Approximation and Streaming Algorithms for Projective Clustering via Random Projections

Michael Kerber* Sharath Raghvendra†

Abstract

Let P be a set of n points in $\mathbb{R}^d$. In the projective clustering problem, given k, q and norm $\rho \in [1, \infty]$, we have to compute a set $\mathcal{F}$ of k q-dimensional flats such that $(\sum_{p\in P} \mathtt{d}(p, \mathcal{F})^\rho)^{1/\rho}$ is minimized; here $\mathtt{d}(p, \mathcal{F})$ represents the (Euclidean) distance of p to the closest flat in $\mathcal{F}$. We let $f_k^q(P, \rho)$ denote the minimal value and interpret $f_k^q(P, \infty)$ to be $\max_{r\in P} \mathtt{d}(r, \mathcal{F})$. When $\rho = 1, 2$ and ∞ and $q = 0$, the problem corresponds to the k-median, k-mean and the k-center clustering problems respectively.

For every $0 < \varepsilon < 1$, $S \subset P$ and $\rho \geq 1$, we show that the orthogonal projection of P onto a randomly chosen flat of dimension $O(((q + 1)^2 \log(1/\varepsilon)/\varepsilon^3) \log n)$ will ε-approximate $f_1^q(S, \rho)$. This result combines the concepts of geometric coresets and subspace embeddings based on the Johnson-Lindenstrauss Lemma. As a consequence, an orthogonal projection of P to an $O(((q + 1)^2 \log((q + 1)/\varepsilon)/\varepsilon^3) \log n)$ dimensional randomly chosen subspace ε-approximates projective clusterings for every k and ρ simultaneously. Note that the dimension of this subspace is independent of the number of clusters k.

Using this dimension reduction result, we obtain new approximation and streaming algorithms for projective clustering problems. For example, given a stream of n points, we show how to compute an ε-approximate projective clustering for every k and ρ simultaneously using only $O((n + d)((q + 1)^2 \log((q + 1)/\varepsilon))/\varepsilon^3 \log n)$ space. Compared to standard streaming algorithms with $\Omega(kd)$ space requirement, our approach is a significant improvement when the number of input points and their dimensions are of the same order of magnitude.

1 Introduction

Consider the *projective clustering problem*: For a set P of n points in $\mathbb{R}^d$, given integers $k, q < n$ and an integer norm $\rho \geq 1$, compute a set $\mathcal{F}$ of k q-dimensional flats (or q-*flats*) such that $(\sum_{p\in P} \mathtt{d}(p, \mathcal{F})^\rho)^{1/\rho}$ is minimized; here $\mathtt{d}(p, \mathcal{F})$ represents the Euclidean distance of p to its closest point on any flat in $\mathcal{F}$. We define

$$f_k^q(P, \rho) := \min_{\mathcal{F}} \Big(\sum_{p\in P} \mathtt{d}(p, \mathcal{F})^\rho\Big)^{1/\rho}$$

and interpret $f_k^q(P, \infty)$ to be $\min_{\mathcal{F}} \max_{p\in P} \mathtt{d}(p, \mathcal{F})$. The projective clustering problem is a generalization of several well-known problems. For example, when $\rho = \infty$, $q = 0$ this problem is the *minimum enclosing ball* (MEB) problem (when $k = 1$) and the *k-center clustering problem* (for arbitrary k). When $\rho = \infty$ 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 the k-means clustering problem (for $\rho = 2$). The projective clustering problem is NP-Hard [5] and, therefore, most research has focused on the design of approximation algorithms. For an error parameter $0 < \varepsilon < 1$, an *ε-approximate projective clustering* is a set of q-flats $\tilde{\mathcal{F}}$ such that $(\sum_{p\in P} \mathtt{d}(p, \tilde{\mathcal{F}})^\rho)^{1/\rho} \leq (1 + \varepsilon) f_k^q(P, \rho)$.

Projective clustering is an important task arising in unsupervised learning, data mining, computer vision and bioinformatics; see [31] for a survey of some of these applications. Given its significance, clustering problems have received much attention leading to new approximation algorithms. The early algorithms for these problems had exponential dependence on d [2, 4] and were well-suited for low-dimensional inputs. However, for many practical problems, the number of input points n and their dimension d are in the same order of magnitude [21].

Badoiu, Indyk and Har-Peled [8] made a breakthrough in the design of high-dimensional clustering algorithms. They designed a *coreset*-based algorithm that quickly constructs a small "most-relevant" subset E of the input points P with the property that an optimal clustering on E is an approximate clustering for P, and use this coreset to compute an approximate clustering. Based on this idea, several coreset-based approximation algorithms for projective clustering were developed, also for the design of *streaming* algorithms for projective clustering[1]; see for example [11, 20, 22]. In recent research, depending on the problem, different definitions of coreset have been used. These definitions vary from weak to strong notions of when a subset is relevant, and therefore yield different size bounds (see for instance [21] for a careful discussion).

Throughout this paper, we use the following definition: a coreset (with respect to ε, q, ρ) is a subset $E \subseteq P$ such that the affine subspace spanned by E contains a

*Max Planck Institute for Informatics, Saarbrücken, Germany, mkerber@mpi-inf.mpg.de

†Virginia Tech, Blacksburg, USA, sharathr@vt.edu

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

q-flat F with $(\sum_{p\in P} \mathtt{d}(p,F)^\rho)^{1/\rho} \leq (1+\varepsilon) f_1^q(P,\rho)$. We let $C_\rho(q,\varepsilon)$ denote the worst-case size of such a coreset for approximating $f_1^q(P,\rho)$. This is a comparably weak version of coresets: we only require that the subspace spanned by E contains some ε-approximate solution; we do *not* require that the optimal solution for E is that ε-approximation. For problems such as MEB, MEC, 1-mean, or 1-median, there are coresets whose size is independent of the number of points and the ambient dimension [7, 8, 23, 35].

Another useful tool for the design of high-dimensional clustering algorithms is the *random projection* method [36]. At its heart is the following well-known lemma [26] which says that an orthogonal projection of any point set to a random $O(\log n/\varepsilon^2)$-dimensional flat ε-approximates pairwise distance between all pairs of points; see [16] for an elementary proof.

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 $\hat{\pi}$ as above as a *random projection* to an m-dimensional flat.

The Johnson-Lindenstrauss Lemma shows that random projections approximate pairwise distances between points. A natural question is what other geometric and structural properties of high-dimensional point cloud are preserved by random projections, and numerous such properties have been identified [1, 12, 24, 28, 32]. Random projection techniques are widely used for clustering problems: ongoing research focuses in particular to the case of k-means clustering [9, 14, 15], although it has also been used for certain projective clustering problems [8, 30].

Our results. We establish a link between coresets and the random projections for the projective clustering problem in Section 2. We show that, for every $0 < \varepsilon < 1$, $q \geq 0$, and $\rho \geq 1$, a random projection to a $O(((q+1)^2\log((q+1)/\varepsilon)/\varepsilon^3 \log n)$-dimensional space ε-approximates $f_1^q(S,\rho)$ for all $S \subseteq P$. 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 $C_\rho(q,\varepsilon)$.Our argument follows the standard proof technique for subspace embeddings (as sketched in [14, 29]) by approximately preserving the lengths of vectors taken from a sufficiently dense ε-net. For a given q and any $\rho \geq 1$, the existence of small-sized coresets with $C_\rho(q,\varepsilon) = O((q+1)^2/\varepsilon\log((q+1)/\varepsilon))$ is known [35]. This leads to the previously mentioned bound on the dimension of the projected space.

As a consequence, we show that by projecting to the same dimension, also $f_k^q(P,\rho)$ is preserved for all k and $\rho \geq 1$. Note that the dimension of the subspace is independent of k and ρ and is only logarithmic in n. Our results imply that improved bounds on the size of the coreset $C_\rho(q,\varepsilon)$ lead to better bounds on the dimension of the random subspace. Interestingly, unlike previous applications of coresets, we do not require a fast method to compute $C_\rho(q,\varepsilon)$. Therefore, we can shoot for even smaller-sized coresets without being restricted by its computation time (Section 3).

Our results has the following applications (Section 4): For a given q and a stream of n points, we give an algorithm that can compute projective clustering of P for every value of k and ρ using only $O(((q+1)^2\log((q+1)/\varepsilon)/\varepsilon^3)(n+d)\log n)$ space. Almost all known (multi-pass) streaming algorithms for projective clustering problems have a linear dependence on the product of k and d, and therefore, they tend to require $\Omega(nd)$ space for when $k = \Theta(n)$. As opposed to this, our algorithm requires $\tilde{O}(n+d)$ space which is particularly useful when n and d are of the same order of magnitude. Also, in many practical scenarios, the number of clusters k and the norm ρ are not known in advance. Our algorithm is also useful in such cases since our dimension reduction technique works for all values of k and ρ simultaneously.

We also generically improve approximation algorithms for projective clustering problems. Again, we project P onto a randomly chosen subspace and compute an approximate solution in the projected subspace. We obtain a solution in the original d-dimensional space by "lifting" each cluster from the 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]; note that k and d are decoupled in our complexity bound.

Finally, 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. Recently, these results have been proven independently by Sheehy [34].

2 Generalized Johnson-Lindenstrauss Lemma

Recall the definition of $f_1^q(P,\rho)$ as the L_ρ-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 cor-

responding 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 under the projection, we can guarantee to preserve $f_1^q(S,\rho)$ approximately as well. We ensure this preservation by the stronger property in Lemma 3 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. 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 $\langle S\rangle$. While the proof of this subspace embedding result has been outlined in previous work [14, 29], we are not aware of a formal proof of the statement.

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 Theorem 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 [32, 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 subsets $S \subset P$ of cardinality c, all q-flats $Q \subset \langle S\rangle$, 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 $d(\hat\pi(p),\hat\pi(Q))$. □

We show our main theorem that random projections preserve $f_1^q(S,\rho)$ for any $S \subseteq P$.

Theorem 4 *Let $0 < \varepsilon < 1$, $P \subset \mathbb{R}^d$ consist of n points, $q \geq 0$ an integer and $\rho \in \mathbb{Z}_{\geq 1} \cup \{\infty\}$. 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 subsets $S \subseteq P$*

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

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 $\varepsilon' = \varepsilon/3$, we get that

$$\begin{aligned} f_1^q(\hat\pi(S),\rho) &\leq \left(\sum_{p\in S} d(\hat\pi(p),\hat\pi(F))^\rho\right)^{1/\rho} \\ &\leq \left(\sum_{p\in S}(1+\varepsilon/3)^\rho d(p,F)^\rho\right)^{1/\rho} \\ &\leq (1+\varepsilon/3)(1+\varepsilon/2)f_1^q(S,\rho) \leq (1+\varepsilon)f_1^q(S,\rho), \end{aligned}$$

where we use $(1+\varepsilon/3)(1+\varepsilon/2) < 1+\varepsilon$ 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

$$\begin{aligned} (1-\varepsilon)f_1^q(S,\rho) &\leq (1-\frac{\varepsilon}{2})(1-\frac{\varepsilon}{3})\left(\sum_{p\in S} d(p,F')^\rho\right)^{1/\rho} \\ &\leq (1-\frac{\varepsilon}{2})\left(\sum_{p\in S} d(\hat\pi(p),G)^\rho\right)^{1/\rho} \\ &\leq (1-\frac{\varepsilon}{2})(1+\frac{\varepsilon}{2})f_1^q(\hat\pi(S),\rho) \leq f_1^q(\hat\pi(S),\rho). \end{aligned}$$

Again, the case $\rho = \infty$ is analogue to $\rho = 1$. □

Theorem 4 implies that $f_k^q(P, \rho)$ 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 $f_1^q(P_i, \rho)$.[2] Using Theorem 4 on the subsets P_i, we get for $\rho < \infty$ that

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

proving the second 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 Coresets for Projective Clustering

Recall that $C_\rho(q, \varepsilon)$ is defined as the coreset size for approximating the L_ρ-optimal q-flat, in the sense that there exists a subset of $C_\rho(q, \varepsilon)$ input points whose span contains an ε-approximate optimal q-flat. Because of space restrictions, we omit the (simple) proofs of the results in this section (see Appendix A).

The case of 0-flats For a point set $P \subset \mathbb{R}^d$, we consider the point that minimizes, for a fixed $\rho \in [1, \infty]$,

$$\delta(q) := \left(\sum_{p \in P} d(p,q)^\rho \right)^{1/\rho}$$

over all $q \in \mathbb{R}^d$. We call the minimizer o in $\mathbb{R}^d$ the *optimal center* and note that $\delta(o) = f_1^0(P, \rho)$. We call o' an *ε-approximate center*, if $\delta(o') \leq (1+\varepsilon)\delta(o)$. Since Theorem 4 only requires a bound on the coreset and no method to compute it, we can free ourselves from algorithmic considerations and concentrate on existential results.

A lower bound of $\Omega(1/\varepsilon)$ can be derived easily by considering the standard simplex. This has been done by Bădoiu and Clarkson [10] for the case $\rho = \infty$.

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

Theorem 6 *There exists a point set such that no subset of less than $1/(2\varepsilon)$ points contains an ε-approximate center, i.e., $C_\rho(0, \varepsilon) = \Omega(1/\varepsilon)$.*

The following result gives an almost tight upper bound for arbitrary ρ. It follows directly from the techniques introduced by Shyamalkumar and Varadarajan [35] for the case of lines through the origin.

Theorem 7 *For any (finite) point set $P \subset \mathbb{R}^d$, there is a set $S \subset P$ of $O(1/\varepsilon \log(1/\varepsilon))$ points such that the subspace spanned by S contains an ε-approximate center. In other words, $C_\rho(0, \varepsilon) = O(1/\varepsilon \log(1/\varepsilon))$.*

Smaller coresets exist for special cases: for $\rho = \infty$, a coreset of size $O(1/\varepsilon)$ (in fact, of size $\lceil 1/\varepsilon \rceil$) exists [10]. It is also known that for $\rho = 2$, the squared distance function $d^2(x, P)$ is a quadratic function in x and can therefore be tackled through sparse greedy optimization in the Frank-Wolfe framework [13, 25].

Theorem 8 *For $\rho = 2$, there is a set of $O(1/\varepsilon)$ points such that their subspace contains an ε-approximate center, i.e., $C_2(0, \varepsilon) = O(1/\varepsilon)$.*

The case of general q The best known bounds for $C_\rho(q, \varepsilon)$ with $q \geq 0$, are again due to Shyamalkumar and Varadarajan. The aforementioned result for lines yields that $C_\rho(1, \varepsilon) = O(1/\varepsilon \log(1/\varepsilon))$, the same bound as for $q = 0$ [35, Lemma 3.2]. They use the line case in an inductive argument to show [35, Lemma 3.3]:

Theorem 9 *For $q \geq 1$, $C_\rho(q, \varepsilon) = O(q^2/\varepsilon \log(q/\varepsilon))$.*

A natural question is to ask about the tightness of the coreset bounds: for the point case $q = 0$, we conjecture that coresets of size $O(1/\varepsilon)$ exist for any norm ρ (currently, this is only established for $\rho \in \{2, \infty\}$). For general q, an improved upper bound of $O(q/\varepsilon)$ would yield a target dimension linear in q in our dimension reduction result.

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 $m := O((q+1)^2/\varepsilon^3 \log n \log((q+1)/\varepsilon))$. In the simplest variant, our streaming initially chooses a $d \times m$ projection matrix uniformly at random, projects every point from the stream to $\mathbb{R}^m$, and stores all points in a set P'. The algorithm, then uses an (offline)-algorithm to approximate $f_k^q(P', \rho)$. The total work space of this

algorithm is $O(dm + nm + S)$, where dm is the space required to store the projection matrix, nm is the size of P', and S is the space required to find approximate clustering of P'. Using any approximation algorithm that computes using linear space, we obtain a streaming algorithm to approximate $f_k^q(P, \rho)$ with a space complexity of $O((q+1)^2(n+d)/\varepsilon^3 \log n \log((q+1)/\varepsilon))$. This is much smaller than the input size of $O(dn)$ and, for small q, not too far from the lower bound of $\Omega(n)$ [3].

In a similar fashion, our results can be used to speed up other streaming approaches: Again, we choose initially 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 its projection $\hat{\pi}(p) \in \mathbb{R}^m$ 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 $O(dm + S)$, with S the space complexity of the m-dimensional streaming algorithm.

Approximate Projective Clustering. Our technique is also useful for the computation of approximate cluster centers: For a set P of n points in $\mathbb{R}^d$, let $T(n, d, q, k, \varepsilon, \rho)$ denote the time complexity to compute k q-flats $\mathcal{F}$ that ε-approximate the optimal solution, that is, $(\sum_{p \in P}(\mathtt{d}(p, \mathcal{F}))^\rho)^{1/\rho} \leq (1+\varepsilon) f_k^q(P, \rho)$. 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 P^i, we compute an ε'-approximation of the best fitting q-flat. We return the collection of these k q-flats as solution. Correctness of this approach follows from Theorem 4 and Corollary 5. 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 A standard tool in capturing topological properties of point cloud data is the *Čech complex*[3]. 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 $O(n^{d+1})$ simplices. Recent work suggests to work instead with an approximation of the Čech complex [27] (or of the closely related Vietoris-Rips complex [33] [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.

Proposition 10 *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 [34] recently showed a slightly stronger result, projecting to $\Theta(\log(n)/\varepsilon^2)$ dimensions.

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. The second author acknowledges support by NSF through grant CCF-1464276.

References

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

[2] P. Agarwal and N. Mustafa. k-means projective clustering. In *Proceedings of the Twenty-third ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 155–165, 2004.

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

[4] P. K. Agarwal, C. M. Procopiuc, and K. R. Varadarajan. Approximation algorithms for a k-line center. *Algorithmica*, 42(3-4):221–230, 2005.

[3] For brevity, we omit a thorough introduction of the topological concepts used in this paragraph. See [18] for more details

[5] D. Aloise, A. Deshpande, P. Hansen, and P. Popat. NP-hardness of euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

[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*, volume 4110 of *LNCS*, pages 272–279. 2006.

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

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

[9] C. Boutsidis, A. Zouzias, and P. Drineas. Random projections for k-means clustering. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 298–306. 2010.

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

[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(3):923–947, 2009.

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

[13] K. Clarkson. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Transactions on Algorithms*, 6(4), 2010.

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

[15] M. B. Cohen, S. Elder, C. Musco, C. Musco, and M. Persu. Dimensionality reduction for k-means clustering and low rank approximation. *CoRR*, abs/1410.6801, 2014.

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

[17] T. Dey, F. Fan, and Y. Wang. Graph induced complex on point data. In *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*, 27:251–275, 2005.

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

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

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

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

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

[25] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 427–435, 2013.

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

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

[28] A. Magen. Dimensionality reductions in ℓ_2 that preserve volumes and distance to affine spaces. *Discrete & Computational Geometry*, 38(1):139–153, 2007.

[29] J. Nelson and H. Nguyen. Lower bounds for oblivious subspace embeddings. In *Automata, Languages, and Programming*, volume 8572 of *Lecture Notes in Computer Science*, pages 883–894. Springer Berlin Heidelberg, 2014.

[30] R. Ostrovsky and Y. Rabani. Polynomial time approximation schemes for geometric k-clustering. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, FOCS '00, pages 349–, 2000.

[31] C. Procopiuc. Projective clustering. In C. Sammut and G. Webb, editors, *Encyclopedia of Machine Learning*, pages 806–811. Springer US, 2010.

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

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

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

[35] N. Shyamalkumar and K. Varadarajan. Efficient subspace approximation algorithms. *Discrete & Computational Geometry*, 47(1):44–63, 2012.

[36] S. Vempala. *The random projection method*, volume 65. AMS Bookstore, 2004.

A Missing proofs of Section 3

Proof of Theorem 6 Consider the standard $(n-1)$-simplex spanned by n points in $\mathbb{R}^n$, namely the points $e_1, \ldots, e_n$. The optimal center for any ρ is the barycenter o given by $(1/n, \ldots, 1/n)$, and we have that

$$\delta(o) = n^{1/\rho}\sqrt{\frac{n-1}{n}}.$$

Choose a subset of size c, w.l.o.g. $e_1, \ldots, e_c$ and let F be the affine subspace spanned by these points. Let o' denote the barycenter of the $(c-1)$-simplex $(e_1, \ldots, e_c)$. Now o' is the point that minimizes $\delta(\cdot)$ over F, since o' is the orthogonal projection on F of any e_i with $i > c$. So, assuming that F contains an ε-approximate center, o' must be an approximate center. On the other hand, $\delta(o')$ is easily computed by noting that $d(o', e_i) = \sqrt{(n-1)/n}$ for any $i \le c$ and $d(o', e_i) = \sqrt{(n+1)/n}$ for any $i > c$. This yields

$$\begin{aligned}&\left(c\left(\sqrt{\frac{c-1}{c}}\right)^\rho + (n-c)\left(\sqrt{\frac{c+1}{c}}\right)^\rho\right)^{\frac{1}{\rho}}\\ =\;& \delta(o') \le (1+\varepsilon)\delta(o) = (1+\varepsilon)n^{1/\rho}\sqrt{\frac{n-1}{n}}.\end{aligned}$$

Raising to the ρ-th power and dividing by n yields that

$$\frac{c}{n}\left(\sqrt{\frac{c-1}{c}}\right)^\rho + \frac{n-c}{n}\left(\sqrt{\frac{c+1}{c}}\right)^\rho \le (1+\varepsilon)^\rho\left(\sqrt{\frac{n-1}{n}}\right)^\rho.$$

Because this bounds has to hold for every n, it must also hold in the limit for $n \to \infty$. That results in

$$\left(\sqrt{\frac{c+1}{c}}\right)^\rho \le (1+\varepsilon)^\rho,$$

and by solving for c yields that $c \ge 1/(2\varepsilon + \varepsilon^2) \ge 1/(3\varepsilon)$ for $\varepsilon \le 1$.

Proof of Theorem 7 We define an iterative procedure which creates points $c_0, c_1, \ldots$ such that c_i is in the subspace spanned by i input points. The initial point c_0 is chosen to be point closest to the optimal center.[4] If some c_i is an ε-approximate center, we are done. Otherwise, we show that we can chose a point c_{i+1} that is significantly closer to o. For that, let s be the point the maximizes

$$\frac{d(c_i, p)}{d(o, p)}$$

over all $p \in P$. By construction, $d(c_i, s) \ge (1+\varepsilon)d(o, s)$ We choose c_{i+1} as the point on the line segment $c_i s$ that is closest to o. It follows easily [35, Lemma 2.1] that $d(c_{i+1}, o) \le (1-\varepsilon/2)d(c_i, o)$. Combined with the triangle inequality and the fact that c_0 is the closest point to o in P, this implies that for any $p \in p$:

$$\begin{aligned}d(c_k, p) &\le d(c_k, o) + d(o, p)\\ &\le (1-\varepsilon/2)^k d(c_0, o) + d(o, p)\\ &\le (1 + (1-\varepsilon/2)^k)d(o, p).\end{aligned}$$

For $k = O(1/\varepsilon \log(1/\varepsilon))$, this means that $d(c_k, p) \le (1+\varepsilon)d(o, p)$ for all $p \in P$, which directly implies that $\delta(c_k) \le (1+\varepsilon)\delta(o)$.

Proof of Theoerem 8 Writing A for the matrix whose columns are the points in P and Δ for the standard simplex with points $e_1, \ldots, e_n$, we can consider the function $g : \Delta \to \mathbb{R}$ defined by

$$\begin{aligned}g(x) &= \sum_{i=1}^n \|Ax - Ae_i\|^2\\ &= \sum_{i=1}^n (x - e_i)^T A^T A(x - e_i)\\ &= x^T M x + x^T b + a\end{aligned}$$

with $M = nA^TA$, $b = -2\sum A^T A e_i$, and $a = \sum e_i^T A^T A e_i$. Therefore, g is a quadratic function. We apply the Frank-Wolfe optimization on the (convex) function g: this method starts in an arbitrary point x_0 in P and improves the approximation quality in every step by moving towards the point in P which the steepest descent. The obtained sequence of iterates $x_0, x_1, \ldots$ converges to the (unique) minimum of g, and by construction, the iterate x_i lies in the span of i points of P.

A crucial quantity in the convergence behavior of Frank-Wolfe is the quantity C_g: this is a scaled form of the Bregman divergence of the function g, measuring the difference between $g(y)$ and the value at y of the tangent plane of g at x, for all pairs x and y. Since g is a quadratic function, [13, Sec. 4.3] asserts that $C_g \le \operatorname{diam}(P)^2$. Writing r for the radius of the MEB of P, this implies $C_g \le 4r^2$.

Slightly abusing notation, we let $o \in \Delta$ denote the point that minimizes g. Using Theorem 2.3 from [13], after running the Franke-Wolfe optimization for $k := 2\lceil 1/\varepsilon \rceil$ steps, we find an iterate x_k on a k-simplex which satisfies

$$g(x_k) - g(o) \le 4\varepsilon C_f \le 16\varepsilon r^2 \le 16\varepsilon g(o),$$

where the last inequality comes from the fact that with p being the furthest point from o, it holds that $g(o) \ge \|o - p\|^2 \ge r^2$. Therefore, we have that $g(x_k) \le (1+16\varepsilon)g(o)$.

[4] Again, since we only care about existence, we can conveniently assume that the center is known to us.