Approximation and Streaming Algorithms for Projective Clustering via Random Projections

Michael Kerber* Sharath Raghvendra†

Abstract

Let $\varepsilon > 0$ be any constant and let P be a set of n points in $\mathbb{R}^d$. We design new streaming and approximation algorithms for clustering points of P. Consider the projective clustering problem: Given $k, q < n$, compute a set $\mathcal{F}$ of k q-flats such that the function $f_k^q(P,\rho) = \sum_{p\in P} \mathtt{d}(p,\mathcal{F})^\rho$ is minimized; here $\mathtt{d}(p,\mathcal{F})$ represents the distance of p to the closest q-flat in $\mathcal{F}$. For $\rho = \infty$, we interpret $f_k^q(P,\rho)$ to be $\max_{r\in P} \mathtt{d}(r,\mathcal{F})$. When $\rho = 1, 2$ and ∞ and $q = 0$, the problem corresponds to the well-known k-median, k-mean and the k-center clustering problems respectively.
Our two main technical contributions are as follows:

(i) Consider an orthogonal projection of P to a randomly chosen $O(C_\rho(q,\varepsilon)\log n/\varepsilon^2)$-dimensional flat. For every subset $S \subseteq P$, we show that such a random projection will ε-approximate $f_1^q(S,\rho)$. This result holds for any integer norm $\rho \geq 1$, including $\rho = \infty$; here $C_\rho(q,\varepsilon)$ is the size of the smallest coreset that ε-approximates $f_1^q(\cdot,\rho)$. For $\rho = 1, 2$ and ∞, $C_\rho(q,\varepsilon)$ is known to be a constant which depends only on q and ε.

(ii) As our second contribution, we improve the size of the coreset when $\rho = \infty$. In particular, we improve the bounds of $C_\infty(q,\varepsilon)$ to $O(q^3/\varepsilon^2)$ from the previously-known $O(q^6/\varepsilon^5 \log 1/\varepsilon)$.

As applications of our work, we obtain better approximation and streaming algorithms for various projective clustering problems over high dimensional point sets. For example, when $\rho = \infty$ and $q \geq 1$, we obtain a streaming algorithm that maintains an ε-approximate solution using $O((d+n)q^3(\log n/\varepsilon^4))$ space. This is better than $O(nd)$ which is the input size and therefore, our algorithm is useful especially when n and d are in the same order of magnitude.

1 Introduction

Clustering high-dimensional data is an important task arising in areas such as machine learning, exploratory data mining, computer vision, pattern recognition and bioinformatics. In this paper, we consider clustering high-dimensional point clouds where the number of points and the dimension of the point cloud are of similar size. In particular, we design new approximation and streaming algorithms for the well-known *projective clustering problem*: For a set P of n points in $\mathbb{R}^d$, given integer parameters $k, q < n$, an error parameter $\varepsilon > 0$ and an integer norm $\rho \geq 1$, we compute a set $\mathcal{F}$ of k q-flats such that the function $f_k^q(P,\rho) = \sum_{p\in P} \mathtt{d}(p,\mathcal{F})^\rho$ is minimized; here $\mathtt{d}(p,\mathcal{F})$ represents the Euclidean distance of p to its closest point on any q-flat in $\mathcal{F}$. When $\rho = \infty$, we interpret $f_k^q(P,\rho)$ to be $\max_{p\in P} \mathtt{d}(p,\mathcal{F})$. Several well-known clustering functions fall in this category. For example, when $\rho = \infty$, $q = 0$ this problem reduces to the *minimum enclosing ball* (MEB) problem (when $k = 1$) and the *k-center clustering problem* (for arbitrary k). For $\rho = \infty$ and $q = 1$, we get the *minimum enclosing cylinder* (MEC) (for $k = 1$) and the *k-cylinder clustering problem* (for arbitrary k). When $q = 0$, we get the k-median clustering problem (for $\rho = 1$) and for k-means clustering problem (for $\rho = 2$).

*Max Planck Institute for Informatics, Saarbrücken, Germany. mkerber@mpi-inf.mpg.de.
†Virginia Tech., Blacksburg, USA. r.sharath@gmail.com.

The paradigm of *coresets* is useful for designing efficient approximation algorithms [4], especially for clustering high-dimensional data. Typically, a coreset is a small "most-relevant" subset C of the input points P with the property that an optimal solution on C is an approximate solution for P. For many problems, coresets can be computed efficiently and therefore they have been used in the design of fast approximation algorithms. Coresets are also useful in the design of *streaming* algorithms[1]; see for example [11, 20, 21].

In the context of projective clustering, a slightly weaker definition of a coreset has been used[2] – for $k = 1$, a coreset is a subset $C \subseteq P$ such that the affine subspace spanned by C contains a q-flat F with $\mathtt{d}(p, F) \leq (1 + \varepsilon) f_1^q(P)$. We let $C_\rho(q, \varepsilon)$ denote the worst-case size of such a coreset for approximating $f_1^q(P)$. For problems such as the MEB, MEC, 1-mean, 1-median, there are coresets whose size is independent of the number of points and the ambient dimension; see [7, 8, 22] for details.

Another tool for handling high-dimensional data is the *random projection* method [31]. At its heart is the following well-known lemma [25] which says that an orthogonal projection of any point set to a random $O(\log n/\varepsilon^2)$-dimensional flat preserves pairwise distances between all pairs of points; see below for a precise statement and see [16] for an elementary proof of the statement.

Theorem 1 (Johnson-Lindenstrauss) *For $0 < \varepsilon < 1$, a set $P \subset \mathbb{R}^d$ of n points, and $m \geq 36 \ln(n)/\varepsilon^2$, there is a map $\hat{\pi} : \mathbb{R}^d \to \mathbb{R}^m$ such that*

$$(1 - \varepsilon)\|u - v\|^2 \leq \|\hat{\pi}(u) - \hat{\pi}(v)\|^2 \leq (1 + \varepsilon)\|u - v\|^2$$

for any $u, v \in P$. Moreover, a randomly chosen map $\hat{\pi}$ of the form $\hat{\pi}(p) = \sqrt{d/m} \cdot \pi(p)$ where π is the orthogonal projection to a m-dimensional subspace of $\mathbb{R}^d$, satisfies that property with probability at least $1/2$.

We abuse notations and refer to $\hat{\pi}$ described above as a *random projection* to an m-dimensional flat.

Our results. We establish a novel link between coresets and the random projections leading to new streaming and approximation algorithms for the projective clustering problem . First, we show that a random projection to $O(C_\rho(q, \varepsilon) \log n/\varepsilon^2)$ dimensional space ε-approximates $f_1^q(S, \rho)$ for every subset $S \subseteq P$. Our result can be seen as a generalization of the JL-Lemma which only handles pairwise distances. The main ingredient of our proof is to show that a random projection to an $O(C_\rho(q, \varepsilon) \log n/\varepsilon^2)$-dimensional subspace "preserves" all flats defined by subsets of size $C_\rho(q, \varepsilon)$. These results show that a very small increase in the dimension of the projected space can preserve a lot more geometric structure than just the pairwise distances – similar observations for different geometric structures have been made in [2, 12, 24, 27, 28]. As a consequence, we argue that projective clustering is preserved under random projections. Our result also provides a new application of coresets – smaller coresets imply better bounds on the dimension of the projected flat. Most algorithmic applications require coresets that can be computed quickly. We only require the existence of a small size coreset. This allows to shoot for smaller bounds on the size of the coreset without optimizing on the construction time.

We highlight this through our second contribution. When $\rho = \infty$, we show the existence of coresets of size $O(q^3/\varepsilon^2)$ (Section 3); the best previously known bound was $O(q^6/\varepsilon^5 \log 1/\varepsilon)$ [22]. Our result significantly improves the dimension of the projected space from $O(q^6/\varepsilon^7 \log 1/\varepsilon \log n)$ to $O(q^3/\varepsilon^4 \log n)$.

Our results have the following applications (Section 4):

- We obtain a generic improvement scheme for streaming algorithms for projective clustering of high-dimensional data that approximates $f_k^q(P, \rho)$ with high probability. Precisely, we reduce the dependence on the dimension by projecting the arriving points and executing a streaming algorithm on these projected points. This simple approach has useful consequences: when $\rho = \infty$, we obtain an $O((d +$

[1] In the streaming setting, algorithms are allowed to make one or few passes over the data and maintain an approximate solution using a very small workspace.

[2] We use this definition throughout the paper

$n)C_\infty(q,\varepsilon)\log n/\varepsilon^2)$ space algorithm for approximating $f_k^q(P,\infty)$ which is significantly better than the $O(nd)$ space required to store the entire input and is close to the lower bound of $\Omega(n)$ known for the case when $q = 0$ [3]. For k-means and k-median, we significantly improve the $O(d^2k^2/\varepsilon^2\log^8 n)$-streaming algorithm by Chen [11] to $O(d\log n/\varepsilon^3 + k^2/\varepsilon^8\log^{10} n)$ for k-means and $O(d\log n/\varepsilon^4 + k^2/\varepsilon^{10}\log 1/\varepsilon\log^{10} n)$ for k-median.

- We also generically improve algorithms that compute an approximate solution for projective clustering problems. Again, we project the input to a lower-dimensional subspace and compute an approximate solution in the projected space. We obtain the solution in d-dimensional space by "lifting" the clusters in projected space separately. For the approximate k-cylinder problem, our approach yields a bound of $O(n\log n 2^{k\log k/\varepsilon} + \frac{dn\log n}{\varepsilon^3})$ which improves the previously known best $O(nd2^{k\log k/\varepsilon})$ [7] in the sense that k and d are decoupled in the complexity bound.
- Since our results imply that, under random projections, the radius of MEB is approximated for every subset of the input, we immediately get an approximation scheme for a d-dimensional Čech complex in Euclidean space by a Čech complex in lower dimensions. In particular, this result bounds the persistence of high-dimensional homology classes of the original Čech complex. Very recently, these results have been proven independently by Sheehy [30].

2 Generalized Johnson-Lindenstrauss Lemma

Recall the definition of $f_1^q(P,\rho)$ as the maximal distance of P to the best fitting q-flat. We show that a random projection to appropriately large subspaces approximately preserves $f_1^q(S,\rho)$ for any subset $S \subseteq P$. What dimension is appropriate for a projection depends on the corresponding coreset size $C := C_\rho(q,\varepsilon)$; precisely, picking a $O(C\log(n)/\varepsilon^2)$-dimensional subspace is enough.

We outline the proof of the statement before giving the technical details in the remainder of the section. For a set $S \subset P$, we let $\langle S\rangle$ denote the *span* of S, that is, the subspace spanned by the points in S. We know that any subset of P has a coreset of size C whose span contains an approximately optimal q-flat F. If the distance of F to any $p \in P$ is preserved, we can guarantee to preserve $f_1^q(S,\rho)$ approximately as well. We ensure this preservation by the stronger property that for any $p \in P$, the distance to *any* q-flat in the span of *any* subset of P of cardinality C is preserved (Lemma 3). Note that the number of such subspaces is bounded by n^C and therefore polynomial in n.

Lemma 3 in turn follows easily from a generalization of the Johnson-Lindenstrauss lemma that we prove first: for an integer $c > 0$, we show that a random projection to roughly $c\log(n)/\varepsilon^2$ dimensions preserves for *all* subset S of c points the distance between any two points in $\langle S\rangle$. Sarlos [28] proves this statement for the span of a fixed subset S (in fact, he proves it for any c-dimensional subspace, even if it is not spanned by points in P). A direct combination of his result with the union bound gives a weaker bound of $O(c^2\log n/\varepsilon^2)$ for the dimension of the projected space; instead, we have to exploit the sharper concentration of the expected length of a vector to obtain a better bound.

Lemma 2 *For $0 < \varepsilon < 1$, a set $P \subset \mathbb{R}^d$ of n points, an integer $c \geq 0$, and $m \geq \lambda \cdot c\log(n)/\varepsilon^2$ for a suitable constant λ, a random projection $\hat\pi$ satisfies with high probability that for any subset $S \subset P$ of cardinality c and for any $u, v \in \langle S\rangle$*

$$(1-\varepsilon)\|u-v\| \leq \|\hat\pi(u)-\hat\pi(v)\| \leq (1+\varepsilon)\|u-v\|.$$

Proof: The proof of Thm. 2.1 in Dasgupta and Gupta [16] implies the following statement: When projecting a unit vector in $\mathbb{R}^d$ to a fixed $m = O(c\log n/\varepsilon^2)$-dimensional subspace, the probability that its squared length does not lie in $((1-\varepsilon)m/d, (1+\varepsilon)m/d)$ is at most

$$2\exp(-\frac{m\varepsilon^2}{4}) \leq 2\exp(-\frac{\lambda c\log n}{4}) \leq n^{-8c}$$

for a suitable constant λ. As they argue, the same bound applies for a fixed unit vector and a uniformly chosen m-dimensional subspace.

A result by Feige and Ofek [19] (see also [6]), translated in geometric terms says that by approximately preserving the pairwise squared distances between a set of at most $\exp(c \ln 18)$ sample points belonging to an c-dimensional subspace, we can approximately preserve the squared length of all unit vectors in the subspace, and thus all pairwise distances; see [28, Proof of Cor. 11] for further explanations. Hence, for a fixed subspace, we need to preserve $\exp(2c \ln 18) \leq \exp(6c)$ distances. Moreover, we want to preserve distances in n^c many subspaces, yielding a total of $\exp(6c)n^c \leq n^{7c}$ distances to be preserved. By the union bound, choosing a m-dimensional subspace uniformly at random, the probability of success is at least $1 - \frac{n^{7c}}{n^{8c}} \geq 1 - 1/n^c$. □

The preservation of point-to-flat distances in low-dimensional subspaces is a simple consequence:

Lemma 3 *Let $0 < \varepsilon < 1$, $P \subset \mathbb{R}^d$ a set of n points and $q < c$ positive integers. With high probability, a random projection to an $O(c \log n/\varepsilon^2)$-dimensional flat satisfies for all subset $S \subset P$ of cardinality c, all q-flat $Q \subset \langle S \rangle$, and all $p \in P$ that*

$$(1-\varepsilon)d(p,Q) \leq d(\hat{\pi}(p), \hat{\pi}(Q)) \leq (1+\varepsilon)d(p,Q).$$

Proof: For any $p \in P$ and any $Q \subset \langle S \rangle$, there exists a space with $c+1$ points that contains both p and Q. Let $t \in Q$ be the point such that $d(p,Q) = \|p - t\|$. Applying Theorem 2 for $c' := c+1$ immediately implies that $d(\hat{\pi}(p), \hat{\pi}(Q)) \leq \|\hat{\pi}(p) - \hat{\pi}(t)\| \leq (1+\varepsilon)d(p,Q)$. The second inequality follows similarly, considering the point $t' \in Q$ that realizes $d(\hat{\pi}(p), \hat{\pi}(Q))$. □

Finally, we show our main theorem that random projections preserve $f_1^q(S,\rho)$ for subset S of P.

Theorem 4 *Let $0 < \varepsilon < 1$, $P \subset \mathbb{R}^d$ consist of n points, $q \geq 0$ an integer and ρ a constant in $\in \mathbb{Z}_{\geq 0} \cup \{\infty\}$. Then with high probability, for $m \geq \lambda \cdot C_\rho(q, \varepsilon/2) \log(n)/\varepsilon^2$ with a suitable constant λ, a random projection $\hat{\pi}$ satisfies for all subset $S \subset P$*

$$(1-\varepsilon)f_1^q(S,\rho) \leq f_1^q(\hat{\pi}(S),\rho) \leq (1+\varepsilon)f_1^q(S,\rho).$$

Proof: Let $S \subseteq P$ arbitrary. We start by showing the second inequality: By the coreset property, there exists a subset $E \subset S$ of $C_\rho(q, \varepsilon/2)$ points such that $\langle E \rangle$ contains a q-flat F that is an $\frac{\varepsilon}{2}$-approximate solution. For $\rho \neq \infty$, applying Lemma 3 with $\varepsilon' = \varepsilon/(4\rho)$ and $c := C_\rho(q, \varepsilon/2)$, we get that

$$f_1^q(\hat{\pi}(S),\rho) \leq \sum_{p \in S} d(\hat{\pi}(p), \hat{\pi}(F))^\rho \leq \sum_{p \in S} (1+\varepsilon/(4\rho))^\rho d(p,F)^\rho \leq (1+\varepsilon/3)(1+\varepsilon/2) f_1^q(S,\rho) \leq (1+\varepsilon) f_1^q(S,\rho),$$

where we use the easy facts that $(1+\varepsilon/(4\rho))^\rho \leq (1+\varepsilon/3)$ and $(1+\varepsilon/3)(1+\varepsilon/2) < 1+\varepsilon$ for $0 \leq \varepsilon \leq 1$. For $\rho = \infty$, the proof for $\rho = 1$ directly carries over.

For the first inequality, we apply the coreset property on the set $\hat{\pi}(S)$: let $\hat{\pi}(E')$ be a coreset for $\hat{\pi}(S)$. Let G denote the approximate solution in $\langle \hat{\pi}(E') \rangle$; it holds that $G = \hat{\pi}(F')$ for some q-flat F' in $\langle E' \rangle$. Using again Lemma 3, we have that

$$(1-\varepsilon)f_1^q(S,\rho) \leq (1-\frac{\varepsilon}{2})(1-\frac{\varepsilon}{3}) \sum_{p \in S} d(p,F')^\rho \leq (1-\frac{\varepsilon}{2}) \sum_{p \in S} ((1-\varepsilon/(4\rho))d(p,F'))^\rho$$

$$\leq (1-\frac{\varepsilon}{2}) \sum_{p \in S} d(\hat{\pi}(p), G)^\rho \leq (1-\frac{\varepsilon}{2})(1+\frac{\varepsilon}{2}) f_1^q(\hat{\pi}(S),\rho) \leq f_1^q(\hat{\pi}(S),\rho)$$

where we use the easily provable fact that $1 - \frac{\varepsilon}{3} \leq (1 - \frac{\varepsilon}{(4\rho)})^\rho$. Again, the case $\rho = \infty$ is analogue to $\rho = 1$. □

As a consequence of Theorem 4, also $f_k^q(P,\rho)$ is preserved for any $k \geq 1$.

Corollary 5 *With the notations of Theorem 4 and $k \geq 1$, a random projection $\hat{\pi}$ satisfies with high probability*

$$(1-\varepsilon)f_k^q(P,\rho) \leq f_k^q(\hat{\pi}(P),\rho) \leq (1+\varepsilon)f_k^q(P,\rho).$$

Proof: Let $\mathcal{F} = \{F_1, \ldots, F_k\}$ denote an optimal collection of q-flats, that is, for any $p \in P$, the closest flat in $\mathcal{F}$ has distance at most $f_k^q(P,\rho)$. Let $P_i \subseteq P$ be the set of points closest to F_i, for $i = 1, \ldots, k$. Note that F_i is the optimal q-flat for P_i, in other words, it realizes $f_1^q(P_i,\rho)$.[3] Using Theorem 4 on the subsets P_i, we get for $\rho < \infty$ that

$$f_k^q(\hat{\pi}(P),\rho) \leq \sum_{i=0}^{k} f_1^q(\hat{\pi}(P_i),\rho) \leq \sum_{i=0}^{k}(1+\varepsilon)f_1^q(P_i,\rho) = (1+\varepsilon)f_k^q(P,\rho),$$

proving the second part of the inequality. The first part follows the same way considering an optimal $\mathcal{F}$ for $\hat{\pi}(P)$. The case $\rho = \infty$ is analogous, replacing all sums by max. □

3 Improved Coreset for Projective Clustering.

In this section, we will show an improved bound for $C_\infty(q,\varepsilon)$ with $q \geq 1$, that is, for the size of a coreset for MEC and its higher-dimensional counterparts. Recall from Theorem 4 that the size of these coresets defines the dimension of the projected space required to preserve the corresponding structure. That implies that knowing good bounds for $C_\infty(q,\varepsilon)$ is useful, even if no algorithm for computing a coreset of that size is available. We free ourselves from algorithmic considerations in this section and prove the following structural result:

Theorem 6 *For any point set $P \subset \mathbb{R}^d$ and $q \geq 1$, there is a set $S \subset P$ of $O(q^3/\varepsilon^2)$ points such that the affine subspace spanned by S contains an ε-approximate q-flat.*

The main difficulty of the proof is the case of lines, namely $q = 1$, and the majority of the remaining section will deal with this case. Therefore, we simplify our notations for the case of lines: Let the distance between a line ℓ and a point p, denoted by $\mathtt{d}(p,\ell)$, be the distance of p to its closest point on ℓ. For a point set $P \subset \mathbb{R}^d$, let $\ell_{\rm opt}$ be the line that minimizes the maximum distance to any point in P. We refer to $\ell_{\rm opt}$ as the *minimum enclosing cylinder, or just MEC*. For any line ℓ, the maximum distance from ℓ to any point in P is referred to as the *radius* of ℓ. Let $r_{\rm opt}$ be the radius of $\ell_{\rm opt}$. We call a line ℓ an *ε-approximate cylinder*, if its maximum distance to any point in P is at most $(1+\varepsilon)r_{\rm opt}$. We re-state Theorem 6 for the case $q = 1$:

Theorem 7 *For any point set $P \subset \mathbb{R}^d$, there is a coreset $S \subseteq P$ of $O(1/\varepsilon^2)$ points such that the affine subspace spanned by S contains an ε-approximate 1-cylinder.*

Notations and preliminary observations. For $p \in P$ and any line ℓ, let $\pi_\ell(p)$ denote the (orthogonal) projection of p onto ℓ. We let $I_\ell(P)$ denote the smallest interval on ℓ which contains the projection of all points, i.e., the set $\{\pi_\ell(p) \mid p \in P\}$. Let $w_\ell(P)$ denote the length of this interval; we refer to $w_\ell(P)$ as *width* of P along the line ℓ. When there is no ambiguity about the point set P, we just write I_ℓ, and w_ℓ. Let $w := w_{\ell_{\rm opt}}$ and $I := I_{\ell_{\rm opt}}$. For any point set P, the diameter of P, $\mathrm{diam}(P)$, denotes the length of the farthest pair of points in P. Clearly, $w_\ell \leq \mathrm{diam}(P)$ for any line. Moreover, triangle inequality implies that $\mathrm{diam}(P) \leq w + 2r_{\rm opt}$ by projecting a diametral pair of points to $\ell_{\rm opt}$. Therefore,

$$w_\ell \leq w + 2r_{\rm opt}. \tag{1}$$

for any line ℓ. The distance of a line ℓ from $\ell_{\rm opt}$ is defined as

$$\mathrm{pdist}(\ell) := \max_{p \in I} \mathtt{d}(p,\ell),$$

[3] For $\rho = \infty$, this is not necessarily true for any optimal solution, but we can replace every q-flat with the optimal one wlog

Note that pdist is defined only with respect to $I \subset \ell_{\mathrm{opt}}$. The following statement follows directly from the definition and triangle inequality:

Proposition 8 *For any line ℓ with $\mathrm{pdist}(\ell) \leq \varepsilon r_{\mathrm{opt}}$, ℓ is an ε-approximate cylinder.*

Moreover, the following simple fact will be used several times in the proofs of this section:

Proposition 9 *For a line segment e and a point $c \in \mathbb{R}^d$, the function $\mathtt{d} : e \to \mathbb{R}, x \mapsto \|x - c\|$ has its maximum at one endpoint of e. Moreover, any point in the interior of e has a function value that is strictly smaller than the maximum.*

Proof Outline. We proof Theorem 7 for the cases $w \leq 6r_{\mathrm{opt}}$ and for $w > 6r_{\mathrm{opt}}$ separately[4]:

When $w > 6r_{\mathrm{opt}}$, we define an iterative procedure to construct the coreset. During each iteration, we maintain not just the subspace, but also a candidate line. After each iteration of the procedure, we show that either the angle between the candidate line and ℓ_{opt} or the distance of the candidate line from I improves significantly. Finally, we show that both events can only happen at most $O(1/\varepsilon^2)$ times. Therefore, after $O(1/\varepsilon^2)$ iterations, the candidate line is within $\varepsilon r_{\mathrm{opt}}$ from I implying that it is an ε-approximate cylinder. This construction procedure is similar to the procedure of [23]. However, we prove a faster convergence rate leading to a small sized coreset.

When $w \leq 6r_{\mathrm{opt}}$, the problem behaves somewhat like the MEB problem. We exploit the fact that there are small MEB-coresets of size $O(1/\varepsilon)$ and derive a MEC-coreset of size $O(1/\varepsilon^2)$ from it. Our construction reveals a high-level analogy between MEB- and MEC-approximations: while the center of an MEB-coreset, as defined by Bădiou and Clarkson [10] can be at a (fairly large) distance of $\sqrt{\varepsilon}\hat{r}$ from the optimal center[5], that approximate center is still within a distance $(1+\varepsilon)\hat{r}$ to all points in P. The same way, the line that we construct does not necessarily form a small angle with the optimal cylinder, but we ensure that the distance of every point in P to this line is at most $(1 + \varepsilon)r_{\mathrm{opt}}$.

Coresets for small w. We assume that $w \leq 6r_{\mathrm{opt}}$. The following statement follows from a lemma by Barany and Füredi [9, p.323]; see also [22, Remark 2.9]: There is a coreset $E_1 \subseteq P$ of $O(1/\varepsilon^2)$ points such that the affine subspace spanned by E_1 contains a point o which is at a distance $(\varepsilon/3)r_{\mathrm{opt}}$ from ℓ_{opt}, i.e.,

$$\mathtt{d}(o, \ell_{\mathrm{opt}}) = (\varepsilon/3)r_{\mathrm{opt}}.$$

We start our construction by assuming that E_1 and o are given to us, and we make o the origin of our coordinate system. We will construct an ε-approximate cylinder through o.

Let $P' \subseteq P$ be the points in P with distance at least $(1 + \varepsilon)r_{\mathrm{opt}}$ from the origin. We find a solution for P' instead of P, because for points in $P \setminus P'$, any line through the origin will be in distance at most $(1 + \varepsilon)r_{\mathrm{opt}}$. Let ℓ'_{opt} be the line through the origin that yields the cylinder with the smallest radius for P'. Note that this cylinder has a radius of at most $(1 + \varepsilon/3)r_{\mathrm{opt}}$. The hyperplane through the origin and orthogonal to ℓ'_{opt} splits $\mathbb{R}^d$ in two halfspaces H^- and H^+. No point of P' can lie on the hyperplane because then, its distance to the origin would be at most $(1 + \varepsilon/3)r_{\mathrm{opt}}$. Next, we construct a point set $Q \subset H^-$ by reflecting points in H^+:

$$Q := \{p \mid p \in P' \cap H^-\} \cup \{-p \mid p \in P' \cap H^+\}$$

Notice that ℓ'_{opt} is the line through the origin that minimizes the maximum distance to all points in Q.

Let h_{high} and h_{low} be hyperplanes orthogonal to ℓ'_{opt} through the endpoints of $I_{\ell'_{\mathrm{opt}}}(Q)$, with h_{low} being the hyperplane with larger distance to the origin. In other words, h_{high} and h_{low} are the hyperplanes with distance equal to the width of Q along ℓ'_{opt}.

[4] We emphasize once more that we need to show only existence of small-sized coresets, so we assume that ℓ_{opt} is given to us.

[5] here $\hat{r}$ is the radius of the optimal minimum enclosing ball.

Let $Q' := \pi(Q, h_{\text{low}})$. The line ℓ'_{opt} is orthogonal to h_{low} and therefore, its projection to h_{low} is a point which we denote by c. In Lemma 11, we show that c is the center of the minimum enclosing ball of Q'. By [10], Q' has an $\frac{\varepsilon^2}{3600}$-coreset for MEB of size $\lceil 3600/\varepsilon^2 \rceil$.[6] Let $E_2 \subset Q$ denote the set of points whose projections define this MEB-coreset, and let c' denote the center of the projected points. Let ℓ' be the line parallel to ℓ'_{opt} through c'. Let o' be the point of intersection of ℓ' with the convex hull of E_2.

The line ℓ through o and o' lies in the affine subspace spanned by $E_1 \cup E_2$. We claim that ℓ is an ε-approximate cylinder. This completes the coreset construction for ε-approximate cylinder for small widths.

Using the facts that all points in Q are at a distance greater than $(1+\varepsilon)r_{\text{opt}}$ from the origin, we get the following properties for the hyperplanes h_{low} and h_{high}.

Lemma 10 *$\mathtt{d}(o, h_{\text{high}}) \geq \sqrt{\varepsilon} r_{\text{opt}}$ and $\mathtt{d}(o, h_{\text{low}}) \leq 8 r_{\text{opt}}$.*

Proof: For the first part, consider a point $q \in Q$ and let q' be its projection on ℓ'_{opt}. Because the triangle oqq' has a right angle at q', we have that

$$\|q'\|^2 = \|q\|^2 - \|q - q'\|^2.$$

Note that $\|q\| \geq (1+\varepsilon)r_{\text{opt}}$ because $q \in Q$, and $\|q - q'\| \leq 1 + \frac{\varepsilon}{3}$ because of the definition of ℓ_{opt}. Therefore,

$$\|q'\|^2 \geq (1+\varepsilon)^2 r_{\text{opt}} - (1 + \frac{\varepsilon}{3})^2 r_{\text{opt}},$$

which leads to $\|q'\| \geq (\sqrt{\frac{4}{3}\varepsilon + \frac{8}{9}\varepsilon^2}) r_{\text{opt}} \geq \sqrt{\varepsilon} r_{\text{opt}}$.

For the second part, recall that the width of ℓ'_{opt} is at most $w + 2r_{\text{opt}}$ by (1). By construction, the intersection point of ℓ'_{opt} and h_{low} is in $I(\ell'_{\text{opt}})$. But the origin is in $I(\ell'_{\text{opt}})$ as well, because it was constructed as a point in the convex hull of P. Therefore, the distance of these two points is at most $w + 2r_{\text{opt}} \leq 8 r_{\text{opt}}$. □

The following lemma establishes that the intersection point of ℓ'_{opt} with h_{low} is the center of MEB of Q'.

Lemma 11 *c is the center of the MEB of Q'.*

Proof: Assume for a contradiction that Q' has an MEB center $\tilde{c} \neq c$. Let $\tilde{\ell}$ denote the line through the origin and $\tilde{c}$. We will show that $\tilde{\ell}$ yields a smaller cylinder than ℓ'_{opt} which is a contradiction to the definition of ℓ'_{opt}.

Let r denote the distance of ℓ'_{opt} from the farthest point in Q. Let $\tilde{r}$ be the radius of the smallest enclosing ball of Q' that is centered at c. Clearly, $\tilde{r} < r$. Fix a point $q \in Q$, let $\tilde{q}$ be its projection to $\tilde{\ell}$ and q' its projection to h_{low}. The distance $d_q := \|q - \tilde{q}\|$ is equal to the distance of q' to some point on the interior of the line segment $c\tilde{c}$. By Proposition 9, the distance of q to $\tilde{\ell}$ is therefore strictly less than $\max\{r, \tilde{r}\} = r$, so the cylinder around $\tilde{\ell}$ with radius $\tilde{r}$ is a smaller cylinder than the one defined by ℓ'_{opt}. □

Lemma 12 *$\|c - c'\| \leq \frac{\varepsilon}{30} r_{\text{opt}}$.*

Proof: Consider the hyperplane through c orthogonal to cc'. There must be a point $q \in Q'$ in the half-space that does not contain c' whose distance to c is the meb radius r (otherwise, we could move the center towards c', decreasing the radius; see [8] for a proof). Let θ be the angle at c in the triangle $cc'q$. By law of cosine,

$$\|c - c'\|^2 = \|q - c'\|^2 - \|q - c\|^2 + 2\|q - c'\|\|q - c\| \cos\theta.$$

Since $\theta \in [\pi/2, \pi]$, $\cos\theta < 0$, so we can remove the last term for an upper bound. Also, by the coreset property, $\|c' - q\| \leq (1 + \frac{\varepsilon^2}{3600}) r$, and we know that $r \leq (1 + \frac{\varepsilon}{3}) r_{\text{opt}} \leq \frac{4}{3} r_{\text{opt}}$ because it equals the radius of the cylinder of ℓ'_{opt}. Therefore,

$$\|c - c'\|^2 \leq (1 + \frac{\varepsilon^2}{3600})^2 r^2 - r^2 = \frac{16}{9}(\frac{\varepsilon^2}{1800} + \frac{\varepsilon^4}{3600^2}) r_{\text{opt}}^2 \leq \frac{\varepsilon^2}{900} r_{\text{opt}}^2,$$

[6] The constant seems unnecessary high, but we have not tried to optimize it in this work.

which proves our claim. □

We can now prove correctness of our construction for the case of small w:

Lemma 13 *Any point in Q has distance at most $(1+\varepsilon)r_{\mathrm{opt}}$ from ℓ.*

Proof: Fix any point $q \in Q$ and let q' be its projection to h_{low}. Let c'' be the point of intersection of ℓ and h_{low}. It is sufficient to show that $\|q' - c''\| \leq (1+\varepsilon)r_{\mathrm{opt}}$: Since $\|q' - c\| \leq (1+\varepsilon/3)r_{\mathrm{opt}}$, and the distance of q to ℓ is realized by the distance of q' to some point in the line segment from c to c''. If the distance of q' to c'' is at most $(1+\varepsilon)r_{\mathrm{opt}}$, the result follows from Proposition 9.

We claim first that $\|c - c''\| \leq \frac{\sqrt{\varepsilon}}{3} r_{\mathrm{opt}}$: Indeed, the slope of ℓ with respect to ℓ'_{opt} is at most $\frac{\sqrt{\varepsilon}}{30}$ because the distance of o' to ℓ'_{opt} equals the distance of c' to c which is at most $\frac{\varepsilon}{30} r_{\mathrm{opt}}$ by Lemma 12 and the distance of c to the projection of o' to ℓ'_{opt} is at least $\sqrt{\varepsilon} r_{\mathrm{opt}}$ by Lemma 10. The distance of the origin to h_{low} is at most $8r_{\mathrm{opt}}$, again by Lemma 10; it follows that the distance of the lines on h_{low} is at most $\frac{8\sqrt{\varepsilon}}{30} r_{\mathrm{opt}} \leq \frac{\sqrt{\varepsilon}}{3} r_{\mathrm{opt}}$. Observe that c, c', and c'' are colinear by construction. We first consider the triangle $q'cc'$, and let θ denote the angle at c. If $\theta \leq \pi/2$, we have that

$$\|q' - c''\|^2 \leq \|q' - c\|^2 + \|c - c''\|^2 \leq (1 + \frac{\varepsilon}{3})^2 r_{\mathrm{opt}}^2 + \frac{\varepsilon}{9} r_{\mathrm{opt}}^2 \leq (1+\varepsilon) r_{\mathrm{opt}}^2$$

and therefore, $\|q' - c''\| \leq (1+\varepsilon)r_{\mathrm{opt}}$ as well. If $\theta \geq \pi/2$, we use the law of cosines to obtain

$$\cos\theta = \frac{\|q' - c'\|^2 - \|q' - c\|^2 - \|c - c'\|^2}{2\|q' - c\|^2 \|c - c'\|}$$

Using the law of cosines again on the triangle $q'cc''$ and replacing $\cos\theta$ by the expression above, we get

$$\begin{aligned} \|q' - c''\|^2 &= \|q' - c\|^2 + \|c - c''\|^2 - \frac{\|c - c''\|}{\|c - c'\|}(\|q' - c'\|^2 - \|q' - c\|^2 - \|c - c'\|^2) \\ &= \|q' - c\|^2 + \|c - c''\|^2 + \frac{\|c - c''\|}{\|c - c'\|}(\|c - c'\|^2 - (\|q' - c'\|^2 - \|q' - c\|^2)) \end{aligned}$$

Because $\theta > \pi/2$, we know that $\|q' - c'\| > \|q' - c\|$, so we can bound

$$\|q' - c''\|^2 \leq \|q' - c\|^2 + \|c - c''\|^2 + \|c - c''\|\|c - c'\| \leq (1 + \frac{\varepsilon}{3})^2 r_{\mathrm{opt}}^2 + \frac{\varepsilon}{9} r_{\mathrm{opt}}^2 + \frac{\varepsilon^{3/2}}{30} r_{\mathrm{opt}}^2 \leq (1+\varepsilon) r_{\mathrm{opt}}^2.$$

□

The case of large w Next, we construct a coreset for the case where $w > 6r_{\mathrm{opt}}$. Given ℓ_{opt}, we give an iterative process to construct a coreset of size $O(1/\varepsilon^2)$ such that the affine subspace spanned by this coreset contains an approximate solution.

Construction. Let $u, v \in P$ be the two points whose projections bound the interval I, that is, they are the extremal points of P with respect to the direction of ℓ_{opt}. We initialize the candidate line $\ell^{(0)}$ to be the line uv. For $i \geq 1$, while $\ell^{(i-1)}$ has distance more than $(1+\varepsilon)r_{\mathrm{opt}}$ from some point in P, we update the candidate line in the i-th step to $\ell^{(i)}$ as follows: Let q be a point of P that is furthest away from $\ell^{(i-1)}$, and let E be the plane spanned by $\ell^{(i-1)}$ and q. We set the projection of ℓ_{opt} to E as our new candidate line $\ell^{(i)}$. After at most $O(1/\varepsilon^2)$ steps, we have our coreset.

Proof Overview: In the ith iteration, let θ_i be the angle between $\ell^{(i)}$ and $\ell^{(i-1)}$ and let ϕ_i be the angle between $\ell^{(i)}$ and $\ell_{\rm opt}$.

We call an iteration of our construction to be a *small angle iteration* if the angle θ_i satisfies $\sin\theta_i \leq \frac{\varepsilon r_{\rm opt}}{4w}$. We refer to all other steps as *large angle iteration*.

The proof proceeds by proving the following statements: For the initial line $\ell^{(0)}$, both its angle and its distance to $\ell_{\rm opt}$ are fairly small, but not small enough to be an approximate cylinder (Lemma 14 and Lemma 18). As i increases, the angle of $\ell^{(i)}$ with $\ell_{\rm opt}$ and distance (pdist) to $\ell_{\rm opt}$ is always non-increasing. (Corollary 16 and Lemma 19). In each large angle iterations, the angle to $\ell_{\rm opt}$ reduces significantly (Corollary 16), so that after $O(1/\varepsilon^2)$ large angle iterations, the angle to $\ell_{\rm opt}$ is so small that no further large angle iteration can happen anymore (Lemma 17). For small angle iterations, the distance (pdist) to $\ell_{\rm opt}$ reduces significantly (Lemma 22), so that after $O(1/\varepsilon^2)$ small angle iterations, the distance drops below $\varepsilon r_{\rm opt}$, and we have constructed an ε-approximate cylinder (Theorem 23).

We continue with filling in the details of our construction. We begin by proving that the initial angle is not too large.

Lemma 14 $\tan\phi_0 \leq \frac{2r_{\rm opt}}{w}$.

Proof: Recall that $\ell^{(0)}$ contains points u and v whose projections onto $\ell_{\rm opt}$ define the interval I. Consider the line $\ell'_{\rm opt}$ parallel to $\ell_{\rm opt}$ that passes through u. Because $\ell_{\rm opt}$ with radius $r_{\rm opt}$ is the optimal cylinder, the distance of $\ell_{\rm opt}$ to $\ell'_{\rm opt}$ is at most $r_{\rm opt}$. The distance of $\ell_{\rm opt}$ to v is also at most $r_{\rm opt}$. By triangle inequality, the distance of $\ell'_{\rm opt}$ and v is at most $2r_{\rm opt}$. Let v' be the projection of v onto $\ell'_{\rm opt}$. The line segment vv' is orthogonal to $\ell'_{\rm opt}$. The result follows from considering the right angled triangle uvv', with $\|uv\| = w$, $vv' = 2r_{\rm opt}$ and has angle ϕ_0 at u. □

This following lemma quantifies the amount of improvement in ϕ_i when θ_i is large. This helps us bound the number of large angle steps in Corollary 16.

Lemma 15 *Let ℓ be a line not orthogonal to a plane E. Let ℓ' denote its projection and let ϕ' denote the angle between ℓ and ℓ'. Let ℓ'' be a line in E obtained by rotating ℓ' by an angle of θ. Let ϕ'' denote the angle of ℓ and ℓ''. Then,*

$$\cos\phi'' = \cos\phi'\cos\theta$$

Proof: Wlog, we can assume that all three lines intersect in a common point o. Let a be a point on ℓ of distance 1 from o, let $b \in \ell'$ be its projection on ℓ'. Within E, the line through b orthogonal to ℓ' intersects ℓ'' in a point which we denote by c. We can determine the side lengths of all edges of the tetrahedron $oabc$: The length of oa is 1. Since oab is orthogonal with right angle at b and angle ϕ' at o, we have that $|ob| = \cos\phi'$ and $|ab| = \sin\phi'$. Since the triangle obc is orthogonal at b and has angle θ at o, we know that $|bc| = \tan\theta\cos\phi'$ and $|oc| = \cos\phi'/\cos\theta$. Finally, the triangle abc is orthogonal at b and therefore $|ac|^2 = |bc|^2 + |ab|^2$.

Now, we consider the triangle oac (without right angle) which has angle ϕ'' at o. By the law of cosines

$$|ac|^2 = |oa|^2 + |oc|^2 - 2|oa||oc|\cos\phi'',$$

and plugging in everything yields

$$\begin{aligned}
\cos\phi'' &= \frac{1+\frac{\cos^2\phi'}{\cos^2\theta}-\tan^2\theta\cos^2\phi'-\sin^2\phi'}{2\frac{\cos\phi'}{\cos\theta}} \\
&= \frac{\cos^2\phi'+\frac{\cos^2\phi'}{\cos^2\theta}-\tan^2\theta\cos^2\phi'}{2\frac{\cos\phi'}{\cos\theta}} \\
&= \frac{\cos\phi'}{2}\left(\cos\theta+\frac{1}{\cos\theta}-\frac{\sin^2\theta}{\cos\theta}\right) \\
&= \frac{\cos\phi'}{2\cos\theta}\left(\cos^2\theta+1-\sin^2\theta\right) \\
&= \cos\phi'\cos\theta
\end{aligned}$$

□

Corollary 16 *In our algorithm, the sequence ϕ_i for $i \geq 0$ is non-increasing. Moreover, if the i-th step is a large angle update, we have that*

$$\cos\phi_i > \sqrt{\frac{1}{1-\left(\frac{\varepsilon r_{\mathrm{opt}}}{4w}\right)^2}}\cos\phi_{i-1}.$$

Proof: Using Lemma 15, it holds in each step of the algorithm that

$$\cos\phi_i = \frac{\cos\phi_{i-1}}{\cos\theta_i}.$$

Since $\cos(\theta_i) \leq 1$, this shows that the $(\cos\phi_i)_{i\geq 0}$ is increasing, so the angles are decreasing. Moreover, if the i-th step is an large angle update, we have by definition that

$$\sin\theta_i > \frac{\varepsilon r_{\mathrm{opt}}}{4w}$$

and this implies the second claim exploiting that $\sin\theta_i = \sqrt{1-\cos^2\theta_i}$. □

The previous lemma shows that if the angle between two consecutive candidates is large, we are making substantial improvement in the angle with respect to the optimal line ℓ_{opt}. A simple calculation shows that because we start with a not too large angle (Lemma 14), the angle becomes very small after $O(1/\varepsilon^2)$ steps.

Lemma 17 *After at most $(64/\varepsilon^2+2)$ large angle updates, the angle ϕ between a candidate ℓ and ℓ_{opt} satisfies*

$$\sin\phi \leq \frac{\varepsilon r_{\mathrm{opt}}}{4w}.$$

In particular, all further iterations are small angle updates.

Proof: Consider the non-increasing sequence (ϕ_i). We know that $\phi_0 \leq \arctan(2r_{\mathrm{opt}}/w)$. Let ϕ denote an element of the sequence after k large angle updates. By Corollary 16, we know that

$$\cos\phi > \left(\sqrt{\frac{1}{1-\left(\frac{\varepsilon r_{\mathrm{opt}}}{4w}\right)^2}}\right)^k \cos\phi_0$$

The condition that $\sin\phi \leq \frac{\varepsilon r_{\mathrm{opt}}}{4w}$ is equivalent to $\cos\phi \geq \sqrt{\frac{1}{1-\left(\frac{\varepsilon r_{\mathrm{opt}}}{4w}\right)^2}}$. Now, since $\cos\arctan(x) = \sqrt{\frac{1}{1+x^2}}$, if k satisfies

$$\sqrt{\frac{1}{1-\left(\frac{\varepsilon r_{\mathrm{opt}}}{4w}\right)^2}} < \left(\sqrt{\frac{1}{1-\left(\frac{\varepsilon r_{\mathrm{opt}}}{4w}\right)^2}}\right)^k \sqrt{\frac{1}{1+\frac{4r_{\mathrm{opt}}^2}{w^2}}},$$

we know that after k iterations, $\sin\phi \leq \frac{\varepsilon r_{\mathrm{opt}}}{4w}$. Solving for k yields that the angle will be sufficiently small after

$$k_0 := \frac{\log 1 + \left(\frac{2r_{\mathrm{opt}}}{w}\right)^2}{\log \frac{1}{1-\left(\frac{\varepsilon r_{\mathrm{opt}}}{4w}\right)^2}} + 2$$

iterations. Using $\log(1+x) \leq x$ for all $x \geq 0$ and $\log\frac{1}{1-x} \geq x$ for all $0 < x < 1$, it follows that

$$k_0 = \frac{\log 1 + \left(\frac{2r_{\mathrm{opt}}}{w}\right)^2}{\log \frac{1}{1-\left(\frac{\varepsilon r_{\mathrm{opt}}}{4w}\right)^2}} + 2 \leq \frac{\frac{4r_{\mathrm{opt}}^2}{w^2}}{\frac{\varepsilon^2 r_{\mathrm{opt}}^2}{16w^2}} + 2 = \frac{64}{\varepsilon^2} + 2.$$

For the second statement, note that Lemma 15 implies that

$$\cos\theta_{i+1} = \frac{\cos\phi_i}{\cos\phi_{i+1}} \geq \cos\phi_i.$$

It follows that θ_{i+1} is a small angle, and the same argument applies for any θ_j with $j \geq i+1$. □

After these angle considerations, we now turn to the distance to ℓ_{opt} with respect to pdist. For convenience, we write $d_i := \mathrm{pdist}(\ell^{(i)})$ for $i \geq 0$.

Lemma 18 $d_0 \leq r_{\mathrm{opt}}$

Proof: Consider the function $f : I \to \mathbb{R}, p \mapsto \mathtt{d}(p, \ell^{(0)})$. By construction, f takes value at most r_{opt} at the boundary of I because $\ell^{(0)}$ goes through the points u and v whose projections on ℓ_{opt} defines I (by construction). By Proposition 9, the function f is smaller than r_{opt} throughout the interval I. □

Lemma 19 *For any $i \geq 0$, $d_{i+1} \leq d_i$.*

Proof: Fix any $p \in I$. Let e denote the plane that $\ell^{(i)}$ and $\ell^{(i+1)}$ are contained in. By construction, the closest point on e to p is on $\ell^{(i+1)}$. It follows that $\mathtt{d}(p, \ell^{(i)}) \geq \mathtt{d}(p, \ell^{(i+1)})$ and therefore, $d_i \geq d_{i+1}$. □

The last step is to show that, in small angle iterations, d_i improves "significantly". In order to do that, we need a preparatory result: Fix a line ℓ. For any $p \in I$, there is a closest point $g(p)$ on ℓ. Define the *interesting region* of ℓ as the convex hull of I_ℓ and all $g(p)$ with $p \in I$.

Lemma 20 *The interesting region of any $\ell^{(i)}$ has length at most $2w$.*

Proof: Set $\ell := \ell^{(i)}$. Now, fix $p \in P$, let $\tilde{p}$ denote its projection to ℓ_{opt} and p^* its projection to ℓ. Moreover, let p' be the point on ℓ with minimum distance to $\tilde{p}$. Note that by triangle inequality, $\|p-p'\| \leq \|p-\tilde{p}\|+\|\tilde{p}-p'\| \leq r_{\mathrm{opt}} + r_{\mathrm{opt}} = 2r_{\mathrm{opt}}$. Since the triangle pp^*p' has a right angle at p^*, it follows that the length of $\|p^* - p'\|$ is also at most $2r_{\mathrm{opt}}$. Therefore, the length of the interesting region is bounded by $w(I) + 4r_{\mathrm{opt}}$ which is bounded $w + 6r_{\mathrm{opt}}$ by (1). Since $w \geq 6r_{\mathrm{opt}}$ is assumed, the statement follows. □

Next, we show that if $\ell^{(i)}$ and $\ell^{(i+1)}$ have a small angle, their interesting regions are well-separated:

Lemma 21 *Let the angle of $\ell^{(i)}$ and $\ell^{(i+1)}$ be small. Let p be a point in the interesting region of $\ell^{(i+1)}$. Then, $\mathtt{d}(p, \ell^{(i)}) \geq \frac{\varepsilon r_{\text{opt}}}{3}$.*

Proof: Since $\ell^{(i)}$ and $\ell^{(i+1)}$ lie in a common plane, we can assume wlog that $\ell^{(i+1)}$ is aligned with the x-axis and the distance to $\ell^{(i)}$ is given by the function $f : x \mapsto mx + n$ induced by $\ell^{(i)}$ where the slope m is given by $\tan\theta$ with θ the angle of the two lines.

By construction of $\ell^{(i+1)}$, there exists a point $q \in P$ that has distance more than $(1+\varepsilon)r_{\text{opt}}$ to $\ell^{(i)}$ and distance at most r_{opt} to $\ell^{(i+1)}$. Let q_{i+1} denote the closest point to q on $\ell^{(i+1)}$. The distance of q_{i+1} to $\ell^{(i)}$ is at least $\varepsilon r_{\text{opt}}$, or equivalently, $f(q_{i+1}) \geq \varepsilon r_{\text{opt}}$. Now, consider an arbitrary point p_{i+1} in the interesting region of $\ell^{(i+1)}$; since q_{i+1} is in the interesting region, $\|p_{i+1} - q_{i+1}\| \leq 2w$. It follows that $f(p_{i+1}) \geq \varepsilon r_{\text{opt}} - \tan\theta \cdot 2w$. Now, since we know that $\theta \leq \arcsin\frac{\varepsilon r_{\text{opt}}}{4w}$, it follows easily that $\cos\theta \geq \frac{3}{4}$ for $w \geq 6r_{\text{opt}}$ and $0 \leq \varepsilon \leq 1$ (using $\cos\arcsin(x) = \sqrt{1-x^2}$). We can thus bound

$$f(p_{i+1}) \geq \varepsilon r_{\text{opt}} - \frac{\sin\theta}{\cos\theta} 2w \geq \varepsilon r_{\text{opt}} - \frac{\frac{\varepsilon r_{\text{opt}}}{4w}}{\frac{3}{4}} 2w = \frac{\varepsilon r_{\text{opt}}}{3}.$$

□

With Lemma 21, showing that the distance drops significantly is just an application of the Pythagorean theorem.

Lemma 22 *If the angle of $\ell^{(i)}$ and $\ell^{(i+1)}$ is small, $d_{i+1} \leq d_i - \frac{\varepsilon^2 r_{\text{opt}}}{36}$.*

Proof: Let $p \in I$ be the point with maximum distance to $\ell^{(i+1)}$. Let p_{i+1} be the point on $\ell^{(i+1)}$ realizing this distance, and let p_i be the point on $\ell^{(i)}$ that realizes the distance $\mathtt{d}(p, \ell^{(i)})$. Since pp_{i+1} is orthogonal to the plane containing $\ell^{(i)}$ and $\ell^{(i+1)}$, we have that

$$\|p - p_i\|^2 = \|p - p_{i+1}\|^2 + \|p_i - p_{i+1}\|^2$$

Note that $\|p - p_{i+1}\| = d_{i+1}$ by construction, and $\|p_i - p_{i+1}\| \geq \frac{\varepsilon r_{\text{opt}}}{3}$ by the Lemma 21, because p_{i+1} is in the interesting region. Therefore, we have that

$$\|p - p_i\| \geq \sqrt{d_{i+1}^2 + \frac{\varepsilon^2 r_{\text{opt}}^2}{9}}.$$

Now, because $\sqrt{x^2 + a} \geq x + \frac{a}{4}$ for any $0 \leq x \leq 1$ and $0 \leq a \leq 1$, we have that

$$d_i \geq \|p - p_i\| \geq d_{i+1} + \frac{\varepsilon^2 r_{\text{opt}}}{36}.$$

□

Theorem 23 *After $k = O(1/\varepsilon^2)$ iterations, $\ell^{(k)}$ is an ε-approximate cylinder.*

Proof: By Lemma 22, every small angle update improves d_i by an additive value of $\frac{\varepsilon^2}{36}$. Because $d_0 \leq r_{\text{opt}}$ and d_i is non-increasing, the algorithm can perform at most $\frac{36}{\varepsilon^2}$ small angle updates. As shown in Lemma 17, the number of updates with non-small angle is also bounded by $O(1/\varepsilon^2)$. □

Extension to q-flats Based on our result for cylinders, we can derive the coreset bound from Theorem 6 following the approach from [22]. For a fixed $q \geq 1$, let F_{opt} denote the optimal q-flat, that is, the q-flat that minimizes the maximal distance to points in P. As before, let r_{opt} denote this distance. We call a q-flat F an *ε-approximate q-flat* if its maximal distance to P is at most $(1+\varepsilon)r_{\text{opt}}$.

As in the cylinder case, we can find a subset of $O(1/\varepsilon^2)$ points whose span contains a point o in distance at most $(\varepsilon/3)r_{\text{opt}}$ from F_{opt} [9, 22]. We define o as the origin of the coordinate system.

Lemma 24 (Lemma 3.4 in [22]) *There is a subset $S \subset P$ of $O(q/\varepsilon^2)$ points such that the linear subspace spanned by S contains a q-flat F through o such that the maximal distance of F to P is $(1+\varepsilon)^q r_{\text{opt}}$.*

Proof: We prove the statement by induction on q, noting that the base case $q = 1$ follows from Theorem 7. For $q > 1$, we let $v_1, \ldots, v_d$ be an orthogonal basis whose first q vectors span F_{opt}. Let $\pi : \mathbb{R}^d \to \mathbb{R}^{d-k+1}$ denote the projection to $v_q, \ldots, v_d$. Note that $\pi(v_q)$ defines a line ℓ in projected space and the maximal distance to a point in P is precisely r_{opt}. Moreover, ℓ defines the optimal cylinder of the projected points because otherwise, we could find a q-flat with a smaller radius. Using Theorem 7, there is a subset of $O(1/\varepsilon^2)$ points containing a line ℓ' with maximal distance $(1+\varepsilon)r_{\text{opt}}$ to $\pi(P)$. It follows that the q-flat F' which is spanned by $v_1, \ldots, v_{d-1}$ and ℓ' has maximal distance $(1+\varepsilon)r_{\text{opt}}$ to P.

We let w_1 denote the unit vector that spans ℓ', and we consider an orthogonal basis $w_1, \ldots, w_d$ next. We define $\sigma : \mathbb{R}^d \to \mathbb{R}^{d-1}$ as the projection to $w_2, \ldots, w_d$. We use the induction hypothesis on the set $P^* := \sigma(P)$: There are $O((q-1)/\varepsilon^2)$ points such that their linear subspace contains a $(q-1)$-flat F^* through o such that its maximal distance to P is $(1+\varepsilon)^{q-1} r^*_{\text{opt}}$, where r^*_{opt} is the radius of the optimal $(q-1)$-flat for P^*. We define F as the q-flat spanned by F^* and w_1. It lies in the span of $O(q/\varepsilon^2)$ points and its distance to P equals the distance of $F^* = \sigma(F)$ to $P^* = \sigma(P)$, which is $(1+\varepsilon)^{q-1} r^*_{\text{opt}}$. We only need to bound r^*_{opt}: Note that the distance of F' to P is equal to the distance of $\sigma(F')$ to $\sigma(P)$ (because F' contains ℓ'). Therefore, r^*_{opt} cannot be more than $(1+\varepsilon)r_{\text{opt}}$ which completes the proof. □

Note that with this lemma, Theorem 6 follows directly by noting that $(1+\varepsilon/(2k))^k \leq 1+\varepsilon$.

4 Applications

Streaming algorithms for projective clustering We consider the projective clustering problem in a streaming context. In this setup, we do not return the cluster centers (the q-flats) but only an ε-approximation of $f_k^q(P, \rho)$. We let $S(n, d, q, k, \varepsilon, \rho)$ be the space complexity for this problem. We assume that n, the size of the stream, is known in advance.

Set $m := C_\rho(q, \varepsilon) \log n/\varepsilon^2$. For our streaming algorithm, we initially choose a $d \times m$ projection matrix uniformly at random which is stored throughout the algorithm. Furthermore, we maintain the workspace of a streaming algorithm that computes an approximation of the considered projective clustering problem in m dimensions. When a new point $p \in \mathbb{R}^d$ arrives, we compute the projection of the point $\hat{\pi}(p) \in \mathbb{R}^m$ and treat this as an input to the m-dimensional streaming algorithm. We return the output value of the m-dimensional streaming algorithm as our result. The correctness of the approach (with high probability) follows from Corollary 5. The space complexity is $O(dm + S(n, m, q, k, \varepsilon, \rho))$.

We consider the case of $\rho = \infty$. Using our improved coreset bound of Theorem 6 for $q \geq 1$, by simply storing all projected points in $O(m)$ dimensions, we get a space complexity of $O((d+n)m) = O((d+n)q^3 \log n/\varepsilon^4)$. Similarly, for $q = 0$, we obtain $O((d+n)\log n/\varepsilon^3)$, which is much smaller than the input size of $O(dn)$ and not too far from the lower bound of $\Omega(n)$ [3].

For k-means ($q = 0$, $\rho = 2$) and k-median ($q = 0$, $\rho = 2$) clustering, there are streaming algorithms that take $O(d^2k^2/\varepsilon^2 \log^8 n)$ work space [11]. Furthermore, we have that $C_2(0, \varepsilon) = O(1/\varepsilon)$; this follows as a straight-forward application of sparse greedy optimization using the Franke-Wolfe algorithm [13], and $C_1(0, \varepsilon) = O(1/\varepsilon^2 \log 1/\varepsilon)$, by a straight-forward modification of the construction in [8, Thm 3.2]

It follows that our algorithm approximates, with high probability, the value of $f_k^0(P,2)$ (k-means) with a space complexity of $O(d\log n/\varepsilon^3 + k^2/\varepsilon^8 \log^{10} n)$ and of $f_k^1(P,2)$ (k-median) with a space complexity of $O(d\log n/\varepsilon^4 + k^2/\varepsilon^{10}\log 1/\varepsilon \log^{10} n)$ Both results are significant improvement when d and n are of the same order of magnitude.

Approximate Projective Clustering. For a set P of n points in $\mathbb{R}^d$, let $T(n,d,q,k,\varepsilon,\rho)$ denote the complexity to compute k q-flats $\mathcal{F}$ that ε-approximate the optimal solution, that is, $\sum_{p\in P}(\mathtt{d}(p,\mathcal{F}))^\rho \leq (1+\varepsilon)f_k^q(P,\rho)$. We design a new algorithm as follows: Set $\varepsilon' := \varepsilon/5$. First, we randomly project the input point set from d to $m := O(C_\rho(q,\varepsilon')\log n/\varepsilon'^2)$ dimensions. Let P' be this set of projected points. Then, we (ε'-approximately) solve the same problem for P' in m dimensions, using some algorithm for this problem as a black box. The computed solution clusters P' in k subsets of points that are closest to a particular q-flat in the solution. We let $P^1,\dots P^k$ be the pre-image of these k clusters and assume wlog that $P^i \cap P^j = \emptyset$. For each of the k clusters computed, say P^i, we compute an ε'-approximation of the best fitting q-flat. We return the collection of these k q-flats as our solution.

We argue why the obtained solution is indeed an ε-approximate optimal solution with high probability. Let $\mathcal{F} = \{F_1,\dots,F_k\}$ be the result of the algorithm. By construction, and with Theorem 4, we know that

$$\sum_{p\in P}(\mathtt{d}(p,\mathcal{F}))^\rho = \sum_{i=0}^{k}\sum_{p\in P_i}(\mathtt{d}(p,F_i))^\rho \leq (1+\varepsilon')\sum_{i=0}^{k} f_1^q(P_i,\rho) \leq (1+\varepsilon')^2\sum_{i=0}^{k} f_1^q(\hat{\pi}(P_i),\rho)$$

with high probability. Moreover, we know that the clustering $\hat{\pi}(P_1),\dots,\hat{\pi}(P_k)$ induces an ε'-approximate optimal clustering. Combined with Corollary 5, we get that

$$(1+\varepsilon')^2\sum_{i=0}^{k} f_1^q(\hat{\pi}(P_i),\rho) \leq (1+\varepsilon')^3 f_k^q(\hat{\pi}(P),\rho) \leq (1+\varepsilon')^4 f_k^q(P,\rho) \leq (1+\varepsilon) f_k^q(P,\rho),$$

which proves correctness. The complexity of the algorithm is bounded by

$$O\left(dmn + T(n,m,q,k,\varepsilon,\rho) + kmn + T(n,d,q,1,\varepsilon,\rho)\right)$$

where the four terms correspond to the complexities of projecting the point set, computing the solution in m dimensions, clustering the projected points, and computing q-flats for each cluster, respectively. As we can see, d and k are decoupled in this bound.

As an example, we get the k-center problem by setting $\rho = \infty$ and $q = 0$. Using the bounds $C_\infty(0,\varepsilon) = 2/\varepsilon$, $T(n,d,0,k,\varepsilon,\infty) = O(nd2^{k\log k/\varepsilon})$ and $T(n,d,0,1,\varepsilon,\infty) = O(\frac{nd}{\varepsilon^2} + \frac{1}{\varepsilon^5})$ from [7], we get a running time of

$$O(n\log n 2^{k\log k/\varepsilon} + \frac{dn\log n}{\varepsilon^3}).$$

Approximating Čech complexes [7] A standard tool in capturing topological properties of topological spaces sampled by point cloud data is the *Čech complex*. It is usually defined to be the *nerve* of balls of some fixed radius α centered at the points from the sample P, and denoted as $\mathcal{C}_\alpha(P)$. An equivalent definition is that a k-simplex $\{p_0,\dots,p_k\}$ is in $\mathcal{C}_\alpha(P)$ if and only if the radius of $\mathrm{meb}(p_0,\dots,p_k)$ is at most α.

The downside of Čech complexes is the size: Their d-skeleton can consist of up to $O(n^d)$ simplices. Recent work suggests to work instead with an approximation of the Čech complex [26] (or of the closely related Vietoris-Rips complex [29][17]). "Approximation" in this context means that the persistence diagrams of the modules induced by the Čech filtration and by the approximate filtration are close to each other. Theorem 4 for $q = 0$, $k = 1$ and $\rho = \infty$ implies that the radius of MEBs is preserved for any subset. That implies immediately that Čech complexes can be approximated by Čech complexes in lower dimensions.

[7] Due to space restrictions, we omit a thorough introduction of the topological concepts used in this paragraph. See [26], or the textbook [18] for more details

Proposition 25 *For $0 < \varepsilon \leq \frac{c-1}{c} < 1$ with $c > 1$ and arbitrary constant, a set $P \subset \mathbb{R}^d$ of n points, and $m = \Theta(\log(n)/\varepsilon^3)$, a random projection $\hat{\pi} : \mathbb{R}^d \to \mathbb{R}^m$ satisfies with high probability that*

$$\mathcal{C}_{(1-c\varepsilon)\alpha}(P) \subseteq \mathcal{C}_\alpha(\hat{\pi}(P)) \subseteq \mathcal{C}_{(1+c\varepsilon)\alpha}(P).$$

An interesting consequence of this statement is that a Čech complex cannot have any significantly persistent features in dimensions higher than m. Independently from our work, Sheehy [30] recently showed a slightly better result, projecting to $\Theta(\log(n)/\varepsilon^2)$ dimensions.

5 Conclusion

In this paper, we demonstrated that random projections not only preserve pairwise-distances but also preserve a wide class of clustering functions. Our proof technique exploits the existence of small-sized coresets for these functions. Interestingly, we needed only an existence of a small-sized coreset without actually being concerned about its computation time. This allowed us to design smaller coreset for the case of the projective clustering problem for the case where $\rho = \infty$.

As applications, we have shown that our techniques improve approximation and streaming algorithms for projective clustering. Our results use the simplest form of random projection, namely, picking a m-dimensional subspace uniformly at random. The disadvantage is that the projection matrix is dense and costly to apply to a point set; several restricted classes of projection matrices have been shown to lead to a better computational behavior, such as random $\{-1, 1\}$-matrices [1], matrices that allow fast multiplication [5] and matrices well suited for sparse inputs [15, 14]. A combination with these results would further strengthen our improvements. We remark that such a combination is not entirely straight-forward because it requires an analysis of the concentration bounds of the underlying methods, similar to Lemma 2 in this work.

Our results have direct implications for the approximation of Čech complexes. However, using the specialized nature of Čech complex, a better bound has been achieved by Sheehy [30]. We pose the question whether our more general technique can be used to derive other approximation guarantees in topological contexts.

Acknowledgments. The authors thank the anonymous referees of an earlier version of this paper whose comments have led to significant simplifications and improvements of our results. The first author acknowledges support by the Max Planck Center for Visual Computing and Communication

References

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

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

[3] P. Agarwal and R. Sharathkumar, Streaming algorithms for extent problems in high dimensions, *Algorithmica*, (2013), 1–16.

[4] P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan, Geometric approximation via coresets, *Combinatorial and computational geometry*, 52 (2005), 1–30.

[5] N. Ailon and B. Chazelle, The fast Johnson-Lindenstrauss transform and approximate nearest neighbors, *SIAM Journal of Computing*, 39 (2009), 302–322.

[6] S. Arora, E. Hazan, and S. Kale, A fast random sampling algorithm for sparsifying matrices, in: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, LNCS*, Vol. 4110, 2006, pp. 272–279.

[7] M. Bâdoiu and K. L. Clarkson, Smaller core-sets for balls, *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2003, pp. 801–802.

[8] M. Bā̆doiu, S. Har-Peled, and P. Indyk, Approximate clustering via core-sets, *Proceedings of the 34th Annual ACM Symposium on Theory of Computing*, 2002, pp. 250–257.

[9] I. Barany and Z. Füredi, Computing the volume is difficult, *Discrete & Computational Geometry*, 2 (1987), 319–326.

[10] M. Bădoiu and K. Clarkson, Optimal core-sets for balls, *Computational Geometry: Theory and Applications*, 40 (2008), 14–22.

[11] K. Chen, On coresets for k-median and k-means clustering in metric and Euclidean spaces and their applications, *SIAM Journal of Computing*, 39 (2009), 923–947.

[12] K. L. Clarkson, Tighter bounds for random projections of manifolds, *Proceedings of the 24th Symposium on Computational Geometry*, 2008, pp. 39–48.

[13] K. L. Clarkson, Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm, *ACM Transactions on Algorithms*, 6 (2010).

[14] K. L. Clarkson and D. P. Woodruff, Low rank approximation and regression in input sparsity time, *Proceedings of the 45th Annual ACM Symposium on Theory of Computing*, 2013, pp. 81–90.

[15] A. Dasgupta, R. Kumar, and T. Sarlos, A sparse Johnson-Lindenstrauss transform, *Proceedings of the 42nd ACM Symposium on Theory of Computing*, 2010, pp. 341–350.

[16] S. Dasgupta and A. Gupta, An elementary proof of a theorem of Johnson and Lindenstrauss, *Random Structures & Algorithms*, 22 (2003), 60–65.

[17] T. Dey, F. Fan, and Y. Wang, Graph induced complex on point data, *Proceedings of the 29th ACM Symposium on Computational Geometry*, 2013.

[18] H. Edelsbrunner and J. Harer, *Computational Topology, An Introduction*, American Mathematical Society, 2010.

[19] U. Feige and E. Ofek, Spectral techniques applied to sparse random graphs, *Random Structures & Algorithms*, 27 (2005), 251–275.

[20] D. Feldman, M. Monemizadeh, C. Sohler, and D. P. Woodruff, Coresets and sketches for high dimensional subspace approximation problems, *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms*, 2010, pp. 630–649.

[21] S. Har-Peled and S. Mazumdar, On coresets for k-means and k-median clustering, *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, 2004, pp. 291–300.

[22] S. Har-Peled and K. R. Varadarajan, Projective clustering in high dimensions using core-sets, *Proceedings of the 18th ACM Symposium on Computational Geometry*, 2002, pp. 312–318.

[23] S. Har-Peled and K. R. Varadarajan, High-dimensional shape fitting in linear time, *Discrete Computational Geometry*, 32 (2004), 269–288.

[24] P. Indyk and A. Naor, Nearest-neighbor-preserving embeddings, *ACM Transactions on Algorithms*, 3 (2007).

[25] W. Johnson and J. Lindenstrauss, Extensions of Lipschitz mappings into a Hilbert space, *Contemporary Mathematics*, 26 (1982), 189–206.

[26] M. Kerber and R. Sharathkumar, Approximate Čech complex in low and high dimensions, *24th International Symposium on Algorithms and Computation*, LNCS 8283, 2013, pp. 666–676.

[27] A. Magen, Dimensionality reductions in ℓ_2 that preserve volumes and distance to affine spaces, *Discrete & Computational Geometry*, 38 (2007), 139–153.

[28] T. Sarlos, Improved approximation algorithms for large matrices via random projections, *Foundations of Computer Science*, 2006, pp. 143–152.

[29] D. Sheehy, Linear-size approximation to the Vietoris-Rips filtration, *Proceedings of the 28th ACM Symposium on Computational Geometry*, 2012, pp. 239–248.

[30] D. R. Sheehy, The persistent homology of distance functions under random projection, *Proceedings of the 30th ACM Symposium on Computational Geometry*, 2014.

[31] S. S. Vempala, *The random projection method*, AMS Bookstore, 2004.