Tight Lower Bound for Approximating (k, ℓ)-Center Clustering under the Fréchet Distance

Kevin Buchin*[1], Jacobus Conradi†[2], Lindsey Deryckere‡[3], Mart Hagedoorn§[1], and Carolin Rehs¶[1]

[1]Technical University Dortmund, Germany
[2]University of Bonn, Germany
[3]University of Sydney, Australia

1 Introduction

Clustering is a fundamental task in computer science that aims to group similar objects while providing a compact representation of the original data. Most existing work assumes inputs in the context of point data [1, 8]. However, recently, effort has been devoted into the clustering of more complex data like curves, which arise naturally in spatial and time series data [2, 4, 6, 5]. With curves, the natural challenge is to additionally bound the complexity of the computed representation. This motivates the study of the (k, ℓ)-center clustering problem under the Fréchet distance, which generalizes the classical k-center problem by restricting center curves to have at most ℓ vertices.

More formally, given a set $\mathcal{G}$ of n polygonal curves of finite size, the (k, ℓ)-center clustering problem asks to find a set $\mathcal{C}$ of k polygonal curves, each of size ℓ, such that the maximum (discrete or continuous) Fréchet distance of any curve in $\mathcal{G}$ to the closest curve in $\mathcal{C}$ is minimized.

Buchin et al. show that a careful adaptation of Gonzalez' algorithm in combination with curve simplification yields a 3-approximation for the discrete Fréchet distance in d dimensions for $d \geq 1$ and for the continuous Fréchet distance in 2 dimensions [3]. A key step in obtaining the 3-approximation is the computation of an optimal simplification of a polygonal curve, which remains unsolved under the continuous Fréchet distance in $\mathbb{R}^d$ for $d > 2$. However, recent work by Cheng and Huang has introduced the first polynomial-time bicriteria approximation scheme for curve simplification in higher dimensions, marking progress towards an optimal simplification [7]. Furthermore, Buchin et al. also show that approximating the problem within a factor close to 2.598 for the discrete Fréchet distance, and $2.25 - \varepsilon$ for the continuous Fréchet distance is NP-hard.

In this paper, we improve upon these results by proving that the (k, ℓ)-center problem under the (discrete and continuous) Fréchet distance is NP-hard to approximate within a factor of $3-\varepsilon$ for any $\varepsilon > 0$. Our results hold for curves in $\mathbb{R}^d$ where $d \geq 2$ as well as the special case when $k = 1$, which implies tightness for the 3-approximation under the continuous Fréchet distance in $\mathbb{R}^2$ and for the 3-approximation under the discrete Fréchet distance in $\mathbb{R}^d$ for $d \geq 2$.

2 Discrete Fréchet

We begin in the setting of discrete Fréchet distance. Here, we show that the (k, ℓ)-center clustering problem cannot be approximated within a factor of $3 - \varepsilon$, for some $\varepsilon > 0$, giving a reduction from the Shortest Common Supersequence (SCS) problem that generalizes the reduction given in [3]. In the SCS problem there is given a set S of n input strings of finite length over a finite alphabet Σ, asking if there exists a string S^* of size t such that each string in S is a subsequence of S^*. Even with a binary alphabet, the SCS problem is shown to be NP-Complete [9]. Furthermore, for two polygonal curves ψ and ϕ, we will denote their discrete Fréchet distance by $d_{DF}(\psi, \phi)$.

Given an instance of the SCS problem, i.e. n input strings over an alphabet $\{A, B\}$, and a value t, which denotes the maximum length of the sought superstring. We construct a corresponding (k, ℓ)-center clustering instance as follows. Let $p_0 = (0, 0)$ be the origin of the Euclidean plane and let c_1^x, c_2^x, c_3^x be regular x-gons centered on p_0 with radii 1, 2, and 3, respectively[1]. The x-gons are oriented such that the coordinate of the first vertex of c_j^x is $(0, j)$ for $1 \leq j \leq 3$. We define point $p_{i,j}^x$ to be the ith vertex of c_j^x for $1 \leq i \leq n$ and $1 \leq j \leq 3$. Next, we define the following point sequence gadgets (see Figures 1 and 2 for an illustrated example with various x):

- G_a^x starts with $p_{1,1}^x$ followed by $p_{i',1}^x$ where $i' := \left(\left(j' \left\lceil \frac{x}{2} \right\rceil\right) \bmod x\right) + 1$ for $1 \leq j' \leq x$.

*Email: kevin.buchin@tu-dortmund.de
†Email: jacobus.conradi@gmx.de
‡Email: lindsey.deryckere@sydney.edu.au
§Email: mart.hagedoorn@tu-dortmund.de
¶Email: carolin.rehs@tu-dortmund.de

[1]Note that when $x = 3$ our reduction coincides with the reduction by Buchin et al. [3]

- C_A^x starts with $p_{1,2}^x$ followed by $p_{i',2}^x$ where $i' := \left(\left(j' \left\lceil \frac{x}{2} \right\rceil\right) \bmod x\right) + 1$ for $1 \leq j' \leq x$.
- G_A^x starts with $p_{1,3}^x$ followed by $p_{i',3}^x$ where $i' := \left(\left(j' \left\lceil \frac{x}{2} \right\rceil\right) \bmod x\right) + 1$ for $1 \leq j' \leq x$.
- G_b^x is the reverse of G_a^x.
- C_B^x is the reverse of C_A^x.
- G_B^x is the reverse of G_A^x.

We refer to G_b^x and G_a^x as *buffer gadgets*.

In order to construct a polygonal curve $\gamma(s)$ for each of the input strings $s \in S$, we use the gadgets defined above to map the letters A and B to the following point sequences.

- $A \to (G_a^x, G_b^x)^t, (G_A^x), (G_a^x, G_b^x)^t$
- $B \to (G_a^x, G_b^x)^t, (G_B^x), (G_a^x, G_b^x)^t$

Here $(G)^t$ means that point sequence G has been repeated and concatenated t times. We further define γ as a function that takes an input string and replaces each letter with a point sequence according to the aforedescribed mapping, concatenates these sequences, and inserts line segments between two consecutive points. Using these mappings for every input string in S we obtain a polygonal curve which is part of the input for the (k, ℓ)-center clustering problem, i.e. $\mathcal{G} = \{\gamma(s_i) | s_i \in S\}$.

Lemma 1 *For any true instance of the SCS decision problem, there exists a center curve of length at most $(x+3)t$ and radius at most $r = 1$ under the discrete Fréchet distance in our construction.*

Proof. If the SCS instance is true, then there exists a common superstring S^* of length at most t. We map this superstring to a polygonal curve β as follows: for every letter A or B we append (p_0, C_A^x, p_0) or (p_0, C_B^x, p_0), respectively, with corresponding line segments between sequential points to β. Therefore, the resulting polygonal curve has at most $(x+3)t$ vertices. Since each input string in S is a substring of the common superstring, whenever the substring omits a letter of the superstring, this letter can be matched to a corresponding buffer gadget. Whenever we need to skip a sequence of buffer gadgets, we match them to point p_0 on the polygonal curve of the superstring. □

Lemma 2 *Let $r < 3\cos(\frac{\pi}{2x})$. There does not exist a polygonal curve ϕ in the plane that satisfies both $d_{DF}(G_A^x, \phi) \leq r$ and $d_{DF}(G_B^x, \phi) \leq r$.*

Proof. Assume that there exists a polygonal curve ϕ such that $d_{DF}(G_A^x, \phi) \leq r$ and $d_{DF}((G_B^x), \phi) \leq r$ for some $r < 3\cos(\frac{\pi}{2x})$. Since the distance between two consecutive points in G_A^x or G_B^x is exactly $6\sin(\frac{\pi}{2} - \frac{\pi}{2x}) = 6\cos(\frac{\pi}{2x})$, there is no point in ϕ that is matched under the Fréchet Distance to two sequential points in G_A^x or G_B^x (see Figure 3a). Furthermore, since G_A^x and G_B^x follow the same polygonal curve but traversed in the opposite order, somewhere a point of ϕ must be matched to two points that have an underlying distance of $6\cos(\frac{\pi}{2x})$. This fact contradicts that there can exist a curve ϕ that satisfies both $d_{DF}(G_A^x, \phi) \leq r$ and $d_{DF}(G_B^x, \phi) \leq r$. □

Lemma 3 *For any instance of an SCS decision problem, consider the above reduction with $x > 3t$. If there exists a center curve of length at most $(x+3)t$ and of radius strictly smaller than $r = 3\cos(\frac{\pi}{2x})$ under the discrete Fréchet distance on the constructed instance, then the original SCS instance is true.*

Proof. Let the center curve be denoted by β. We claim we can transform this polygonal curve into a superstring S^* of S from the instance of the SCS decision problem. We define disks D_i^x of radius r centered at points $p_{i,3}^x$ for $1 \leq i \leq x$.

First, remove all points at the start and end from β that do not lie in D_1^x and, for any consecutive points in β that both lie in D_1^x, remove the latter point. Next, split β by points that lie in D_1^x. This gives us the set of subsequences B from which we will build the superstring S^*. For each subsequence $\beta' \in B$ we calculate the discrete Fréchet distance to G_A^x and G_B^x and by Lemma 2 the following three options are possible:

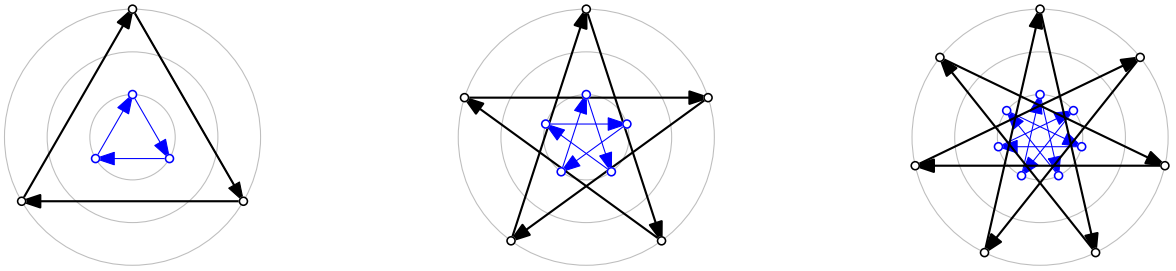
- $d_{DF}((G_A^x), \beta') \leq r$ and $d_{DF}((G_B^x), \beta') > r$; the letter A is added to S^*.
- $d_{DF}((G_A^x), \beta') > r$ and $d_{DF}((G_B^x), \beta') \leq r$: the letter B is added to S^*.
- $d_{DF}((G_A^x), \beta') > r$ and $d_{DF}((G_B^x), \beta') > r$: this subsequence is ignored and no letters are added to S^*.

We claim that the resulting string S^* is a common superstring of the set of strings S. Since β is a center curve of radius r, it is within Fréchet distance r to all input strings. By their construction, every letter A and B in s_i generates a subsequence G_A^x and G_B^x, respectively, in $\gamma(s_i)$. In order to match each subsequence G_A^x or G_B^x, center β needs to visit disks D_i^x in the correct order and have a vertex in each disk, always starting and ending in disk D_1^x.

The size of the generated string S^* is at most

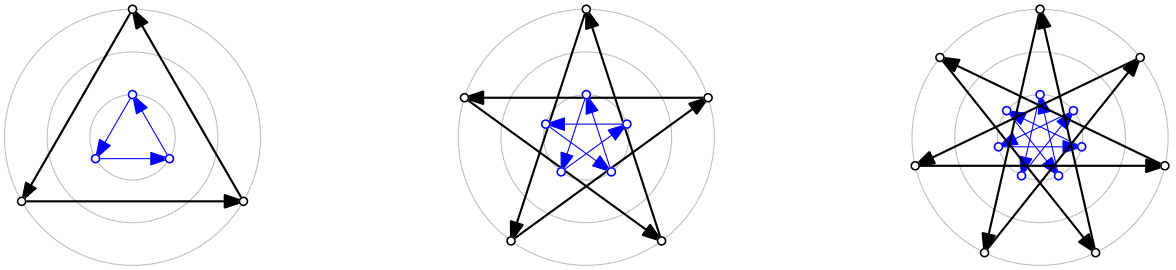
$$\left\lfloor \frac{(x+3)t}{x} \right\rfloor = t + \left\lfloor \frac{3t}{x} \right\rfloor$$

Hence, for any $x > 3t$ the length of S^* is at most t, implying the SCS instance is true. □



(a) $x = 3$ (b) $x = 5$ (c) $x = 7$

Figure 1: Three point sequences for various x, the blue sequence represents G_a^x and the black sequence represents G_A^x.



(a) $x = 3$ (b) $x = 5$ (c) $x = 7$

Figure 2: Three point sequences for various x, the blue sequence represents G_b^x and the black sequence represents G_B^x.

Then, from Lemma 1 and 3, by choosing $x := 2 \cdot \max(3t, \frac{1}{\varepsilon}) + 1$, the following result directly follows.

Theorem 4 *The (k, ℓ)-center problem under the discrete Fréchet distance is NP-hard to approximate within a factor of $3 - \varepsilon$ for any $\varepsilon > 0$ for curves in $\mathbb{R}^d$ with $d \geq 2$, even if $k = 1$.*

Proof. From Lemma 1 and 3 it follows that for some $x > 3t$, the (k, ℓ)-center clustering problem is NP-hard to approximate within a factor of $3\cos(\frac{\pi}{2x}) \geq 3 - \frac{3\pi}{2x}$, for $x > 0$. Hence, if we choose $x := \max(\frac{3\pi}{2\varepsilon}, 3t)+1$, for some $\varepsilon > 0$, the (k, ℓ)-center clustering problem under the discrete Fréchet distance is NP-hard to approximate within a factor of $3 - \varepsilon$. Trivially, this result generalizes to any dimension $d > 2$ using the same construction. □

3 Continuous Fréchet

For the continuous Fréchet distance we prove an analogous result as shown in Section 2. In the following, we denote $d_{CF}(\psi, \phi)$ as the continuous Fréchet distance between polygonal curves ψ and ϕ.

Lemma 5 *Let $r < \frac{3}{2}\cos(\frac{\pi}{x}) + \frac{3}{2}$. There does not exist a polygonal curve ϕ in the plane that satisfies both $d_{CF}(G_A^x, \phi) \leq r$ and $d_{CF}(G_B^x, \phi) \leq r$.*

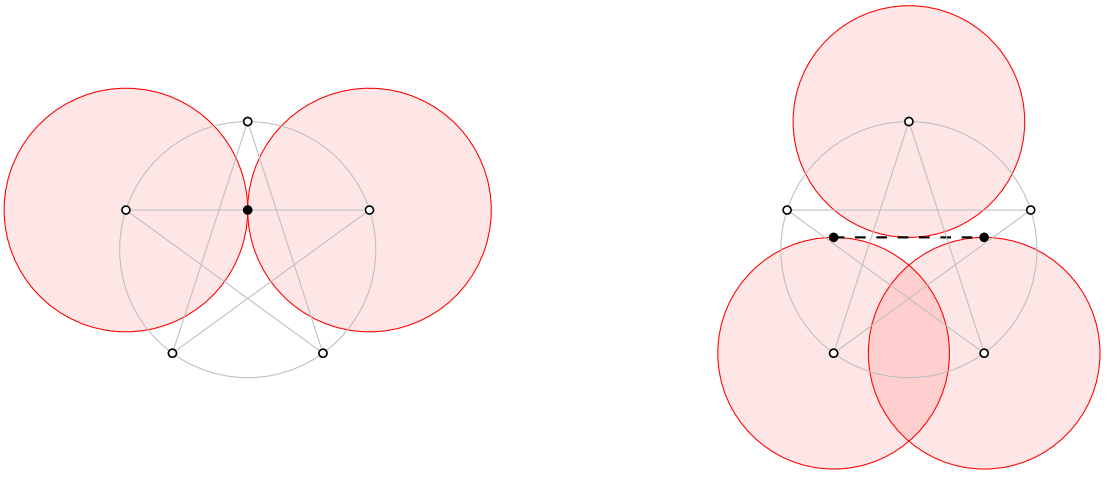
Proof. Assume there exists a curve ϕ with $d_{CF}(G_A^x, \phi) \leq r$ and $d_{CF}(G_B^x, \phi) \leq r$. We start by showing that for each vertex $v \in G_A^x$ there must exist a vertex $w \in \phi$ such that the distance between v and w is at most r. Supposing this is not the case, some v must be matched to an edge of ϕ. However, since three consecutive points of G_A^x form an isosceles triangle, when $r < \frac{3}{2}\cos(\frac{\pi}{x}) + \frac{3}{2}$ this is not possible (see Figure 3b). Analogously, the same can be shown for G_B^x.

Furthermore, with a similar argument as was used in the proof of Lemma 2, each vertex of ϕ cannot be matched to two consecutive vertices of G_A^x or G_B^x. Hence, since $\frac{3}{2}\cos(\frac{\pi}{x}) + \frac{3}{2} < 3\cos(\frac{\pi}{2x})$ for $x \geq 3$, curve ϕ cannot exist. □

The proofs of the following lemmas are omitted, as they are analogous to the proofs of Lemmas 1 and 3.

Lemma 6 *For any true instance of the SCS decision problem, there exists a center curve of length at most $(x+3)t$ and radius at most $r = 1$ under the continuous Fréchet distance in our construction.*

Lemma 7 *For any instance of an SCS decision problem, consider the above reduction with $x > 3t$. If there exists a center curve of length at most $(x + 3)t$ and of radius strictly smaller than $r = \frac{3}{2}\cos(\frac{\pi}{x}) + \frac{3}{2}$ under the continuous Fréchet distance, then the instance is true.*



(a) $r = 3\cos(\frac{\pi}{2x})$ (b) $r = \frac{3}{2}\cos(\frac{\pi}{x}) + \frac{3}{2}$

Figure 3: Critical Fréchet distances for the discrete and continuous case.

Again, from these results, we can finally conclude with the following theorem.

Theorem 8 *The (k, ℓ)-center problem under the continuous Fréchet distance is NP-hard to approximate within a factor of $3 - \varepsilon$ for any $\varepsilon > 0$ for curves in $\mathbb{R}^d$ with $d \geq 2$, even if $k = 1$.*

Proof. From Lemma 1 and 3 it follows that for some $x > 3t$, the (k, ℓ)-center clustering problem is NP-hard to approximate within a factor of $\frac{3}{2}\cos(\frac{\pi}{x}) + \frac{3}{2} \geq 3 - \frac{3\pi}{2x}$, for $x > 0$. Hence, if we choose $x := \max(\frac{3\pi}{2\varepsilon}, 3t) + 1$, for some $\varepsilon > 0$, the (k, ℓ)-center clustering problem under the continuous Fréchet distance is NP-hard to approximate within a factor of $3 - \varepsilon$. Trivially, this result generalizes to any dimension $d > 2$ using the same construction. □

References

[1] Pankaj Agarwal and Cecilia Magdalena Procopiuc. Exact and approximation algorithms for clustering. *Algorithmica*, 33:201–226, 2002.

[2] Milutin Brankovic, Kevin Buchin, Koen Klaren, André Nusser, Aleksandr Popov, and Sampson Wong. (k, l)-medians clustering of trajectories using continuous dynamic time warping. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, page 99–110. Association for Computing Machinery, 2020.

[3] 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 Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '19, page 2922–2938. Society for Industrial and Applied Mathematics, 2019.

[4] Kevin Buchin, Anne Driemel, Natasja van de L'Isle, and André Nusser. klcluster: Center-based clustering of trajectories. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 496–499. Association for Computing Machinery, 2019.

[5] Maike Buchin, Anne Driemel, and Dennis Rohde. Approximating (k,ℓ)-median clustering for polygonal curves. *ACM Trans. Algorithms*, 19(1), February 2023.

[6] Maike Buchin and Dennis Rohde. Coresets for (k, ℓ)-median clustering under the fréchet distance. In *Conference on Algorithms and Discrete Applied Mathematics*, pages 167–180. Springer, 2022.

[7] Siu-Wing Cheng and Haoqiang Huang. Curve simplification and clustering under fréchet distance. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '23, pages 1414–1432. Society for Industrial and Applied Mathematics, 2023.

[8] Wen-Lian Hsu and George Nemhauser. Easy and hard bottleneck location problems. *Discrete Applied Mathematics*, 1(3):209–215, 1979.

[9] 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.