Random Projections for Curves in High Dimensions

Ioannis Psarros
Athena Research Center, Marousi, Greece

Dennis Rohde
Universität Bonn, Germany

Abstract

Modern time series analysis requires the ability to handle datasets that are inherently high-dimensional; examples include applications in climatology, where measurements from numerous sensors must be taken into account, or inventory tracking of large shops, where the dimension is defined by the number of tracked items. The standard way to mitigate computational issues arising from the high dimensionality of the data is by applying some dimension reduction technique that preserves the structural properties of the ambient space. The dissimilarity between two time series is often measured by "discrete" notions of distance, e.g. the dynamic time warping or the discrete Fréchet distance. Since all these distance functions are computed directly on the points of a time series, they are sensitive to different sampling rates or gaps. The continuous Fréchet distance offers a popular alternative which aims to alleviate this by taking into account all points on the polygonal curve obtained by linearly interpolating between any two consecutive points in a sequence.

We study the ability of random projections à la Johnson and Lindenstrauss to preserve the continuous Fréchet distance of polygonal curves by effectively reducing the dimension. In particular, we show that one can reduce the dimension to $O(\varepsilon^{-2} \log N)$, where N is the total number of input points while preserving the continuous Fréchet distance between any two determined polygonal curves within a factor of $1 \pm \varepsilon$. We conclude with applications on clustering.

2012 ACM Subject Classification Theory of computation → Computational geometry; Theory of computation → Random projections and metric embeddings

Keywords and phrases polygonal curves, time series, dimension reduction, Johnson-Lindenstrauss lemma, Fréchet distance

Digital Object Identifier 10.4230/LIPIcs.SoCG.2023.53

Related Version *Full Version*: https://arxiv.org/abs/2207.07442

Funding *Ioannis Psarros*: The author was partially supported by the EU's Horizon 2020 Research and Innovation programme, under the grant agreement No. 957345: "MORE".
Dennis Rohde: The author was partially supported by the Hausdorff Center for Mathematics.

1 Introduction

Time series analysis lies in the core of various modern applications. Typically, a time series consists of various (physical) measurements over time. Formally, it is a finite sequence of points in $\mathbb{R}^d$. Depending on the use case, the ambient space may be extremely high-dimensional, for example $d \in 2^{\Omega(\log n)}$, or even $d \in 2^{\Omega(n)}$, where n is the number of given sequences. For example, large facilities nowadays supervise their production lines using a plethora of sensors. Another concrete example are climatology applications, where data consist of measurements from multiple sensors, each one corresponding to a different dimension.

Many analysis techniques are based on (dis-)similarity between time series, c.f. [27]. This is often measured by distance functions such as the Euclidean distance, which however requires the time series to be of same length and does not include any form of alignment between the sequences. This is of course less expressive than distances which are indeed defined over an optimal alignment, e.g. the dynamic time warping, or the discrete Fréchet distance, which are based on Euclidean distances between the points but enable compensation of differences in phase. A common downside of these distances is that they take into account solely the points of a time series. Hence, they are sensitive to differences in sampling rates or data gaps. Here, the continuous Fréchet distance offers a popular alternative which aims to alleviate this issue by assuming that time series are discretizations of continuous functions of time. It is an extension of the discrete Fréchet distance that takes into account all points on the polygonal curves obtained by linearly interpolating between any two consecutive points in a sequence (where the interpolation is carried out only implicitly).

Two main parameters typically govern computational tasks associated with the Fréchet distance: the lengths of the time series and the number of dimensions of the ambient space. In this paper, we study the problem of compressing the input with respect to the latter parameter using a dimension reducing linear transform that preserves Euclidean distances within a factor of $(1 \pm \varepsilon)$. These transforms, which are usually named Johnson-Lindenstrauss (JL) transforms or embeddings, are a popular tool in dimensionality reduction. The preservation of pairwise distances within a factor of $(1 \pm \varepsilon)$ is sometimes called JL guarantee. Recent work has provided various probability distributions over JL transforms [14, 21, 1, 28, 24], which are efficient to sample from and which yield the JL guarantee with at least constant positive probability while the target dimension is only $O(\varepsilon^{-2} \log n)$, where n is the size of the input point set. Towards applying this result on time series, one can easily guarantee that all Euclidean distances between points of the time series are preserved. While this has direct implications on "discrete" notions of distances between time series, the case of the continuous Fréchet distance is far more intriguing.

1.1 Related Work

In their seminal paper [23], Johnson and Lindenstrauss proved the following statement, which is commonly known as the Johnson-Lindenstrauss lemma and coined the term JL embedding.

▶ **Theorem 1** ([23]). *For any $n \in \mathbb{N}$ and $\varepsilon \in (0,1)$ there exists a probability distribution over linear maps $f \colon \mathbb{R}^d \to \mathbb{R}^{d'}$, where $d' \in O(\varepsilon^{-2} \log n)$, such that for any n-point set $X \subset \mathbb{R}^d$ the following holds with high probability over the choice of f:*

$$\forall p, q \in X: \ (1-\varepsilon)\|p-q\| \leq \|f(p)-f(q)\| \leq (1+\varepsilon)\|p-q\|.$$

In their proof, Johnson and Lindenstrauss [23] show that this can be achieved by orthogonally projecting the points onto a random linear subspace of dimension $O(\varepsilon^{-2} \log n)$ – and indeed there are point sets that require $\Omega(\varepsilon^{-2} \log n)$ dimensions [5, 25, 26]. Several proofs of their statement followed, these however don't require a proper projection but only a multiplying the points with a certain random matrix, cf. [22, 14, 1, 24, 4].

The impact of a JL embedding on higher-dimensional objects other than points has already been studied. Magen [29, 30] shows that applying a (scaled) JL embedding not only to a given set $P \subset \mathbb{R}^d$ of points, but to $P \cup W$, where $W \subset \mathbb{R}^d$ is a well-chosen set of points determined by P, approximately preserves the height and angles of all triangles determined by any three points in P. Magen even extends this result and shows that by a clever choice of W, the volume (Lebesgue measure) of the convex hull of any $k-1$ points from P is approximately preserved when the target dimension is in $\Theta(\varepsilon^{-2} k \log|P|)$. Furthermore, in this case the distance of any point from P to the affine hull of any $k-1$ other points from P is also approximately preserved. Furthermore, JL embeddings can even be utilized to preserve all pairwise Euclidean and geodesic distances on a smooth manifold [8].

Fréchet distance preserving embeddings are a relatively unexplored topic. Recently Driemel and Krivosija [17] studied the first Fréchet distance preserving embedding for c-packed curves, which are curves whose intersections with any ball of radius r are of length at most cr. This class of curves was introduced by Driemel et al. [16] and has so far been considered a viable assumption for realistic curves, see e.g. [3, 9, 15]. Driemel and Krivosija consider projections on random lines, where curves are orthogonally projected on a vector which is sampled uniformly at random from the unit sphere. They observed that in any case (even if the curves are not c-packed), the discrete Fréchet distance between the curves decreases. Furthermore, they show that with high probability the discrete Fréchet distance between two curves σ and τ, of complexity (number of vertices of the curve) at most m, decreases by a factor in $O(m)$. Finally, they proved that there exist c-packed curves such that the discrete Fréchet distance decreases by a factor in $\Omega(m)$. The latter also holds for the continuous Fréchet distance and for the dynamic time warping distance.

More recently, Meintrup et al. [32] studied JL embeddings in the context of preserving the Fréchet distance to facilitate k-median clustering of curves in a high-dimensional ambient space. They show that when the dimension is reduced to $\Theta(\varepsilon^{-2} \log N)$, where N is the total number of vertices of the given curves, the Fréchet distances are preserved up to a combined multiplicative error of $(1 \pm \varepsilon)$ and additive error of $\pm\varepsilon L$, where L is the largest arclength of any input curve. For their proof, they only use the JL guarantee, i.e., the $(1 \pm \varepsilon)$-preservation of Euclidean distances, and properties of the polygonal curves and the Fréchet distance, while linearity is not taken into account. In this setting, it seems that the additive error is possible – Meintrup et al. give a simple example where some vertex-to-vertex distances expand and others contract, which induces an additive error to the Fréchet distance. Meintrup et al. complement their results with experimental evaluation showing that in real world data and using a JL transform (which is a *linear* map), the Fréchet distance is preserved within the multiplicative error only in almost any case. The other cases can not be distinguished between a failed attempt to obtain a JL embedding (recall the probabilistic nature) and a successful attempt to obtain a JL embedding with the additive error occuring.

1.2 Our Contributions

We study the ability of random projections à la Johnson and Lindenstrauss to preserve the continuous Fréchet distances among a given set of n polygonal curves, each of complexity (number of vertices of the curve) at most m. We show that there exists a set X of vectors (in $\mathbb{R}^d$), of size polynomial in n and m and depending only on the given curves, such that any JL transform for the curves vertices and X also preserves the continuous Fréchet distance between any two of the given polygonal curves within a factor of $(1 \pm \varepsilon)$, without additional additive error. This effectively extends the JL guarantee to pairwise Fréchet distances. By plugging in any known JL transform from one of [14, 21, 1, 28, 24] we obtain our main dimension reduction result which states that one can reduce the number of dimensions to $O(\varepsilon^{-2} \log(nm))$. We achieve our result using a completely different approach than Meintrup et al. [32]. Our approach relies on Fréchet distance predicates originating from [2]. These allow a reduction from deciding the continuous distance to a finite set of events occurring. Using only the predicates, it is relatively easy to prove that the Fréchet distance between two curves does not expand by more than a factor of $(1 + \varepsilon)$ under a (*linear*) JL transform. To prove that the Fréchet distance does not contract by more than a factor of $(1 - \varepsilon)$ is however much more challenging. We achieve this by proving that *all* distances between one fixed point and any point on a fixed line do not contract by more than a factor of $(1 - \varepsilon)$ when a JL transform is applied to a well-chosen set of four vectors determined by the point and the line, which is then applied to any vertex of any curve and any line determined by an edge of any curve. We note that this result is comparable to a result by Magen [29, 30], but our statement is stronger since it takes into account *all* distances between the fixed point and the line and not only the affine distance, i.e., the distance between the point and its orthogonal projection onto the line.

Our motivation is that distance preserving dimensionality reductions imply improved algorithms for various tasks. Best-known algorithms for many proximity problems under the continuous Fréchet distance have exponential dependency on the dimension, in at least one of their performance parameters. Such algorithms either directly employ the continuous Fréchet distance, e.g. the approximation algorithms for k-clustering problems [13], or approximate it with the discrete Fréchet distance by resampling the time series to a higher granularity. For example, to the best of our knowledge, the best solution for the approximate near neighbor (ANN) problem in general dimensions derives from building the data structure of Filtser et al. [20], which originally solves the problem for the discrete Fréchet distance, on a modified input. The idea is that a new dense set of vertices can be added to each input polygonal curve so that the discrete Fréchet distance of the resulting curves approximates the continuous Fréchet distance of the original curves. Under the somewhat restrictive assumption that the arclength of each curve is short, a small number of new vertices suffices. Even in this case though, the space and preprocessing time of the data structure depends exponentially on the number of dimensions. Obviously, polynomial-time algorithms (e.g. [10]) can also benefit from reducing the number of dimensions, especially when it comes to real applications.

Our embedding naturally inherits desired properties of the JL transforms like the fact that they are oblivious to the input. This makes it directly applicable to data structure problems like the above-mentioned ANN problem. Moreover, we show that our embedding is also applicable to estimating clustering costs. First, we show that one can approximate the optimal k-center cost within a constant factor, with an algorithm that has no dependency on the original dimensionality apart from an initial step of randomly projecting the input curves. Second, we show that one can use any algorithm for computing the k-median cost in the dimensionality-reduced space to get a constant factor approximation of the k-median cost in the original space.

1.3 Organization

The paper is organized as follows. In Section 2 we introduce the necessary notation, definitions and the concept of Fréchet distance predicates. In Section 3 we prove our main result in two steps. First, as a warm-up, we prove that an application of any JL transformation for the given curves vertices and a polynomial-sized set X determined by these does not increase Fréchet distances by more than a factor of $(1+\varepsilon)$. In Section 3.1 we prove the challenging part that this also does not decrease Fréchet distances by less than a factor of $(1-\varepsilon)$. Interestingly, here a different polynomial-sized set X' is used. In Section 3.2 we combine both to our main result. Finally, in Section 4 we apply our main result to clustering of curves; we modify an existing approximation algorithm for the (k,ℓ)-center problem (see [10]) which has negligibly decreased approximation quality compared to the original and we prove that applying *any* algorithm for the (k,ℓ)-median problem (such as the one from [13]) on the embedded curves leads to a constant factor approximation in terms of clustering cost. Section 5 concludes the paper.

2 Preliminaries

For $n \in \mathbb{N}$ we define $[n] = \{1, \dots, n\}$. By $\|\cdot\|$ we denote the Euclidean norm, by $\langle \cdot, \cdot \rangle$ we denote the Euclidean dot product and by $\mathbb{S}^{d-1} = \{p \in \mathbb{R}^d \mid \|p\| = 1\}$ we denote the unit sphere in $\mathbb{R}^d$. We define line segments, the building blocks of polygonal curves.

▶ **Definition 2.** *A line segment between two points $p_1, p_2 \in \mathbb{R}^d$, denoted by $\overline{p_1 p_2}$, is the set of points $\{(1-\lambda)p_1 + \lambda p_2 \mid \lambda \in [0,1]\}$. For $\lambda \in \mathbb{R}$ we denote by $\operatorname{lp}(\overline{p_1 p_2}, \lambda)$ the point $(1-\lambda)p_1 + \lambda p_2$, lying on the line supporting the segment $\overline{p_1 p_2}$.*

We formally define polygonal curves.

▶ **Definition 3.** *A (parameterized) curve is a continuous mapping $\tau \colon [0,1] \to \mathbb{R}^d$. A curve τ is polygonal, if and only if, there exist $v_1, \dots, v_m \in \mathbb{R}^d$, no three consecutive on a line, called τ's vertices and $t_1, \dots, t_m \in [0,1]$ with $t_1 < \dots < t_m$, $t_1 = 0$ and $t_m = 1$, called τ's instants, such that τ connects every two consecutive vertices $v_i = \tau(t_i), v_{i+1} = \tau(t_{i+1})$ by a line segment.*

We call the line segments $\overline{v_1 v_2}, \dots, \overline{v_{m-1} v_m}$ the edges of τ and m the complexity of τ, denoted by $|\tau|$. Sometimes we will argue about a sub-curve $\tau[i,j]$ of a given curve τ, which is the polygonal curve determined by the vertices $v_i, \dots, v_j$. We define two notions of continuous Fréchet distances. We note that the weak Fréchet distance is used only rarely.

▶ **Definition 4.** *Let σ, τ be curves. The weak Fréchet distance between σ and τ is*

$$\mathrm{d_{wF}}(\sigma,\tau) = \inf_{\substack{f\colon [0,1]\to[0,1] \\ g\colon [0,1]\to[0,1]}} \max_{t\in[0,1]} \|\sigma(f(t)) - \tau(g(t))\|,$$

where f and g are continuous functions with $f(0) = g(0) = 0$ and $f(1) = g(1) = 1$. The Fréchet distance between σ and τ is

$$\mathrm{d_F}(\sigma,\tau) = \inf_{\substack{f\colon [0,1]\to[0,1] \\ g\colon [0,1]\to[0,1]}} \max_{t\in[0,1]} \|\sigma(f(t)) - \tau(g(t))\|,$$

where f and g are continuous bijections with $f(0) = g(0) = 0$ and $f(1) = g(1) = 1$.

We define the type of embedding we are interested in. Since we want to keep our results general, we do not specify the target number of dimensions. As a consequence, we drop the JL-terminology and call these $(1 \pm \varepsilon)$-embeddings.

▶ **Definition 5.** *Given a set $P \subset \mathbb{R}^d$ of points and $\varepsilon \in (0,1)$, a function $f \colon \mathbb{R}^d \to \mathbb{R}^{d'}$ is a $(1 \pm \varepsilon)$-embedding for P, if it holds that*

$$\forall p, q \in P : (1-\varepsilon)\|p-q\| \le \|f(p) - f(q)\| \le (1+\varepsilon)\|p-q\|.$$

We note that if f is linear and $0 \in P$, then $\forall p \in P$: $(1-\varepsilon)\|p\| \leq \|f(p)\| \leq (1+\varepsilon)\|p\|$.

We now define valid sequences with respect to two polygonal curves. Such a sequence can be seen as a discrete skeleton in deciding the continuous Fréchet distance and is derived from the free space diagram concept used in Alt and Godau's algorithm [6].

▶ **Definition 6.** *Let σ, τ be polygonal curves with vertices $v_1^\sigma, \ldots, v_{|\sigma|}^\sigma$, respectively $v_1^\tau, \ldots, v_{|\tau|}^\tau$. A valid sequence with respect to σ and τ is a sequence $\mathcal{F} = (i_1, j_1), \ldots, (i_k, j_k)$ with*

- $i_1 = j_1 = 1$, $i_k = |\sigma| - 1$, $j_k = |\tau| - 1$,
- $(i_l, j_l) \in [|\sigma| - 1] \times [|\tau| - 1]$,
- $(i_l - i_{l-1}, j_l - j_{l-1}) \in \{(0,1), (1,0), (0,-1), (-1,0)\}$ *for all* $1 < l < k$ *and*
- *any pair* $(i_l, j_l) \in [|\sigma| - 1] \times [|\tau| - 1]$ *appears at most once in* $\mathcal{F}$.

A valid sequence is said to be monotone if $(i_l - i_{l-1}, j_l - j_{l-1}) \in \{(0,1), (1,0)\}$ for all $1 < l < k$.

Further decomposing the free space diagram concept, any valid sequence for two curves σ, τ and any radius $r \geq 0$ induces a set of predicates which truth values in conjunction determine whether $\mathrm{d_F}(\sigma, \tau) \leq r$, respectively $\mathrm{d_{wF}}(\sigma, \tau) \leq r$.

▶ **Definition 7** ([2, 19])**.** *Let σ, τ be polygonal curves with vertices $v_1^\sigma, \ldots, v_{|\sigma|}^\sigma$, respectively $v_1^\tau, \ldots, v_{|\tau|}^\tau$ and $r \in \mathbb{R}_{\geq 0}$. We define the Fréchet distance predicates for σ and τ with respect to r.*

- $(P_1)^{\sigma,\tau,r}$: *This predicate is true, iff* $\|\sigma_1 - \tau_1\| \leq r$.
- $(P_2)^{\sigma,\tau,r}$: *This predicate is true, iff* $\|\sigma_{|\sigma|} - \tau_{|\tau|}\| \leq r$.
- $(P_3)_{(i,j)}^{\sigma,\tau,r}$: *This predicate is true, iff there exists a point* $p \in \overline{v_i^\sigma v_{i+1}^\sigma}$ *with* $\|p - v_j^\tau\| \leq r$.
- $(P_4)_{(i,j)}^{\sigma,\tau,r}$: *This predicate is true, iff there exists a point* $p \in \overline{v_j^\tau v_{j+1}^\tau}$ *with* $\|p - v_i^\sigma\| \leq r$.
- $(P_5)_{(i,j,k)}^{\sigma,\tau,r}$: *This predicate is true, iff there exist* $p_1 = \mathrm{lp}\left(\overline{v_j^\sigma v_{j+1}^\sigma}, t_1\right)$ *and* $p_2 = \mathrm{lp}\left(\overline{v_j^\sigma v_{j+1}^\sigma}, t_2\right)$ *with* $\|v_i^\tau - p_1\| \leq r$, $\|v_k^\tau - p_2\| \leq r$ *and* $t_1 \leq t_2$.
- $(P_6)_{(i,j,k)}^{\sigma,\tau,r}$: *This predicate is true, iff there exist* $p_1 = \mathrm{lp}\left(\overline{v_i^\tau v_{i+1}^\tau}, t_1\right)$ *and* $p_2 = \mathrm{lp}\left(\overline{v_i^\tau v_{i+1}^\tau}, t_2\right)$ *with* $\|v_j^\sigma - p_1\| \leq r$, $\|v_k^\sigma - p_2\| \leq r$ *and* $t_1 \leq t_2$.

The following two theorems state the aforementioned facts. These will be one of our main tools in obtaining our main results. We note that these are rephrased here to fit our needs.

▶ **Theorem 8** ([19])**.** *Let σ, τ be polygonal curves and $r \in \mathbb{R}_{\geq 0}$. There exists a valid sequence $\mathcal{F}$ with respect to σ and τ, such that $(P_1)^{\sigma,\tau,r} \wedge (P_2)^{\sigma,\tau,r} \wedge \Psi_w^{\sigma,\tau,r}(\mathcal{F})$ is true, where*

$$\Psi_w^{\sigma,\tau,r}(\mathcal{F}) = \bigwedge_{\substack{(i,j)\in[|\sigma|]\times[|\tau|] \\ (i,j-1),(i,j)\in\mathcal{F}}} (P_3)_{(i,j)}^{\sigma,\tau,r} \bigwedge_{\substack{(i,j)\in[|\tau|]\times[|\sigma|] \\ (i-1,j),(i,j)\in\mathcal{F}}} (P_4)_{(i,j)}^{\sigma,\tau,r},$$

if, and only if, $\mathrm{d_{wF}}(\sigma, \tau) \leq r$.

▶ **Theorem 9** ([2, 19])**.** *Let σ, τ be polygonal curves and $r \in \mathbb{R}_{\geq 0}$. There exists a monotone valid sequence $\mathcal{F}$ with respect to σ and τ, such that $(P_1)^{\sigma,\tau,r} \wedge (P_2)^{\sigma,\tau,r} \wedge \Psi^{\sigma,\tau,r}(\mathcal{F})$ is true, where*

$$\Psi^{\sigma,\tau,r}(\mathcal{F}) = \bigwedge_{\substack{(i,j)\in[|\sigma|]\times[|\tau|] \\ (i,j-1),(i,j)\in\mathcal{F}}} (P_3)_{(i,j)}^{\sigma,\tau,r} \bigwedge_{\substack{(i,j)\in[|\tau|]\times[|\sigma|] \\ (i-1,j),(i,j)\in\mathcal{F}}} (P_4)_{(i,j)}^{\sigma,\tau,r} \bigwedge_{\substack{(i,j,k)\in[|\tau|]\times[|\sigma|]\times[|\tau|] \\ (i,j-1),(i,k)\in\mathcal{F} \\ j<k}} (P_5)_{(i,j,k)}^{\sigma,\tau,r} \bigwedge_{\substack{(i,j,k)\in[|\tau|]\times[|\sigma|]\times[|\sigma|] \\ (i-1,j),(k,j)\in\mathcal{F} \\ i<k}} (P_6)_{(i,j,k)}^{\sigma,\tau,r},$$

if, and only if, $\mathrm{d_F}(\sigma, \tau) \leq r$.

3 Linear Embeddings preserve Fréchet Distances

In this section we prove our main results on embeddings of polygonal curves that approximately preserve the Fréchet distance. In the following Lemma 10, we show that linear $(1 \pm \varepsilon)$-embeddings for a polynomial number of points determined by the input polygonal curves imply embeddings for the curves that are not expansive by a factor greater than $(1+\varepsilon)$. Similarly, in Section 3.1, we show that linear $(1 \pm \varepsilon)$-embeddings for a polynomial number of points determined by the curves, imply embeddings for the curves that are not contractive by a factor smaller than $(1-\varepsilon)$. Combining these two bounds yields our main results in Section 3.2. Our main dimensionality reduction result states that one can embed a set of n polygonal curves of complexity at most m into a Euclidean space of dimensions $d' \in O(\varepsilon^{-2} \log(nm))$, so that all Fréchet distances are preserved within a factor of $(1 \pm \varepsilon)$. The embedding is implemented by mapping the vertices of each polygonal curve with a JL transform. The image of each input curve is a curve in $\mathbb{R}^{d'}$ having as vertices the images of the original vertices.

▶ **Lemma 10.** *Let σ, τ be polygonal curves with vertices $v_1^\sigma, \ldots, v_{|\sigma|}^\sigma$, respectively $v_1^\tau, \ldots, v_{|\tau|}^\tau$, and let f be a linear $(1 \pm \varepsilon)$-embedding for $P = \{v_1^\sigma, \ldots, v_{|\sigma|}^\sigma, v_1^\tau, \ldots, v_{|\tau|}^\tau\} \cup P'$, where P' is a set of points determined by σ and τ with $|P'| \in O(|\sigma|^2 \cdot |\tau| + |\tau|^2 \cdot |\sigma|)$. Let σ' and τ' be polygonal curves with vertices $f(v_1^\sigma), \ldots, f(v_{|\sigma|}^\sigma)$, respectively $f(v_1^\tau), \ldots, f(v_{|\tau|}^\tau)$. It holds that*

- $\mathrm{d_{wF}}(\sigma', \tau') \le (1+\varepsilon)\, \mathrm{d_{wF}}(\sigma, \tau)$ *and*
- $\mathrm{d_F}(\sigma', \tau') \le (1+\varepsilon)\, \mathrm{d_F}(\sigma, \tau)$.

The proof follows by an application of the $(1 \pm \varepsilon)$-embedding to all points determined by the (weak) Fréchet distance predicates. The proof can be found in the full paper [35].

3.1 Lower Bound

In this section, we show that we can use linear $(1 \pm \varepsilon)$-embeddings for a polynomial number of points determined by the input polygonal curves to define embeddings for the curves that are not contractive with respect to their Fréchet distance by a factor smaller than $(1-\varepsilon)$.

We first introduce a few necessary technical lemmas and then we proceed with the main result. We make use of the following lemma, which indicates that inner products are (weakly) concentrated in $(1 \pm \varepsilon)$-embeddings. Slightly different versions of this lemma have been used before (see e.g. [7, 34, 36]). Our statement is a bit more generic, it holds for any linear $(1 \pm \varepsilon)$-embedding, and we make use of the involved scaling factors, we include a proof in [35].

▶ **Lemma 11.** *Let f be a linear $(1 \pm \varepsilon)$-embedding for a finite set $P \subset \mathbb{R}^d$ with $0 \in P$. For all $p, q \in P$ it holds that*

$$\langle p, q \rangle - 16\varepsilon(\|p\| \cdot \|q\|) \le \langle f(p), f(q) \rangle \le \langle p, q \rangle + 14\varepsilon(\|p\| \cdot \|q\|).$$

Next, we prove that $(1 \pm \varepsilon)$-embeddings for a specific point set do not contract distances between any point on a fixed ray starting from the origin and a fixed point lying in a certain halfspace by a factor smaller than $(1-3\varepsilon)$.

▶ **Lemma 12.** *Let $x \in \mathbb{R}^d$ and $u \in \mathbb{S}^{d-1}$ such that $\langle x, u \rangle \le 0$. Let f be a linear $(1 \pm \varepsilon/16)$-embedding for $\{0, x, u\}$. For any $\lambda \ge 0$, we have*

$$\|f(x) - \lambda \cdot f(u)\| \ge (1-3\varepsilon)\|x - \lambda u\|.$$

Proof. By Lemma 11 and Definition 5: i) $\langle f(x), f(u)\rangle \in \langle x,u\rangle \pm \varepsilon\|x\|$, ii) $\|f(x)\| \in (1\pm\varepsilon)\|x\|$ and iii) $\|f(u)\| \in (1\pm\varepsilon)$. For any $\lambda \geq 0$ we have:

$$\begin{aligned}\|f(x)-\lambda\cdot f(u)\|^2 &= \|f(x)\|^2+\lambda^2\cdot\|f(u)\|^2-2\lambda\cdot\langle f(x),f(u)\rangle\\ &\geq (1-\varepsilon)^2\|x\|^2+(1-\varepsilon)^2\lambda^2-2\lambda\langle x,u\rangle-2\lambda\varepsilon\|x\| \qquad (1)\\ &\geq (1-\varepsilon)^2\|x\|^2+(1-\varepsilon)^2\lambda^2-(1-\varepsilon)^2\cdot 2\lambda\cdot\langle x,u\rangle-2\lambda\varepsilon\cdot\|x\| \qquad (2)\\ &\geq (1-\varepsilon)^2\|x-\lambda u\|^2-2\varepsilon\lambda\|x\|\\ &\geq (1-\varepsilon)^2\|x-\lambda u\|^2-2\varepsilon\|x-\lambda u\|^2 \qquad (3)\\ &\geq (1-3\varepsilon)^2\|x-\lambda u\|^2,\end{aligned}$$

where the last inequality holds, since $\varepsilon/16 \in (0, 1/4]$. In Equation (1) we use events i), ii), iii), in Equation (2) we use the fact that $\langle x,u\rangle \leq 0$, and in Equation (3) we use the fact that $\langle x,u\rangle \leq 0$ and $\lambda \geq 0$ implies that $\|x-\lambda u\| \geq \lambda$ and $\|x-\lambda u\| \geq \|x\|$. ◀

We now prove our main technical lemma. This says that given a fixed line and a fixed point p, there is a set P of points such that any linear $(1\pm\varepsilon)$-embedding for P does not contract distances between p and any point on the line by a factor smaller than $(1-3\varepsilon)$. A somewhat similar statement appears in [29] which however focuses on the distortion of point-line distances, i.e., how the distance between a point and its orthogonal projection onto the line changes after the embedding.

▶ **Lemma 13.** *Let $x,y,z \in \mathbb{R}^d$ and $\ell = \{\mathrm{lp}(\overline{yz},\lambda) \mid \lambda \in \mathbb{R}\}$ be the line supporting $\overline{yz}$. Let f be a linear $(1\pm\varepsilon/16)$-embedding for $\{0,u,-u,x-(t+\langle x,u\rangle\cdot u)\}$, where $u \in \mathbb{S}^{d-1}$ and $t \in \mathbb{R}^d$, such that $\langle u,t\rangle = 0$ and $\{t+\lambda u \mid \lambda \in \mathbb{R}\} = \ell$. For all $\lambda \in \mathbb{R}$ it holds that*

$$\|f(x)-f(t+\lambda u)\| \geq (1-3\varepsilon)\|x-(t+\lambda u)\|.$$

Proof. We first note that such an element t exists, namely the orthogonal projection of 0 onto ℓ. Let $p = t+\langle x-t,u\rangle\cdot u = t+\langle x,u\rangle\cdot u$ be the projection of x onto ℓ and let $x' = x-p$. Notice that

$$\langle x',u\rangle = \langle x,u\rangle - \langle t+\langle x,u\rangle\cdot u,u\rangle = \langle x,u\rangle-\langle t,u\rangle-\langle x,u\rangle = 0.$$

We apply Lemma 12 on the vectors x', u. This implies that for any $\lambda \geq 0$,

$$\begin{aligned} &\|f(x')-\lambda f(u)\| && \geq (1-3\varepsilon)\|x'-\lambda u\|\\ \Longleftrightarrow\ &\|f(x-p)-\lambda f(u)\| && \geq (1-3\varepsilon)\|x-p-\lambda u\|\\ \Longleftrightarrow\ &\|f(x-(t+\langle x,u\rangle\cdot u))-\lambda f(u)\| && \geq (1-3\varepsilon)\|x-(t+\langle x,u\rangle\cdot u)-\lambda u\|\\ \Longleftrightarrow\ &\|f(x)-f(t)-\langle x,u\rangle\cdot f(u)-\lambda f(u)\| && \geq (1-3\varepsilon)\|x-t-\langle x,u\rangle\cdot u-\lambda u\|\\ \Longleftrightarrow\ &\|f(x)-f(t+(\langle x,u\rangle+\lambda)\cdot u)\| && \geq (1-3\varepsilon)\|x-(t+(\langle x,u\rangle+\lambda)\cdot u)\|.\end{aligned}$$

Now by reparametrizing $\lambda' \leftarrow \langle x,u\rangle+\lambda$, we conclude that for any $\lambda' \geq \langle x,u\rangle$,

$$\|f(x)-f(t+\lambda'\cdot u)\| \geq (1-3\varepsilon)\|x-(t+\lambda'\cdot u)\|. \qquad (4)$$

Finally, we apply Lemma 12 on the vectors $x', -u$. Notice that $\langle x',-u\rangle = -\langle x',u\rangle = 0$. This implies that for any $\lambda \geq 0$,

$$\begin{aligned} &\|f(x')-\lambda f(-u)\| && \geq (1-3\varepsilon)\|x'-\lambda(-u)\|\\ \Longleftrightarrow\ &\|f(x-p)-\lambda f(-u)\| && \geq (1-3\varepsilon)\|x-p-\lambda(-u)\|\end{aligned}$$

$$\iff \|f(x-(t+\langle x,u\rangle\cdot u))-\lambda f(-u)\| \geq (1-3\varepsilon)\|x-(t+\langle x,u\rangle\cdot u)-\lambda(-u)\|$$
$$\iff \|f(x)-f(t)-\langle x,u\rangle\cdot f(u)-\lambda f(-u)\| \geq (1-3\varepsilon)\|x-t-\langle x,u\rangle\cdot u-\lambda(-u)\|$$
$$\iff \|f(x)-f(t+(\langle x,u\rangle-\lambda)\cdot u)\| \geq (1-3\varepsilon)\|x-(t+(\langle x,u\rangle-\lambda)\cdot u)\|.$$

Now by reparametrizing $\lambda' \leftarrow \langle x,u\rangle-\lambda$, we conclude that for any $\lambda' \leq \langle x,u\rangle$,

$$\|f(x)-f(t+\lambda'\cdot u)\| \geq (1-3\varepsilon)\|x-(t+\lambda'\cdot u)\|. \tag{5}$$

Equation (4) and Equation (5) conclude the lemma. ◀

Using the lemma above we can finally prove the main result of this section.

▶ **Lemma 14.** *Let σ,τ be polygonal curves with vertices $v_1^\sigma,\ldots,v_{|\sigma|}^\sigma$, respectively $v_1^\tau,\ldots,v_{|\tau|}^\tau$, and let f be a linear $(1\pm\varepsilon/48)$-embedding for $P=\{v_1^\sigma,\ldots,v_{|\sigma|}^\sigma,v_1^\tau,\ldots,v_{|\tau|}^\tau\}\cup P'$, where P' is a set of points determined by σ and τ with $|P'|\in O(|\sigma|\cdot|\tau|)$. Let σ' and τ' be polygonal curves with vertices $f(v_1^\sigma),\ldots,f(v_{|\sigma|}^\sigma)$, respectively $f(v_1^\tau),\ldots,f(v_{|\tau|}^\tau)$. It holds that*
- $\mathrm{d_{wF}}(\sigma',\tau') \geq (1-\varepsilon)\,\mathrm{d_{wF}}(\sigma,\tau)$ *and*
- $\mathrm{d_F}(\sigma',\tau') \geq (1-\varepsilon)\,\mathrm{d_F}(\sigma,\tau)$.

Proof. For the first claim, let $r=\mathrm{d_{wF}}(\sigma,\tau)$, for the second claim let $r=\mathrm{d_F}(\sigma,\tau)$. In both cases, let $r'=(1-\varepsilon)r$.

In the following, we prove that for any (monotone) valid sequence $\mathcal{F}$ and any $\delta>0$ we have that $(P_1)^{\sigma',\tau',r'-\delta}\wedge(P_2)^{\sigma',\tau',r'-\delta}\wedge\Psi_w^{\sigma',\tau',r'-\delta}(\mathcal{F})$, respectively $(P_1)^{\sigma',\tau',r'-\delta}\wedge(P_2)^{\sigma',\tau',r'-\delta}\wedge\Psi^{\sigma',\tau',r'-\delta}(\mathcal{F})$, is false and therefore $\mathrm{d_{wF}}(\sigma',\tau')>r'-\delta$, respectively $\mathrm{d_F}(\sigma',\tau')>r'-\delta$ by Theorem 8, respectively Theorem 9.

Now, let $\mathcal{F}$ be an arbitrary (monotone) valid sequence. By definition of r and Theorem 8, respectively Theorem 9, we know that for any $\delta>0$ it holds that $(P_1)^{\sigma,\tau,r-\delta}\wedge(P_2)^{\sigma,\tau,r-\delta}\wedge\Psi_w^{\sigma,\tau,r-\delta}(\mathcal{F})$, respectively $(P_1)^{\sigma,\tau,r-\delta}\wedge(P_2)^{\sigma,\tau,r-\delta}\wedge\Psi^{\sigma,\tau,r-\delta}(\mathcal{F})$, is false. If $(P_1)^{\sigma,\tau,r-\delta}$ or $(P_2)^{\sigma,\tau,r-\delta}$ is false then clearly $(P_1)^{\sigma',\tau',r'-\delta}$ or $(P_2)^{\sigma',\tau',r'-\delta}$ is also false by Definitions 5 and 7. In the following, we assume that $\Psi_w^{\sigma,\tau,r-\delta}(\mathcal{F})$, respectively $\Psi^{\sigma,\tau,r-\delta}(\mathcal{F})$ is false.

Since the arguments for predicates of type P_3 and P_4 are analogous, we focus on the former type. Assume that $\Psi_w^{\sigma,\tau,r-\delta}(\mathcal{F})$ is false because a predicate $(P_3)_{(i,j)}^{\sigma,\tau,r-\delta}$ is false. This means that there does not exist a point $p\in\overline{v_i^\sigma v_{i+1}^\sigma}$ with $\|p-v_j^\tau\|\leq r-\delta$. At this point, recall that since f is linear, any points $\mathrm{lp}\left(\overline{pq},t_1\right),\ldots,\mathrm{lp}\left(\overline{pq},t_n\right)$, where $p,q\in\mathbb{R}^d$, are still collinear when f is applied and the relative order on the directed lines supporting $\overline{pq}$ is preserved, which is immediate since $f(\mathrm{lp}\left(\overline{pq},t_i\right))=\mathrm{lp}\left(\overline{f(p)f(q)},t_i\right)$. By Lemma 13 for any $t\in\mathbb{R}$ and the determined point $p=\mathrm{lp}\left(\overline{v_i^\sigma v_{i+1}^\sigma},t\right)$ on the line supporting $\overline{v_i^\sigma v_{i+1}^\sigma}$ it holds that $\|f(v_j^\tau)-f(p)\|\geq(1-\varepsilon)\|p-v_j^\tau\|$. Thus, for any $f(p)\in\overline{f(v_i^\sigma)f(v_{i+1}^\sigma)}$ we have $p\in\overline{v_i^\sigma v_{i+1}^\sigma}$ and $\|f(v_j^\tau)-f(p)\|\geq(1-\varepsilon)\|p-v_j^\tau\|$, which in conclusion is larger than $r'-\delta$, hence $(P_3)_{(i,j)}^{\sigma',\tau',r'-\delta}$ is false and therefore $\Psi_w^{\sigma',\tau',r'-\delta}(\mathcal{F})$ is false. The first claim follows by Theorem 8.

Now, since again the arguments for predicates of type P_5 and P_6 are also analogous, we focus on the former. Assume that $\Psi^{\sigma,\tau,r-\delta}(\mathcal{F})$ is false, because a predicate $(P_5)_{(i,j,k)}^{\sigma,\tau,r-\delta}$ is false. This means that for any two $t_1,t_2\in\mathbb{R}$ with $t_1\leq t_2$, the points $p_1=\mathrm{lp}\left(\overline{v_j^\sigma v_{j+1}^\sigma},t_1\right)$ and $p_2=\mathrm{lp}\left(\overline{v_j^\sigma v_{j+1}^\sigma},t_2\right)$ do not satisfy $\|v_i^\tau-p_1\|\leq r-\delta$ or $\|v_k^\tau-p_2\|\leq r-\delta$. Since by Lemma 13 we have $\|f(v_i^\tau)-f(p_1)\|\geq(1-\varepsilon)\|v_i^\tau-p_1\|$ and $\|f(v_k^\tau)-f(p_2)\|\geq(1-\varepsilon)\|v_k^\tau-p_2\|$, one of these distances must be larger than $r'-\delta$ and it follows that $(P_5)_{(i,j,k)}^{\sigma',\tau',r'-\delta}$ is false. Therefore, $\Psi^{\sigma',\tau',r'-\delta}(\mathcal{F})$ is false and the second claim follows by Theorem 9.

Finally, for the above statements to hold, the set P' contains 0, both directions $u, -u \in \mathbb{S}^{d-1}$ determined by an edge of σ or τ and all points $x - (t + \langle x, u\rangle \cdot u)$, where x is a vertex of a curve σ or τ, and t, u determine a line supporting an edge of τ or σ. ◀

3.2 Main Result

We now prove our main result which combines the upper and lower bounds on the distortion and Lemma 10.

▶ **Theorem 15.** *Let $T = \{\tau_1, \ldots, \tau_n\}$ be a set of polygonal curves in $\mathbb{R}^d$, each of complexity at most m. There exists a probability distribution over linear maps $f\colon \mathbb{R}^d \to \mathbb{R}^{d'}$, where $d' \in O(\varepsilon^{-2}\log(nm))$, such that with high probability over the choice of f, the following is true for all $\sigma, \tau \in T$:*

- $|\mathrm{d_{wF}}(\sigma,\tau) - \mathrm{d_{wF}}(F(\sigma),F(\tau))| \le \varepsilon \cdot \mathrm{d_{wF}}(\sigma,\tau)$ *and*
- $|\mathrm{d_{F}}(\sigma,\tau) - \mathrm{d_{F}}(F(\sigma),F(\tau))| \le \varepsilon \cdot \mathrm{d_{F}}(\sigma,\tau)$*,*

where for any $\tau \in T$ with vertices $v_1^\tau, \ldots, v_{|\tau|}^\tau$ we let $F(\tau)$ be the curve with vertices $f(v_1^\tau), \ldots, f(v_{|\tau|}^\tau)$.

Proof. We apply Theorem 1 on the set P of $O(n^2m^3)$ points determined by an application of Lemmas 10 and 14 on all pairs of curves in T. ◀

4 Application to Clustering

In this section, we study the effect of randomized $(1\pm\varepsilon)$-embeddings on the cost of k-clustering of polygonal curves. In particular, we show that a constant factor approximation of the cost of the optimal k-center solution can be computed with an algorithm, which, except for the time needed to embed the input curves, runs in time independent of the input dimensionality. Moreover, we show that the optimal cost of the k-median problem is preserved within a constant factor in the target space. This means that running any algorithm for the k-median problem in the target space, yields an algorithm for estimating the cost in the original space.

This effectively reduces the computational effort required for approximating the clustering cost, and it directly assists analytical tasks like estimating the optimal number of clusters – where cost estimations for multiple values of k are typically performed.

4.1 Clustering under the Fréchet Distance

In 2016, Driemel et al. [18] introduced clustering under the Fréchet distance, for the purpose of clustering (one-dimensional) time series. The objectives, named (k,ℓ)-center and (k,ℓ)-median, are derived from the well-known k-center and k-median objectives in Euclidean k-clustering. Both are NP-hard [18, 10, 11], even if $k = 1$ and $d = 1$, and the (k,ℓ)-center problem is even NP-hard to approximate within a factor of $(2.25 - \varepsilon)$ in general dimensions [10]. One particularity of these clustering approaches is that the obtained center curves should be of low complexity. In detail, while the given curves have complexity at most m each, the centers should be of complexity at most ℓ each, where $\ell \ll m$ is a constant. The idea behind is that due to the linear interpolation, a *compact* summary of the cluster members through an aggregate center curve is enabled. A nice side effect is that overfitting, which may occur without the complexity restriction, is suppressed. For further details see [18].

We now present a modification of the constant factor approximation algorithm for (k,ℓ)-center clustering from [10]. We note that due to its appealing complexity, this algorithm is used vastly in practice (c.f. [12]) and therefore constitutes a prime candidate to be combined with dimensionality reduction.

4.2 (k,ℓ)-Center Clustering

We formally define the (k,ℓ)-center clustering objective.

▶ **Definition 16.** *The (k,ℓ)-center clustering problem is to compute a set C of k polygonal curves in $\mathbb{R}^d$, of complexity at most ℓ each, which minimizes the cost $\max_{\tau\in T}\min_{c\in C} \mathrm{d_F}(\tau,c)$, where $T=\{\tau_1,\dots,\tau_n\}$ is a given set of polygonal curves in $\mathbb{R}^d$ of complexity at most m each, and $k\in\mathbb{N}, \ell\in\mathbb{N}_{\geq 2}$ are constant parameters of the problem.*

The following algorithm largely makes use of simplifications of input curves. We formally define this concept.

▶ **Definition 17.** *An α-approximate minimum-error ℓ-simplification of a curve τ in $\mathbb{R}^d$ is a curve $\sigma = \mathrm{simpl}(\tau)$ in $\mathbb{R}^d$ with at most ℓ vertices, where $\ell\in\mathbb{N}_{\geq 2}$ and $\alpha\geq 1$ are given parameters, such that $\mathrm{d_F}(\tau,\sigma)\leq \alpha\cdot \mathrm{d_F}(\tau,\sigma')$ for all other curves σ' with ℓ vertices.*

A simplification $\sigma = \mathrm{simpl}(\tau)$ is vertex-restricted if the sequence of its vertices is a subsequence of the sequence of τs vertices. Crucial in our modification of the algorithm by Buchin et al. [10] is that we want to compute simplifications in the dimensionality-reduced ambient space to spare running time. In the following, we give a thorough analysis of the effect of dimensionality reduction before simplification. The proof can be found in the full paper [35].

▶ **Theorem 18.** *Let F be the embedding of Theorem 15 with parameter $\varepsilon\in(0,1/2]$, for a given set T of n polygonal curves in $\mathbb{R}^d$ of complexity at most m each, all segments $\overline{v_i^\tau v_j^\tau}$, all subcurves $\tau[i,j]$ as well as all vertex-restricted ℓ-simplifications of all $\tau\in T$ (where $v_1^\tau,\dots,v_{|\tau|}^\tau$ are the vertices of τ and $i,j\in[|\tau|]$ with $i<j$). For each $\tau\in T$, a 4-approximate minimum-error ℓ-simplification $\mathrm{simpl}(F(\tau))$ of $F(\tau)$ can be computed in time $O(d'\cdot|\tau|^3\log|\tau|)$ and for all $\sigma\in T$ it holds that*

$$(1-\varepsilon)\,\mathrm{d_F}(\sigma,\mathrm{simpl}(\tau)) \leq \mathrm{d_F}(F(\sigma),\mathrm{simpl}(F(\tau)) \leq (1+\varepsilon)\,\mathrm{d_F}(\sigma,\mathrm{simpl}(\tau)),$$

where $\mathrm{simpl}(\tau)$ denotes a $(4+16\varepsilon)$-approximate minimum-error ℓ-simplification of τ.

We now present our modification of the algorithm. Let F denote the embedding from Theorem 15 for $T\cup T'\cup C^*$, where T' is the set of all segments $\overline{v_i^\tau v_j^\tau}$, all subcurves $\tau[i,j]$ as well as all vertex-restricted ℓ-simplifications of all $\tau\in T$ (where $v_1^\tau,\dots,v_{|\tau|}^\tau$ are the vertices of τ and $i,j\in[|\tau|]$ with $i<j$), and C^* is an optimal set of k centers for T.

The algorithm first sets $C=\{\mathrm{simpl}(F(\tau))\}$ for an arbitrary $\tau\in T$. Then, until $|C|=k$ it computes a curve $\tau\in T$ that maximizes $\min_{c\in C}\mathrm{d_F}(F(\tau),c)$ and sets $C=C\cup\{\mathrm{simpl}(F(\tau))\}$. Finally, it returns C.

We now prove the approximation guarantee and analyse the running time of this algorithm, thereby we adapt parts of the analysis in [10]. The proof can be found in the full paper [35].

▶ **Theorem 19.** *Given a set T of n polygonal curves in $\mathbb{R}^d$ of complexity at most m each, and a parameter $\varepsilon\in(0,1/2]$, the above algorithm returns a solution C to the (k,ℓ)-center clustering problem, consisting of k curves in $\mathbb{R}^{O(\varepsilon^{-2}\ell\log(knm))}$ of complexity at most ℓ each, such that*

$$(1-3\varepsilon)r^* \leq \max_{\tau\in T}\min_{c\in C}\mathrm{d_F}(F(\tau),c)\leq(6+38\varepsilon)r^*,$$

where r^* denotes the cost of an optimal solution. The algorithm has running time

$$O(\varepsilon^{-2}k\ell\log(nm+k)m^3\log m+\varepsilon^{-2}\ell\log(nm+k)k^2nm\log m).$$

4.3 (k, ℓ)-Median Clustering

In this section, we show that the cost of the optimal (k, ℓ)-median solution is preserved within a constant factor, when projecting the input curves as described in Section 3. We first define the (k, ℓ)-median clustering problem.

▶ **Definition 20.** *The (k, ℓ)-median clustering problem is to compute a set C of k polygonal curves in $\mathbb{R}^d$ of complexity at most ℓ each, which minimizes the cost $\sum_{\tau \in T} \min_{c \in C} \mathrm{d_F}(\tau, c)$, where $T = \{\tau_1, \ldots, \tau_n\}$ is a given set of polygonal curves in $\mathbb{R}^d$ of complexity at most m each, and $k \in \mathbb{N}, \ell \in \mathbb{N}_{\geq 2}$ are constant parameters of the problem.*

In Section 4.3.1, we focus on the case $\ell \geq m$, and we bound the distortion of the optimal cost by a factor of $2 + O(\varepsilon)$. In Section 4.3.2, we discuss case $\ell < m$, and we bound the distortion of the optimal cost by a factor of $6 + O(\varepsilon)$.

4.3.1 Unrestricted Medians

In this section, we present our results on the (k, ℓ)-median clustering problem, when $\ell \geq m$. Computing medians of complexity $\ell = m$ is a widely accepted scenario following, for example, from the wide acceptance of local search methods for clustering, which explore candidate solutions from the set of input curves. The proof follows a similar reasoning as in Section 4.3.2 and is diverted to the full paper [35]. Comparing to Section 4.3.2, we obtain an improved bound on the approximation factor. This is mainly because simplifications are no longer needed in order to obtain a meaningful bound.

▶ **Theorem 21.** *Let $T = \{\tau_1, \ldots, \tau_n\}$ be a set of polygonal curves in $\mathbb{R}^d$ of complexity at most m each and let $\ell \geq m$. There exists a probability distribution over linear maps $f \colon \mathbb{R}^d \to \mathbb{R}^{d'}$, where $d' \in O(\varepsilon^{-2} \log(n\ell))$, such that with high probability over the choice of f, the following is true. For any polygonal curve τ with vertices $v_1^\tau, \ldots, v_{|\tau|}^\tau$, we define $F(\tau)$ to be the curve with vertices $f(v_1^\tau), \ldots, f(v_{|\tau|}^\tau)$. Then,*

$$\frac{1-\varepsilon}{2} \cdot r^* \leq r_f^* \leq (1+\varepsilon) \cdot r^*,$$

where r^* is the cost of an optimal solution to the (k, ℓ)-median problem on T, and r_f^* is the cost of an optimal solution to the (k, ℓ)-median problem on $F(T)$.

4.3.2 Restricted Medians

To bound the cost of the optimal (k, ℓ)-median in the projected space, we use the notion of simplifications which was introduced in Section 4.2. By an averaging argument, for each cluster, there exists an input curve σ_i which is within distance $\frac{1}{|T_i|} \cdot \sum_{\tau \in T_i} \mathrm{d_F}(F(\tau), c_i^f)$ from the optimal median c_i^f, where T_i is the input curves associated with the i^{th} cluster in the projected space. To lower bound the optimal cost in the projected space, we repeatedly apply the triangle inequality on distances involving a vertex-restricted ℓ-simplification of σ_i and a vertex-restricted ℓ-simplification of $F(\sigma_i)$. The upper bound simply follows by the non-contraction guarantee of JL transforms, on distances between input curves and the optimal medians in the original space. The proof can be found in the full paper [35].

▶ **Theorem 22.** *Let $T = \{\tau_1, \ldots, \tau_n\}$ be a set of polygonal curves in $\mathbb{R}^d$ of complexity at most m each and let $\ell < m$. There exists a probability distribution over linear maps $f \colon \mathbb{R}^d \to \mathbb{R}^{d'}$, where $d' \in O(\varepsilon^{-2} \ell \log(nm))$, such that with high probability over the choice of f, the following* *is true. For any polygonal curve* τ *with vertices* $v_1^\tau, \dots, v_{|\tau|}^\tau$*, we define* $F(\tau)$ *to be the curve with vertices* $f(v_1^\tau), \dots, f(v_{|\tau|}^\tau)$*. Then,*

$$\frac{1-\varepsilon}{6\cdot(1+\varepsilon)}\cdot r^* \leq r_f^* \leq (1+\varepsilon)\cdot r^*,$$

where r^* *is the cost of an optimal solution to the* (k,ℓ)*-median problem on* T*, and* r_f^* *is the cost of an optimal solution to the* (k,ℓ)*-median problem on* $F(T)$*.*

5 Conclusion

Our results are in line with the results by Magen [29, 30] in the sense that by increasing the constant hidden in the O-notation specifying the number of dimensions of the dimensionality-reduced space, JL transforms become more powerful and do not only preserve pairwise Euclidean distances but also affine distances, angles and volumes, and as we have proven, Fréchet distances.

Concerning JL transforms we have improved the work by Meintrup et al. [32] by proving that no additive error is involved in the resulting Fréchet distances. To facilitate this result, we had to incorporate the linearity of these transforms, which is not done in [32]. Interestingly, this shows that when one uses a terminal embedding instead (see e.g. [33]) – for example to handle a dynamic setting involving queries – this may induce an additive error to the Fréchet distance, as the results by Meintrup et al. [32] can still be applied but ours can not since terminal embeddings are non-linear. Consequently, in contrast to Euclidean distances where a terminal embedding constitutes a proper extension of a JL embedding, this may not be the case when it comes to Fréchet distances.

One open question of practical importance is whether one can improve our result for polygonal curves that satisfy some realistic structural assumption, e.g., c-packness [16]. Moreover, it is possible that our implications on clustering can be improved. One question there is whether one can reduce (or eliminate) the dependence on n from the target dimension, in the same spirit as with the analogous results for the Euclidean distance [31].

References

1 Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003.

2 Peyman Afshani and Anne Driemel. On the complexity of range searching among curves. In Artur Czumaj, editor, *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*, pages 898–917. SIAM, 2018.

3 Pankaj K. Agarwal, Kyle Fox, Jiangwei Pan, and Rex Ying. Approximating Dynamic Time Warping and Edit Distance for a Pair of Point Sequences. In Sándor P. Fekete and Anna Lubiw, editors, *32nd International Symposium on Computational Geometry, SoCG, June 14-18, Boston, MA, USA*, volume 51 of *LIPIcs*, pages 6:1–6:16. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2016.

4 Nir Ailon and Bernard Chazelle. The Fast Johnson–Lindenstrauss Transform and Approximate Nearest Neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009.

5 Noga Alon. Problems and results in extremal combinatorics–I. *Discrete Mathematics*, 273(1-3):31–53, 2003.

6 Helmut Alt and Michael Godau. Computing the Fréchet Distance between two Polygonal Curves. *International Journal of Computational Geometry & Applications*, 5:75–91, 1995.

7 Rosa I. Arriaga and Santosh S. Vempala. An Algorithmic Theory of Learning: Robust Concepts and Random Projection. In *40th Annual Symposium on Foundations of Computer Science, FOCS '99, 17-18 October, 1999, New York, NY, USA*, pages 616–623. IEEE Computer Society, 1999.

8 Richard G. Baraniuk and Michael B. Wakin. Random projections of smooth manifolds. *Found. Comput. Math.*, 9(1):51–77, 2009.

9 Karl Bringmann. Why Walking the Dog Takes Time: Frechet Distance Has No Strongly Subquadratic Algorithms Unless SETH Fails. In *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS, Philadelphia, PA, USA, October 18-21*, pages 661–670. IEEE Computer Society, 2014.

10 Kevin Buchin, Anne Driemel, Joachim Gudmundsson, Michael Horton, Irina Kostitsyna, Maarten Löffler, and Martijn Struijs. Approximating (k, ℓ)-center clustering for curves. In Timothy M. Chan, editor, *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 2922–2938. SIAM, 2019.

11 Kevin Buchin, Anne Driemel, and Martijn Struijs. On the Hardness of Computing an Average Curve. In Susanne Albers, editor, *17th Scandinavian Symposium and Workshops on Algorithm Theory*, volume 162 of *LIPIcs*, pages 19:1–19:19. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

12 Kevin Buchin, Anne Driemel, Natasja van de L'Isle, and André Nusser. klcluster: Center-based Clustering of Trajectories. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 496–499, 2019.

13 Maike Buchin, Anne Driemel, and Dennis Rohde. Approximating (k, ℓ)-Median Clustering for Polygonal Curves. In Dániel Marx, editor, *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, SODA, Virtual Conference, January 10 - 131*, pages 2697–2717. SIAM, 2021.

14 Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22(1):60–65, January 2003.

15 Anne Driemel and Sariel Har-Peled. Jaywalking Your Dog: Computing the Fréchet Distance with Shortcuts. *SIAM Journal on Computing*, 42(5):1830–1866, 2013.

16 Anne Driemel, Sariel Har-Peled, and Carola Wenk. Approximating the Fréchet Distance for Realistic Curvesin Near Linear Time. *Discrete & Computational Geometry*, 48(1):94–127, 2012.

17 Anne Driemel and Amer Krivosija. Probabilistic Embeddings of the Fréchet Distance. In *Proceedings of the 16th International Workshop on Approximation and Online Algorithms (WAOA)*, pages 218–237, 2018.

18 Anne Driemel, Amer Krivosija, 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, 2016.

19 Anne Driemel, André Nusser, Jeff M. Phillips, and Ioannis Psarros. The VC Dimension of Metric Balls under Fréchet and Hausdorff Distances. *Discrete & Computational Geometry*, 66(4):1351–1381, 2021.

20 Arnold Filtser, Omrit Filtser, and Matthew J. Katz. Approximate Nearest Neighbor for Curves - Simple, Efficient, and Deterministic. In Artur Czumaj, Anuj Dawar, and Emanuela Merelli, editors, *47th International Colloquium on Automata, Languages, and Programming, ICALP 2020, July 8-11, 2020, Saarbrücken, Germany (Virtual Conference)*, volume 168 of *LIPIcs*, pages 48:1–48:19. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

21 Péter Frankl and Hiroshi Maehara. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory, Series B*, 44(3):355–362, 1988.

22 Piotr Indyk and Rajeev Motwani. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In Jeffrey Scott Vitter, editor, *Proceedings of the Thirtieth Annual*

ACM Symposium on the Theory of Computing, Dallas, Texas, USA, May 23-26, 1998, pages 604–613. ACM, 1998.

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

24 Daniel M. Kane and Jelani Nelson. Sparser Johnson-Lindenstrauss Transforms. *Journal of the ACM*, 61(1):4:1–4:23, 2014.

25 Kasper Green Larsen and Jelani Nelson. The Johnson-Lindenstrauss Lemma Is Optimal for Linear Dimensionality Reduction. In Ioannis Chatzigiannakis, Michael Mitzenmacher, Yuval Rabani, and Davide Sangiorgi, editors, *43rd International Colloquium on Automata, Languages, and Programming, ICALP, July 11-15, Rome, Italy*, volume 55 of *LIPIcs*, pages 82:1–82:11. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2016.

26 Kasper Green Larsen and Jelani Nelson. Optimality of the Johnson-Lindenstrauss Lemma. In Chris Umans, editor, *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS, Berkeley, CA, USA, October 15-17*, pages 633–638. IEEE Computer Society, 2017.

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

28 Nathan Linial, Eran London, and Yuri Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, June 1995.

29 Avner Magen. Dimensionality Reductions That Preserve Volumes and Distance to Affine Spaces, and Their Algorithmic Applications. In José D. P. Rolim and Salil P. Vadhan, editors, *Randomization and Approximation Techniques, 6th International Workshop, RANDOM, Cambridge, MA, USA, September 13-15, Proceedings*, volume 2483 of *Lecture Notes in Computer Science*, pages 239–253. Springer, 2002.

30 Avner Magen. Dimensionality Reductions in l_2 that Preserve Volumes and Distance to Affine Spaces. *Discrete & Computational Geometry*, 38(1):139–153, 2007.

31 Konstantin Makarychev, Yury Makarychev, and Ilya P. Razenshteyn. Performance of Johnson-Lindenstrauss transform for k-means and k-medians clustering. In Moses Charikar and Edith Cohen, editors, *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, pages 1027–1038. ACM, 2019.

32 Stefan Meintrup, Alexander Munteanu, and Dennis Rohde. Random Projections and Sampling Algorithms for Clustering of High-Dimensional Polygonal Curves. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS, December 8-14, Vancouver, BC, Canada*, pages 12807–12817, 2019.

33 Shyam Narayanan and Jelani Nelson. Optimal terminal dimensionality reduction in Euclidean space. In Moses Charikar and Edith Cohen, editors, *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, pages 1064–1069. ACM, 2019.

34 Christos H. Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh S. Vempala. Latent Semantic Indexing: A Probabilistic Analysis. *Journal of Computer and System Sciences*, 61(2):217–235, 2000.

35 Ioannis Psarros and Dennis Rohde. Random projections for curves in high dimensions. *CoRR*, abs/2207.07442, 2022. `arXiv:2207.07442`.

36 Tamás Sarlós. Improved Approximation Algorithms for Large Matrices via Random Projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 143–152, 2006.