On the complexity of the middle curve problem*

Maike Buchin[1], Nicole Funk[2], and Amer Krivošija[3]

1 **Ruhr-Universität Bochum, Germany**
maike.buchin@rub.de
2 **Department of Computer Science, TU Dortmund, Germany**
nicole.funk@tu-dortmund.de
3 **Department of Computer Science, TU Dortmund, Germany**
amer.krivosija@tu-dortmund.de

Abstract

For a set of curves, Ahn et al. [1] introduced the notion of a *middle curve* and gave algorithms computing these with run time exponential in the number of curves. Here we study the computational complexity of this problem: we show that it is NP-complete and give approximation algorithms.

1 Introduction

Consider a group of birds migrating together. Several of these birds are GPS-tagged to analyze their behavior. The resulting data is a set of sequences of their positions. Such a sequence of data points can be interpreted as a polygonal curve. We want to represent the movement of the whole group, for instance to compare it to other groups or species. For this, we use a representative curve. Such a representative curve is also useful in other applications, such as the analysis of handwritten text or speech recognition.

There have been a few different approaches of defining such a representative curve. Buchin *et al.* [4] defined the median level of curves as only using parts input curves, where the median can change directions where two input curves cross paths. Har-Peled and Raichel [10] define a mean curve, which can be chosen freely and minimizes the distance to the input curves. They give an algorithm exponential in the number of curves for computing this.

Another approach is a version of the (k, ℓ)-center problem, which asks for a set of k center curves of complexity at most ℓ for which the distance of each input curve to its nearest center is minimized. In particular, the $(1, \ell)$-center problem asks for only one such center curve. The (k, ℓ)-center problem for curves was first introduced by Driemel *et al.* [8] and further analyzed by Buchin *et al.* [5] and Buchin *et al.* [6].

However, none of these representative curves use only actual data points of the GPS tracks. This could lead to the representative curves containing positions that the moving entities (e.g. birds) could not have visited. As the data points in the input curves are more reliable Ahn *et al.* [1] defined the *middle curve* to only use these points. For a more accurate representation of the original curves, Ahn *et al.* [1] define three variants of the middle curve. We use their definition of a middle curve in this paper.

Related work Ahn *et al.* [1] presented algorithms for all three variants of the middle curve problems, whose running time is exponential in the number of input curves. For several

* This work is based on the student research project by the second author N. Funk. A. Krivošija was supported by the German Science Foundation (DFG) Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis", project A2. A full version of this work is available at http://arxiv.org/abs/2001.10298

36th European Workshop on Computational Geometry, Würzburg, Germany, March 16–18, 2020.
This is an extended abstract of a presentation given at EuroCG'20. 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.

representative curve problems it is known that they are NP-hard, such as (k, ℓ)-center [5, 8], minimum enclosing ball [5], (k, ℓ)-median [8], 1-median under Fréchet and dynamic time warping distance [6, 7]. Some problems are NP-hard even to approximate better than a constant factor, e.g. (k, ℓ)-center problem [5]. Similarly, Buchin *et al.* [3] showed, that assuming the Strong Exponential Time Hypothesis (SETH) the Frechet distance of k curves of complexity n each cannot be computed significantly faster than $O(n^k)$ time.

Our results We prove NP-completeness of the Middle Curve problem presented by Ahn *et al.* [1]. Next we define a parameterized version of the problem, and present a simple exact algorithm as well as an $(2 + \varepsilon)$-approximation algorithm for the parameterized problem.

2 Preliminaries

A polygonal curve P is given by a sequence of vertices $\langle p_1, \ldots, p_m \rangle$ with p_i in $\mathbb{R}^d$, $1 \leq i \leq m$, and for $1 \leq i < m$ the pair of vertices (p_i, p_{i+1}) is connected by the straight line segment $\overline{p_i p_{i+1}}$. We call the number of vertices m of the curve its *complexity*. Let the input consist of n polygonal curves $\mathcal{P} = \{P_1, \ldots, P_n\}$, each of complexity m.

Fréchet Distance We define the discrete Fréchet distance of two curves $P' = \langle p'_1, \ldots, p'_{m'} \rangle$ and $P'' = \langle p''_1, \ldots, p''_{m''} \rangle$ as follows: we call a *traversal* T of P' and P'' a sequence of pairs of indices (i, j) of vertices $(p'_i, p''_j) \in P' \times P''$ such that

i) the traversal T begins with $(1, 1)$ and ends with (m', m''), 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 note that every traversal is monotone. Denote $\mathcal{T}$ the set of all traversals T of P' and P''. The discrete Fréchet distance between P' and P'' is defined as:

$$d_{dF}(P', P'') = \min_{T \in \mathcal{T}} \max_{(i,j) \in T} \|p_i - q_j\|_2.$$

We call the set of pairs of vertices $(p', p'') \in P' \times P''$ that realize $d_{dF}(P', P'')$ a *matching*, and say that these pairs of vertices are matched. A related similarity measure is the continuous Fréchet distance d_F, we refer to [2] and the full version for details. Both d_{dF} and d_F are metrics.

Middle Curve Given a set of n polygonal curves $\mathcal{P}$, a value $\delta \geq 0$, and a distance measure γ for polygonal curves. We use $\gamma = d_{dF}$ as in [1], for the continuous Fréchet distance d_F the definitions hold verbatim. A **middle curve** at distance δ to $\mathcal{P}$ is a curve $M = \langle m_1, \ldots, m_\ell \rangle$ with vertices $m_i \in \bigcup_{P_j \in \mathcal{P}} \bigcup_{p \in P_j} \{p\}$, $1 \leq i \leq \ell$, s.t. $\max\{d_{dF}(M, P_j) : P_j \in \mathcal{P}\} \leq \delta$ holds.

If the vertices of a middle curve M respect the order given by the curves of $\mathcal{P}$, then we call M an **ordered middle curve**. Formally, for all $1 \leq j \leq n$, if the vertex $m_i \in M$ is matched to $p_k \in P_j$ realizing $d_{dF}(M, P_j)$, then for the vertices $m_{i'} \in M$, $i < i'$, it holds that $m_{i'} \in \left(\bigcup_{P_x \in \mathcal{P} \setminus P_j} \bigcup_{p \in P_x} \{p\}\right) \cup (\bigcup \{p_{k'} : p_{k'} \in P_j, k' > k\})$. If the vertices of M are matched to themselves in their original curves $P \in \mathcal{P}$ in the matching realizing $d_{dF}(M, P) \leq \delta$, we have a **restricted middle curve**. Note that an ordered middle curve is a middle curve, and a restricted middle curve is ordered as well.

We define the decision problem corresponding to finding such a curve. Given a set of polygonal curves $\mathcal{P} = \{P_1, \ldots, P_n\}$ and a $\delta \geq 0$ as parameters. Unordered Middle Curve problem returns TRUE iff there exists a middle curve M at distance δ to $\mathcal{P}$. The Ordered

Middle Curve and Restricted Middle Curve returns true iff there exists an ordered and a restricted middle curve respectively at distance δ to $\mathcal{P}$.

Ahn *et al.* [1] presented dynamic programming algorithms for each variant of the middle curve problem. The running times of these algorithms for $n \geq 2$ curves of complexity at most m are $O(m^n \log m)$ for the unordered case, $O(m^{2n})$ for the ordered case, and $O(m^n \log^n m)$ for the restricted middle curve case. All three cases have running time exponential in n, yielding the question if there is a lower bound for these problems. In the following section we prove that the Middle Curve problem is NP-complete.

3 NP-completeness

The technique for the proof that all variants of the Middle curve are NP-hard is based on the proof by Buchin *et al.* [5] and Buchin, Driemel, and Struijs [6] for the NP-hardness of the Minimum enclosing ball and 1-median problems for curves under Fréchet distance. Their proof is a reduction from the Shortest Common Supersequence (SCS), which is known to be NP-hard [11]. SCS problem gets as input a set $S = \{s_1, \ldots, s_n\}$ of n sequences over a binary alphabet $\Sigma = \{A, B\}$ and $t \in \mathbb{N}$. SCS returns true iff there exists a sequence s^* of length at most t, that is a supersequence of all sequences in S.

Our NP-hardness proof differs from the proof of [5, 6] in three aspects. First, the mapping of the characters of the sequence is extended by additional points. Second, in order to validate all three variants of our problem, the conditions of the restricted middle curve have to be fulfilled, i.e. each vertex has to be matched to itself. Third, our representative curve is limited to the vertices of the input curves. Due to the hierarchy of the middle curve problems we show the reductions from SCS to the Restricted Middle Curve, and from Unordered Middle Curve to SCS. We get the following theorem.

▶ **Theorem 3.1.** *Every variant of Middle Curve problem for the discrete and the continuous Fréchet distance is NP-hard.*

With this we can prove the NP-completeness of the Middle Curve decision problem. Given a Middle Curve instance $(\mathcal{P}, \delta)$ with $\mathcal{P}$ containing n curves of complexity m, we guess non-deterministically a middle curve M of complexity ℓ. We can decide whether the Fréchet distance between M and a curve $P \in \mathcal{P}$ is at most δ in $\mathcal{O}(m\ell)$ time using the algorithm by Alt and Godau [2] for the continuous, and by Eiter and Mannila [9] for the discrete Fréchet distance. We note that the algorithm by Alt and Godau [2] has to be modified a bit, as it uses a random access machine instead of a Turing machine, as this allows the computation of square roots in constant time. But comparing the distances is possible by comparing the squares of the square roots, thus this results in a non-deterministic $\mathcal{O}(nm\ell)$-time algorithm for the Middle Curve problem.

In order to decide the Ordered Middle Curve problem, it is necessary to compare the middle curve to the input curves, which is possible in $O(nm)$ time. For the restricted Restricted Middle Curve problem the matching corresponding to the Frechet distance $\leq \delta$ has to be known. This matching is a result of the decision algorithm by Alt and Godau [2]. Given this matching it can be checked in $\mathcal{O}(m + \ell)$ time if a vertex is matched to itself. This yields the following theorem.

▶ **Theorem 3.2.** *Every variant of the Middle Curve problem for the discrete or continuous Fréchet distance is NP-complete.*

If the SCS problem is parameterized by the number of input sequences n, it is known to be W[1]-hard [7]. In our reduction from SCS the number of input curves in the constructed

MIDDLE CURVE instance is $n+2$. Thus the shown reduction is also a parameterized reduction from SCS with the parameter n to the MIDDLE CURVE problem parameterized by the number of input curves, yielding the following theorem.

▶ **Theorem 3.3.** *The* MIDDLE CURVE *problem for the discrete and continuous Fréchet distance parameterized by the number of input curves n is W[1]-hard.*

4 Approximation algorithm

A different way of parameterizing the MIDDLE CURVE problem is to use the complexity of the middle curve. Given a set of polygonal curves $\mathcal{P}$, a $\delta \geq 0$, and a parameter $\ell \in \mathbb{N}$. We define the PARAMETERIZED MIDDLE CURVE decision problems, that return TRUE iff a middle curve of complexity $\leq \ell$ with corresponding conditions exists (for each of the three variants).

It is clear that there exists a simple brute force optimization algorithm for the PARAMETERIZED MIDDLE CURVE instance $(\mathcal{P}, \delta, \ell)$, that tests all ℓ-tuples of the vertices from the curves in $\mathcal{P}$ in $\mathcal{O}\big((mn)^{\ell} m\ell \log m\ell\big)$. This holds for all three versions of the problem.

We want to give an approximation algorithm for PARAMETERIZED MIDDLE CURVE optimization problem for the discrete Frechet distance. For this we use an approximation of the (k, ℓ)-center optimization problem on curves. The (k, ℓ)-center problem for curves was introduced by Driemel *et al.* [8]. Given a set $\mathcal{P} = \{P_1, \ldots, P_n\}$ of polygonal curves of complexity at most m, it looks for a set of curves $\mathcal{C} = \{C_1, \ldots, C_k\}$, each of complexity at most ℓ, that minimizes $\max_{P \in \mathcal{P}} \min_{i=1}^{k} \gamma(C_i, P)$ for a distance measure γ. The unordered PARAMETERIZED MIDDLE CURVE optimization problem is a $(1, \ell)$-center problem, where the curve C_1 is limited to vertices from the input curves and the distance measure γ is a variant of the Fréchet distance.

Given a set $\mathcal{P}$ of n curves of complexity m in $\mathbb{R}^d$, let C be the $(1, \ell)$-center curve returned by some α-approximation algorithm for the discrete Fréchet distance. Let $\delta = \max_{P \in \mathcal{P}} d_{dF}(C, P)$. We construct d-dimensional balls centered at vertices of the curve C with radius δ. It holds that $d_{dF}(C, P) \leq \delta$, $\forall P \in \mathcal{P}$, thus in each ball centered at the vertices of C there has to be a vertex of each curve from $\mathcal{P}$. We choose at random one vertex from each of the ℓ balls, and connect them with line segments in the order of the vertices along C. We denote the curve we got with M, and claim that it is a good approximation of an unordered parameterized middle curve. See Figure 1 for an illustration of the algorithm.

Let C^* be an optimal $(1, \ell)$-center curve (for the discrete Fréchet distance) for the given input set $\mathcal{P}$. Let $\delta^* = \max_{P \in \mathcal{P}} d_{dF}(C^*, P)$. It holds that $\delta \leq \alpha\delta^*$. For each $P \in \mathcal{P}$ and each vertex of P, there is a vertex in M, that is at distance at most 2δ (diameter of the ball both of them lie in). Thus there is a traversal of P and M with pairwise distance of the vertices at most 2δ, implying $d_{dF}(M, P) \leq 2\delta$. We have $d_{dF}(M, P) \leq 2\delta \leq 2\alpha\delta^*$.

Let the optimal parameterized middle curve with complexity ℓ be M^*. By definition it holds that $\delta^* = \max_{P \in \mathcal{P}} d_{dF}(C^*, P) \leq \max_{P \in \mathcal{P}} d_{dF}(M^*, P)$. Thus

$$d_{dF}(M, P) \leq 2\alpha \max_{P \in \mathcal{P}} d_{dF}(M^*, P)\,,$$

and M is a 2α-approximation to the optimal parameterized middle curve. This implies:

▶ **Lemma 4.1.** *Given a set of n curves $\mathcal{P}$ each with complexity at most m, a $\delta > 0$ and an α-approximation algorithm for (k, ℓ)-center with running time T, we can compute a 2α-approximation of the* PARAMETERIZED MIDDLE CURVE *optimization problem for discrete Fréchet distance in $O(\ell mn + T)$ time.*

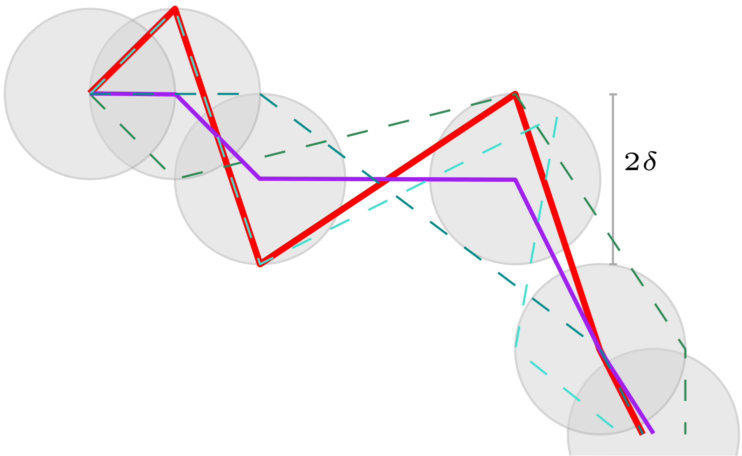

Figure 1 Illustration of the approximation algorithm. The input curves are dashed and in shades of green, while the $(1, \ell)$-center approximation with distance δ is the full purple curve. The constructed middle curve is the red fat curve.

Plugging the $(1+\varepsilon)$-approximation algorithm of Buchin *et al.* [6] for (k, ℓ)-center for discrete Fréchet distance into Lemma 4.1, we get

▶ **Theorem 4.2.** *Given a set of n curves $\mathcal{P}$ each of complexity at most m, and a $\delta > 0$, we can compute a $(2+\varepsilon)$-approximation of the* Parameterized Middle Curve *optimization problem for discrete Fréchet distance in $\mathcal{O}\big(((c\ell)^{\ell} + \log(\ell+n))\ell mn\big)$ time, with $c = (\frac{4\sqrt{d}}{\varepsilon}+1)^d$.*

5 Conclusion

We showed that the Middle Curve problem is NP-complete and gave a $(2+\varepsilon)$-approximation for the Parameterized Middle Curve problem, parameterized in the complexity of the middle curve. It would be interesting to gain further insight into the complexity of the parameterized problem. Fixing the parameter in the brute-force algorithm gives an XP-algorithm, however it remains open whether Parameterized Middle Curve is in FPT.

References

1 H.-K. Ahn, H. Alt, M. Buchin, E. Oh, L. Scharf, and C. Wenk. A middle curve based on discrete Fréchet distance. In E. Kranakis, G. Navarro, and E. Chávez, editors, *LATIN 2016: Theoretical Informatics - 12th Latin American Symposium*, pages 14–26, 2016.

2 H. Alt and M. Godau. Computing the Fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 05:75–91, 1995.

3 K. Buchin, M. Buchin, M. Konzack, W. Mulzer, and A. Schulz. Fine-grained analysis of problems on curves. In *Proceedings of the 32nd European Workshop on Computational Geometry*, 2016.

4 K. Buchin, M. Buchin, M. van Kreveld, M. Löffler, R. I. Silveira, C. Wenk, and L. Wiratma. Median trajectories. *Algorithmica*, 66(3):595–614, 2013.

5 K. Buchin, A. Driemel, J. Gudmundsson, M. Horton, I. Kostitsyna, M. Löffler, and M. Struijs. Approximating (k, ℓ)-center clustering for curves. In T. M. Chan, editor, *Proceedings of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 2922–2938, 2019.

6 K. Buchin, A. Driemel, and M. Struijs. On the hardness of computing an average curve. *CoRR*, abs/1902.08053, 2019.

7 L. Bulteau, F. Hüffner, C. Komusiewicz, and R. Niedermeier. Multivariate algorithmics for NP-hard string problems. *Bulletin of the EATCS*, 114, 2014.

8 A. Driemel, A. Krivošija, and C. Sohler. Clustering time series under the Fréchet distance. In R. Krauthgamer, editor, *Proceedings of the 27th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 766–785, 2016.

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

10 S. Har-Peled and B. Raichel. The Fréchet distance revisited and extended. *ACM Transactions on Algorithms*, 10(1):3:1–3:22, 2014.

11 K. 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.