that $1/\varepsilon$ can be bounded by a polynomial function in n, since p, q are constants, so β can be polynomially bounded) and the number of input curves is bounded by a linear function in $|S|$, we get the following result:

▶ **Theorem 7.** *The average curve problem for the (p,q)-DTW distance is NP-hard, for any $p, q \in \mathbb{N}$. When parametrised in the number of input curves m, this problem is W[1]-hard. There exists no $f(n) \cdot n^{o(m)}$ time algorithm for this problem unless ETH fails.*

We can apply our construction to the $(1,\ell)$-median problem for (p,q)-DTW with ℓ part of the input. This simplifies the proof of Lemma 6, since can ensure there are at most $i+j$ signal vertices by setting $\ell = i+j$, and therefore no longer have to find an upper bound for $\mathrm{DTW}_p^q(c^*, A_i) + \mathrm{DTW}_p^q(c^*, B_j)$. As a consequence, we can extend our analysis to solutions of cost γr for some fixed $\gamma \geq 1$ and observe that we can pick β large enough to make the reduction work. This means that for this variant, it is hard to find a constant factor approximation for any constant factor (for all forms of hardness in Theorem 7).

4 Algorithms for (k,ℓ)-center curve clustering

4.1 $(1+\varepsilon)$-approximation for (k,ℓ)-center clustering for discrete Frechét in $\mathbb{R}^d$

In this section, we develop a $(1+\varepsilon)$-approximation algorithm for the (k,ℓ)-center problem under the discrete Fréchet distance that runs in $O(mn\log(n))$ time for fixed k, ℓ, ϵ. To do this, we use the following Lemma.

▶ **Lemma 8.** *Let $k, \ell \in \mathbb{N}$, $\delta \in \mathbb{R}$ and $X > 0$. Suppose there are two sets $C = \{c_1, \ldots, c_k\}$ and $C^* = \{c_1^*, \ldots, c_k^*\}$, both containing k curves in $\mathbb{R}^d$ of complexity ℓ, where the elements of C are known, but those of C^* are not. Additionally, suppose that for all $\gamma \in C \cup C^*$ and for all curves $c^* \in C^*$, there exists a curve $c \in C$ such that $\mathrm{d}_{dF}(c, c^*) \leq \delta$. Then we can compute a set of vertices $V \subset \mathbb{R}^d$ of size at most $k\ell(\lceil \frac{\delta\sqrt{d}}{X} \rceil + 1)^d$ in $O(|V|)$ time such that there exists a set $\widetilde{C} = \{\tilde{c}_1, \ldots, \tilde{c}_k\}$ with $\mathrm{d}_{dF}(c_i^*, \tilde{c}_i) \leq X$, $|\tilde{c}_i| = \ell$ and the vertices of $\tilde{c}_i$ are in V, for all $1 \leq i \leq k$.*

Proof. For each vertex v of a curve in C, we construct an axis-parallel d-dimensional hypercube centred at v of sidelength 2δ. We then divide this hypercube into smaller hypercubes of side-length $2\frac{X}{\sqrt{d}}$. The grid $L(v)$ is the set of all vertices of these smaller hypercubes and so $|L(v)| = (\lceil \frac{\delta\sqrt{d}}{X} \rceil + 1)^d$. Let p be a point such that $\|p - v\| \leq \delta$. Then p lies inside one of the small hypercubes and so there is a vertex $p' \in L(u)$ (a vertex of the small hypercube) such that $\|p - p'\| \leq \frac{\sqrt{d}}{2} \cdot \frac{2X}{\sqrt{d}} = X$. Construct V as the union of $L(v)$ over all vertices v of curves in C. Since the curves in C have at most $k\ell$ unique vertices, $|V| \leq k\ell|L(v)| = k\ell(\lceil \frac{\delta\sqrt{d}}{X} \rceil + 1)^d$.

Let $c^* \in C^*$. There exists a curve $c \in C$ with $\mathrm{d}_{dF}(c, c^*) \leq \delta$, which means that each vertex u of c^* has distance at most δ to some vertex v of c. So, there exists a vertex $v' \in L(v) \subseteq V$ such that $\|u - v'\| \leq X$. Construct the curve $\tilde{c}$ by connecting all such vertices v' by line segments. By construction, $\mathrm{d}_{dF}(\tilde{c}, c^*) \leq \delta$, $|\tilde{c}| = \ell$, and all vertices of $\tilde{c}$ are in V. So, we can take $\widetilde{C} = \{\tilde{c} \mid c^* \in C^*\}$. ◀

▶ **Corollary 9.** *Suppose we are given an α-approximation to the (k,ℓ)-center problem. Given a set G of m input curves in $\mathbb{R}^d$, each of complexity at most n, and positive integers k,ℓ and some $0<\varepsilon\leq 1$, we can compute an $(1+\varepsilon)$-approximation to the (k,ℓ)-center problem for the discrete Fréchet distance in $O\left((Ck\ell)^{k\ell}\cdot k\ell\cdot mn\right)$ time, with $C\leq\left(\lceil\frac{2\alpha\sqrt{d}}{\varepsilon}\rceil+1\right)^d$.*

Proof. Let $\mathcal{C}$ be the solution of the α-approximation algorithm, $\mathcal{C}^*$ an optimal solution, with costs Δ and O, respectively. Let $c^*\in\mathcal{C}^*$, then (assuming without loss of generality that its cluster is non-empty) there is curve $g\in G$ such that $\mathrm{d}_{dF}(c^*,g)\leq O$. Since $\mathcal{C}$ attains cost Δ, there is a $c\in\mathcal{C}$ such that $\mathrm{d}_{dF}(c,g)\leq\Delta$. So, $\mathrm{d}_{dF}(c,c^*)\leq\mathrm{d}_{dF}(c,g)+\mathrm{d}_{dF}(g,c^*)\leq 2\Delta$, and we can apply Lemma 8 with $\delta=2\Delta$ and $X=\varepsilon\cdot\Delta/\alpha\leq\varepsilon O$ to obtain the set vertices V of size at most $k\ell(\lceil\frac{2\alpha\sqrt{d}}{\varepsilon}\rceil+1)^d$. We test the discrete Fréchet distance to the input curves of each of those sets of k curves with ℓ vertices in V to find a set $\widetilde{\mathcal{C}}$ such that for all $c^*\in\mathcal{C}^*$, there is a curve $\tilde{c}\in\widetilde{\mathcal{C}}$ with $\mathrm{d}_{dF}(c^*,\tilde{c})\leq\varepsilon O$. (For each set, this takes $O(mn\ell k)$ time) Since for any $g\in G$, there is a curve $c^*\in\mathcal{C}^*$ such that $\mathrm{d}_{dF}(g,c^*)\leq O$, there is a $\tilde{c}\in\widetilde{\mathcal{C}}$ such that $\mathrm{d}_{dF}(g,\tilde{c})\leq\mathrm{d}_{dF}(g,c^*)+\mathrm{d}_{dF}(\tilde{c},c^*)\leq(1+\varepsilon)O$. ◀

Given a set G of m input curves of complexity at most n each, we first use the algorithm by Buchin et al. [4] to compute a 3-approximation for the (k,ℓ)-center problem in $O(km\cdot\ell n\log(\ell+n))$ time. (The specific algorithm is not so important, we can use any efficient constant factor approximation algorithm and achieve the same result.) Using Corollary 9, we first compute a 2-approximation in $O(((6\sqrt{d})^dk\ell)^{k\ell}k\ell mn)$ time and apply Corollary 9 with this 2-approximation to obtain an $(1+\varepsilon)$-approximation in $O(((\frac{4\sqrt{d}}{\varepsilon})^dk\ell)^{k\ell}k\ell mn)$ time. We conclude the following theorem:

▶ **Theorem 10.** *Given m input curves in $\mathbb{R}^d$, each of complexity at most n, and positive integers k,ℓ and some $0<\varepsilon\leq 1$, we can compute an $(1+\varepsilon)$-approximation to the (k,ℓ)-center problem for the discrete Fréchet distance in $O\left(((Ck\ell)^{k\ell}+\log(\ell+n))\cdot k\ell\cdot mn\right)$ time, with $C=\left(\frac{4\sqrt{d}}{\varepsilon}+1\right)^d$.*

4.2 $(1+\varepsilon)$-approximation for (k,ℓ)-median clustering for the discrete Frechét distance in $\mathbb{R}^d$

Before we discuss $(1+\varepsilon)$-approximations, we first look at methods to approximate within a constant factor. For an arbitrary metric distance d, the 1-median problem has a simple $(2-\frac{2}{m})$-approximation: let β be an optimal solution to the 1-median problem with cost O, and take the curve $g_0\in G$ that has minimum distance to β over the curves in G as the centre. To see that the approximation ratio holds, note that $\sum_{g\in G}d(g_0,g)\leq\sum_{g\in G\setminus\{g_0\}}d(g_0,\beta)+d(\beta,g)=(m-2)d(g_0,\beta)+O\leq(\frac{m-2}{m}+1)O=(2-\frac{2}{m})O$, so the achieved approximation factor is indeed $(2-\frac{2}{m})$. If we want to approximate the $(1,\ell)$-median for a metric distance d over curves, we can get an approximation factor of $(2+c-\frac{2}{m})$ by finding a c-approximation of the minimum error ℓ-simplification of g_0. So, we can find a $(3-\frac{2}{m})$-approximation for $(1,\ell)$-median of discrete Fréchet and a $(6-\frac{2}{m})$-approximation for continuous Fréchet in polynomial time.

For the discrete k-median of an arbitrary metric distance d we can find a 4-approximation in polynomial time [14]. This gives a 8-approximation for the unrestricted k-median problem. Using the same technique as for the $(1,\ell)$-median, we get a 9-approximation for (k,ℓ)-median under the discrete Fréchet distance and a 12-approximation for (k,ℓ)-median under the continuous Fréchet distance.

For each $g \in G$, denote their minimum error ℓ-simplification under the discrete Fréchet distance by $\bar{g}$. To solve the problem when $k = 1$, note that there exists a $g_0 \in G$ such that $\mathrm{d}_{dF}(\bar{g_0}, \beta) \leq \frac{1}{m}\sum_{g\in G} \mathrm{d}_{dF}(\beta, \bar{g}) \leq \frac{1}{m}\sum_{g\in G} \mathrm{d}_{dF}(\beta, g) + \mathrm{d}_{dF}(g, \bar{g}) \leq \frac{1}{m}\sum_{g\in G} 2\,\mathrm{d}_{dF}(\beta, g) = 2O/m \leq 2O/m$, where the first inequality holds since $\bar{\gamma}$ is closer to β than $\bar{g}$ for all $g \in G$, the second is the triangle inequality of d_{dF}, and the third holds since $\bar{\gamma}$ is a minimum error ℓ-simplification of γ and $|\beta| \leq \ell$. It is not easy to determine g_0 without knowing β, but we can run our algorithm for all m options. This means that we can use Lemma 8 to find an $(1+\varepsilon)$-approximation as in Corollary 9 by searching grids of $(\lceil \frac{4\sqrt{d}}{\varepsilon} \rceil + 1)^d$ vertices.

For the (k, ℓ)-median problem with $k > 1$ we can also find an $(1+\varepsilon)$-approximation by searching in balls around the vertices of a center curve that realises a 9-approximation with cost Δ, but we need to take a different radius. The total cost inside a single cluster is at most the total cost of the clustering, so $\mathrm{d}_{dF}(\bar{\gamma}, \beta) \leq \sum_{g\in G} \mathrm{d}_{dF}(\bar{g}, \beta)$ and so we take radius 2Δ. We cannot get a tighter bound unless we can select a curve for each cluster in the optimal solution, in which case we can improve the bound by the factor $\frac{1}{2}$ (since every optimal cluster with non-zero cost contains at least 2 curves). However, we cannot do much better: take for instance $k > 1$, with $m - 2$ curves divided over $k - 1$ clusters with internal distance $< \varepsilon$ of each other and 2 curves in a cluster with internal distance 2δ and distances of $> \delta$ in between clusters. Then the minimal distance to the optimal center of the cluster with 2 curves is at least δ (triangle inequality) and the total optimal cost is $2\delta + (m-2)\varepsilon$, so as $\varepsilon \to 0$, the fraction of the minimal distance of this cluster over the total cost goes to $\frac{1}{2}$. We summarize our results in the following theorem:

▶ **Theorem 11.** *Given m input curves in $\mathbb{R}^d$, each of complexity at most n, and positive integers k, ℓ and some $0 < \varepsilon \leq 1$, we can compute an $(1+\varepsilon)$-approximation to the (k, ℓ)-center problem for the discrete Fréchet distance in $O\left((Ck\ell)^{k\ell} \cdot k\ell \cdot m^2 n\right)$ time when $k = 1$ with $C = \left(\frac{4\sqrt{d}}{\varepsilon}\right)^d$. When $k > 1$, we require $O\left((Ck\ell)^{k\ell} \cdot k\ell \cdot mn + m^3 + mn\ell \log n \log(n/\ell) + \ell^2 m^2\right)$ time, where $C = \left(\frac{4m\sqrt{d}}{\varepsilon}\right)^d$.*

4.3 Exact algorithm for (k, ℓ)-center under discrete Fréchet in 2D

We give an algorithm that solves the (k, ℓ)-center problem for the discrete Fréchet distance in 2D in polynomial time for fixed k and ℓ. We first show how to solve the decision version of this problem and use it as a subroutine to solve the optimisation problem.

The main idea of the algorithm for the decision version is based on the following observation: for a given r, we have $\min_{c\in\mathcal{C}} \mathrm{d}_{dF}(c, g) \leq r$ for all $g \in G$ if and only if each vertex p of a curve in $\mathcal{C}$ lies in the intersection of the disks of radius r around all vertices q from curves in G that p is matched with. Furthermore, it does not matter where the vertex p lies within the intersection region. This means we can select a vertex for each maximal overlapping region (i.e. each region such that the set of disks intersecting the region is not contained in another region) and exhaustively test all sets with k curves of ℓ vertices that can be constructed by using only the selected vertices to determine if there exists a set of curves $\mathcal{C}$ such that $\min_{c\in\mathcal{C}} \mathrm{d}_{dF}(c, g) \leq r$ for all $g \in G$.

To find all maximal intersection regions, we first compute the planar graph $\mathcal{G} = (V, E)$, where V is the set of all intersection points between boundaries of disks centred around a vertex from our input curves with radius r and E is the set of arcs on the boundary of those disks ending at two intersection points. This graph has $O((nm)^2)$ vertices and arcs and can be computed in $O((nm)^2)$ time [6], see Figure 3 for an example. By traversing the intersection points and arcs on the boundary, we can find the at most $O((nm)^2)$ maximal

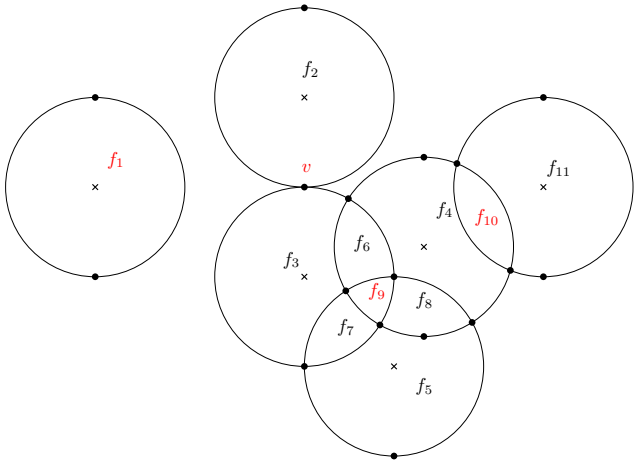

Figure 3 An example configuration of $\mathcal{G} = (V, E)$. Crosses indicate the vertices from the curves in G, dots indicate vertices from V and all bounded faces are numbered. The maximal intersection regions are the faces f_1 and f_9 and the vertex v (in red). Note that while all arcs on the boundary of f_2 are convex for that face, f_2 is not maximal, since its boundary intersects the boundary of f_3 only at vertex v.

intersection regions. So, we test $O((mn)^{2k\ell})$ sets of center curves, for which we can test whether an input curve has discrete Fréchet distance less than r to a single curve among the k center curves in $O(m\ell)$. This means the algorithm for the decision version takes $O((mn)^{2k\ell}k\ell m)$ time.

To find a minimum r such that a (k, ℓ)-center exists, note that we only have to consider the decision problem for those r where the topology of the intersection regions in $\mathcal{G}$ is different. If we start with $r = 0$ and gradually increase it, the topology of $\mathcal{G}$ changes only when a new maximal intersection is created, which then consists of exactly one point p. This means that there is a subset of our disks such that point p is the earliest point where all disks have a non-empty intersection. So, p must be the center of the minimum enclosing disk for this subset of disks. Since a minimum enclosing disk is determined by at most 3 points, there can be at most one unique point for every triple in set of vertices of the input curves which give at most $O((mn)^3)$ distinct values of r where the topology of $\mathcal{G}$ changes. By performing a binary search on these values, we can find the optimal value in $O(\log(mn))$ calls to the algorithm for the decision version and we get the following result:

▶ **Theorem 12.** *Given a set of n curves G in the plane with at most m vertices each, we can find a solution to the (k, ℓ)-center problem for the discrete Fréchet distance in $O((mn)^{2k\ell}k\ell m \log(mn))$ time.*

5 Conclusion

In this paper, we have shown that the 1-median problem is computationally hard under the discrete Fréchet, continuous Fréchet, and DTW distance. A natural question is whether this problem is hard to approximate. Efficient constant factor approximation algorithms are known for the Fréchet distance (see Section 4.2), but not for DTW. If we extend our analysis in Lemma 3 to a solution with cost γr for some $\gamma \geq 1$, the construction works when $\gamma = 1 + O(\frac{1}{m^2})$ (where the constant is independent of other input parameters), which gives a lower bound of $1 + O(\frac{1}{m^2})$ on the approximation factor. If we do the same for

Lemma 6, we get that it is hard to approximate 1-median under (p, q)-DTW for any factor $< 1 + 2((1 + \frac{1}{\min(i,j)})^{q/p} - 1)$. So, it remains an open problem to find a constant lower bound for approximating 1-median for this distance measures. However, note that for the $(1, \ell)$-median problem with ℓ part of the input, our construction implies that it is hard to approximate this problem within *any* constant factor (See the end of Section 3).

On the positive side, we have given $(1 + \varepsilon)$-approximation algorithms for (k, ℓ)-center and (k, ℓ)-median problems under discrete Fréchet in Euclidean space and an exact algorithm for the (k, ℓ)-center problem under discrete Fréchet in 2D that all run in polynomial time for fixed k, ℓ, ε. It would be interesting to see if these algorithms can be adapted to the DTW or continuous Fréchet settings. Our approximation algorithms rely on the fact that good approximations have small distance to some optimal solution and that we can search a bounded space (the set of balls surrounding the vertices) for better approximations. The first property does not hold for DTW, since it is non-metric and the second property does not hold for continuous Fréchet, since the vertices of a curve with small continuous Fréchet distance do not have to be near the vertices of the other curve. The latter property is also crucial for the exact algorithm.

References

1 Helmut Alt and Michael Godau. Computing the Fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 5:75–91, 1995. doi:10.1142/S0218195995000064.

2 Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristics for k-median and facility location problems. *SIAM Journal of Computing*, 33(3):544–562, 2004. doi:10.1137/S0097539702416402.

3 Markus Brill, Till Fluschnik, Vincent Froese, Brijnesh Jain, Rolf Niedermeier, and David Schultz. Exact mean computation in dynamic time warping spaces. *Data Mining and Knowledge Discovery*, 33(1):252–291, 2019. doi:10.1007/s10618-018-0604-8.

4 Kevin Buchin, Anne Driemel, Joachim Gudmundsson, Michael Horton, Irina Kostitsyna, Maarten Löffler, and Martijn Struijs. Approximating (k, ℓ)-center clustering for curves. In *Proceedings of the 30th ACM-SIAM Symposium on Discrete Algorithms*, pages 2922–2938, 2019. doi:10.1137/1.9781611975482.181.

5 Laurent Bulteau, Vincent Froese, and Rolf Niedermeier. Tight hardness results for consensus problems on circular strings and time series. *arXiv preprint arXiv:1804.02854*, 2018. URL: http://arxiv.org/abs/1804.02854.

6 Bernard Marie Chazelle and Der-Tsai Lee. On a circle placement problem. *Computing*, 36(1-2):1–16, 1986.

7 Anne Driemel, Amer Krivošija, and Christian Sohler. Clustering time series under the Fréchet distance. In *Proceedings of the 27th ACM-SIAM Symposium on Discrete Algorithms*, pages 766–785. Society for Industrial and Applied Mathematics, 2016.

8 Thomas Eiter and Heikki Mannila. Computing discrete Fréchet distance. Technical Report CD-TR 94/64, Christian Doppler Laboratory for Expert Systems, TU Vienna, Austria, 1994.

9 Kaspar Fischer, Bernd Gärtner, and Martin Kutz. Fast smallest-enclosing-ball computation in high dimensions. In *Proceedings of the 11th Annual European Symposium on Algorithms*, pages 630–641, 2003. doi:10.1007/978-3-540-39658-1_57.

10 Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985. doi:10.1016/0304-3975(85)90224-5.

11 Lalit Gupta, Dennis L Molfese, Ravi Tammana, and Panagiotis G Simos. Nonlinear alignment and averaging for estimating the evoked potential. *IEEE Transactions on Biomedical Engineering*, 43(4):348–356, 1996.

12 Ville Hautamäki, Pekka Nykänen, and Pasi Fränti. Time-series clustering by approximate prototypes. In *Proceedings of the 19th International Conference on Pattern Recognition*, pages 1–4, 2008. doi:10.1109/ICPR.2008.4761105.

13 Kamal Jain, Mohammad Mahdian, and Amin Saberi. A new greedy approach for facility location problems. In *Proceedings of the 34th ACM Symposium on Theory of Computing*, pages 731–740, 2002. doi:10.1145/509907.510012.

14 Kamal Jain and Vijay V. Vazirani. Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and lagrangian relaxation. *J. ACM*, 48(2):274–296, 2001. doi:10.1145/375827.375845.

15 Shi Li and Ola Svensson. Approximating k-median via pseudo-approximation. *SIAM Journal of Computing*, 45(2):530–547, 2016. doi:10.1137/130938645.

16 Nimrod Megiddo. Linear-time algorithms for linear programming in r^3 and related problems. *SIAM Journal of Computing*, 12(4):759–776, 1983. doi:10.1137/0212052.

17 Nimrod Megiddo and Kenneth J. Supowit. On the complexity of some common geometric location problems. *SIAM Journal of Computing*, 13(1):182–196, 1984. doi:10.1137/0213014.

18 François Petitjean and Pierre Gançarski. Summarizing a set of time series by averaging: From Steiner sequence to compact multiple alignment. *Theoretical Computer Science*, 414(1):76 – 91, 2012. doi:10.1016/j.tcs.2011.09.029.

19 Krzysztof Pietrzak. On the parameterized complexity of the fixed alphabet shortest common supersequence and longest common subsequence problems. *Journal of Computer and System Sciences*, 67(4):757–771, 2003.

20 Kari-Jouko Räihä and Esko Ukkonen. The shortest common supersequence problem over binary alphabet is NP-complete. *Theoretical Computer Science*, 16(2):187 – 198, 1981. doi:10.1016/0304-3975(81)90075-X.