On the hardness of finding an average curve

Kevin Buchin[1], Anne Driemel[2], and Martijn Struijs[1]

1 **Department of Mathematics and Computing Science, TU Eindhoven, The Netherlands**
{k.a.buchin,m.a.c.struijs}@tue.nl
2 **University of Bonn, Hausdorff Center for Mathematics, Germany**
driemel@cs.uni-bonn.de

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. Moreover, closing some gaps in the literature, we show positive results for the (k, ℓ)-center variant under the discrete Fréchet distance. In particular, we give an $\widetilde{O}(mn)$-time $(1 + \varepsilon)$-approximation algorithm and a polynomial-time exact algorithm for fixed k, ℓ and ε.

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) [9], continuous and discrete Fréchet distance [1, 6].

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 $\max_{g \in \mathcal{G}} \min_{i=1}^{k} \delta(c_i, g)$.

We call the 1-median problem the *average curve* problem. Clustering on points for general k in the plane or higher dimension is often NP-hard [8] and clustering curves tends to be hard even when $k = 1$ and the curves lie in 1D. For instance, Buchin et. al. [2] 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 a ratio better than 2. In this paper, 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.

Denote the set of all warping paths (or alignments, see also [9]) between curves x and y by $\mathcal{W}_{x,y}$. For any integers $p, q \geq 1$, define $\mathrm{DTW}_p^q(x, y) := \left(\min_{W \in \mathcal{W}_{x,y}} \sum_{(i,j) \in W} |x_i - y_j|^p\right)^{q/p}$. We call DTW_p^q the (p, q)-DTW distance.

The average curve problem for the $(2, 2)$-DTW distance has resisted efficient algorithms so far, which motivated several heuristic approaches [7, 9]. A formal proof of NP-hardness has only recently been given by Bulteau et. al. [3], who additionally show the $(2, 2)$-DTW

35th European Workshop on Computational Geometry, Utrecht, The Netherlands, March 18–20, 2019.
This is an extended abstract of a presentation given at EuroCG'19. 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 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. In this paper, we prove the same hardness results of the average curve problem for the (p, q)-DTW distance for any $p, q \in \mathbb{N}$, with a different method. While Bulteau et. al. [3] note at the end of Section 5 their method might generalise to more variants of the DTW distance, when $p \neq q$, the (p, q)-DTW distance does not fit in their framework since then q/p is a non-trivial exponent. Furthermore, when $p = q = 1$, the variant has the form required in their framework, but the condition required for their hardness proof of an intermediate problems fails.

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. So, the (k, ℓ)-center problem is to find a set of curves $\mathcal{C} = \{c_1, \ldots c_k\}$, each with at most ℓ vertices that minimizes $\max_{g \in \mathcal{G}} \min_{i=1}^{k} \delta(c_i, g)$ and 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. [5], 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 [2], we 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. In this paper, 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.

2 Hardness of finding average curves

To show the hardness of the average curve problem for the Fréchet and DTW distance, we reduce from a variant of the NP-hard *Shortest Common Supersequence* (SCS) problem [10, 11], 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 $s \in S$?

▶ **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.*

The proof idea is to reduce from SCS: 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's in s by B and vice versa. We reduce to the instance $(S', t + 2, 3t)$. If s^* is a common supersequence of length t for S, then $s^* + AB^{2t}A + c(s^*)$ is a supersequence of S' with the correct character count. Optimal supersequences of S' can be decomposed into this form. ◀

2.1 Complexity of the average curve under the Fréchet distance

We will show the hardness of finding the average curve under the discrete and continuous Fréchet distance d_{dF} and d_F via the following reduction from FCCS. Given an instance

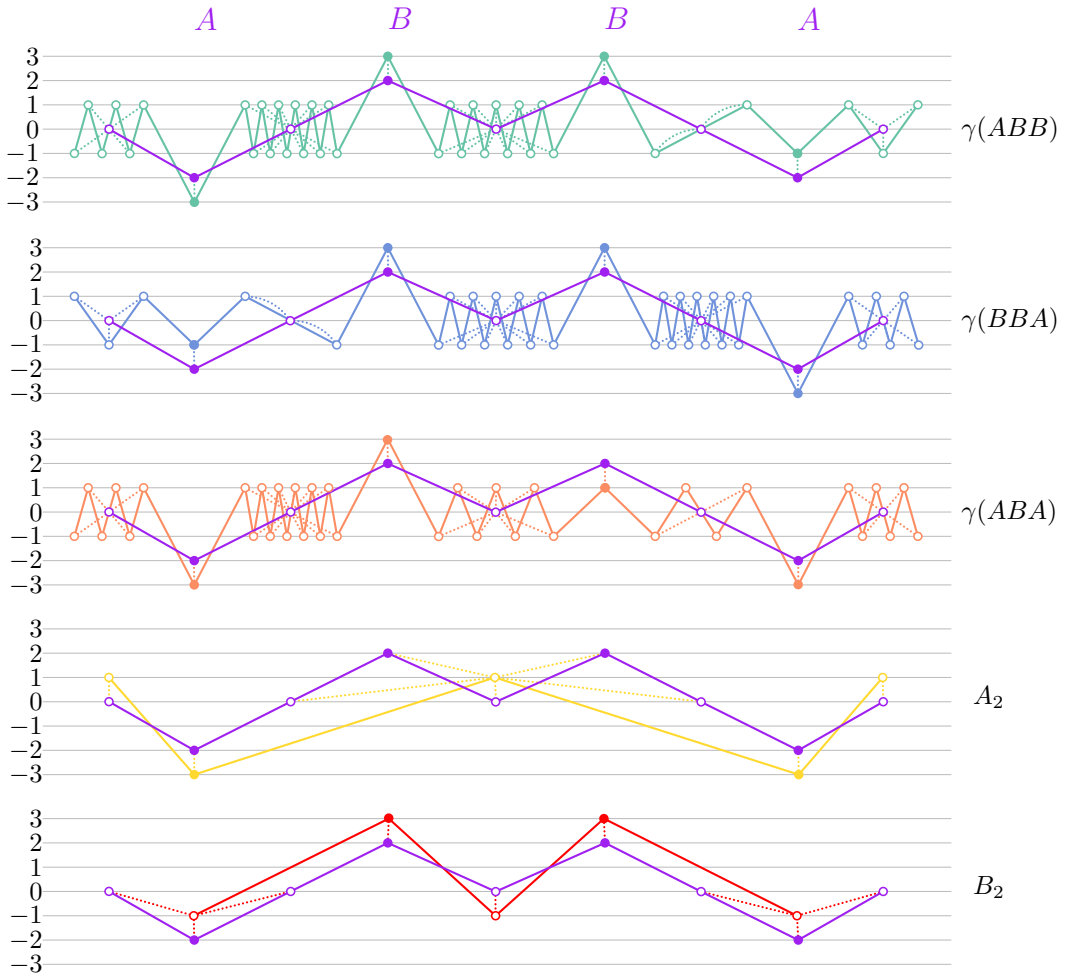

Figure 1 Five 1D-curves from $G \cup R_{i,j}$ in the reduction for the Fréchet average curve problem and a center curve constructed from $ABBA$ (purple) as in Lemma 2. Matchings are indicated by dotted lines. Note that each of these matchings achieves a (discrete) Fréchet distance of 1.

(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 a 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, see Figure 1.

▶ **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. Since $\mathrm{d}_F(x, y) \leq \mathrm{d}_{dF}(x, y)$ for all curves x, y, considering the discrete version suffices.

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 $l \in \{1, \ldots, 2|s^*| + 1\}$. Note s^* is a supersequence of the sequence of letter gadgets in any curve $g \in G \cup R_{i,j}$ and therefore we can match all letter gadgets from g within distance 1 such that we get $\mathrm{d}_{dF}(c, g) \leq 1$. This means $\sum_{g \in G \cup R_{i,j}} \mathrm{d}_{dF}(c, g) \leq |S| + 2\alpha = r$. ◀

For the converse, we can show that if there is a curve c^* with $\sum_{g \in G \cup R_{i,j}} \mathrm{d}_F(c^*, g) \leq r$, then $\mathrm{d}_F(c^*, g) < 2$ for all $g \in G \cup R_{i,j}$. This means we can apply the hardness proof for the 1-center problem under the Fréchet distance from [2] to partition c^* into A-parts, B-parts and buffer parts and construct a supersequence for S from the sequence of A/B-parts in c^*.

▶ **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.*

Since the reduction runs in polynomial time and the number of input curves is bounded by a quadratic function in $|S|$, we get the following result.

▶ **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.*

2.2 Complexity of the average curve under the DTW distance

We will show that the average curve problem for (p, q)-DTW is NP-hard for all $p, q \in \mathbb{N}$. We use the same reduction from Section 2.1, but now map the characters of $s \in S$ to

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

use the curves $A_i = g_0^\beta (g_a^\beta g_0^\beta)^i$ and $B_j = g_0^\beta (g_b^\beta g_0^\beta)^j$, and set $r = \sum_{s \in S} (i + j - |s|)^{q/p} + \alpha(i^{q/p} + j^{q/p})$, $\beta = \lceil r/\varepsilon^q \rceil + 1$, $\alpha = |S|$, where $\varepsilon > 0$ is chosen sufficiently small and depends only on i, j, p, q. See Figure 2 for an example 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. This is analogous to Lemma 2. ◀

For the converse, we identify vertices in a satisfying curve c^* that are close to g_a or g_b, such that g_a^β and g_b^β subcurves must be matched to them and construct a supersequence s' out of them. The curves A_i and B_j are used to show that $\#_A(s') = i$ and $\#_B(s') = j$.

▶ **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.*

Since the reduction runs in polynomial time 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.*

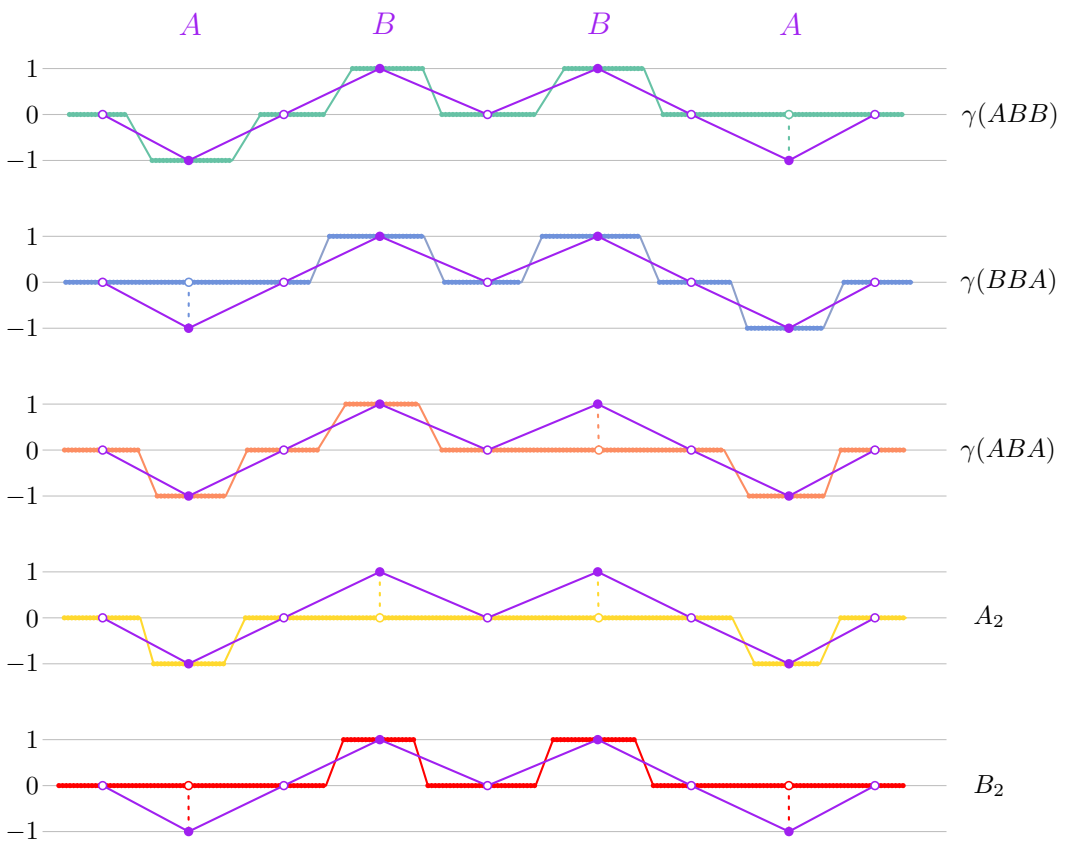

Figure 2 Five 1D-curves from $G \cup R_{i,j}$ and a center curve constructed from $ABBA$ (purple) as in Lemma 5. Fat horizontal lines indicate β consecutive vertices. Vertices that match at distance 0 touch, those matching at distance 1 are indicated by dotted lines. The center has 1 mismatch with the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

3 $(1+\varepsilon)$-approximation for (k,ℓ)-center clustering for the discrete Frechet distance in $\mathbb{R}^d$

In this section, we develop a $(1+\epsilon)$-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.

Given a set $\mathcal{G}$ of m input curves in $\mathbb{R}^d$ of complexity at most n each, use the algorithm by Buchin et. al. [2] to compute a set $\mathcal{C}$ of k curves that forms a 3-approximation for the (k,ℓ)-center problem in $O(km\cdot \ell n\log(\ell+n))$ time. Call the cost of these centers Δ. Let $\mathcal{C}^*$ be an optimal solution that achieves cost O. For each vertex p^* in $\mathcal{C}^*$, there is a vertex q on an input curve with $\|p^*-q\|\leq O$ and there is a vertex p in $\mathcal{C}$ with $\|p-q\|\leq\Delta$. So, by the triangle inequality, all vertices of $\mathcal{C}^*$ lie within a ball of radius 2Δ centred at a vertex of $\mathcal{C}$.

We can cover these balls with a regular grid of $O(\varepsilon^{-d})$ vertices with distance of $\varepsilon\cdot 2\Delta/(3\sqrt{d})$, so that there exists a vertex $g(p^*)$ on such a grid with $\|p^*-g(p^*)\|\leq \varepsilon\Delta/3=\varepsilon O$. So, for every curve $c^*\in\mathcal{C}^*$, there exists a single curve $g(c^*)$ of gridpoints with $\mathrm{d}_{dF}(g(c^*),c^*)\leq\varepsilon O$, which means that for all $g\in\mathcal{G}$, there exists a curve c^* such that $\mathrm{d}_{dF}(g,g(c^*))\leq(1+\varepsilon)O$. This means the set $\{g(c^*)\mid c^*\in\mathcal{C}^*\}$ gives a $(1+\varepsilon)$-approximation, which we can find by iterating over all curves using the gridpoints. We conclude with the following theorem:

▶ **Theorem 8.** *Given m curves in $\mathbb{R}^d$, each of complexity at most n, and $k,\ell\in\mathbb{N}$ 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 $C=\left(\frac{6\sqrt{d}}{\varepsilon}\right)^d$.*

4 Exact algorithm for (k,ℓ)-center for 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.

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 $g\in\mathcal{G}$ 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 $\mathcal{G}$ that p is matched with. Furthermore, it does not matter where the vertex p lies within the intersection region. This means we can select one vertex for each 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 $g\in\mathcal{G}$.

The corresponding arrangement of circles has complexity $O((nm)^2)$, and can be computed in that time [4], see Figure 3 for an example. We solve the optimisation version by performing a binary search over the at most $O((mn)^3)$ values of r at which the arrangement changes combinatorially, which occurs only when some disks intersect at a single point.

▶ **Theorem 9.** *Given a set of m curves G in the plane with at most n vertices each, we can find a solution to the (k,ℓ)-center problem in $O((mn)^{2k\ell+1}k\ell\log(mn))$ time.*

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 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.

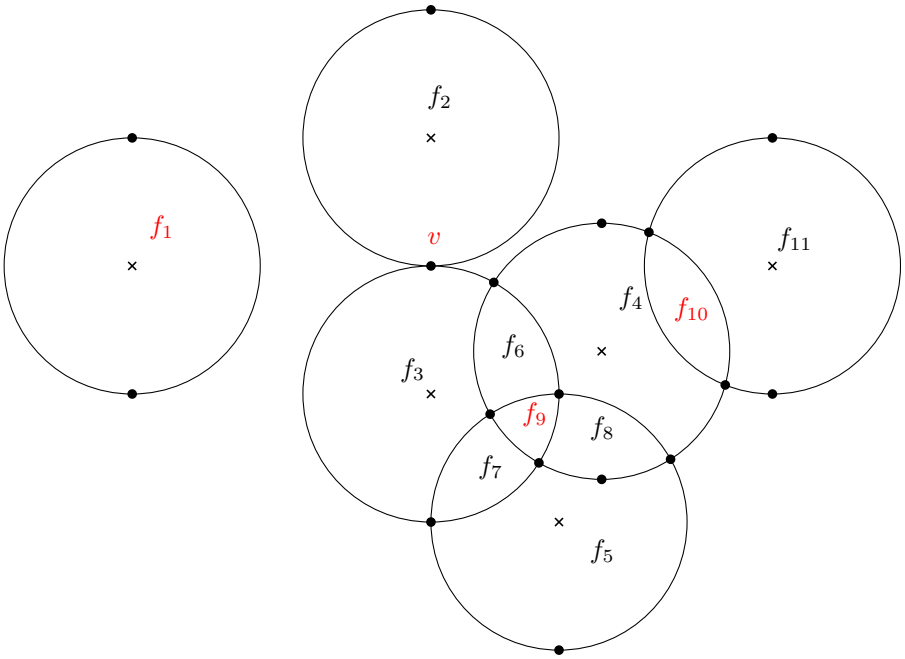

Figure 3 A possible arrangement of circles. Crosses indicate the vertices from the curves in $\mathcal{G}$, all bounded faces are numbered. The relevant intersection regions for the Fréchet distance are in red.

3 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`.

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

5 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.

6 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.

7 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.

8 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`.

9 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`.

10 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.

11 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`.