Coresets for (k, l)-Clustering under the Fréchet Distance

Maike Buchin and Dennis Rohde

TU Dortmund University
maike.buchin@tu-dortmund.de dennis.rohde@cs.tu-dortmund.de

Abstract

We investigate the problem of clustering a set T of n polygonal curves in $\mathbb{R}^d$ under the Fréchet distance, with respect to the (k, l)-center and the (k, l)-median objective functions. These were recently defined by Driemel et al. as an adaption of the well-known k-center and k-median objectives with the restriction that the center-curves are composed of up to l line segments. Driemel et al. already developed approximation-schemes for these objectives, for $d = 1$. Recently Buchin et al. developed a constant-factor approximation algorithm for the (k, l)-center objective for general d. Further they provide hardness results for that objective. We tie in with these results by providing construction-techniques for small size ε-coresets for the (k, l)-center objective, if the given curves are of well-behaved structure, and for the discrete k-median objective. That is, we restrict the possible center-sets to all subsets of T of cardinality k and thus ignore the restriction on the complexity of the center-curves.

1 Introduction

Clustering is a thoroughly studied topic that has a great impact in the field of data analysis. Every problem in this topic has an intrinsic property: Given a collection P of n objects and an integer k, one wants to divide P into k pieces, the so called clusters, such that the objects in those clusters are some kind of related, cf. [4]. In many problem-formulations each cluster is induced by a representative object. In our setting, these representatives are given by an objective function over which one optimizes. There are three such objective functions that are well-known: k-means, k-median and k-center. Initially these functions were defined in the context of clustering points in the Euclidean space. There are also definitions of the k-median and the k-center in the context of clustering points in general metric spaces.

In our setting, we are given a set T of n polygonal curves in $\mathbb{R}^d$ endowed with the Fréchet distance and an integer k, as well as an integer l. Again we want to divide T into k clusters, i.e., we are looking for a partition of T of cardinality k. Driemel et al. [2] already studied this setting for $d = 1$. They introduce two restrictions, one on the input-curves and one on the representatives, namely the input-curves are composed of up to m line segments each and the representative curves of the clusters are composed of up to l line segments each. The respective objective functions that enforce these restrictions are called (k, l)-center and (k, l)-median. The authors develop quasi linear-time approximation-schemes for these objectives. Recently, Buchin et al. [1] developed a 3-approximate algorithm for the (k, l)-center objective for $d \in \mathbb{N}$, as well as hardness-results, i.e., for $d = 1$ the (k, l)-center is hard to approximate within a factor of $1.5 - \delta$ and for $d > 1$ within a factor of $2.25 - \delta$, for $\delta > 0$.

Let f be one of the objective functions and C be a set of k representative curves. A set S is an ε-coreset for f, if for all choices of C it holds that $|f(T, C) - f(S, C)| \leq \varepsilon \cdot f(T, C)$. Such a coreset is particularly important when clustetring queries shall be answered efficiently, i.e., return the cost for a given center-set C. In this work we give an overview of our results on small cardinality ε-coresets for the (k, l)-center objective and the discrete k-median objective,

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.

i.e., we restrict all possible center-sets to the subsets of T of cardinality k and therefore ignore the restriction on the complexity of the center-curves. For a set of line segments we provide ε-coresets of cardinality dependent on $\frac{1}{\varepsilon^{2d}}$, with respect to the (k,l)-center objective. For a set of polygonal curves of complexity at most m each we provide ε-coresets of cardinality dependent on $\frac{l^m}{\varepsilon^{dm}} + m^m$ and the ratio $\frac{\delta}{\alpha}$, where α is the value of a c-approximate solution and δ is the length of a longest line segment of any center-curve associated with that solution, with respect to the (k,l)-center objective, but only if $\frac{\delta}{\alpha} \in \mathcal{O}(\sqrt[2m]{n})$. Finally, for a set of n polygonal curves we provide ε-coresets of cardinality dependent on $\frac{\ln(n)}{\varepsilon^2}$, with respect to the discrete k-median objective. All results presented here stem from the Master thesis of the second author [6] (available on arXiv) and all proofs can be found there.

Related Work To the best of our knowledge, clustering polygonal curves under the (k,l)-center or the (k,l)-median objective has only been studied in [2] and [1], in that order. As it was already mentioned, Driemel et al. introduce the (k,l)-center and the (k,l)-median objective functions. Additionally, they develop $(1+\varepsilon)$-approximation algorithms for these objectives under the restriction that $d = 1$ and ε, k and l are fixed. The algorithms have running-time $\widetilde{\mathcal{O}}(n \cdot m)$. They also provide first hardness results for the (k,l)-center and the (k,l)-median objectives. Finally, they prove that the Fréchet space (Δ, d_F) (a formal definition follows) has unbounded doubling dimension. The 3-approximation algorithm for the (k,l)-center that is developed by Buchin et al. has running-time $\mathcal{O}\left(km(nl\log(l+m)) + m^2\log(m)\right)$. Additional to this algorithm and the already mentioned hardness results they provide similar hardness results for the discrete Fréchet distance and on the minimum enclosing ball problem for polygonal curves under the Fréchet distance.

2 Preliminaries

▶ **Definition 2.1.** A polygonal curve with vertices $v_1, \ldots, v_m \in \mathbb{R}^d$ is defined as the parametric curve connecting each contiguous pair of vertices by a line segment, which we call the edges of the curve. The number of vertices is called complexity of the curve. By Δ_m we denote the equivalence class of polygonal curves of complexity at most m and by $\Delta := \cup_{m\in\mathbb{N}_{\geq 2}}\Delta_m$ we denote the equivalence class of all polygonal curves.

▶ **Definition 2.2.** Let $\mathcal{F}$ be the set of all continuous, injective and non-decreasing functions $f\colon [0,1] \to [0,1]$ with $f(0) = 0$ and $f(1) = 1$. The Fréchet distance between polygonal curves τ and σ is defined as $d_F(\tau,\sigma) = \inf_{f\in\mathcal{F}} \max_{t\in[0,1]} \|\tau(f(t)) - \sigma(t)\|$, where $\|\cdot\|$ is the Euclidean norm.

▶ **Definition 2.3.** Given a set T of n polygonal curves of complexity at most m each and two integers k and l, the (k,l)-center objective is to return the optimal cost of $\min_{C\subset\Delta_l,|C|=k} \max_{\tau\in T} \min_{c\in C} d_F(\tau,c)$. The discrete k-median objective is to return the optimal cost of $\min_{C\subset T,|C|=k} \sum_{\tau\in T} \min_{c\in C} d_F(\tau,c)$.

▶ **Definition 2.4.** Let T be a given set of n polygonal curves. Also, let f be an objective function and C be a set of k cluster-representatives. A set S of polygonal curves is called ε-coreset for T with respect to f, if for all possible choices of C it holds that $|f(T,C) - f(S,C)| \leq \varepsilon \cdot f(T,C)$. S is called *weighted* ε-coreset if every $s \in S$ is assigned a weight $w_s \in \mathbb{R}$, that flows into the value of f.

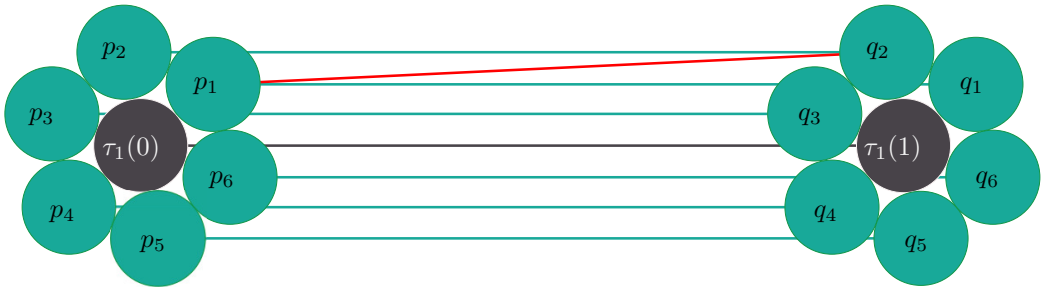

Figure 1 This is the construction used for Theorem 3.1, for $d = 2$. The curves are defined with respect to the center points of the balls. The set T, which cannot be embedded into $(\mathbb{R}^d, d_E)$, where d_E is the Euclidean distance, consists of $\overline{\tau(0)\tau(1)}$ (the common nearest neighbor), the line segments $\overline{p_i q_i}$, for $i \in \{1, \ldots, 6\}$, plus $\overline{p_1 q_2}$. The segment $\overline{p_1 q_2}$ breaks every possible isometric embedding.

3 What is the Difference between Points and Curves?

At first, we investigate whether we can transform the given polygonal curves into points in the Euclidean space through an isometric embedding. Such a transformation would have multiple benefits: When constructing an ε-coreset for any application, often one simply thins out the input-set as much as possible. When such an embedding is available we could apply existing construction-techniques, track which points are thrown out and then throw out all curves that map to these points. If only the value of a clustering is needed this would give us the opportunity to directly obtain the value through one of the numerous algorithms for points in $\mathbb{R}^d$. Unfortunately, such an isometric embedding does not exist for every possible set of polygonal curves, even if we restrict ourselves to line segments.

▶ **Theorem 3.1.** *For any $d \in \mathbb{N}$, there exists a set of polygonal curves in $\mathbb{R}^d$ that cannot be isometrically embedded into $(\mathbb{R}^d, d_E)$, where d_E is the Euclidean distance.*

This result is achieved by proving that an isometric embedding, if existent, would violate the d-dimensional kissing number, given certain sets of polygonal curves, cf. Fig. 1. The d-dimensional kissing number is the maximum number of points in $\mathbb{R}^d$ that can share a common nearest neighbor point (cf. [7]). We note that a similar result is implied due to the fact that certain four-point graphs endowed with the shortest-path metric cannot be embedded into $\mathbb{R}^d$, for any d, while they can be embedded into (Δ_{13}, d_F) with ambient space $\mathbb{R}$, cf. [3]. Nevertheless, our result is stronger because it holds for (Δ_m, d_F), for any $m \geq 2$.

4 Coresets for the (k, l)-center Objective

There is a common technique for constructing ε-coresets for the k-center objective, given a set P of points in the Euclidean space: Run a c-approximate algorithm on P to obtain a value α of the objective function. Let C be the center-set associated with this value. By the structure of the objective function we have that $P \subseteq \cup_{q \in C} \{p \in \mathbb{R}^d \mid \|p - q\| \leq \alpha\} =: E$. Additional, if α^* is the optimal cost for P under the k-center objective, then $\frac{\alpha}{c}$ is a lower bound on this number.

Now around every $q \in C$ we place a grid G_q of edge length $2 \cdot \alpha$, thus $E \subseteq \cup_{q \in C} G_q$. We set the edge-length of the cells of every grid to $\varepsilon \cdot \frac{1}{\sqrt{d}} \cdot \frac{\alpha}{c}$, thus we cannot move a point contained in a cell more than $\varepsilon \cdot \alpha^*$ without leaving the cell. At last we go through every cell of all G_q and if it contains more than one point from P we remove all but one of those from P. The resulting set P' is an ε-coreset for the k-center objective of cardinality $\mathcal{O}\left(\frac{1}{\varepsilon^d}\right)$.

This scheme can easily be adapted for a set T of polygonal curves, utilizing the 6-approximation algorithm by Buchin et al. [1]. For line segments this is particularly easy: Place such a grid around each end point of every center-curve. Again, by the structure of the objective function the end points of the input-curves are contained in those grids. For a curve τ call $\tau(0)$ its initial point and $\tau(1)$ its end point, further call the grid around $\tau(0)$ the initial point grid and the grid around $\tau(1)$ the end point grid. Now, successively for each center-curve, we go through every pair of a cell of the initial point grid and a cell of the end point grid and remove all but one line segment from T that have their initial point, respective end point in those cells. The resulting set T' is an ε-coreset for the $(k,2)$-center objective of cardinality $\mathcal{O}\left(\frac{1}{\varepsilon^{2d}}\right)$.

▶ **Theorem 4.1.** *There exists an algorithm that, given a set of n line segments in d-dimensional Euclidean space and a parameter $\varepsilon \in (0,1)$, computes an ε-coreset for the $(k,2)$-center objective of cardinality $\mathcal{O}\left(\frac{1}{\varepsilon^{2d}}\right)$, in time $\mathcal{O}\left(\frac{n}{\varepsilon^{2d}}\right)$.*

For polygonal curves of complexity at least 3 this scheme has a flaw: The vertices of the input-curves do not necessarily lie within distance α to any vertex of a center-curve. Thus, we have to cover the whole center-curves with grids and therefore the cardinality of the resulting ε-coreset depends on the ratio of roughly the length δ of a longest edge of any center-curve and the value α of the c-approximate solution. For the ε-coreset to have sublinear cardinality we have to check if $\frac{\delta}{2\alpha}$ exceeds, say $\sqrt[2m]{n}$, in advance. If this is the case we are not able to provide an ε-coreset utilizing this technique. If this is not the case we now have to consider any combination of m cells of the grids that cover a center-curve, therefore the cardinality of the resulting ε-coreset is exponential in m.

▶ **Theorem 4.2.** *There exists an algorithm that, given a set of n polygonal curves of complexity at least 3 and at most m each, in d-dimensional Euclidean space and a parameter $\varepsilon \in (0,1)$, computes an ε-coreset for the (k,l)-center objective of cardinality*

$\mathcal{O}\left(2^{3m} \cdot \sqrt{n} \cdot \frac{l^{12d^2m}}{\varepsilon^{dm}} + 2^m m^m\right)$ *in time*

$\mathcal{O}\left(\left(2^{3m} \cdot n^{1.5} \cdot \frac{l^{12d^2m} m}{\varepsilon^{dm}} + 2^m m^{m+1} n\right) + nm\log(m) + m^3\log(m)\right)$, *if* $\frac{\delta}{\alpha} \in \mathcal{O}(\sqrt[2m]{n})$. *Otherwise, the algorithm fails and then has running-time* $\mathcal{O}\left(nm\log(m) + m^3\log(m)\right)$.

5 Coresets for the discrete k-median Objective

For the (discrete) k-median objective we use standard-techniques for approximating sums. In [5] Langberg and Schulman define a sensitivity sampling framework and show how it can be used to approximate the value of sum-based clustering objectives such as the k-median or the k-means. The proofs are formulated with respect to point-sets from the Euclidean space and some arbitrary norm. Nevertheless, they also work for polygonal curves under the Fréchet distance.

However, to the best of our knowledge, there are no results on the VC dimension of the Fréchet space (Δ, d_F) yet[1]. Therefore, to bound the probability that a sample is an ε-coreset,

[1] Though there are results to appear at SoCG '19 by Driemel et al.

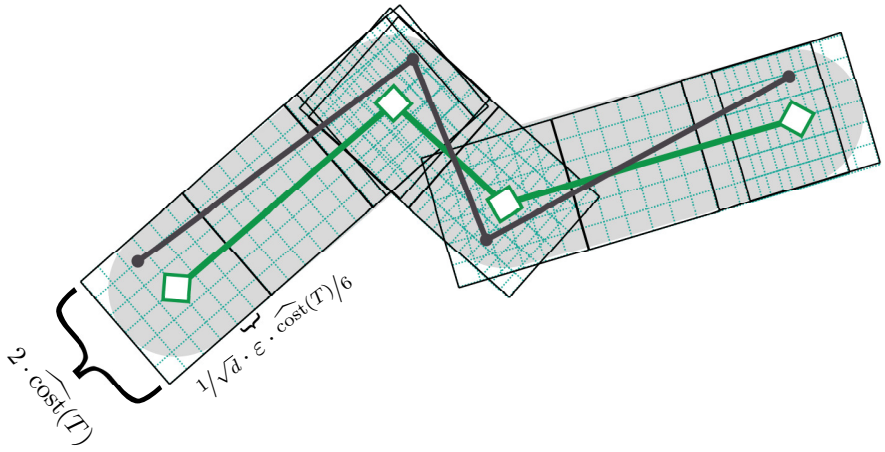

Figure 2 Exemplary grid-cover that is used for the algorithm of Theorem 4.2, for general polygonal curves and $d = 2$. A center-curve is depicted in green with cubes in black and associated grids in light blue. A curve with Fréchet distance less than $\widehat{\text{cost}}(T) := \alpha$ is also depicted. It can be observed that the vertices of this curve lie in at least one cell of a grid.

in particular that it can be used to approximate the value of the objective function for *all* possible center-sets, we restricted ourselves to the *discrete* k-median objective.

The sensitivity sampling framework works as follows: We are given a set T of n polygonal curves and run a 6-approximation algorithm to obtain a value α on the k-median objective function and the associated center-set C. The approximation algorithm we use is a local-search heuristic that uses a solution from the 6-approximation algorithm for the (k, l)-center objective by Buchin et al. as initial guess and thus has running-time $\mathcal{O}(n^2m^2 \log(m))$. We use α and C to assign every $\tau \in T$ a sensitivity value s_τ. These sensitivities and the total sensitivity $S := \sum_{\tau \in T} s_t$ suffice to build a probability distribution $\psi \colon T \to [0, 1]$, such that a sample of cardinality $\ell(\varepsilon, n) \in \Omega\left(\frac{\ln(n)}{\varepsilon^2}\right)$ from T with respect to ψ is a weighted ε-coreset for the discrete k-median objective with probability at least $\frac{2}{3}$, where every member of the sample is weighted by $\frac{n}{\ell(\varepsilon,n)}$. This is due to the fact that curves which have high impact on the value of the objective function for at least one center-set are sampled with higher probability, i.e., the probability to sample a curve is proportional to its "importance".

▶ **Theorem 5.1.** *There exists an algorithm that, given a set of n polygonal curves of complexity at most m each, and a parameter $\varepsilon \in (0, 1)$, computes an ε-coreset for the discrete k-median objective of cardinality $\mathcal{O}\left(\frac{\ln(n)}{\varepsilon^2}\right)$ in time $\mathcal{O}\left(n^2 \cdot m^2 \log(m) + \frac{\ln^2(n)}{\varepsilon^2}\right)$, with probability at least $\frac{2}{3}$.*

References

1 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*, pages 2922–2938. 10.1137/1.9781611975482.181.

2 Anne Driemel, Amer Krivošija, and Christian Sohler. Clustering Time Series Under the Fréchet Distance. In *Proceedings of the Twenty-seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 766–785. Society for Industrial and Applied Mathematics, 2016. ISBN 978-1-611974-33-1.

3 Piotr Indyk, Piotr Indyk, and Jiri Matousek. Low-Distortion Embeddings of Finite Metric Spaces. *Handbook of Discrete and Computational Geometry*, pages 177—-196, 2004.

4 Anil K. Jain. Data Clustering: 50 Years Beyond k-means. *Pattern Recognition Letters*, 31 (8):651–666, 2010. ISSN 0167-8655. 10.1016/j.patrec.2009.09.011.

5 Michael Langberg and Leonard J. Schulman. Universal Epsilon-Approximators for Integrals. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 598–607, 2010. 10.1137/1.9781611973075.50.

6 Dennis Rohde. Coresets for (k, l)-Clustering under the Fréchet Distance. Master's thesis, TU Dortmund University, December 2018.

7 Kenneth Zeger and Allen Gersho. Number of Nearest Neighbors in a Euclidean Code. *IEEE Transactions on Information Theory*, 40(5):1647–1649, 1994. ISSN 0018-9448. 10.1109/18.333884.