Random projections and sampling algorithms for clustering of high-dimensional polygonal curves

Stefan Meintrup
Faculty of Computer Science
TU Dortmund University
Dortmund, Germany
stefan.meintrup@tu-dortmund.de

Alexander Munteanu*
Dortmund Data Science Center
TU Dortmund University
Dortmund, Germany
alexander.munteanu@tu-dortmund.de

Dennis Rohde
Chair of Efficient Algorithms and Complexity Theory
TU Dortmund University
Dortmund, Germany
dennis.rohde@cs.tu-dortmund.de

Abstract

We study the center and median clustering problems for high-dimensional polygonal curves with finite but unbounded complexity. We tackle the computational issue that arises from the high number of dimensions by defining a Johnson-Lindenstrauss projection for polygonal curves. We analyze the resulting error in terms of the Fréchet distance, which is a natural dissimilarity measure for curves. Our algorithms for the median clustering achieve sublinear dependency on the number of input curves via subsampling. For the center clustering we utilize Buchin et al. (2019a) algorithm that achieves linear running-time in the number of input curves. We evaluate our results empirically utilizing a fast, CUDA-parallelized variant of the Alt and Godau algorithm for the Fréchet distance. Our experiments show that our clustering algorithms have fast and accurate practical implementations that yield meaningful results on real world data from various physical domains.

1 Introduction

Time-series are sequences of measurements taken at certain instants of time. They arise in numerous applications, e.g., in the physical, geo-spatial, technical or financial domains (Zhang et al., 2007; Chapados and Bengio, 2008; Zimmer et al., 2018). Often there are multiple measurements per time instant, e.g., when there are numerous synchronized sensors. While the analysis of time-series is a well-studied topic, cf. Hamilton (1994); Liao (2005); Aghabozorgi et al. (2015), there are only few approaches that take high-dimensional multivariate time-series into account. In this work we build upon Driemel et al. (2016), who developed the first \((1 + \varepsilon)\)-approximation algorithms for clustering univariate time-series under the Fréchet distance. Their idea is that due to environmental circumstances—time-series often have heterogeneous lengths and their measurements are taken at different time-intervals. Common approaches, where univariate time-series are represented by a point in a high-dimensional space, each dimension corresponding to one instant of time, become hard or even impossible to apply. Rather than focusing on concrete instants of time, Driemel et al. (2016) suggest considering only the monotonous order of the measurements. An implicit linear interpolation between every pair of consecutive measurements yields a polygonal curve. The Fréchet

*is supported jointly by the faculties of Statistics and Computer Science within the Dortmund Data Science Center (DoDSc)
distance is a natural distance measure for the resulting curves. Unfortunately, it is rarely used due to its computational cost. It intuitively measures the maximum distance one must traverse, when continuously and monotonously walking along two curves under an optimal speed adjustment.

We extend this further to the multivariate case. When synchronized sensors are available, we construct high-dimensional curves from multiple univariate time-series, i.e., the number of dimensions equals the number of simultaneously measured attributes. In this work we focus on a large number of dimensions, say \( d \in \Omega(n) \) (\( n \): number of input-curves) and high complexity of the curves, i.e., the number of their vertices is bounded by \( m \in \Omega(n) \) each. Since the Alt and Godau algorithm for computing the Fréchet distance by Alt and Godau (1995) has running-time \( O(d \cdot m^2 \log(m)) \), we end up with a super-cubic complexity in the number of input curves. We tackle this issue by parallelizing the Alt and Godau algorithm via CUDA-enabled GPUs and reducing the number of dimensions. For the latter, SVD-based feature-selection approaches are common, cf. Billsus and Pazzani (1998); Hong (1991). Instead, we focus on Gaussian random projections via the seminal Johnson-Lindenstrauss embedding (Johnson and Lindenstrauss, 1984). These directly yield explicit error-guarantees for the discrete Fréchet distance since it is based on a finite number of Euclidean distances. But if we restrict to this discrete distance measure, we would loose the aforementioned linear interpolation between every two consecutive measurements, which is desirable and implicitly included in the continuous Fréchet distance.

We thus study how the error from the Johnson-Lindenstrauss embedding propagates in the continuous case. In our theoretical analysis we show the first explicit error bound for the continuous Fréchet distance by extending the Johnson-Lindenstrauss embedding to polygonal curves. We project the vertices of the curve down from \( d \) to \( O(\varepsilon^{-2} \log n) \) dimensions and re-connect their images in the low-dimensional space in the given order. The error is bounded by an \( \varepsilon \)-fraction relative to the Fréchet distance and to the length of the largest edge of the input curves. This gives a combined multiplicative and additive approximation guarantee, similar to the lightweight coresets of Bachem et al. (2018).

Just as Driemel et al. (2016), we study the center and median clustering problems. We restrict the search space of feasible solutions to the input. Those problems have polynomial-time exhaustive-search algorithms: calculate the cost of each possible curve by maximizing or summing –depending on the objective– over all other input curves. In our setting, this is prohibitive since it takes \( O(n^2) \) distance evaluations. We evaluate the \( k \)-center approximation algorithm by Buchin et al. (2019a) based on Gonzalez’ algorithm. We analyze various data sets from machine learning databases and a real-world data set from the DELTA particle accelerators and storage ring facility. Our experiments show promising results concerning the approximation of the Fréchet distance under the Johnson-Lindenstrauss embedding and a massive improvement of the running-time, as well as the robustness of the resulting clusters.

Further we propose and analyze a sampling-scheme for the discrete median under the Fréchet distance when the number of input curves is also high. We show that a sample of constant size already yields a \((2 + \varepsilon)\)-approximation in the worst case. Under reasonable assumptions on the distribution of the data, the same algorithm yields a \((1 + \varepsilon)\)-approximation. To this end we introduce a natural parameter that quantifies the fraction of outliers as a function of \( n \), setting this approach in the light of beyond worst-case analysis, cf. Roughgarden (2019). The number of samples needed depends on this parameter and is almost always constant unless the fraction of outliers tends to 1/2 at a high rate, depending on \( n \). Finally, we note that our techniques do not only apply to multivariate time-series, but to high-dimensional polygonal curves in general and thus may be valuable to the communities of computational geometry as well as the field of machine learning.

Our contributions We advance the study of clustering high-dimensional polygonal curves under the Fréchet distance both, in theory and in practice. Specifically,

1) we show an extension of the Gaussian random projections of Johnson-Lindenstraus to polygonal curves and provide rigorous bounds on the distortion of their continuous Fréchet distance.

2) we provide a highly efficient CUDA-parallelized implementation of the algorithm by Alt and Godau (1995) for computing the Fréchet distance.

3) we experimentally assess how the developed techniques perform in the context of approximation algorithms for the \( k \)-center clustering of time series resp. polygonal curves under the Fréchet distance.
4) we provide sublinear sampling algorithms for the 1-median clustering of time series resp. polygonal curves under the Fréchet distance.

5) we evaluate the proposed methods on benchmark and real-world data from various physical domains.

1.1 Related work

**Clustering under the Fréchet distance** Driemel et al. (2016) developed the first \(k\)-center and \(k\)-median clustering algorithms for one-dimensional polygonal curves under the Fréchet distance, which provably achieve an approximation factor of \((1 + \varepsilon)\). The resulting centers are curves from a discretized family of simplified curves, whose complexity is parameterized by a parameter \(\ell\). Their algorithms have near-linear running time in the input size for constant \(\varepsilon, k\) and \(\ell\) but are exponential in the latter quantities. The first extension of \(k\)-center to higher dimensional curves was done in Buchin et al. (2019a). In that paper, however it was shown that there is no polynomial-time approximation scheme unless P=NP. In the case of the discrete Fréchet distance on two-dimensional curves, the hardness of approximation within a factor close to 2.598 was established even for \(k = 1\). Finally, Gonzalez’ algorithm yields a \(3\)-approximation in any dimension. Even more recently Buchin et al. (2019b) showed that the \(k\)-median problem is also NP-hard for \(k = 1\) and improved upon the aforementioned \((1 + \varepsilon)\)-approximations. Open problems thus include dimensionality reduction for high-dimensional curves and practical algorithms that do not depend exponentially on the involved parameters.

**Algorithm engineering for the Fréchet distance** Bringmann et al. (2019) describe an improved version of one of the best algorithms that was developed by the participants of the GIS Cup 2017. The goal of the cup was to answer Fréchet queries as fast as possible, i.e., given a set of curves \(T\), a query curve \(q\) and a positive real \(r\), return all curves from \(T\) that are within distance \(r\) to \(q\). Roughly speaking, all top algorithms (see also Baldus and Bringmann (2018); Buchin et al. (2017); Dütsch and Vahrenhold (2017)) utilized heuristics to filter out all \(\tau \in T\), that are certainly within distance \(r\) to \(q\) or certainly not. In the best case, the common algorithm by Alt and Godau only served as a relapse option when no clear decision could be found in advance. Since the heuristics mostly have sublinear running-time, the Fréchet distance computation is speed up massively in the average case. The Alt and Godau algorithm is also improved by simplifying the resulting free-space diagram.

**Random projections for problems in computational geometry** Random projections have several applications as embedding techniques in computational geometry. One of the most influential work was Agarwal et al. (2013) who applied the Johnson-Lindenstrauß embedding, among others, to surfaces, curves for the sake of tracking moving points. Only recently Driemel and Krivosija (2018) studied the first probabilistic embeddings of the Fréchet distance by projecting the curves on a random line. Another work that inspired our dimensionality reduction approach is due to Sheehy (2014). He noticed that a Johnson-Lindenstrauß embedding of points yields an embedding for their entire convex hull with additive error. We conjecture that reducing the complexity \(m\), i.e., the number of vertices, of the curves with multiplicative error is not achievable for the Fréchet distance, since Braverman et al. (2019) recently showed a lower bound of \(\Omega(n)\) for sketching, i.e., compressing the strongly related Dynamic Time Warping distance of sequences via linear embeddings.

**Beyond-worst-case and relaxations** A common assumption is that “Clustering is difficult only when it does not matter” (Daniely et al., 2012). Similarly it has been noted for many other problems that while being particularly hard to solve in the worst-case, they are relatively simple to solve for typical or slightly perturbed inputs. Beyond-worst-case-analysis tries to parametrize the notion of typical and to derive better bounds in terms of this parameter assuming its value is small. See Munteanu et al. (2018) for a recent contribution in machine learning. These assumptions are usually weaker than the norm in statistical machine learning which is closer to average-case analysis, for example when data points are modeled as i.i.d. samples from some distribution. See (Roughgarden, 2019) for an extensive overview and more details. Another complementary recent approach is weakening the usual multiplicative error guarantees by an additional additive error term in favor of a computational speedup. Those relaxations still perform competitively well in practice (Bachem et al., 2018).
2 Dimensionality Reduction for Polygonal Curves

We begin with the basic definitions, all missing proofs can be found in Appendix A in the supplement. Polygonal curves are composed of consecutive line segments which we define as follows.

**Definition 1** (line segment). A line segment between two points \( p_1, p_2 \in \mathbb{R}^d \), denoted by \( p_1p_2 \), is the set of points \( \{(1 - \lambda)p_1 + \lambda p_2 \mid \lambda \in [0, 1]\} \). For \( \lambda \in [0, 1] \) we denote by \( l_p(p_1p_2, \lambda) \) the point \((1 - \lambda)p_1 + \lambda p_2, \) lying on \( p_1p_2 \).

We next define polygonal curves. Thereby we need an exact parametrization of the points on the individual line segments to express any point on the curve in terms of its segments’ vertices. This unusually complicates the definition but simplifies the notation and will later be needed in the context of Johnson-Lindenstrauss embeddings.

**Definition 2** (polygonal curve). A parameterized curve is a continuous mapping \( \tau : [0, 1] \rightarrow \mathbb{R}^d \). Let \( \mathcal{H} \) be the set of all continuous, injective and non-decreasing functions \( h : [0, 1] \rightarrow [0, 1] \) with \( h(0) = 0 \) and \( h(1) = 1 \), which we call reparameterizations.

A curve \( \tau \) is polygonal, if there exist \( h \in \mathcal{H}, v_1, \ldots, v_m \in \mathbb{R}^d \), no three consecutive on a line, called \( \tau \)’s vertices and \( t_1, \ldots, t_m \in [0, 1] \) with \( t_1 < \cdots < t_m, t_1 = 0 \) and \( t_m = 1 \), called \( \tau \)’s instants, such that

\[
\tau(h(t)) = \begin{cases} 
  l_p(v_1v_2, h(t) - t_1) / (t_2 - t_1), & \text{if } h(t) \in [0, t_2) \\
  \vdots \\
  l_p(v_{m-1}v_m, h(t) - t_{m-1}) / (t_m - t_{m-1}), & \text{if } h(t) \in [t_{m-1}, 1]
\end{cases}
\]

In the following we will assume that \( h \) is the identity function, because the Fréchet distance, which is subsequently defined, is invariant under reparameterizations. We only need \( h \) to keep our definition general. Further, we call \( m \) the complexity of \( \tau \), denoted by \( |\tau| \). We are now ready to define the (continuous) Fréchet distance.

**Definition 3** (continuous Fréchet distance). The Fréchet distance between polygonal curves \( \tau \) and \( \sigma \) is defined as

\[
d_F(\tau, \sigma) := \inf_{h \in \mathcal{H}} \max_{t \in [0, 1]} \| \tau(t) - \sigma(h(t)) \|,
\]

where \( \| \cdot \| \) is the Euclidean norm.

We give a standard definition of the median and center clustering.

**Definition 4** (clustering). Given a set of \( T \) of polygonal curves, the \( k \)-median, resp. \( k \)-center clustering problem is to find a set \( C \) of \( k \) centers such that the sum of the distances from the curves in \( T \) to the nearest center in \( C \), respective the maximum distance of a curve in \( T \) to the nearest center in \( C \) is minimized.

In our setting we restrict the center-set to be a subset of the input, because some arbitrary curves may minimize the objective value but not every curve makes sense in the given physical context.

We next give a basic definition of the seminal Johnson-Lindenstrauss embedding result, cf. Johnson and Lindenstrauss (1984). Specifically, they showed that a properly rescaled Gaussian matrix mapping from \( d \) to \( d' \in O(\epsilon^{-2} \log n) \) dimensions satisfies the following definition with positive constant probability.

**Definition 5** (Johnson-Lindenstrauss embedding). Given a set \( P \subset \mathbb{R}^d \) of points, a function \( f : \mathbb{R}^d \rightarrow \mathbb{R}^{d'} \) is a \((1 \pm \epsilon)\)-Johnson-Lindenstrauss embedding for \( P \), if it holds that

\[
\forall p, q \in P : (1 - \epsilon)\|p - q\| \leq \|f(p) - f(q)\| \leq (1 + \epsilon)\|p - q\|,
\]

with constant probability at least \( \rho \in (0, 1] \) over the random construction of \( f \).

In Definition 6 we extend the mapping \( f \) from Definition 5 to polygonal curves by applying it to the vertices of the curves and re-connecting their images in the given order.

**Definition 6** \((1 \pm \epsilon)\)-Johnson-Lindenstrauss embedding for polygonal curves. Let \( \tau \) be a polygonal curve, \( t_1, \ldots, t_m \) be its instants and \( v_1, \ldots, v_m \) be its vertices. Let \( f \) be a \((1 \pm \epsilon)\)-Johnson-Lindenstrauss embedding for \( \{v_1, \ldots, v_m\} \). By \( F(\tau) \) we define the \((1 \pm \epsilon)\)-Johnson-Lindenstrauss
embedding for \( \tau \) as follows:

\[
F(\tau)(t) := \begin{cases} 
lp \left( f(v_1), \frac{t-t_1}{t_2-t_1} \right), & \text{if } t \in [0, t_2) \\
\vdots \\
lp \left( f(v_{m-1}), \frac{t-t_{m-1}}{t_m-t_{m-1}} \right), & \text{if } t \in [t_{m-1}, 1] 
\end{cases}
\]

For a set \( T := \{\tau_1, \ldots, \tau_n\} \) of polygonal curves we define

\[
F(T) := \{F(\tau) \mid \tau \in T \}.
\]

In this case we require the function \( f \) to be a \((1 \pm \varepsilon)\)-Johnson-Lindenstrauss embedding for the set of all vertices of all \( \tau \in T \).

We next give an explicit bound on the resulting error for the Fréchet distance after applying the map of Definition 6 to the input curves. Note that the previously mentioned approach by Sheehy (2014) for the convex hull of points is not directly applicable since two curves might be drawn apart from each other making the error arbitrary large. Our additive error will depend only on the length of line segments between consecutive points of a curve, which is usually bounded. We first express the distance between two points on two distinct line segments using their relative positions on the respective line segment.

**Proposition 7.** Let \( s_1 := \overrightarrow{p_1p_2} \) and \( s_2 := \overrightarrow{q_1q_2} \) be line segments between two points \( p_1 := (p_{1,1}, \ldots, p_{1,d}), p_2 := (p_{2,1}, \ldots, p_{2,d}) \in \mathbb{R}^d \), respective \( q_1 := (q_{1,1}, \ldots, q_{1,d}), q_2 := (q_{2,1}, \ldots, q_{2,d}) \in \mathbb{R}^d \). For any \( \lambda_p, \lambda_q \in [0, 1] \) and \( p := \lp (\overrightarrow{p_1p_2}, \lambda_p) \) lying on \( s_1 \), as well as \( q := \lp (\overrightarrow{q_1q_2}, \lambda_q) \) lying on \( s_2 \), it holds that

\[
\|p - q\|^2 = -(\lambda_p - \lambda_p^2)\|p_1 - p_2\|^2 - (\lambda_q - \lambda_q^2)\|q_1 - q_2\|^2 + (1 - \lambda_p - \lambda_q + \lambda_p \lambda_q)\|p_1 - q_1\|^2 + (\lambda_q - \lambda_p \lambda_q)\|p_1 - q_1\|^2 + (\lambda_p - \lambda_p \lambda_q)\|p_2 - q_1\|^2 + (\lambda_p \lambda_q)\|p_2 - q_2\|^2.
\]

Proposition 7 can be proven using the law of cosines, the geometric and algebraic definition of the dot product and tedious algebraic manipulations.

Using Proposition 7, our calculation yields an explicit error-bound when applying Definition 6 to both line-segments. This is formalized in Lemma 8.

**Lemma 8.** Let \( P := \{p_1, \ldots, p_n\} \subset \mathbb{R}^d \) be a set of points and \( f \) be a \((1 \pm \varepsilon)\)-Johnson-Lindenstrauss embedding for \( P \). Let \( p_1, p_2, q_1, q_2 \in P \), for arbitrary \( \lambda_p, \lambda_q \in [0, 1] \) and \( p := \lp (\overrightarrow{p_1p_2}, \lambda_p) \), \( p' := \lp \left( f(p_1), f(p_2), \lambda_p \right) \), as well as \( q := \lp (\overrightarrow{q_1q_2}, \lambda_q) \), \( q' := \lp \left( f(q_1), f(q_2), \lambda_q \right) \) it holds that

\[
(1 - \varepsilon)^2\|p - q\|^2 - \varepsilon(\|p_1 - p_2\|^2 + \|q_1 - q_2\|^2) \leq \|p' - q'\|^2 \leq (1 + \varepsilon)^2\|p - q\|^2 + \varepsilon(\|p_1 - p_2\|^2 + \|q_1 - q_2\|^2)
\]

is satisfied with probability at least \( \rho \in (0, 1] \) over the random construction of \( f \).

This finally yields our main theorem which states the desired error guarantee for the Fréchet distance of a set of polygonal curves.

**Theorem 9.** Let \( T := \{\tau_1, \ldots, \tau_n\} \) be a set of polygonal curves, \( \gamma \) be the maximum distance of two consecutive vertices of a curve in \( T \) and \( F \) be a \((1 \pm \varepsilon)\)-Johnson-Lindenstrauss embedding for \( T \). With constant probability at least \( \rho \in (0, 1] \) it holds for all \( \tau, \sigma \in T \) that

\[
\sqrt{(1 - \varepsilon)^2d_\rho^2(\tau, \sigma) - 2\varepsilon \gamma^2} \leq d_F(F(\tau), F(\sigma)) \leq \sqrt{(1 + \varepsilon)^2d_\rho^2(\tau, \sigma) + 2\varepsilon \gamma^2},
\]

where the exact value for \( \rho \) stems from the technique used for obtaining \( f \).

Let us first note that these bounds tend to \( d_F(\tau, \sigma) \) as \( \varepsilon \) tends to 0. The multiplicative error bounds are similar to coresets which are popular data reduction techniques in clustering, cf. (Feldman et al., 2010, 2013; Sohler and Woodruff, 2018). The additional additive error is in line with the relaxation given by lightweight coresets (Bachem et al., 2018).
3 Experiments

The main motivation for our work is that any application which is utilizing the Fréchet distance suffers from its computational cost. In general, there are three parameters on which the running-time depends: the number of curves, their complexity and the dimension of the ambient space. We tackle the first utilizing subsampling schemes, the second by parallelizing the Alt and Godau algorithm and the third using the Johnson-Lindenstrauss embedding. We now seek to answer the following questions:

Q1 What is the quality of the Fréchet distance computation after applying the Johnson-Lindenstrauss embedding?

Q2 What is the impact of our techniques on the running-time of the Fréchet distance computation?

Q3 Does the approximation quality suffice to obtain meaningful and robust results when solving clustering problems?

Before we answer these questions based on our experimental results, we describe our data sets, the modifications we applied to the Alt and Godau algorithm, and our setup.

Data sets Our first data set stems from the UCI machine learning repository. It was taken by monitoring a hydraulic test rig via multiple sensors (cf. Helwig et al. (2015)), including six pressure sensors PS1, . . . , PS6. In a total of 2204 test-cycles, each sensor measured 6000 values in each cycle. We chose to build six polygonal curves with 2204 points each in the 6000-dimensional Euclidean space for the sole purpose of answering Q1 and Q2. The second and most important data set stems from the DELTA electron accelerator (TU Dortmund University, 2019). In about twenty weeks every year the DELTA storage ring only serves as supplier of synchrotron radiation for numerous affiliated experiments. During these weeks the focus of the facility lies on very stable and reliable operation. In the last year sporadic jumps in beam orbit measurements disturbed the stable operation. Since these jumps occurred at different measurement points with different time-intervals in-between, they could not be explained nor correlated so far. We built 16 curves out of 16 weeks of operation, where each dimension corresponds to one of the various pressure-, temperature-, voltage-, current- or position-sensors, 616 in total. Since these sensors are not synchronized, we only combine them at instants, where every sensor has a value not older than ten minutes. The resulting curves have a number of points varying from about 200 to about 600. Informally, the curves describe the “movement” of the storage ring in the parameter-space throughout the respective week of operation. In the following we try to exhibit unknown structures in the data set by clustering utilizing our techniques, thus answering Q3. For comparison we conduct the same experiments using another data set from the UCI repository. Here, we use weather simulations (Lucas et al., 2015) to construct 2922 curves with 15 vertices each in 327-dimensional Euclidean space.

Algorithm modifications Contrary to the participants in the GIS Cup 2017 (cf. Section 1.1), we decided to parallelize the Alt and Godau algorithm utilizing CUDA-enabled graphic cards. The improvement scales linearly in the number of CUDA cores of the available graphic cards. We thus improve the worst-case behavior of the algorithm. This is crucial in our setting, because due to the high number of vertices per curve, every exact evaluation of the Fréchet distance breaks the advantage of filtering. The option to use heuristic filters in advance remains and is part of ongoing work.

Setup We ran our experiments on a high performance linux cluster, which has twenty GPU nodes with two Intel Xeon E5-2640v4 CPUs, 64 GB of RAM and two Nvidia K40 GPUs each. This makes 2880 CUDA cores per card. To minimize interference, each experiment was run on an exclusive core and both GPUs, with 30 GB of RAM guaranteed. Each experiment was run ten times for each parametrization (ε, k).

Q1 Concerning all data sets we can say that the quality of the Fréchet distance computation after applying the Johnson-Lindenstrauss embedding is very good. In Fig. 1 we depict the results of the Fréchet distance computations vs. the chosen values for ε. It can be observed that even for larger values of ε, the effective error does not exceed the ten percent error margin. There are few outliers, though. In most of the cases, up to ε = 0.5 the median of the results is very close to the exact Fréchet distance, which we consider successful.

Q2 In Fig. 2 we depict the running-times of the experiments that lead to the values depicted in Fig. 1. While the naïve implementation of the Alt and Godau algorithm took about roughly three hours, we
Figure 1: Fréchet distance values under the \((1 \pm \varepsilon)\)-Johnson-Lindenstrauss embedding. The green horizontal line represents the exact Fréchet distance and the red horizontal lines represent the ten percent error-margin. The lateral axis shows the values for \(\varepsilon\) plugged into the embedding.

Figure 2: Running-time of the algorithms, where sequential is a naïve implementation of the Alt and Godau algorithm, parallel is our CUDA-enabled variant and the suffix "\_rp" means that the data was randomly projected before.
were able to lower the running-time to about 40 seconds for $\varepsilon = 0.5$. For small values of $\varepsilon$ the target dimension did increase, which explains the higher running-time. Also, when running the experiments with PS3 and PS6, sometimes the upper and lower bounds that our implementation is using were matching. Thus, the algorithm would terminate without even constructing the free space diagram (cf. Alt and Godau (1995)) once, which explains the high variance.

Q3 The short answer is yes. In Fig. 3 the objective values of the DELTA and weather simulation clustering experiments are depicted. We see that the number of clusters in the DELTA data set can be identified successfully. Further in about 99.75% of the 400 experiments for the DELTA data set the same center-set was identified using the embedded curves and the original ones, which we consider robust. It was confirmed that the clusters are meaningful. For the weather simulation data set things seem different. The plot of the objective function shows that the data is likely to be very homogenous, which is confirmed by the nature of the data. Here, only in 27.25% of the 400 experiments the same center-set is identified using the embedded curves and the original ones. The deviating center-sets are of similar cost, though.

We also tried to conduct experiments using the local-search algorithm by Arya et al. (2004), which is also applicable for $k$-median clustering under the Fréchet distance. Unfortunately, these experiments exceeded the running-time limit of the cluster, which motivates the following section and lead to new experiments.

4 Fréchet 1-median

Here we study the median problem and aim to find an input curve that minimizes the sum of distances to all other given curves. Instead of exhaustively trying all curves and computing all pairwise distances that may occur, we aim to find a small candidate set of possible 1-median and another small witness set which serves as a proxy to sum over the distances instead of summing over all given curves. We will use the following theorem of Indyk (2000) which was developed for general discrete metric spaces and reduces the effort to evaluating the sum of distances only to the few witnesses and only for a few candidates, as described.

**Theorem 10.** (Indyk, 2000, Theorem 31) Let $\varepsilon \in (0,1]$ be a constant and $T$ be a set of polygonal curves. Further let $W$ be a non-empty uniform sample from $T$. For $\tau, \sigma \in T$ with $\sum_{\tau' \in T} d_F(\tau, \tau') > (1 + \varepsilon) \sum_{\tau' \in T} d_F(\sigma, \tau')$ it holds that $\Pr[\sum_{\tau' \in W} d_F(\tau, \tau') \leq \sum_{\tau' \in W} d_F(\sigma, \tau')] < \exp(-\varepsilon^2|W|/64)$.

Using Theorem 10 this yields $(2 + \varepsilon)$-approximation.

**Theorem 11.** Given parameters $\varepsilon, \delta \in (0,1/2)$ and a non-empty set $T$ of polygonal curves, we can use a uniform sample $S := \{s_1, \ldots, s_{\ell_S}\}$ of cardinality $\mathcal{O}\left(\ln(1/\delta)/\varepsilon\right)$ of candidates and a uniform sample $W := \{w_1, \ldots, w_{\ell_W}\}$ of cardinality $\mathcal{O}\left(\ln(\ell_S/\delta)/\varepsilon^2\right)$ of witnesses, to obtain a $(2 + \varepsilon)$-approximate 1-median $c_S \in S$ with probability at least $1 - \delta$. 
Figure 4: Running-times and objective values for the Fréchet 1-median. The lateral axis shows the values for \( \varepsilon \) plugged into the sampling algorithm. The red horizontal lines represent the ten percent error-margin and the green horizontal line represents the optimal objective value. \( \text{epsilon}_{rp} \) is the value for \( \varepsilon \) that is plugged into the embedding.

Under natural assumptions, setting our analysis in the Beyond-Worst-Case regime (Roughgarden, 2019), we can even get a \((1 + \varepsilon)\)-approximation on a subsample of sublinear size. The high-level idea behind is that the curves are usually not equidistant to the optimal median. Relative to the average cost, there will be some outliers, some curves at medium distance and also some curves very close to the optimal. Now if there are quite a good number of outliers, but also not too many, they make up a good share of the total cost. This implies that the number of curves at medium distance is bounded by a constant fraction of the curves. Finally this implies that the number of curves that are close to the optimal median, is not too small such that a small sample will include at least one of them.

**Theorem 12.** Let \( \varepsilon, \delta \in (0, 1/2) \), and \( T \) be a non-empty set of polygonal curves with at least \((1 - \varepsilon)\gamma(n)\) outliers, for \( 0 < \gamma(n) < 1/2 \). We can use a uniform sample \( S := \{s_1, \ldots, s_{\ell_S}\} \) of cardinality \( \ell_S = \mathcal{O}\left(\frac{\ln(1/\delta)}{1/2 - \gamma(n)}\right) \) of candidates and a uniform sample \( W := \{w_1, \ldots, w_{\ell_W}\} \) of cardinality \( \mathcal{O}\left(\ln(\ell_S/\delta)/\varepsilon^2\right) \) of witnesses, to obtain a \((1 + \varepsilon)\)-approximate \( 1 \)-median \( c_S \in S \) with probability at least \( 1 - \delta \).

Note in particular, that the samples in Theorem 12 still have constant size unless the fraction of outliers \( \gamma(n) \) tends arbitrarily close to \( 1/2 \), in which case the usual notion of an outlier is not met for two reasons: first, half of the curves would be considered outliers, and second their distance to the optimal median is not much larger than the medium curves implying that basically all curves are in a narrow annulus around the average distance. Both observations make the notion of outliers highly questionable. The details are in the proof and Fig. 5 in the supplement.

**1-median experiment** We conducted additional experiments on the weather simulation dataset. Fig. 4 shows that employing the subsampling schemes yields substantial improvements in terms of running-times while the approximation error remains robust to the choices of the approximation parameters \( \varepsilon_s \) and \( \varepsilon_{rp} \) plugged into the subsampling scheme and the embedding, respectively.

**5 Conclusion**

We provided a practical implementation and theoretical approaches to make clustering under the Fréchet distance viable. We developed a dimensionality reduction for high-dimensional polygonal curves, a parallelized variant of the Alt and Godau algorithm, and subsampling schemes for computing the Fréchet 1-median. We demonstrated the efficiency, accuracy and robustness of our approach.
References

Agarwal, P. K., Har-Peled, S., and Yu, H. (2013). Embeddings of surfaces, curves, and moving points in Euclidean space. *SIAM Journal on Computing*, 42(2):442–458.

Aghabozorgi, S., Shirkhoshidi, A. S., and Wah, T. Y. (2015). Time-series clustering – a decade review. *Information Systems*, 53:16 – 38.

Alt, H. and Godau, M. (1995). Computing the Fréchet Distance between two Polygonal Curves. *International Journal of Computational Geometry & Applications*, 05:75–91.

Arya, V., Garg, N., Khandekar, R., Meyerson, A., Munagala, K., and Pandit, V. (2004). Local Search Heuristics for k-Median and Facility Location Problems. *SIAM Journal on Computing*, 33(3):544–562.

Bachem, O., Lucic, M., and Krause, A. (2018). Scalable k-means clustering via lightweight coresets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, (KDD)*, pages 1119–1127.

Baldus, J. and Bringmann, K. (2018). A fast implementation of near neighbors queries for Fréchet distance (GIS Cup). *CoRR*, abs/1803.00806.

Billsus, D. and Pazzani, M. J. (1998). Learning collaborative information filters. In *Proceedings of the Fifteenth International Conference on Machine Learning*, (ICML), pages 46–54.

Braverman, V., Charikar, M., Kusmaul, W., Woodruff, D. P., and Yang, L. F. (2019). The one-way communication complexity of dynamic time warping distance. *CoRR*, abs/1903.03520. to appear in proceedings of the 35th International Symposium on Computational Geometry (SoCG).

Bringmann, K., Künnemann, M., and Nusser, A. (2019). Walking the dog fast in practice: Algorithm engineering of the Fréchet distance. *CoRR*, abs/1901.01504. to appear in proceedings of the 35th International Symposium on Computational Geometry (SoCG).

Buchin, K., Diez, Y., van Diggelen, T., and Meulemans, W. (2017). Efficient trajectory queries under the Fréchet distance (GIS Cup). In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL, pages 101:1–101:4.

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

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

Chapados, N. and Bengio, Y. (2008). Augmented functional time series representation and forecasting with gaussian processes. In *Advances in Neural Information Processing Systems 20*, pages 265–272.

Daniely, A., Linial, N., and Saks, M. E. (2012). Clustering is difficult only when it does not matter. *CoRR*, abs/1205.4891.

Driemel, A. and Krivosija, A. (2018). Probabilistic embeddings of the Fréchet distance. In *Proceedings of the 16th International Workshop on Approximation and Online Algorithms (WAOA)*, pages 218–237.

Driemel, A., Krivosija, A., and Sohler, C. (2016). Clustering time series under the Fréchet distance. In *Proceedings of the 27th ACM-SIAM Symposium on Discrete Algorithms*, (SODA), pages 766–785.

Dütsch, F. and Vahrenhold, J. (2017). A filter-and-refinement-algorithm for range queries based on the Fréchet distance (GIS Cup). In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL, pages 100:1–100:4.

Feldman, D., Momemizadeh, M., Sohler, C., and Woodruff, D. P. (2010). Coresets and sketches for high dimensional subspace approximation problems. In *Proceedings of the 21st Annual ACM-SIAM symposium on Discrete Algorithms*, (SODA), pages 630–649.

Feldman, D., Schmidt, M., and Sohler, C. (2013). Turning Big Data into tiny data: Constant-size coresets for k-means, PCA and projective clustering. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms*, (SODA), pages 1434–1453.

Hamilton, J. (1994). *Time Series Analysis*. Princeton University Press.
Helwig, N., Pignanelli, E., and Schütze, A. (2015). Condition monitoring of a complex hydraulic system using multivariate statistics. In 2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings, pages 210–215.

Hong, Z.-Q. (1991). Algebraic feature extraction of image for recognition. Pattern Recognition, 24(3):211 – 219.

Indyk, P. (2000). High-dimensional Computational Geometry. PhD thesis, Stanford University, CA, USA.

Johnson, W. B. and Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. Contemporary Mathematics, 26(1):189–206.

Liao, T. W. (2005). Clustering of time series data—a survey. Pattern Recognition, 38(11):1857 – 1874.

Lucas, D. D., Yver Kwok, C., Cameron-Smith, P., Graven, H., Bergmann, D., Guilderson, T. P., Weiss, R., and Keeling, R. (2015). Designing optimal greenhouse gas observing networks that consider performance and cost. Geoscientific Instrumentation, Methods and Data Systems, 4(1):121–137.

Munteanu, A., Schwiegelshohn, C., Sohler, C., and Woodruff, D. P. (2018). On coresets for logistic regression. In Advances in Neural Information Processing Systems 31, (NeurIPS), pages 6562–6571.

Roughgarden, T. (2019). Beyond worst-case analysis. Commun. ACM, 62(3):88–96.

Sheehy, D. R. (2014). The persistent homology of distance functions under random projection. In 30th Annual Symposium on Computational Geometry, (SoCG), page 328.

Sohler, C. and Woodruff, D. P. (2018). Strong coresets for k-median and subspace approximation: Goodbye dimension. In 59th IEEE Annual Symposium on Foundations of Computer Science, (FOCS), pages 802–813.

TU Dortmund University (2019). DELTA (Dortmund electron accelerator).

Zhang, J., Roy, D., Devadiga, S., and Zheng, M. (2007). Anomaly detection in modis land products via time series analysis. Geo-spatial Information Science, 10(1):44–50.

Zimmer, C., Meister, M., and Nguyen-Tuong, D. (2018). Safe active learning for time-series modeling with gaussian processes. In Advances in Neural Information Processing Systems 31, pages 2730–2739.
A Omitted material

Proof of Proposition 7. We have:

\[ \|p - q\|^2 = \sum_{i=1}^{d} [(p_{1,i} - \lambda_p(p_{1,i} - p_{2,i})) - (q_{1,i} - \lambda_q(q_{1,i} - q_{2,i}))]^2 \quad (I) \]

\[ = \sum_{i=1}^{d} [(p_{1,i} - \lambda_p(p_{1,i} - p_{2,i}))^2 - 2(p_{1,i} - \lambda_p(p_{1,i} - p_{2,i}))(q_{1,i} - \lambda_q(q_{1,i} - q_{2,i})) + (q_{1,i} - \lambda_q(q_{1,i} - q_{2,i}))^2] \quad (II) \]

\[ = \sum_{i=1}^{d} [p_{1,i}^2 - 2\lambda_p p_{1,i}(p_{1,i} - p_{2,i}) + \lambda_p^2(p_{1,i} - p_{2,i})^2 \]

\[- 2(p_{1,i} - \lambda_p(p_{1,i} - p_{2,i}))(q_{1,i} - \lambda_q(q_{1,i} - q_{2,i})) + q_{1,i}^2 - 2\lambda_q q_{1,i}(q_{1,i} - q_{2,i}) + \lambda_q^2(q_{1,i} - q_{2,i})^2] \quad (III) \]

\[ = \sum_{i=1}^{d} [(1 - 2\lambda_p + \lambda_p^2)p_{1,i}^2 + \lambda_p^2p_{2,i}^2 + (1 - 2\lambda_q + \lambda_q^2)q_{1,i}^2 + \lambda_q^2q_{2,i}^2 \]

\[ + 2\lambda_p(1 - \lambda_p)p_{1,i}p_{2,i} + 2\lambda_qp_{1,i}q_{2,i} + 2\lambda_pq_{1,i}q_{2,i} + 2\lambda_q(1 - \lambda_q)q_{1,i}q_{2,i}] \quad (V) \]

\[ = (1 - \lambda_p)^2\|p_1\|^2 + \lambda_p^2\|p_2\|^2 + (1 - \lambda_q)^2\|q_1\|^2 + \lambda_q^2\|q_2\|^2 \]

\[ + 2\lambda_p(1 - \lambda_p)(p_1, q_2) + 2(\lambda_p + \lambda_q - \lambda_p\lambda_q - 1)(q_1, q_2) \]

\[ + 2(\lambda_p\lambda_q - \lambda_q)(p_2, q_1) - 2\lambda_p(1 - \lambda_q)(q_1, q_2) \quad (VI) \]

\[ = (1 - \lambda_p)^2\|p_1\|^2 + \lambda_p^2\|p_2\|^2 + (1 - \lambda_q)^2\|q_1\|^2 + \lambda_q^2\|q_2\|^2 \]

\[ + 2\lambda_p(1 - \lambda_p)(p_1, q_2) + 2(\lambda_p + \lambda_q - \lambda_p\lambda_q - 1)(q_1, q_2) \]

\[ + 2(\lambda_p\lambda_q - \lambda_q)(p_2, q_1) - 2\lambda_p(1 - \lambda_q)(q_1, q_2) \quad (VII) \]

\[ = (1 - \lambda_p)^2\|p_1\|^2 + \lambda_p^2\|p_2\|^2 + (1 - \lambda_q)^2\|q_1\|^2 + \lambda_q^2\|q_2\|^2 \]

\[ + (\lambda_p - \lambda_p^2)(\|p_1\|^2 + \|q_2\|^2 - \|p_1 - p_2\|^2) \]

\[ + (\lambda_q - \lambda_q^2)(\|q_1\|^2 + \|q_2\|^2 - \|q_1 - q_2\|^2) \]

\[ - (1 - \lambda_p - \lambda_q + \lambda_p\lambda_q)(\|p_1\|^2 + \|q_1\|^2 - \|p_1 - q_1\|^2) \]

\[ - (\lambda_q - \lambda_p\lambda_q)(\|p_1\|^2 + \|q_2\|^2 - \|p_1 - q_2\|^2) \]

\[ - (\lambda_p - \lambda_p\lambda_q)(\|q_1\|^2 + \|q_2\|^2 - \|q_1 - q_2\|^2) \]

\[ + \lambda_p\lambda_q(\|p_2\|^2 + \|q_2\|^2 - \|p_2 - q_2\|^2) \quad (VIII) \]

\[ = (1 - 1 - 2\lambda_p + \lambda_p + \lambda_p + \lambda_p^2 - \lambda_p^2 + \lambda_q - \lambda_q + \lambda_p\lambda_q - \lambda_p\lambda_q)\|p_1\|^2 \]

\[ + (\lambda_p^2 - \lambda_p^2 + \lambda_p - \lambda_p + \lambda_p\lambda_q - \lambda_p\lambda_q)\|p_2\|^2 \]

12
We obtain Eq. (I) to Eq. (V) using only algebraic manipulations, Eq. (VI) is obtained using the definition of the Euclidean norm and the algebraic definition of the dot product, in Eq. (VII) we use the geometric definition of the dot product and finally in Eq. (VIII) we apply the law of cosines. Eq. (IX) follows by algebraic manipulations.

**Proof of Lemma 8.** First note that the construction of \( f \) succeeds with probability \( \rho \in (0, 1] \) by Definition 5. We condition the remaining proof on this event.

From Proposition 7 we now know that

\[
\|p - q\|^2 = - (\lambda_p - \lambda_p^2)\|p_1 - p_2\|^2 - (\lambda_q - \lambda_q^2)\|q_1 - q_2\|^2 + (1 - \lambda_p - \lambda_q + \lambda_p \lambda_q)\|p_1 - q_1\|^2 + (\lambda_q - \lambda_p \lambda_q)\|p_1 - q_2\|^2 + (\lambda_p - \lambda_p \lambda_q)\|p_2 - q_1\|^2 + (1 - \lambda_p - \lambda_q + \lambda_p \lambda_q)\|p_2 - q_2\|^2
\]

and

\[
\|p' - q'\|^2 = - (\lambda_p - \lambda_p^2)\|f(p_1) - f(p_2)\|^2 - (\lambda_q - \lambda_q^2)\|f(q_1) - f(q_2)\|^2 + (1 - \lambda_p - \lambda_q + \lambda_p \lambda_q)\|f(p_1) - f(q_1)\|^2 + (\lambda_q - \lambda_p \lambda_q)\|f(p_1) - f(q_2)\|^2 + (\lambda_p - \lambda_p \lambda_q)\|f(p_2) - f(q_1)\|^2 + (1 - \lambda_p - \lambda_q + \lambda_p \lambda_q)\|f(p_2) - f(q_2)\|^2
\]

Because every coefficient is non-negative, it can be observed that this sum is maximized under \( f \) when

\[
\|f(p_1) - f(p_2)\|^2 = (1 - \varepsilon)^2\|p_1 - p_2\|^2,
\]

\[
\|f(q_1) - f(q_2)\|^2 = (1 - \varepsilon)^2\|q_1 - q_2\|^2,
\]

\[
\|f(p_1) - f(q_1)\|^2 = (1 + \varepsilon)^2\|p_1 - q_1\|^2,
\]

\[
\|f(p_1) - f(q_2)\|^2 = (1 + \varepsilon)^2\|p_1 - q_2\|^2,
\]

\[
\|f(p_2) - f(q_1)\|^2 = (1 + \varepsilon)^2\|p_2 - q_1\|^2
\]

and

\[
\|f(p_2) - f(q_2)\|^2 = (1 + \varepsilon)^2\|p_2 - q_2\|^2.
\]

Using the facts that \((1 + \varepsilon)^2 - (1 - \varepsilon)^2 = 4\varepsilon\), \((\lambda_q - \lambda_q^2) \leq \frac{1}{4}\) and \((\lambda_p - \lambda_p^2) \leq \frac{1}{4}\), we get that \(\|p' - q'\|^2 \leq (1 + \varepsilon)^2\|p - q\|^2 + \varepsilon(\|p_1 - p_2\|^2 + \|q_1 - q_2\|^2)\). The lower bound follows analogously.

**Proof of Theorem 9.** First note that the construction of \( f \) and thus also \( F \) succeeds with probability \( \rho \in (0, 1] \) by Definition 5. We condition the remaining proof on this event.

Let \( \tau, \sigma \in T \) be arbitrary polygonal curves and \( v^\tau_1, \ldots, v^\tau_{|\tau|} \), respective \( v^\sigma_1, \ldots, v^\sigma_{|\sigma|} \) be their vertices, as well as \( t^\tau_1, \ldots, t^\tau_{|\tau|} \), respective \( t^\sigma_1, \ldots, t^\sigma_{|\sigma|} \) be their instants. Further let

\[
g \in \arg \inf_{h \in H} \max_{t \in [0, 1]} \|\tau(t) - \sigma(h(t))\|
\]

and

\[
g' \in \arg \inf_{h \in H} \max_{t \in [0, 1]} \|F(\tau)(t) - F(\sigma)(h(t))\|
\]

Let \( t_1 \in \arg \max_{t \in [0, 1]} \|F(\tau)(t) - F(\sigma)(g(t))\| \), there exists an \( i \in \{1, \ldots, |\tau|\} \) and \( j \in \{1, \ldots, |\sigma|\} \) with \( t_i^\tau \leq t_1 \leq t_{i+1}^\tau \) and \( t_j^\sigma \leq g(t_1) \leq t_{j+1}^\sigma \), such that we can write

\[
F(\tau)(t_1) = L p \left( f(v^\tau_{i+1}) - f(v^\tau_i), \frac{t_{i+1}^\tau - t_i^\tau}{t_{i+1}^\tau - t_i^\tau} \right)
\]

and

\[
F(\sigma)(g(t_1)) = L q \left( f(v^\sigma_j) - f(v^\sigma_{j+1}), \frac{t_{j+1}^\sigma - t_j^\sigma}{t_{j+1}^\sigma - t_j^\sigma} \right)
\]

where \( L p \) and \( L q \) are linear mappings.
\[ F(\sigma)(g(t_1)) = \text{lp} \left( f(v_j^\sigma) f(v_j^{*+1}), \frac{g(t_1) - t_j^\sigma}{t_j^{*+1} - t_j^\sigma} \right), \]
\[ \tau(t_1) = \text{lp} \left( v_j^t v_j^{*+1}, \frac{t_1 - t_j^*}{t_j^{*+1} - t_j^*} \right), \]
and
\[ \sigma(g(t_1)) = \text{lp} \left( \frac{g(t_1) - t_j^*}{t_j^{*+1} - t_j^*}, v_j^t v_j^{*+1} \right). \]

For each \( t_1' \in \arg\max_{t \in [0,1]} \| F(\tau)(t) - F(\sigma)(g(t)) \| \) we obtain:
\[ d_F^2(F(\tau), F(\sigma)) = \| F(\tau)(t_1') - F(\sigma)(g(t')) \|^2 \] (I)
\[ \leq \| F(\tau)(t_1) - F(\sigma)(g(t_1)) \|^2 \] (II)
\[ \leq (1 + \epsilon)^2 \| \tau(t_1) - \sigma(g(t_1)) \|^2 + \epsilon (\| v_j^t - v_i^{t+1} \|^2 + \| v_j^* - v_i^{*+1} \|^2) \] (III)
\[ \leq (1 + \epsilon)^2 \max_{t \in [0,1]} \| \tau(t) - \sigma(g(t)) \|^2 + \epsilon (\| v_j^t - v_i^{t+1} \|^2 + \| v_j^* - v_i^{*+1} \|^2) \]
\[ \leq (1 + \epsilon)^2 d_F^2(\tau, \sigma) + 2\epsilon \gamma^2 \]

Eq. (I) follows by definition of \( t_1' \) and \( g' \), Eq. (II) follows from the fact that \( g' \) is an infimum, Eq. (III) follows from an application of Lemma 8 and the last inequality follows from Definition 3 and the definition of \( \gamma \).

Let \( t_2 \in \arg\max_{t \in [0,1]} \| \tau(t) - \sigma(g(t)) \| \), again, there exists an \( i \in \{1, \ldots, |\tau|\} \) and \( j \in \{1, \ldots, |\sigma|\} \) with \( t_i^2 \leq t_2 \leq t_i^{*+1} \) and \( t_j^* \leq g'(t_2) \leq t_j^{*+1} \), such that we can write
\[ F(\tau)(t_2) = \text{lp} \left( f(v_j^t f(v_j^{*+1}), \frac{t_2 - t_j^*}{t_j^{*+1} - t_j^*} \right), \]
\[ F(\sigma)(g'(t_2)) = \text{lp} \left( f(v_j^* f(v_j^{*+1}), \frac{g'(t_2) - t_j^*}{t_j^{*+1} - t_j^*} \right), \]
\[ \tau(t_2) = \text{lp} \left( v_j^t v_j^{*+1}, \frac{t_2 - t_j^*}{t_j^{*+1} - t_j^*} \right), \] and
\[ \sigma(g'(t_2)) = \text{lp} \left( \frac{g'(t_2) - t_j^*}{t_j^{*+1} - t_j^*}, v_j^t v_j^{*+1} \right). \]

For each \( t_1' \in \arg\max_{t \in [0,1]} \| F(\tau)(t) - F(\sigma)(g'(t)) \| \) we obtain:
\[ d_F^2(F(\tau), F(\sigma)) = \| F(\tau)(t_1') - F(\sigma)(g(t')) \|^2 \] (IV)
\[ \geq \| F(\tau)(t_2) - F(\sigma)(g'(t_2)) \|^2 \] (V)
\[ \geq (1 - \epsilon)^2 \| \tau(t_2) - \sigma(g'(t_2)) \|^2 - \epsilon (\| v_j^t - v_i^{t+1} \|^2 + \| v_j^* - v_i^{*+1} \|^2) \] (VI)
\[ \geq (1 - \epsilon)^2 \max_{t \in [0,1]} \| \tau(t) - \sigma(g(t)) \|^2 - \epsilon (\| v_j^t - v_i^{t+1} \|^2 + \| v_j^* - v_i^{*+1} \|^2) \]
\[ \geq (1 - \epsilon)^2 d_F^2(\tau, \sigma) - 2\epsilon \gamma^2 \]

Here Eq. (IV) follows by the definition of \( g' \) and \( t_1' \), Eq. (V) follows, because the term is maximized for \( t_1' \), Eq. (VI) follows from an application of Lemma 8 and the last inequality follows from Definition 3 and the definition of \( \gamma \).

**Proof of Theorem 11.** Let \( c \in \arg\min_{\tau \in T} \sum_{\tau' \in T} d_F(\tau, \tau') \) be an optimal 1-median for \( T \) and let \( X(\tau) := d_F(\tau, c) \) be a uniformly distributed random variable. By the uniform distribution and linearity \( E[X] = \frac{1}{|T|} \sum_{\tau \in T} d_F(\tau, c) \). Now, let
\[ B_{1+\epsilon} := \left\{ \tau \in T \mid d_F(\tau, c) \leq \frac{(1 + \epsilon)}{|T|} \sum_{\tau' \in T} d_F(\tau', c) \right\} \].

14
We define a parameter 

$$0$$

way to bound the number of curves with medium contribution. We have $\epsilon T$.

Rearranging yields the desired bound

$$\left| \text{cost}(\tau) \right| \leq (2 + \epsilon) \left| \text{cost}(\tau) \right|,$$

for this event is bounded by $|S| \exp \left( \frac{-\epsilon^2 \ell_W}{64} \right) \leq |S| \exp \left( \frac{-\epsilon^2 \ell_W}{64} \right) \leq |S| \exp \left( -\ln (2|S|) \right) \leq \frac{\delta}{2}\right.$.

Let $c_S \in \arg \min_{\tau \in S} \sum_{\tau' \in T} d_F(\tau, \tau')$. We do not want any bad sample $s \in S$ with $\sum_{\tau \in T} d_F(s, \tau) > (1 + \epsilon) \sum_{\tau \in T} d_F(c_S, \tau)$ to have lower cost with respect to $W$ than $c_S$. Using Theorem 10 and a union bound over the elements of $S$ and $\ell_W = \frac{\Delta}{\epsilon}$, the probability for this event is bounded by

$$\sum_{s \in S} \exp \left( \frac{-\epsilon^2 \ell_W}{64} \right) \leq |S| \exp \left( \frac{-\epsilon^2 \ell_W}{64} \right) \leq |S| \exp \left( -\ln (2|S|) \right) \leq \frac{\delta}{2}\right.$.

Now, if we take the $s \in S$ that minimizes $\sum_{\tau' \in W} d_F(s, \tau')$, by an additional application of the union bound, with probability at least $1 - \delta$ it holds that

$$\sum_{\tau' \in T} d_F(s, \tau') \leq (1 + \epsilon) \sum_{\tau' \in T} d_F(c_S, \tau') \leq (1 + \epsilon)(2 + \epsilon) \sum_{\tau' \in T} d_F(c, \tau') \leq (2 + 4\epsilon) \sum_{\tau' \in T} d_F(c, \tau').$$

The claim follows by rescaling $\epsilon$ by $\frac{1}{4}$.

**Proof of Theorem 12.** Let $c^* \in \arg \min_{\tau \in T} \sum_{\tau' \in T} d_F(\tau, \tau')$ be an optimal Fréchet 1-median for $T$. For any non-empty set $A$ of curves and a curve $c$ let $\text{cost}(A, c) = \sum_{\tau \in A} d_F(\tau, c)$ denote the cost, i.e., the sum of Fréchet distances to $c$. Let $\Delta = \text{cost}(T, c^*)$ denote the optimal cost. We define a parameter $0 < \gamma(n) = \frac{1}{2} - \nu(n) < \frac{1}{2}$ which specifies the fraction of outliers as a function of $n$. We choose the radius $r_1 = \frac{\Delta}{\gamma(n)n}$ which parametrizes the distance of the outliers from the optimal median. Similarly, let $r_2 = 2e \frac{\Delta}{n}$. Note that indeed $r_1 > r_2$ as desired, since $\gamma(n), \epsilon < \frac{1}{2}$. We partition the curves in $T$ according to their contribution relative to the average distance into disjoint sets $T = F \cup M \cup C$ where $F = \{ \tau \in T \ | \ d_F(\tau, c^*) > r_1 \}$ are the curves far from $c^*$, $M = \{ \tau \in T \ | \ r_2 < d_F(\tau, c^*) \leq r_1 \}$ are the curves with medium distance, and $C = \{ \tau \in T \ | \ d_F(\tau, c^*) \leq r_2 \}$ are the curves that are close to the optimal median.

Note that if $|F| > n \cdot \gamma(n)$ then $\text{cost}(F, c^*) \geq |F| \cdot r_1 > n\gamma(n) \cdot \frac{1}{\gamma(n)} \frac{\Delta}{n} = \Delta$. Together with our assumption this means that we have $(1 - \epsilon)n\gamma(n) \leq |F| \leq n\gamma(n)$.

Similarly, $\text{cost}(F, c^*) \geq (1 - \epsilon)n\gamma(n)|F| \frac{1}{\gamma(n)} \frac{\Delta}{n} = (1 - \epsilon)\Delta$, which means that the outliers make up a constant fraction of the optimal cost.

Now this implies that $\text{cost}(T \setminus F, c^*) \leq \Delta - (1 - \epsilon)\Delta = \epsilon \Delta$, which we can leverage in the following way to bound the number of curves with medium contribution. We have

$$\epsilon \Delta \geq \text{cost}(T \setminus F, c^*) = \text{cost}(M \cup C, c^*) \geq \text{cost}(M, c^*) \geq |M| \cdot r_2 = |M| \cdot 2\epsilon \frac{\Delta}{n}.$$ 

Rearranging yields the desired bound $|M| \leq \frac{\epsilon \Delta}{2} \cdot n = n$.

Let $A$ be the event that a uniform sample from $T$ is contained in $C$. By the disjoint union, the probability for this event can be bounded by

$$\Pr[A] = \frac{|C|}{n} = \frac{|T| - |M| - |F|}{n} \geq \frac{n - \frac{n}{2} - n\gamma(n)}{n} = 1 - \gamma(n) := \nu(n)$$
As previously we sample a logarithmic number of witnesses $\ell_W = \frac{64}{\varepsilon^2} \ln \left( \frac{2 \ln n}{\delta} \right)$ such that by Proposition 10 and an application of the union bound the probability that any center that is worse than $c$ by a factor of more than $(1 + \varepsilon)$ has lower cost than $c$ with respect to $W$ is bounded by

$$\sum_{s \in S} \exp \left( - \frac{\varepsilon^2 \ell_W}{64} \right) \leq |S| \cdot \frac{\delta}{2\ell_S} = \frac{\delta}{2}.$$ 

Thus with probability at least $1 - \delta$ we have that both, our sample $S$ contains a $(1 + 2\varepsilon)$-approximate solution $\tilde{c}$ and any point $c' \in S$ that evaluates equal or better than $c$ on the sample $W$ is within $(1 + \varepsilon)$ of the cost of $\tilde{c}$. Thus $\text{cost}(T, c') \leq (1 + 2\varepsilon)(1 + \varepsilon) \text{cost}(T, c^*) \leq (1 + 4\varepsilon) \text{cost}(T, c^*)$.

We conclude the proof by rescaling $\varepsilon$. 

---

**Figure 5:** Distributions of curves (for simplicity represented as points) around their median. The red circles represent the radii $r_1, r_2$ defining the sets of far, medium and close curves (cf. proof of Theorem 12, best viewed in color). The left plot shows a “typical” distribution where the median yields a good representative of the data that is robust against outliers. There is a reasonable but not too large number of outliers, that are far away from the center and many curves are close to the optimal median. Such distributions typically arise in physical domains. In such a situation, the sampling algorithm of Theorem 12 yields a $(1 + \varepsilon)$-approximation. In the right plot we see a distribution which is much more uniform. Most points are in an annulus about the average distance, there are no far away outliers, and few curves close to the optimal. To find one of the latter, the $(1 + \varepsilon)$-approximation needs too many samples. Note however, that the same algorithm yields a $(2 + \varepsilon)$-approximation via Theorem 11 that works in general for all inputs.