Downloaded 05/02/16 to 109.230.67.85. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

A Near-Linear Algorithm for Projective Clustering Integer Points*

Kasturi Varadarajan† Xin Xiao‡

Abstract

We consider the problem of projective clustering in Euclidean spaces of non-fixed dimension. Here, we are given a set P of n points in $\mathbb{R}^m$ and integers $j \geq 1$, $k \geq 0$, and the goal is to find j k-subspaces so that the sum of the distances of each point in P to the nearest subspace is minimized. Observe that this is a shape fitting problem where we wish to find the best fit in the L_1 sense. Here we will treat the number j of subspaces we want to fit and the dimension k of each of them as constants. We consider instances of projective clustering where the point coordinates are integers of magnitude polynomial in m and n. Our main result is a randomized algorithm that for any $\varepsilon > 0$ runs in time $O(mn \text{ polylog}(mn))$ and outputs a solution that with high probability is within $(1+\varepsilon)$ of the optimal solution.

To obtain this result, we show that the fixed dimensional version of the above projective clustering problem has a small *coreset*. We do that by observing that in a fairly general sense, shape fitting problems that have small coresets in the L_∞ setting also have small coresets in the L_1 setting, and then exploiting an existing construction for the L_∞ setting. This observation seems to be quite useful for other shape fitting problems as well, as we demonstrate by constructing the first "regular" coreset for the circle fitting problem in the plane.

1 Introduction

A shape fitting problem is specified by a pair $(\mathbb{R}^d, \mathcal{F})$, where $\mathbb{R}^d$ denotes the d-dimensional Euclidean space and $\mathcal{F}$ is a family of shapes in $\mathbb{R}^d$. For example, $\mathcal{F}$ can be the family of all hyperplanes in $\mathbb{R}^d$; that is, each element of $\mathcal{F}$ is a hyperplane in $\mathbb{R}^d$. Two more examples that are of considerable interest to our work are:

1. The (j,k) *projective clustering problem*, where for some $j \geq 1$ and $0 \leq k \leq d-1$, $\mathcal{F}$ is the family of shapes with each shape being a union of j k-subspaces.
2. The *circle fitting problem* where $d = 2$, and $\mathcal{F}$ is the family of all circles in the plane.

An *instance* of a shape fitting problem $(\mathbb{R}^d, \mathcal{F})$ is specified by a finite set of points $P \subseteq \mathbb{R}^d$, and the goal is to find the shape $F \in \mathcal{F}$ that best fits P. To define this goal formally, let us assume that there is a function $\text{dist} : \mathbb{R}^d \times \mathcal{F} \to \mathbb{R}^+$ that given a point $p \in \mathbb{R}^d$ and shape $F \in \mathcal{F}$ specifies the "distance" of point p from shape F. In this article, we will take this to be the minimum Euclidean distance from p to a point in the shape F, *i.e.* $\min_{q \in F} \| p - q \|_2$. Let $\text{cost}(P,F) = \sum_{p \in P} \text{dist}(p,F)$ denote the cost of fitting point set P with shape F.

The shape fitting problem $(\mathbb{R}^d, \mathcal{F})$ then is the following: given a finite set $P = \{p_1, \ldots, p_n\} \subseteq \mathbb{R}^d$ of points, find $F \in \mathcal{F}$ that minimizes $\text{cost}(P,F)$. We will refer to this problem as shape fitting in the L_1 sense, since the goal is to find the shape F that minimizes the L_1 norm of the vector $[\text{dist}(p_1,F), \text{dist}(p_2,F), \ldots, \text{dist}(p_n,F)]$. In contrast, the shape problem $(\mathbb{R}^d, \mathcal{F})$ in the L_∞ sense takes as input a finite $P \subseteq \mathbb{R}^d$ as before, but seeks to find the shape $F \in \mathcal{F}$ that minimizes $\max_{p \in P} \text{dist}(p,F)$.

One special case of the (j,k) projective clustering problem that is considered here is the *integer* (j,k) *projective clustering problem*. Here, the coordinates of each of the input points in P are restricted to be integers of magnitude at most $\Delta = (nd)^c$, where $c > 0$ is a constant that can be chosen to be arbitrarily large.

In the context of a shape fitting problem $(\mathbb{R}^d, \mathcal{F})$, a *coreset* for a point set P is informally a smaller, possibly weighted set Q so that for any shape $F \in \mathcal{F}$, the cost of fitting P with F is approximately the same as the cost of fitting Q with F. There are variants of this notion that have different flavors, but a small *coreset* is of interest because solving the shape fitting problem near-optimally for the coreset yields a near-optimal solution for the original point set. Another related application is in the context of small-space streaming algorithms, where a coreset compactly represents the point set in the context of the shape fitting problem.

Formally, for a shape fitting problem $(\mathbb{R}^d, \mathcal{F})$ and an approximation parameter $0 \leq \varepsilon < 1$, an L_∞ ε-coreset for point set $P \subseteq \mathbb{R}^d$ is a subset $Q \subseteq P$ such that for any shape $F \in \mathcal{F}$, $\max_{q \in Q} \text{dist}(q,F) \geq (1-\varepsilon) \max_{p \in P} \text{dist}(p,F)$, or equivalently $\max_{p \in P} \text{dist}(p,F) - \max_{q \in Q} \text{dist}(q,F) \leq \varepsilon \cdot \max_{p \in P} \text{dist}(p,F)$. The *size* of the coreset Q is defined to be $|Q|$.

*This material is based upon work supported by the National Science Foundation under Grant No. 0915543.

†Department of Computer Science, University of Iowa, Iowa City, IA 52242. Email: kasturi-varadarajan@uiowa.edu

‡Department of Computer Science, University of Iowa, Iowa City, IA 52242. Email: xin-xiao@uiowa.edu

Copyright © SIAM.
Unauthorized reproduction of this article is prohibited.

Downloaded 05/02/16 to 109.230.67.85. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

An L_1 ε-coreset for P is a subset $S \subseteq \mathbb{R}^d$ of points, with each point $p \in S$ associated with a weight $w_p \geq 0$, such that for any $F \in \mathcal{F}$, we have $|\text{cost}(P, F) - \text{cost}(S, F)| \leq \varepsilon \cdot \text{cost}(P, F)$. Here, we abuse notation slightly to let $\text{cost}(S, F) = \sum_{p \in S} w_p \cdot \text{dist}(p, F)$. The size of the coreset is defined to be $|S|$.

1.1 Previous Work. There has been a vast amount of work on the (j, k) projective clustering problem, falling under two categories. In the *fixed dimensional setting*, the dimension d is considered a constant, whereas in the *high dimensional setting*, d is part of the input. To give a concise review of the previous work that is relevant to this article, we will focus on the task of finding a shape whose fit is within $(1 + \varepsilon)$ of the optimal, where $\varepsilon > 0$ is an arbitrarily small constant. In this context, we are interested in (a) algorithms with running times near-linear in the input size, which can be taken to be n, the number of input points, in the fixed dimensional setting, and $n \cdot d$ in the high dimensional setting; and (b) coresets whose size is bounded by a polynomial in $\log n$ in the fixed dimensional setting and by a polynomial in $d \cdot \log n$ in the high dimensional setting. With this focus, we may restrict our attention to the case where j and k are both (arbitrarily large) constants.

When $j = 1$, the shape that we want to fit is a single k-dimensional subspace. Near-linear algorithms are known here in both the L_1 and L_∞ settings even when the dimension is part of the input; we refer the reader to [26, 30, 21, 19]. The recent work of Feldman and Langberg [19] constructs small L_1 coresets in such high dimensional contexts.

Turning to $j > 1$, the $(j, 0)$ projective clustering problem in the L_∞ and L_1 senses are better known as the j-center and j-median problem, respectively. When the dimension is a constant, some early constructions [1, 25] describe near-linear algorithms and small coresets. In high dimensions, Badoiu et al. [9] present a near-linear algorithm and a weaker type of coreset for shape fitting in the L_∞ sense. For shape fitting in the L_1 sense, early near-linear algorithms can be found in [9, 27]; later works not only improved the running time but gave increasingly smaller coreset constructions [11, 20, 28, 19].

In the $(j, 1)$ projective clustering, the shape that we want to fit is a union of j lines. In fixed dimension, a near-linear algorithm and a coreset construction was given by [5] for the L_∞ context and by [17, 23] for the L_1 context. In high dimension, a near-linear algorithm for the L_1 context given by [15]; a coreset construction was recently given by [19].

Such a pleasant state of affairs does not persist for the (j, k) projective clustering problem for $k \geq 2$. No near-linear algorithms are known even in fixed dimension, and Har-Peled [22] gives fixed-dimensional examples that demonstrate that small coresets need not exist. To address this situation, some of the research presents near-linear bicriteria approximation algorithms [18, 19], where the output shape can have more than j subspaces, each with dimension k or somewhat larger. Another direction starts with the observation that the points in the example of Har-Peled [22] have coordinates that when viewed as integers are exponentially large. Thus, Edwards and Varadarajan [16] consider the integer (j, k) projective clustering problem, where these coordinates are only polynomially large. For this problem, they give small coresets and near-linear algorithms in the L_∞ sense in fixed dimension. This article extends this line of research, as described below.

On the practical side, there are several heuristics for versions of the (j, k) projective clustering problem, including CLIQUE [8], ENCLUS [12], DOC [29], PROCLUS [6], ORCLUS [7], and [4].

The circle fitting problem. The problem of fitting a circle to a set of points in the plane and a cylinder to a set of points in $\mathbb{R}^3$ has received considerable attention, see [2] for some earlier references. Near-linear algorithms and small coresets were first discovered for the L_∞ setting [1]. Subsequently, Har-Peled [24] was able to obtain near-linear algorithms for the L_1 setting.

1.2 Our Results and Techniques. Our first result is a near-linear algorithm for integer (j, k) projective clustering in the L_1 sense when the dimension is part of the input. Recall that in this problem we are given a set P of n points in $\mathbb{R}^m$ and integers $j \geq 1$, $k \geq 0$, and the goal is to find j k-subspaces so that the sum of the distances of each point in P to the nearest subspace is minimized; the point coordinates are integers of magnitude polynomial in m and n. Our randomized algorithm , for any parameter $\varepsilon > 0$, runs in time $O(mn \text{ polylog}(mn))$ and outputs a solution that with constant probability is within $(1 + \varepsilon)$ of the optimal solution.

To obtain this result, we observe that in a fairly general sense, shape fitting problems that have small coresets in the L_∞ setting also have small coresets in the L_1 setting. Using this observation, and the coreset construction of [16] for the L_∞ setting in fixed dimension, we are able to obtain a small coreset for the L_1 setting in fixed dimension. To solve the problem when the dimension is part of the input, we use a known dimension reduction result of Deshpande and Varadarajan [15].

Copyright © SIAM.
Unauthorized reproduction of this article is prohibited.

Downloaded 05/02/16 to 109.230.67.85. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Thus, we give the first near-linear algorithm for an interesting case of (j, k) projective clustering in high dimensions, when j and k are arbitrarily large constants. Another way of stating our result is that we have a near-linear approximation for the general (not integer) (j, k) projective clustering problem, provided the optimal fit is only polynomially smaller than the diameter of the input point set.

Our observation that shape fitting problems that have small coresets in the L_∞ setting also have small coresets in the L_1 setting appears to be useful beyond the projective clustering problem. We demonstrate this by using it to present a small L_1 coreset for the circle fitting problem, thus answering a question posed by Har-Peled [24]. The near-linear algorithm of Har-Peled for this problem does work via a compact representation of the input point set, but this representation is not an L_1 ε-coreset as defined here.

The connection between L_∞ and L_1 coresets builds on a sampling scheme due to Langberg and Schulman [28] for constructing small L_1 coresets. Their sampling scheme, which is a low-variance sampling scheme like some earlier ones [11, 13, 14, 20], is based on the notion of *sensitivity*. Roughly speaking, they show that shape fitting problems with small sensitivity have small L_1 coresets. What we observe is that shape fitting problems with small L_∞ coresets have small sensitivity. What our paper hopefully argues is that the resulting connection between L_∞ and L_1 coresets is a conceptually useful one.

Organization of the article. We have already defined the shape fitting problems we will consider, together with the notions of L_1 and L_∞ coresets. In Section 2, we describe the sampling scheme of Langberg and Schulman [28] that is based on their notion of sensitivity. In Section 3, we show that shape fitting problems with small L_∞ coresets have small sensitivity. As a consequence, we have a shape-oblivious sampling scheme for integer projective clustering and circle fitting that with high probability is good for any single shape in the family $\mathcal{F}$ of shapes. To be able to say that the sample is good for *every* shape in $\mathcal{F}$, we need to (a) argue that there is a polynomial-sized subfamily $\mathcal{F}' \subseteq \mathcal{F}$ that is a good "discretization" of the whole family $\mathcal{F}$, and (b) just apply a union bound over the subfamily $\mathcal{F}'$. We do this in Section 4 to derive a small L_1 coreset for circle fitting and integer projective clustering in fixed dimension. It is worth pointing out that while the discretization step is a standard feature in many constructions of L_1 coresets, the nature of the shape fitting problems we consider forces us to adapt a relatively recent discretization due to Vigneron [31]. Finally, in Section 5, we obtain our near-linear algorithm for integer projective clustering in high dimensions.

2 Preliminaries: Sampling Scheme Using Sensitivity

Langberg and Schulman [28] defined the notion of sensitivity in the context of shape fitting problems, and demonstrated the usefulness of sampling using sensitivity for constructing small L_1 coresets. We describe their sampling scheme, which is a key ingredient of the results reported here. Let $P \subset \mathbb{R}^d$ be a point set corresponding to some shape fitting problem $(\mathbb{R}^d, \mathcal{F})$. For any point $p \in P$, the sensitivity of p is defined to be

$$\sigma_P(p) := \sup_{F \in \mathcal{F}} \frac{\operatorname{dist}(p, F)}{\sum_{q \in P} \operatorname{dist}(q, F)}$$

The total sensitivity is defined to be

$$\mathfrak{S}_n := \sup_{P \subset \mathbb{R}^d, |P| = n} \sum_{p \in P} \sigma_P(p)$$

From the above definition, it worth noting that the sensitivity of p, $\sigma_P(p)$, depends only on the input point set P. In other words, for the input point set P, we can compute the sensitivity $\sigma_P(\cdot) : P \to [0, 1]$, and $\sigma_P(p)$ does not depend on the choice of any shape F. The total sensitivity, $\mathfrak{S}_n$, only depends on the size of the input point set P, n; it does not depend on any particular choice of P. A trivial upper bound for $\mathfrak{S}_n$ is n, as $\sigma_P(p) \leq 1$ for every $p \in P$. However, as the sampling scheme below requires that the size of a good sample depends on (the upper bound of) $\mathfrak{S}_n$, it is desirable to obtain better bounds on $\mathfrak{S}_n$.

Let s_P be a (point-wise) upper bound of $\sigma_P(\cdot)$ (*i.e.* $\forall p \in P$, $\sigma_P(p) \leq s_P(p)$), then $\sum_{p \in P} s_P(p)$ is an upper bound of $\sum_{p \in P} \sigma_P(p)$. We drop the subscript "P" of s_P and $\sigma_P(\cdot)$ in the following discussion when it is clear from context that the input point set is P. The sampling scheme is the following: define a probability distribution on P, where the probability of picking $p \in P$ is

$$\Pr(p \text{ is chosen}) = \frac{s(p)}{\sum_{p \in P} s(p)} \tag{2.1}$$

Independently pick a (multi)set R of r points from P (with replacement) according to the above probability distribution, the final output is the weighted point set S, where the weight w_p of a point $p \in S$ is

$$w_p := \frac{1}{r} \cdot \frac{1}{\Pr(p \text{ is chosen})} = \frac{1}{r} \cdot \frac{\sum_{p \in P} s(p)}{s(p)}$$

Copyright © SIAM.
Unauthorized reproduction of this article is prohibited.

Downloaded 05/02/16 to 109.230.67.85. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

LEMMA 2.1. *For a fixed shape F, we have*

$$\mathbb{E}(\mathit{cost}(S,F)) = \mathit{cost}(P,F),$$
$$\mathrm{Var}\left(\mathit{cost}(S,F)\right) \leq \frac{1}{r}\cdot\left(\sum_{p\in P} s(p)\right)\cdot\left(\mathit{cost}(P,F)\right)^2.$$

where (as we recall)

$$\mathit{cost}(S,F) = \sum_{p\in S} w_p\, \mathit{dist}(p,F),$$
$$\mathit{cost}(P,F) = \sum_{p\in P} \mathit{dist}(p,F).$$

LEMMA 2.2. *Fix a shape F. Let $\epsilon \in (0,1)$.*

$$\Pr\left(|\mathit{cost}(S,F) - \mathit{cost}(P,F)| \leq \epsilon\, \mathit{cost}(P,F)\right) \geq 1 - 2\exp\left(-r\cdot\frac{\epsilon^2}{2(1+\sum_{p\in P} s(p))^2}\right) \tag{2.2}$$

Proof. Denote the cost of assigning the weighted point in i^{th} draw by X_i. The expectation of X_i is $\mathrm{cost}(P,F)/r$:

$$\mathbb{E}(X_i) = \sum_{p\in P}\left(w_p \mathrm{dist}(p,F)\right)\cdot \Pr(\text{p is picked}) = \frac{1}{r}\cdot\sum_{p\in P}\mathrm{dist}(p,F) = \frac{1}{r}\mathrm{cost}(P,F)$$

Moreover, X_i is bounded from above by $(\sum_{p\in P} s(p)/r)\cdot \mathrm{cost}(P,F)$: if an arbitrary point p is picked, the cost of assigning the weighted point p to F is $w_p\mathrm{dist}(p,F)$; plugging in the definition of w_p, we have

$$w_p\mathrm{dist}(p,F) = \frac{1}{r}\cdot\frac{1}{\Pr(\text{p is picked})}\cdot\mathrm{dist}(p,F) = \frac{1}{r}\cdot\frac{\sum_{p\in P} s(p)}{s(p)}\cdot\mathrm{dist}(p,F) \leq \frac{1}{r}\cdot\left(\sum_{p\in P} s(p)\right)\cdot\mathrm{cost}(P,F)$$

The last inequality follows from the definition of sensitivity of point p:

$$\frac{\mathrm{dist}(p,F)}{\mathrm{cost}(P,F)} \leq \sigma_P(p) \leq s(p)$$

Note that we have

$$|\mathrm{cost}(S,F) - \mathrm{cost}(P,F)| = \left|\sum_{i=1}^{r} X_i - r\cdot\frac{1}{r}\mathrm{cost}(P,F)\right| = \left|\sum_{i=1}^{r}\left(X_i - \frac{1}{r}\mathrm{cost}(P,F)\right)\right|$$

Consider the random variable $X_i - \frac{1}{r}\mathrm{cost}(P,F)$. We have

$$\mathbb{E}\left(X_i - \frac{1}{r}\mathrm{cost}(P,F)\right) = 0,$$
$$|X_i - \frac{1}{r}\mathrm{cost}(P,F)| \leq \frac{\sum_{p\in P} s(p)}{r}\cdot\mathrm{cost}(P,F) + \frac{1}{r}\cdot\mathrm{cost}(P,F) = \frac{1+\sum_{p\in P} s(p)}{r}\cdot\mathrm{cost}(P,F)$$

An application of Azuma-Hoeffding implies that

$$\Pr\left(|\mathrm{cost}(S,F) - \mathrm{cost}(P,F)| \geq \epsilon\mathrm{cost}(P,F)\right) \leq 2\exp\left(-\frac{(\epsilon\mathrm{cost}(P,F))^2}{2\cdot r\cdot\left(\frac{1}{r}\cdot\left(1+\sum_{p\in P} s(p)\right)\cdot\mathrm{cost}(P,F)\right)^2}\right) \leq 2\exp\left(-r\cdot\frac{\epsilon^2}{2\left(1+\sum_{p\in P} s(p)\right)^2}\right)$$

□

The Lemma suggests that for the sampling scheme is effective for any fixed shape $F \in \mathcal{F}$ provided r is significantly larger than $\frac{\left(1+\sum_{p\in P} s(p)\right)^2}{\varepsilon^2}$. It is therefore natural to identify shape fitting problems for which $\sum_{p\in P} s(p)$ is $o(\sqrt{n})$, where $n = |P|$. We turn to this question next.

3 L_∞ Coresets to Sensitivity

In this section, we describe a key observation that shape fitting problems with small L_∞ coresets have small sensitivity.

LEMMA 3.1. *Consider a shape fitting problem $(\mathbb{R}^d, \mathcal{F})$. Suppose that for some $0 \leq \delta < 1$, there is non-decreasing function $f_\delta(n)$ so that any point set $P' \subseteq \mathbb{R}^d$ of size n admits an L_∞ δ-coreset of size at most $f_\delta(n)$. Then for any $P \subseteq \mathbb{R}^d$ of size n, we can compute an upper bound $s(p)$ on the sensitivity $\sigma_P(p)$ for each $p \in P$, so that $\sum_{p\in P} s(p) \leq \frac{f_\delta(n)\log n}{1-\delta}$.*

Proof. We construct a sequence of subsets $P = P_1 \supseteq P_2 \supseteq P_3 \cdots P_m$, where $m \leq n$ and $|P_m| \leq f_\delta(n)$. P_{i+1} is constructed from P_i as follows. If $|P_i| \leq f_\delta(n)$, the sequence ends. Otherwise, we compute an L_∞ δ-coreset Q_i of P_i whose size is at most $f_\delta(n)$, and let $P_{i+1} = P_i \setminus Q_i$. This finishes the description of the construction.

Let Q_m denote the set P_m. Now, the sets $Q_1, Q_2, \ldots, Q_m$ partition P. We claim that for any $q \in Q_i$, its sensitivity $\sigma_P(q)$ can be upper bounded by

Copyright © SIAM.
Unauthorized reproduction of this article is prohibited.

Downloaded 05/02/16 to 109.230.67.85. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

$s(q) = \frac{1}{(1-\delta)i}$. To show this, consider an arbitrary shape $F \in \mathcal{F}$. Consider any $1 \leq j \leq i$. Observe that $q \in P_j$; let $q_j \in Q_j$ be the point in the δ-coreset Q_j of P_j such that $\text{dist}(q_j, F) = \max_{p \in Q_j} \text{dist}(p, F)$. We have

$$\begin{aligned}\text{dist}(q_j, F) &= \max_{p \in Q_j} \text{dist}(p, F) \geq (1-\delta) \cdot \max_{p \in P_j} \text{dist}(p, F) \\ &\geq (1-\delta) \cdot \text{dist}(q, F).\end{aligned}$$

Thus $\frac{\text{dist}(q,F)}{\sum_{p \in P} \text{dist}(p,F)} \leq \frac{\text{dist}(q,F)}{\sum_{1 \leq j \leq i} \text{dist}(q_j,F)} \leq \frac{1}{(1-\delta)i}$. Therefore, $\sigma_P(q) \leq s(q) = \frac{1}{(1-\delta)i}$.

Finally, $\sum_{p \in P} s(p) = \sum_{i=1}^{m} \frac{|Q_i|}{(1-\delta)i} \leq f_\delta(n) \sum_{i=1}^{m} \frac{1}{(1-\delta)i} \leq \frac{f_\delta(n) \log n}{1-\delta}$. □

The construction in the proof has some resemblence to constructions of L_∞ coresets with outliers [3]. We note that the proof actually yields an algorithm for computing the bound $s(p)$ for each $p \in P$, provided we have at hand an algorithm for computing the δ-coreset Q_i of P_i. Instead of computing the δ-coreset from scratch for each P_i, we can use dynamic algorithms for maintaining coresets under insertion and deletion. We initialize the structure by inserting points in P_1. For each i, we delete the points in the δ-coreset Q_i of P_i; after deleting every point in Q_i, the dynamic structure will hold our δ-coreset Q_{i+1} of P_{i+1}.

As an example of an application of Lemma 6, let us consider the $(j, 0)$-projective clustering problem $(\mathbb{R}^d, \mathcal{F})$. The L_∞ version of this is better known as the j-center problem, and its L_1 version as the j-median. It is well known that for this shape fitting problem, any $P \subset \mathbb{R}^d$ admits an L_∞ $(2/3)$-coreset of size $j+1$. In fact, such a coreset $\{p_1, \ldots, p_{j+1}\}$ is obtained by choosing $p_1 \in P$ arbitrarily, and for each $1 \leq i \leq j$, letting p_{i+1} be the point in P furthest from $\{p_1, \ldots, p_i\}$. Thus Lemma 6 yields a bound of $O(j \log n)$ on the total sensitivity of P. Comparing this with the bound of $O(j)$ on the total sensitivity by Langberg and Schulman [28], we see that the utility of Lemma 6 is not that it yields the best possible bounds on the total sensitivity, but that it yields pretty good bounds with relative ease. This is useful for more complicated shape fitting problems, to which we turn next. In the remainder of this section, and throughout Section 4, we treat the dimension d as a constant.

THEOREM 3.1. *Let $P \subseteq \mathbb{R}^d$ be an n-point instance of a shape fitting problem $(\mathbb{R}^d, \mathcal{F})$ that is either (a) circle fitting, (b) $(j, 1)$ projective clustering, or (c) integer (j, k) projective clustering. We can compute in $O(n(\log n)^{O(1)})$ time an upper bound $s(p)$ on the sensitivity $\sigma_P(p)$ for each $p \in P$ so that $\sum_{p \in P} s(p) \leq (\log n)^{O(1)}$. For the $(j, 1)$ projective clustering problem, the constant in the exponent of the logarithm depends on j and d, and for the integer (j, k) projective clustering problem, it depends on j, k, and d.*

Proof. **Circle Fitting**: An L_∞ 1/2-coreset of size $O(1)$ can be computed for any n-point set can be computed in time $O(n)$, see for example [2] and [1]. Using the dynamization technique described in these papers, such a 1/2-coreset can be maintained in $(\log n)^{O(1)}$ time per insert or delete. The result follows using Lemma 6 and the remarks following its proof on the implied algorithm and its dynamization.

$(j, 1)$ **projective clustering**: An L_∞ 1/2-coreset of size $O(1)$ (with the constant depending on j) exists for any n-point set [5], but the construction in that paper does not describe an efficient enough algorithm for constructing such a coreset. Nevertheless, using techniques that are now standard, a 1/2-coreset of size $(\log n)^{O(1)}$ can be computed in $O(n(\log n)^{O(1)})$ time. The dynamization technique described in [1] allows us to maintain a 1/2-coreset in $(\log n)^{O(1)}$ time per insertion and deletion.

Integer (j, k) **projective clustering**: An L_∞ 1/2-coreset of size $(\log \Delta \cdot \log n)^{O(1)}$ can be computed in time $n(\log \Delta \cdot \log n)^{O(1)}$ for any n-point set with integer coordinates and diameter Δ [16]. The dynamization technique in [1] allows us to maintain a 1/2-coreset in $(\log \Delta \log n)^{O(1)}$ time per insertion and deletion. The result follows by recalling that Δ is $(nd)^{O(1)}$ for any input to the integer projective clustreing problem with n points. □

4 Discretization and Coresets

Theorem 3.1 gives a way of obtaining good bounds on the sensitivities of each of the points in the input P. If these bounds are used in the sampling scheme described in Section 2, then Lemma 2.2 tells us that for a high enough sample size, the sample approximates P with respect to a *fixed* shape $F \in \mathcal{F}$ with high probability. We would like the approximation to hold with respect to *every* shape in $\mathcal{F}$. To do this, it is convenient to show, roughly speaking, the existence of a polynomial sized subfamily $\mathcal{F}' \subseteq \mathcal{F}$ with the property that if the sample approximates P with respect to every shape in $\mathcal{F}'$, it approximates P with respect to every shape in $\mathcal{F}$. We call such an $\mathcal{F}'$ a *discretization* for $\mathcal{F}$.

Discretization is a tool that is used in many coreset constructions, but the construction here achieving it is different from those in most of the previous papers because of the actual shape fitting problems to which it is applied. Our construction has some resemblance to the cover code construction in [28]. We first stating our result on the discretization for the circle fitting problem,

Copyright © SIAM.
Unauthorized reproduction of this article is prohibited.

Downloaded 05/02/16 to 109.230.67.85. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

and then for the projective clustering problem.

THEOREM 4.1. *Let $P \subseteq \mathbb{R}^d$ be an n-point instance of the circle fitting problem $(\mathbb{R}^2, \mathcal{F})$, and $\frac{1}{n} \leq \varepsilon < 1$ be a parameter. There exists a set $\mathcal{C}$ of $O(n^{12})$ circles with the following guarantee: let $S \subseteq P$ be any subset, with a weight $w_p \geq 0$ for each $p \in S$, satisfying the properties that (a) $|cost(S,C) - cost(P,C)| \leq \varepsilon cost(P,C)$ for every circle $C \in \mathcal{C}$; and (b) the overall weights of points in S, $\sum_{p\in S} w_p$, is at most $2n$. Then such an S is an L_1 5ε-coreset for P, that is, $|cost(S,C) - cost(P,C)| \leq 5\varepsilon cost(P,C)$ for every circle $C \in \mathcal{F}$.*

The discretization theorem for projective clustering is proved in a similar way, but has to handle one complication: while a circle is a "nice" shape, a union of j k-subspaces is only a union of "nice" shapes.

THEOREM 4.2. *Let $P \subseteq \mathbb{R}^d$ be an n-point instance of the (j,k) projective clustering problem problem $(\mathbb{R}^d, \mathcal{F})$, and $\frac{1}{n} \leq \varepsilon < 1$ be a parameter. There exists a subset $\mathcal{F}' \in \mathcal{F}$ of size $n^{O(1)}$ with the following guarantee: let $S \subseteq P$ be any subset, with a weight $w_p \geq 0$ for each $p \in S$, and $\frac{1}{n} \leq \varepsilon < 1$ be a parameter satisfying the properties that (a) $|cost(S,F) - cost(P,F)| \leq \varepsilon cost(P,F)$ for every shape $F \in \mathcal{F}'$; and (b) the overall weight of points in S, $\sum_{p\in S} w_p$, is at most $2n$. Then such an S is an L_1 5ε-coreset for P, that is, $|cost(S,F) - cost(P,F)| \leq 5\varepsilon cost(P,F)$ for every shape $F \in \mathcal{F}$. The constant in the exponent of the polynomial bounding the size of $\mathcal{F}'$ depends on j, k, and d.*

We first use the discretization theorems to derive the existence of coresets for the circle fitting and projective clustering problems, and then present their proofs.

THEOREM 4.3. *Let $P \subseteq \mathbb{R}^d$ be an n-point instance of a shape fitting problem $(\mathbb{R}^d, \mathcal{F})$ that is either (a) circle fitting, (b) $(j,1)$ projective clustering, or (c) integer (j,k) projective clustering. Let $\frac{5}{n} \leq \varepsilon < 1$ be a parameter. There is a randomized algorithm that runs in $n\frac{(\log n)^{O(1)}}{\varepsilon^2}$ time and outputs with probability at least $1/6$ an ε-coreset S of size $\frac{(\log n)^{O(1)}}{\varepsilon^2}$.*

Proof. We describe the proof in detail for the circle fitting problem, and then make a brief remark on the very similar proofs for the projective clustering problems. The algorithm is to first compute an upper bound $s(p)$ on the sensitivity $\sigma_P(p)$ for each $p \in P$ using Theorem 3.1. Fix the set $\mathcal{C}$ implied by Theorem 4.1. (Note that we don't need to actually compute $\mathcal{C}$, we just need it for the analysis.)

Using the upper bounds $s(\cdot)$, we constructed a weighted sample $S \subseteq P$ of size r as described in Section 2. We only need to set the number of samples r sufficiently large so that with probability at least $1/6$, the following two conditions hold for S: (a) $|\text{cost}(P,C) - \text{cost}(S,C)| \leq \varepsilon \cdot \text{cost}(P,C)$ for every circle $C \in \mathcal{C}$, and (b) $\sum_{p\in S} w_p \leq 2n$.

We first consider condition (a). We set r large enough so that the following inequality holds for a fixed circle $C \in \mathcal{C}$:

$$\Pr\left(|\text{cost}(P,C) - \text{cost}(S,C)| \leq \epsilon \text{cost}(P,C)\right) \geq 1 - \frac{1}{3}\cdot\frac{1}{|\mathcal{C}|}.$$

Using Lemma 2.2, it suffices to set

$$r = O\left(\frac{\left(\sum_{p\in P} s(p)\right)^2}{\epsilon^2} \cdot \ln|\mathcal{C}|\right).$$

The choice of r guarantees that condition (a) in Theorem 4.1 holds with probability at least $2/3$, which can be shown by an application of union bound to the set of circles in $\mathcal{C}$.

Consider condition (b). The expectation of $\sum_{p\in S} w_p$ is the cardinality of P: the expected weight of the point in each draw is n/r, by the following calculation,

$$\sum_{p\in P} w_p \cdot \Pr(\text{p is selected}) =$$
$$\sum_{p\in P}\left(\frac{1}{r}\cdot\frac{1}{\Pr(\text{p is selected})}\right)\cdot\Pr(\text{p is selected}) = \frac{n}{r}$$

And we totally draw r points from P, hence the overall expectation $\mathbb{E}\left(\sum_{p\in S} w_p\right)$ is n by linearity of expectation. Using Markov inequality, we have

$$\Pr\left(\sum_{p\in S} w_p \geq 2n\right) \leq 1/2.$$

Thus condition (a) and condition (b) holds simultaneously with probability at least $1/6$. Theorem 4.1 then tells us that S is an L_1 5ϵ-coreset of P with probability at least $1/6$. Substitute the upper bounds of $|\mathcal{C}|$ and $\sum_{p\in P} s(p)$, the number of samples we need to draw is

$$r = \frac{(\log n)^{O(1)}}{\varepsilon^2}.$$

The proof for the projective clustering problems is very similar – the only change is the use of Theorem 4.2 instead of Theorem 4.1. □

Copyright © SIAM.
Unauthorized reproduction of this article is prohibited.

Downloaded 05/02/16 to 109.230.67.85. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

We note that for the $(j,1)$ projective clustering problem, small L_1 coresets are already know to exist [17, 23]; the derivation in Theorem 4.3 is different from these earlier ones and is arguably simpler at the conceptual level.

4.1 Discretization for Circle Fitting. In this section, we prove Theorem 4.1 by presenting a discretization for the circle fitting problem. Given the point set $P = \{p_1, \ldots, p_n\}$, we show that there exists a small set of circles $\mathcal{C}$ which satisfies the requirements stated in Theorem 4.1. We follow a similar approach as [31].

For any $\delta \in \mathbb{R}$ and any function $f : \mathbb{R}^3 \to \mathbb{R}$, the δ-level set of f, denoted by $\text{lev}(f,\delta)$, is the set of points in $\mathbb{R}^3$ satisfying $f(x,y,r) = \delta$:

$$\text{lev}(f,\delta) := \{(x,y,r) | f(x,y,r) = \delta\}.$$

Let $C_{x,y,r}$ denote the circle in the plane with center (x,y) and radius r. Let $\delta_i = i/n^2$, where $i = 1, \cdots, n^2+1$. For each pair of points $p_i = (x_i, y_i)$ and $p_j = (x_j, y_j)$, define a function $f_{i,j} : \mathbb{R}^3 \to \mathbb{R}$:

$$\begin{aligned} f_{i,j}(x,y,r) &:= \frac{\text{dist}(p_i, C_{x,y,r})}{\text{dist}(p_j, C_{x,y,r})} \\ &= \frac{|\sqrt{(x-x_i)^2+(y-y_i)^2} - r|}{|\sqrt{(x-x_j)^2+(y-y_j)^2} - r|}, \end{aligned}$$

which is the ratio of the distance from point p_i to the circle $C_{x,y,r}$ to the distance from point p_j to this circle. For each point p_i in P, we define a function $f_i : \mathbb{R}^3 \to \mathbb{R}$:

$$\begin{aligned} f_i(x,y,r) &:= \text{dist}(p_i, C_{x,y,r}) \\ &= |\sqrt{(x-x_i)^2+(y-y_i)^2} - r|, \end{aligned}$$

which is the distance from point p_i to the circle $C_{x,y,r}$. The level sets for the circle fitting problem are:

$$\begin{aligned} G :=& \{\text{lev}(f_{i,j}, \delta_k) \,|\, 1 \le i,j \le n, 1 \le k \le n^2+1\} \cup \\ & \{\text{lev}(f_i, 0) \,|\, 1 \le i \le n\}. \end{aligned}$$

Note that only points (x,y,r) with $r \ge 0$ correspond to circles as r is the radius of the circle $C_{x,y,r}$, thus it suffices to consider $\mathbb{R}^2 \times \mathbb{R}^+$. The level sets in G partitions this space. We denote the arrangement of level sets in G by $\mathcal{A}(G)$. Set $\delta_0 = 0$ and $\delta_{n^2+2} = \infty$. Consider a cell A in $\mathcal{A}(G)$ and suppose $p_j \in P$ is a point such that $\text{dist}(p_j, C_{x,y,r}) > 0$ for every $(x,y,r) \in A$. Let $p_i \in P$ be any point. It is not hard to see that there exists $0 \le k \le n^2+1$, such that $f_{i,j}(x,y,r) \in [\delta_k, \delta_{k+1})$ for all $(x,y,r) \in A$. This property of the arrangement is what we are after. Note however that a level set in G is not necessarily a zero set of a constant-degree polynomial, and hence we are not aware of ways of showing good upper bounds on the number of cells in $\mathcal{A}(G)$.

As we will see, each level set in G is closely related to the zero set of a constant degree polynomial. In the following, we compute a set G' that includes such a constant-degree polynomial, with variables x, y, r, corresponding to each level set in G. For technical reasons, G' will also include some other related polynomials. The arrangement (restricted to $\mathbb{R}^2 \times \mathbb{R}^+$) of the zero set of polynomials in G', denoted by $\mathcal{A}(G')$, has the property stated in Lemma 4.1. For the sake of exposition, the Lemma is stated before describing the actual polynomials in G'.

LEMMA 4.1. *Let C be a cell in $\mathcal{A}(G')$. Assume that $p_j \in P$ satisfies that $dist(C_{x,y,r}, p_j) > 0$, $\forall (x,y,r) \in C$. Fix p_i and p_j. Let $\delta_0 = 0$ and $\delta_{n^2+2} = \infty$. There exists an integer k, $0 \le k \le n^2+1$, such that $f_{i,j}(x,y,r) \in [\delta_k, \delta_{k+1})$, $\forall (x,y,r) \in C$.*

Once we have G', we compute the set $\mathcal{C}$ in Theorem 4.1 so that it includes at least one point from each cell of $\mathcal{A}(G')$; the size of $\mathcal{C}$ is $O(|G'|^3)$ [10]. In the following section, we describe the set of polynomials G' and prove Lemma 4.1. Subsequently, we show that $\mathcal{C}$ provides the guarantee in the statement of Theorem 4.1.

4.1.1 Computing the set G' of polynomials. Fix i, j and k. Observe that set $\text{lev}(f_{i,j}, \delta_k)$ is a subset of the zero set of a constant-degree polynomial in x, y and r (x_i, y_i, x_j, y_j and δ_k are constants): we have

$$\frac{|\sqrt{(x-x_i)^2+(y-y_i)^2} - r|}{|\sqrt{(x-x_j)^2+(y-y_j)^2} - r|} = \delta_k. \tag{4.3}$$

Multiply both sides by $|\sqrt{(x-x_j)^2+(y-y_j)^2} - r|$, we obtain

$$\begin{aligned} &|\sqrt{(x-x_i)^2+(y-y_i)^2} - r| = \\ &\delta_k |\sqrt{(x-x_i)^2+(y-y_i)^2} - r|. \end{aligned} \tag{4.4}$$

Squaring both sides, we obtain

$$\begin{aligned} &\left(\sqrt{(x-x_i)^2+(y-y_i)^2} - r\right)^2 = \\ &\delta_k^2 \left(\sqrt{(x-x_j)^2+(y-y_j)^2} - r\right)^2, \end{aligned} \tag{4.5}$$

Arrange the terms, we have

$$\begin{aligned} &(x-x_i)^2+(y-y_i)^2+r^2-\delta_k^2\left((x-x_j)^2+(y-y_j)^2+r^2\right) \\ &= 2r\sqrt{(x-x_i)^2+(y-y_i)^2} - 2\delta_k^2 r\sqrt{(x-x_j)^2+(y-y_j)^2} \end{aligned} \tag{4.6}$$

Copyright © SIAM.
Unauthorized reproduction of this article is prohibited.

Downloaded 05/02/16 to 109.230.67.85. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Squaring both sides, we have

$$\begin{aligned}&\Big[(x-x_i)^2+(y-y_i)^2+r^2-\\&\delta_k^2\left((x-x_j)^2+(y-y_j)^2+r^2\right)\Big]^2=\\&\Big(2r\sqrt{(x-x_i)^2+(y-y_i)^2}-\\&2\delta_k^2 r\sqrt{(x-x_j)^2+(y-y_j)^2}\Big)^2\end{aligned}\tag{4.7}$$

Arrange the terms, we have

$$\begin{aligned}&\Big[(x-x_i)^2+(y-y_i)^2+r^2-\\&\delta_k^2\left((x-x_j)^2+(y-y_j)^2+r^2\right)\Big]^2-\\&\Big[4r^2\left((x-x_i)^2+(y-y_i)^2\right)+\\&4\delta_k^4 r^2\left((x-x_j)^2+(y-y_j)^2\right)\Big]=\\&-8\delta_k^2 r^2\sqrt{(x-x_i)^2+(y-y_i)^2}\sqrt{(x-x_j)^2+(y-y_j)^2}\end{aligned}\tag{4.8}$$

Squaring both sides, we have

$$\begin{aligned}&\Big(\big[(x-x_i)^2+(y-y_i)^2+r^2-\\&\delta_k^2\left((x-x_j)^2+(y-y_j)^2+r^2\right)\big]^2-\\&\Big[4r^2\left((x-x_i)^2+(y-y_i)^2\right)+\\&4\delta_k^4 r^2\left((x-x_j)^2+(y-y_j)^2\right)\Big]\Big)^2=\\&64\delta_k^4 r^4\left((x-x_i)^2+(y-y_i)^2\right)\left((x-x_j)^2+(y-y_j)^2\right),\end{aligned}\tag{4.9}$$

Hence $\mathrm{lev}\,(f_{i,j},\delta_k)$ is a subset of the zero set of the polynomial $g_{i,j,k}(x,y,r)$:

$$\begin{aligned}&g_{i,j,k}(x,y,r)=\\&\Big(\big[(x-x_i)^2+(y-y_i)^2+r^2-\\&\delta_k^2\left((x-x_j)^2+(y-y_j)^2+r^2\right)\big]^2-\\&\big[4r^2\left((x-x_i)^2+(y-y_i)^2\right)+\\&4\delta_k^4 r^2\left((x-x_j)^2+(y-y_j)^2\right)\big]\Big)^2-\\&64\delta_k^4 r^4\left((x-x_i)^2+(y-y_i)^2\right)\left((x-x_j)^2+(y-y_j)^2\right)\end{aligned}$$

We add to G' the polynomials $g_{i,j,k}$ for each $1\le i,j\le n$ and $1\le k\le n^2+1$. From the derivation of $g_{i,j,k}$, it is easy to see that it is possible that some point $(x,y,r)\in\mathrm{zer}\,(g_{i,j,k})$, while $(x,y,r)\notin\mathrm{lev}\,(f_{i,j},\delta_k)$. As an example, consider Eq (4.8) and Eq (4.9): points satisfying Eq (4.8) also satisfy Eq (4.9); however, points satisfying the following equation satisfy Eq (4.9), do not satisfy Eq (4.8) unless $x=x_i$ and $y=y_i$ or $x=x_j$ and $y=y_j$.

$$\begin{aligned}&\big[(x-x_i)^2+(y-y_i)^2+r^2-\\&\delta_k^2\left((x-x_j)^2+(y-y_j)^2+r^2\right)\big]^2-\\&\big[4r^2\left((x-x_i)^2+(y-y_i)^2\right)+\\&4\delta_k^4 r^2\left((x-x_j)^2+(y-y_j)^2\right)\big]=\\&8\delta_k^2 r^2\sqrt{(x-x_i)^2+(y-y_i)^2}\sqrt{(x-x_j)^2+(y-y_j)^2},\end{aligned}$$

For this technical reason, we now add to G' an extra set of polynomials associated with the level set $\mathrm{lev}\,(f_{i,j},\delta_k)$.

$$\begin{aligned}g'_j(x,y,r):=&(x-x_j)^2+(y-y_j)^2-r^2.\\g''_{i,j,k}(x,y,r):=&(x-x_i)^2+(y-y_i)^2+r^2-\\&\delta_k^2\left((x-x_j)^2+(y-y_j)^2+r^2\right).\\g'''_{i,j,k}(x,y,r):=&r^2\left((x-x_i)^2+(y-y_i)^2\right)-\\&\delta_k^4\left((x-x_j)^2+(y-y_j)^2\right).\\g''''_{i,j,k}(x,y,r):=&\Big[(x-x_i)^2+(y-y_i)^2+r^2-\\&\delta_k^2\left((x-x_j)^2+(y-y_j)^2+r^2\right)\Big]^2-\\&\Big[4r^2\left((x-x_i)^2+(y-y_i)^2\right)+\\&4\delta_k^4 r^2\left((x-x_j)^2+(y-y_j)^2\right)\Big]\end{aligned}$$

Note that $\mathrm{lev}\,(f_j,0)=\mathrm{zer}\left(g'_j\right)$, hence we do not need to add polynomials for level sets in $\{\mathrm{lev}\,(f_i,0)\,|1\le i\le n\}$ again.

Now we prove Lemma 4.1.

Proof. We prove the lemma by contradiction. Suppose the statement is false. Then there exists k, such that

$$f_{i,j}(x,y,r)<\delta_k\le f_{i,j}(x',y',r'),$$

for some (x,y,r) and (x',y',r') in C. Since $f_{i,j}$ is a continuous function in this cell C, there exists a point $(x'',y'',r'')\in C$, such that $f_{i,j}(x'',y'',r'')=\delta_k$. Then $g_{i,j,k}(x'',y'',r'')=0$. Since (x,y,r) and (x'',y'',r'') are in the same cell, $g_{i,j,k}(x,y,r)=0$. Since $f_{i,j}(x'',y'',r'')=\delta_k$, $g''''_{i,j,k}(x'',y'',r'')\le 0$. Thus $g''''_{i,j,k}(x,y,r)\le 0$. Thus (x,y,r) satisfies Eq (4.8), thus also Eq (4.7). According to Eq (4.6), $g''_{i,j,k}(x'',y'',r'')$ and $g'''_{i,j,k}(x'',y'',r'')$ are both nonnegative or negative, thus $g''_{i,j,k}(x,y,r)$ and $g'''_{i,j,k}(x,y,r)$ are also both

Copyright © SIAM.
Unauthorized reproduction of this article is prohibited.

Downloaded 05/02/16 to 109.230.67.85. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

nonnegative or negative, hence (x, y, r) also satisfies Eq (4.6), which implies (x, y, r) satisfies Eq (4.4). Since $\text{dist}((x, y, r), p_j) \neq 0$, we have $f_{i,j}(x, y, r) = \delta_k$, which contradicts the assumption that $f_{i,j}(x, y, r) < \delta_k$. □

The number of polynomials $g_{i,j,k}$, $g''_{i,j,k}$, $g'''_{i,j,k}$ and $g''''_{i,j,k}$ is $O(n^4)$; there are $n(n-1)$ pairs of distinct points, and for each pair, there are n^2+1 choices of δ_k. The number of polynomials of g'_j is n. Hence, the total number of polynomials in G' is $O(n^4)$. An application of the results in [10] implies that there exists a subset $\mathcal{C}$ of $\mathbb{R}^3$ of cardinality $O(n^{12})$, such that each cell of the arrangement $\mathcal{A}(G')$ contains at least one point from $\mathcal{C}$.

4.1.2 $\mathcal{C}$ is a good discretization. Let $S \subset P$ be a weighted subset satisfying both condition (a) and (b) in Theorem 4.1, that is, (a) $|\text{cost}(S, C) - \text{cost}(P, C)| \leq \varepsilon\text{cost}(P, C)$ for every circle $C \in \mathcal{C}$; (b) the overall weights of points in S, $\sum_{p\in S} w_p$, is at most $2n$. We now show that S approximates P with respect to every circle on the plane. Consider a circle C' with center (x', y') and radius r'. Suppose (x', y', r') is in a cell A of the arrangement $\mathcal{A}(G')$. Suppose $C \in \mathcal{C}$ with center (x, y) and radius r is the circle such that (x', y', r') and (x, y, r) are in the same cell A. In other words, C and C' are in the same set of circles corresponding to the cell A. We show that if the sample S approximates P with respect to C, then S also approximates P with respect to C', as long as the overall weights of points in the sample S is $O(n)$.

If for every point p in the input point set P, $\text{dist}(p, C') = 0$, that is, all the points are incident on the circle C', then trivially the sampling error $|\text{cost}(P, C') - \text{cost}(S, C')|$ is zero, since $\text{cost}(P, C')$ and $\text{cost}(S, C')$ are both zero. Therefore, we may assume that there exists some point not incident on C'. In particular, let p_{i^*} be a furthest point from C', then we have

$$\text{dist}(p_{i^*}, C') = \max_{p\in P} \text{dist}(p, C') > 0.$$

We have

$$\begin{aligned}(4.10)\qquad &\left|\frac{\text{cost}(P,C')}{\text{dist}(p_{i^*},C')} - \frac{\text{cost}(S,C')}{\text{dist}(p_{i^*},C')}\right| \leq \\ &\left|\frac{\text{cost}(P,C')}{\text{dist}(p_{i^*},C')} - \frac{\text{cost}(P,C)}{\text{dist}(p_{i^*},C)}\right| + \\ &\left|\frac{\text{cost}(P,C)}{\text{dist}(p_{i^*},C)} - \frac{\text{cost}(S,C)}{\text{dist}(p_{i^*},C)}\right| + \\ &\left|\frac{\text{cost}(S,C)}{\text{dist}(p_{i^*},C)} - \frac{\text{cost}(S',C)}{\text{dist}(p_{i^*},C)}\right|,\end{aligned}$$

where by definition,

$$\text{cost}(P, C) = \sum_{p\in P} \text{dist}(p, C),$$
$$\text{cost}(S, C) = \sum_{p\in S} w_p \text{dist}(p, C).$$

Consider the first addend on the right-hand side of Eq (4.10).

$$\begin{aligned}(4.11)\qquad &\left|\frac{\text{cost}(P,C')}{\text{dist}(p_{i^*},C')} - \frac{\text{cost}(P,C)}{\text{dist}(p_{i^*},C)}\right| \\ &\leq \sum_{i=1}^{n} \left|\frac{\text{dist}(p_i,C')}{\text{dist}(p_{i^*},C')} - \frac{\text{dist}(p_i,C)}{\text{dist}(p_{i^*},C)}\right| \\ &= \sum_{i=1}^{n} |f_{i,i^*}(x', y', r') - f_{i,i^*}(x, y, r)| \leq n \cdot \frac{1}{n^2} = \frac{1}{n}.\end{aligned}$$

The last inequality follows from the fact that (x, y, r) and (x', y', r') are in the same cell of $\mathcal{A}(G')$ and Lemma 4.1.

Following a similar argument, we have the following upper bound of the third addend:

$$\begin{aligned}&\left|\frac{\text{cost}(S,C)}{\text{dist}(p_{i^*},C)} - \frac{\text{cost}(S,C')}{\text{dist}(p_{i^*},C')}\right| \\ &\leq \sum_{p\in S} w_p \cdot \left|\frac{\text{dist}(p,C)}{\text{dist}(p_{i^*},C)} - \frac{\text{dist}(p,C')}{\text{dist}(p_{i^*},C')}\right| \\ &\leq \left(\sum_{p\in S} w_p\right) \cdot \frac{1}{n^2} \leq \frac{2}{n}.\end{aligned}$$

The last inequality follows from the assumption that $\sum_{p\in S} w_p \leq 2n$.

Consider the second addend in Eq (4.10). By assumption, S approximates P with respect to C as $C \in \mathcal{C}$, that is,

$$|\text{cost}(P, C) - \text{cost}(S, C)| \leq \epsilon\text{cost}(P, C).$$

Hence, dividing both sides by $\text{dist}(p_{i^*}, C)$, we obtain

$$\left|\frac{\text{cost}(P,C)}{\text{dist}(p_{i^*},C)} - \frac{\text{cost}(S,C)}{\text{dist}(p_{i^*},C)}\right| \leq \epsilon \cdot \frac{\text{cost}(P,C)}{\text{dist}(p_{i^*},C)}$$

Further, since

$$\frac{\text{cost}(P,C)}{\text{dist}(p_{i^*},C)} \leq \frac{\text{cost}(P,C')}{\text{dist}(p_{i^*},C')} + n \cdot \frac{1}{n^2},$$

we have

$$(4.12)\qquad \left|\frac{\text{cost}(P,C)}{\text{dist}(p_{i^*},C)} - \frac{\text{cost}(S,C)}{\text{dist}(p_{i^*},C)}\right| \leq \epsilon \cdot \frac{\text{cost}(P,C')}{\text{dist}(p_{i^*},C')} + \epsilon \cdot \frac{1}{n}$$

Copyright © SIAM.
Unauthorized reproduction of this article is prohibited.

Downloaded 05/02/16 to 109.230.67.85. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Eq (4.10), Eq (4.11) and Eq (4.12) together imply that

$$\left|\frac{\text{cost}(P,C')-\text{cost}(S,C')}{\text{dist}(p_{i^*},C')}\right| \le \epsilon\cdot\frac{\text{cost}(P,C')}{\text{dist}(p_{i^*},C')}+\frac{3+\epsilon}{n}$$
$$\le \epsilon\cdot\frac{\text{cost}(P,C')}{\text{dist}(p_{i^*},C')}+\frac{4}{n}$$

Multiply both sides by $\text{dist}(p_{i^*},C')$, we obtain

$$|\text{cost}(P,C')-\text{cost}(S,C')| \le \epsilon\text{cost}(P,C')+\frac{4\text{dist}(p_{i^*},C')}{n}$$
$$\le \epsilon\text{cost}(P,C')+4\epsilon\cdot\text{cost}(P,C')=5\epsilon\text{cost}(P,C'),$$

where the second inequality follows because $1/n \le \epsilon$.

4.2 Discretization for Projective Clustering. In this section, we present the proof of Theorem 4.2, the discretization theorem for projective clustering.

The family of shapes are j k-flats, where a k-flat is a subspace spanned by k vectors in $\mathbb{R}^d$ translated by a point in $\mathbb{R}^d$. To be precise, the k-flat determined by $k+1$ vectors $\mathbf{v_1}, \mathbf{v_1}, \cdots, \mathbf{v_{k+1}}$ is

$$F_{\mathbf{v_1},\cdots,\mathbf{v_{k+1}}} = \mathbf{v_{k+1}} + \text{span}\,(\mathbf{v_1},\cdots,\mathbf{v_k})$$
$$= \left\{\mathbf{v_{k+1}} + \sum_{i=1}^{k} a_i\mathbf{v_i}\middle| a_i\in\mathbb{R}, i=1,\cdots,k\right\},$$

where $\mathbf{v_i}$, $i=1,\cdots,k+1$ are vectors in $\mathbb{R}^d$. The set of k-flats is

$$\mathcal{K}=\{F_{\mathbf{v_1},\cdots,\mathbf{v_{k+1}}}|\langle\mathbf{v_i},\mathbf{v_j}\rangle=0, 1\le i<j\le k, \langle\mathbf{v_i},\mathbf{v_i}\rangle=1, i=1,\cdots,k\}.$$

Observe that we may parameter each k-flat with $(k+1)d$ real numbers by considering each vector $\mathbf{v_i}$ as a sequence of d reals numbers: given a sequence of real numbers $(x_{u,v})$, $1\le u\le d$, $1\le v\le k+1$, the vector $\mathbf{v_i}$ is $(x_{1,i},\cdots,x_{d,i})$, $i=1,\cdots,k+1$. The k-flat corresponding to this sequence $(x_{u,v})$ is the k-flat corresponding to the $k+1$ vectors obtained from this sequence. Since we parameter each k-flat with k orthogonal unit vectors, we only need to consider a subset U of $\mathbb{R}^{(k+1)d}$. The family of shapes is $\mathcal{K}^j$, which consists of all j-tuples of k-flat. Therefore, we may parameter each shape, $\{F_1,\cdots,F_j\}$, $F_i\in\mathcal{K}$ by a sequence of $j(k+1)d$ real numbers in U^j.

For convenience of notation, we denote the point in $\mathbb{R}^{j(k+1)d}$ as $(x_{u,v,w})$, where $1\le u\le d$, $1\le v\le k+1$, $1\le w\le j$. The i^{th} k-flat, denote by by $h^i_{x_{1,1,1},\cdots,x_{d,k+1,j}}$, is determined by $x_{u,v,i}$, where $1\le u\le d$ and $1\le v\le k+1$: we obtain $k+1$ vectors in $\mathbb{R}^d$, which are

$$\begin{bmatrix}x_{1,1,i}\\x_{2,1,i}\\\vdots\\x_{d,1,i}\end{bmatrix}, \begin{bmatrix}x_{1,2,i}\\x_{2,2,i}\\\vdots\\x_{d,2,i}\end{bmatrix}, \cdots, \begin{bmatrix}x_{1,k+1,i}\\x_{2,k+1,i}\\\vdots\\x_{d,k+1,i}\end{bmatrix}.$$

The k-flat $h^i_{x_{1,1,1},\cdots,x_{d,k+1,j}}$ is

$$h^i_{x_{1,1,1},\cdots,x_{d,k+1,j}} = \left\{\begin{bmatrix}x_{1,k+1,i}\\x_{2,k+1,i}\\\vdots\\x_{d,k+1,i}\end{bmatrix} + \sum_{v=1}^{k} a_v\begin{bmatrix}x_{1,v,i}\\x_{2,v,i}\\\vdots\\x_{d,v,i}\end{bmatrix}\middle| a_v\in\mathbb{R}\right\}.$$

The distance from a point $p=(p_1,\cdots,p_d)$ in $\mathbb{R}^d$ to the above k-flat is

$$\text{dist}(p,h^i_{x_{1,1,1},\cdots,x_{d,k+1,j}}) = \left\|\begin{bmatrix}p_1-x_{1,k+1,i}\\p_2-x_{2,k+1,i}\\\vdots\\p_d-x_{d,k+1,i}\end{bmatrix} - \sum_{v=1}^{k}\left\langle\begin{bmatrix}p_1-x_{1,k+1,i}\\p_2-x_{2,k+1,i}\\\vdots\\p_d-x_{d,k+1,i}\end{bmatrix},\begin{bmatrix}x_{1,v,i}\\x_{2,v,i}\\\vdots\\x_{d,v,i}\end{bmatrix}\right\rangle\begin{bmatrix}x_{1,v,i}\\x_{2,v,i}\\\vdots\\x_{d,v,i}\end{bmatrix}\right\|_2$$

The distance from a point $p\in\mathbb{R}^d$ to j k-flats determined by $\{x_{u,v,w}\}$ is the minimum distance from the point to one of the flats:

$$\min_{1\le i\le j}\text{dist}(p,h^i_{x_{1,1,1},\cdots,x_{d,k+1,j}}).$$

Therefore, once we obtain a partition of the space $\mathbb{R}^{j(k+1)d}$, such that for each cell $A\cap U^j$, there exists $p,q\in P$ and $0\le l\le n^2+1$, with

$$\frac{\min_{1\le i\le j}\text{dist}(p,h^i_{x_{1,1,1},\cdots,x_{d,k+1,j}})}{\min_{1\le i'\le j}\text{dist}(q,h^{i'}_{x_{1,1,1},\cdots,x_{d,k+1,j}})}\in[\delta_l,\delta_{l+1}),$$
$$\forall(x_{1,1,1},\cdots,x_{d,k+1,j})\in A\cap U^j,$$

then we can obtain the discretization $\mathcal{F}'$ in Theorem 4.2 by picking one point from each $A\cap U^j$. The set $\mathcal{F}'$ is a good discretization, following a similar reasoning as discretization of the circle fitting problem.

It seems difficult to find a polynomial with variables $x_{1,1,1},\cdots,x_{d,k+1,j}$ whose zero set contains $\text{lev}\,(g_{p,q},\delta_k)$, where $g_{p,q}$ is defined as

$$g_{p,q}(x_{1,1,1},\cdots,x_{d,k+1,j}) := \frac{\min_{1\le i\le j}\text{dist}(p,h^i_{x_{1,1,1},\cdots,x_{d,k+1,j}})}{\min_{1\le i'\le j}\text{dist}(q,h^{i'}_{x_{1,1,1},\cdots,x_{d,k+1,j}})}.$$

Hence we do not use level sets defined by

$$g_{p,q}(x_{1,1,1},\cdots,x_{d,k+1,j})=\delta_k, k=1,\cdots,n^2+1.$$

Instead, we consider the family of functions $g_{p,q,i,i'}:\mathbb{R}^{j(k+1)d}\to\mathbb{R}$, $p,q\in P$, $1\le i,i'\le j$, which is defined

Copyright © SIAM.
Unauthorized reproduction of this article is prohibited.

Downloaded 05/02/16 to 109.230.67.85. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

below:

$$g_{p,q,i,i'}(x_{1,1,1},\cdots,x_{d,k+1,j}) := \frac{\mathrm{dist}(p, h^i_{x_1,\cdots,x_{j(k+1)d}})}{\mathrm{dist}(q, h^{i'}_{x_1,\cdots,x_{j(k+1)d}})}.$$

The level sets are $\{\mathrm{lev}\,(g_{p,q,i,i'}, \delta_k)\,|p,q \in P, 1 \le i,i' \le j\}$. The intuition is that in a subset of U^j, if $g_{p,q,i,i'}$ only changes slightly for each pair of flats, then the ratio $g_{p,q}$ also changes slightly. This fact is proved in the Lemma 4.2.

LEMMA 4.2. *Define $\delta_0 = 0$ and $\delta_{n^2+2} = \infty$. Consider a subset A of U^j, and a $q \in P$ such that the distance from q to any j-tuple of k-flats determined by points in A is nonzero (that is, q is not contained in any shape determined by points in A). Let $p \in P$ be any point distinct from q. Suppose that for every $1 \le i,i' \le j$, it holds that for some integer $l = l_{i,i'}$, where $0 \le l \le n^2+1$,*

$$g_{p,q,i,i}(x_{1,1,1},\cdots,x_{d,k+1,j}) \in [\delta_l, \delta_{l+1})$$
$$\forall (x_{1,1,1},\cdots,x_{d,k+1,j}) \in A$$

(note that l depends on the choice of i,i'). Then there exists an integer $0 \le l \le n^2+1$ so that, for every point $(x_1,\cdots,x_{j(k+1)d})$ in A,

$$g_{p,q}(x_{1,1,1},\cdots,x_{d,k+1,j}) = \frac{\min_{1\le i\le j} dist(p, h^i_{x_1,\cdots,x_{j(k+1)d}})}{\min_{1\le i\le j} dist(q, h^i_{x_1,\cdots,x_{j(k+1)d}})} \in [\delta_l, \delta_{l+1}).$$

Proof. We prove the lemma by contradiction. Suppose there does not exist such an l, then there exists an integer t, $1 \le t \le n^2+1$, such that for some two points $\mathbf{x} = (x_1,\cdots,x_{j(k+1)d})$ and $\mathbf{y} = (y_1,\cdots,y_{j(k+1)d})$ in A, it holds that $g_{p,q}(\mathbf{x}) < \delta_t$ while $g_{p,q}(\mathbf{y}) \ge \delta_t$. Our goal is to derive the fact that there would exist two integers l' and l, such that

$$\frac{\mathrm{dist}(p, h^{l'}(\mathbf{x}))}{\mathrm{dist}(q, h^l(\mathbf{x}))} < \delta_t,$$

while

$$\frac{\mathrm{dist}(p, h^{l'}(\mathbf{y}))}{\mathrm{dist}(q, h^l(\mathbf{y}))} \ge \delta_t,$$

which violates the assumption that $g_{p,q,l,l'}(x_{1,1,1},\cdots,x_{d,k+1,j}) \in [\delta_s, \delta_{s+1})$ for some integer s.
Choose the index l' so that

$$\min_{1\le i\le j} \mathrm{dist}(p, h^i_{\mathbf{x}}) = \mathrm{dist}(p, h^{l'}_{\mathbf{x}}).$$

and choose l so that

$$\min_{1\le i\le j} \mathrm{dist}(q, h^i_{\mathbf{y}}) = \mathrm{dist}(q, h^l_{\mathbf{y}}).$$

Then

$$\frac{\mathrm{dist}(p, h^{l'}_{\mathbf{x}})}{\mathrm{dist}(q, h^l_{\mathbf{x}})} \le \frac{\mathrm{dist}(p, h^{l'}_{\mathbf{x}})}{\min_{1\le i\le j} \mathrm{dist}(q, h^i_{\mathbf{x}})} = \frac{\min_{1\le i\le j} \mathrm{dist}(p, h^i_{\mathbf{x}})}{\min_{1\le i\le j} \mathrm{dist}(q, h^i_{\mathbf{x}})} < \delta_t,$$

$$\frac{\mathrm{dist}(p, h^{l'}_{\mathbf{y}})}{\mathrm{dist}(q, h^l_{\mathbf{y}})} \ge \frac{\min_{1\le i\le j} \mathrm{dist}(p, h^i_{\mathbf{y}})}{\mathrm{dist}(q, h^l_{\mathbf{y}})} = \frac{\min_{1\le i\le j} \mathrm{dist}(p, h^i_{\mathbf{y}})}{\min_{1\le i\le j} \mathrm{dist}(q, h^i_{\mathbf{y}})} \ge \delta_t.$$

□

We now describe the complete collection of level sets. For each point p in P and i, define the function $g_{p,i} : \mathbb{R}^{j(k+1)d} \to \mathbb{R}$:

$$g_{p,i}(x_{1,1,1},\cdots,x_{d,k+1,j}) := \mathrm{dist}(p, h^i_{x_1,\cdots,x_{j(k+1)d}}).$$

Define $\delta_k = k/n^2$, $k = 1,\cdots,n^2+1$. The set of level sets are:

$$G = \{\mathrm{lev}\,(g_{p,q,i,i'}, \delta_k)\,|p,q \in P, p \ne q, 1 \le i,i' \le j, 1 \le k \le n^2+1\} \cup \{\mathrm{lev}\,(g_{p,i}, 0)\,|p \in P, 1 \le i \le j\}.$$

Following a similar approach as the discretization for circle fitting problem, we do not consider the arrangement of the level sets in G directly; instead we compute a collection G' of constant-degree polynomials, with variables $x_{1,1,1},\cdots,x_{d,k+1,j}$, such that Lemma 4.3 holds. Let $\mathcal{A}(G')$ denote the arrangement (restricted to U^j) of the zero sets of the polynomials in G'.

LEMMA 4.3. *Let C be a cell in $\mathcal{A}(G')$. Assume that $q \in P$ and an integer $i' \in [1,j]$ satisfies that $g_{q,i'}(x_{1,1,1},\cdots,x_{d,k+1,j}) > 0$, $\forall (x_{1,1,1},\cdots,x_{d,k+1,j}) \in C$. Fix $p \in P$ and $i \in [1,j]$. Let $\delta_0 = 0$ and $\delta_{n^2+2} = \infty$. There exists an integer l, $0 \le l \le n^2+1$, such that $g_{p,q,i,i'}(x_{1,1,1},\cdots,x_{d,k+1,j}) \in [\delta_l, \delta_{l+1})$, $\forall (x_{1,1,1},\cdots,x_{d,k+1,j}) \in C$.*

We now compute the collection of polynomials in G'. By definition,

$$g_{p,q,i,i'}(x_{1,1,1},\cdots,x_{d,k+1,j}) = g_{p,i}(x_{1,1,1},\cdots,x_{d,k+1,j})/g_{q,i'}(x_{1,1,1},\cdots,x_{d,k+1,j}).$$

Note that $(g_{p,i}(x_{1,1,1},\cdots,x_{d,k+1,j}))^2$ and $(g_{q,i'}(x_{1,1,1},\cdots,x_{d,k+1,j}))^2$ are polynomials of $x_{u,v,w}$, $1 \le u \le d$, $1 \le v \le k+1$, $1 \le w \le j$, the level set $\mathrm{lev}\,(g_{p,q,i,i'}, \delta_k)$ is a subset of the zero set of a polynomial of $j(k+1)d$ variables:

$$P_{p,q,i,i',k}(x_{1,1,1},\cdots,x_{d,k+1,j}) := (g_{p,i}(x_{1,1,1},\cdots,x_{d,k+1,j}))^2 - \delta_k^2 (g_{q,i'}(x_{1,1,1},\cdots,x_{d,k+1,j}))^2.$$

Copyright © SIAM.
Unauthorized reproduction of this article is prohibited.

The level set $\text{lev}(g_{p,i}, 0)$ is the zero set of the polynomial $(g_{p,i}(x_{1,1,1}, \cdots, x_{d,k+1,j}))^2$. Let

$$G' := \left\{P_{p,q,i,i',k} | p, q \in P, 1 \le i, i' \le j, 1 \le k \le n^2 + 1\right\} \cup \left\{(g_{p,i}(x_{1,1,1}, \cdots, x_{d,k+1,j}))^2 | p \in P, 1 \le i \le j\right\}.$$

Following an analogous argument as the proof of Lemma 4.1[1], Lemma 4.3 can be shown. There are $O(n^4)$ polynomials of constant degree, hence the number of cells in the arrangement of $\mathcal{A}(G')$ is $O(n^{4j(k+1)d})$ [10].

The rest of the proof of Theorem 4.2 – showing that any weighted subset $S \subseteq P$ satisfying conditions (a) and (b) in the statement of the theorem is an L_1 5ε-coreset of P – follows by a similar argument as for the circle fitting problem, using Lemma 4.2.

5 Projective Clustering in High Dimensions

We consider the integer (j, k) projective clustering problem $(\mathbb{R}^m, \mathcal{F})$. Let $P \subseteq \mathbb{R}^m$ be an input instance of n points with integer coordinates of magnitude at most $\Delta = (mn)^{10}$, and $0 < \varepsilon < 1$ be a parameter. We describe an algorithm that runs in $O(mn(\log(mn))^{O(1)})$ and returns a shape $F \in \mathcal{F}$ (a union of j k-flats) that with probability at least a constant is nearly optimal: $\text{cost}(P, F) \le (1 + \varepsilon)\text{cost}(P, F')$ for any $F' \in \mathcal{F}$. Note that we consider j and k constants but the dimension m as part of the input. We have used m rather than d to denote the dimension of the host space to emphasize that here, unlike in the last two sections, it is not a constant. For simplicity, we assume that the shape we are trying to fit is a union of j linear k-subspaces in $\mathbb{R}^m$, as opposed to a union of affine subspaces.

The result is obtained in three steps. First, we use a known dimension reduction result to reduce the problem to a (j, k) projective clustering in constant dimension. To solve the projective clustering problem in constant dimension, we compute a small coreset using essentially Theorem 4.3. In the third step, we solve the projective clustering problem on the coreset nearly optimally in time polynomial in the size of the coreset.

5.1 Dimension reduction. Using the algorithm of Deshpande and Varadarajan [15], we compute in time $nm\left(\frac{kj}{\varepsilon}\right)^{O(1)}$ a linearly independent subset $\{a_1, a_2, \ldots, a_{d'}\} \subseteq P$ whose span contains (with probability at least 0.9) a shape $F \in \mathcal{F}$ such that $\text{cost}(P, F) \le (1 + \varepsilon)\text{cost}(P, F')$ for any $F' \in \mathcal{F}$. Here, $d' = \left(\frac{kj}{\varepsilon}\right)^{O(1)}$ is a constant. Let V denote the subspace spanned by $\{a_1, a_2, \ldots, a_{d'}\}$. It now suffices to solve the following problem nearly optimally: among the shapes in $\mathcal{F}$ that are *contained* in V, find the one that minimizes $\text{cost}(P, \cdot)$.

Fix $b \in \mathbb{R}^m$ orthogonal to V. For $p \in P$, let $\bar{p}$ denote the orthogonal projection of p onto V and $p^\perp$ the projection of p onto the orthogonal complement of V. For $p \in P$, let $p' = \bar{p} + ||p^\perp||_2 b$, and let $P' = \{p' \mid p \in P\}$. Observe that $\text{cost}(P, F) = \text{cost}(P', F)$ for any $F \in \mathcal{F}$ that is contained in V. It therefore suffices to solve the following problem nearly optimally: among the shapes in $\mathcal{F}$ that are *contained* in V, find the one that minimizes $\text{cost}(P', \cdot)$. This is a (j, k) projective clustering problem in $d' + 1$ dimensions, except for the additional constraint that the shape must lie in the d'-dimensional subspace V.

5.2 Computing a Coreset. Our next step is to compute an L_1 ε-coreset Q for P' using Theorem 4.3, treating P' as a point set in $d' + 1$ dimensions. For any $p' \in P'$, we have $||p'||_2 = ||p||_2 \le \sqrt{m}\Delta$; however, the coordinates of p' when expressed in terms of an orthonormal basis for the subspace spanned by V and b are not necessarily integers. So we have to address this technicality before applying Theorem 4.3. This is not hard to do given the following lemma.

LEMMA 5.1. *Let F be an optimal solution for the (j, k) projective clustering problem on the point set P. If $\text{cost}(P, F) > 0$, then $\text{cost}(P, F) > \frac{1}{(m\Delta)^c}$, for some constant c that depends only on k.*

Proof. We first need the following observation.

FACT 5.1. *Let $\{p_1, p_2, \ldots, p_{k+1}\}$ be any linearly independent subset of P. The $(k+1)$-dimensional volume of the simplex spanned by this subset is at least $\frac{1}{((k+1)!)^2}$.*

Proof. Let A be the $(k + 1) \times m$ matrix whose rows are the vectors p_i. Then the volume of the simplex in question is $\frac{1}{((k+1)!)^2}\det(AA^T)$. The matrix AA^T has entries that are all integers. □

Suppose that F, the optimal solution is a union of the j k-subspaces $f_1, f_2, \ldots, f_j$. Let $P_1, \ldots, P_j$ be the partition of P obtained by assigning each point in P to the nearest of these j subspaces. Assuming $\text{cost}(P, F) > 0$, at least one of the sets, say P_i, contains $(k + 1)$ linearly independent points $\{q_1, \ldots, q_{k+1}\}$. Let

[1] Note that $\text{zer}\left(P_{p,q,i,i',k}\right)$ indeed contains points that are not in $\text{lev}\left(g_{p,q,i,i'}, \delta_k\right)$, which are points satisfying $g_{p,i}(x_{1,1,1}, \cdots, x_{d,k+1,j}) = -\delta_k g_{q,i'}(x_{1,1,1}, \cdots, x_{d,k+1,j})$, or points satisfying $g_{p,i}(x_{1,1,1}, \cdots, x_{d,k+1,j}) = g_{q,i'}(x_{1,1,1}, \cdots, x_{d,k+1,j}) = 0$. Since $g_{p,i}$ and $g_{q,i'}$ are non-negative functions, and $\delta_k > 0$, if $g_{p,i}(x_{1,1,1}, \cdots, x_{d,k+1,j}) = -\delta_k g_{q,i'}(x_{1,1,1}, \cdots, x_{d,k+1,j})$, then $g_{p,i}(x_{1,1,1}, \cdots, x_{d,k+1,j}) = g_{q,i'}(x_{1,1,1}, \cdots, x_{d,k+1,j}) = 0$, which is in the zero set of $(g_{q,i'})^2$.

Copyright © SIAM.
Unauthorized reproduction of this article is prohibited.

Downloaded 05/02/16 to 109.230.67.85. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 05/02/16 to 109.230.67.85. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

f_i' be a k-subspace in the span of $\{q_1, \ldots, q_{k+1}\}$ that contains the projection of f_i on this span. Then the set $\{q_1, \ldots, q_{k+1}\}$ is contained in a $(k+1)$-dimensional box, k of whose sides have length $2\max_{t=1}^{k+1} ||q_t||_2$ and whose $(k+1)$-th side has length $2\max_{t=1}^{k+1} \text{dist}(q_t, f_i')$. This box must contain the simplex spanned by $\{q_1, \ldots, q_{k+1}\}$, so we have:

$$\frac{1}{((k+1)!)^2} \leq 2^{k+1}(\max_{t=1}^{k+1} ||q_t||_2)^k \cdot \max_{t=1}^{k+1} \text{dist}(q_t, f_i')$$
$$\leq (2\Delta m)^{k+1} \max_{t=1}^{k+1} \text{dist}(q_t, f_i').$$

The lemma follows from the above inequality by observing that $\text{cost}(P, F) \geq \max_{t=1}^{k+1} \text{dist}(q_t, f_i') \geq \max_{t=1}^{k+1} \text{dist}(q_t, f_i)$. □

If $\text{cost}(P, F) = 0$ for the optimal $F \in \mathcal{F}'$, then this must be true for some $F' \in \mathcal{F}$ that is contained in V as well. This means that P itself must be contained in V. In this case, such an F' can be found by applying the method of [16] for shape fitting in the L_∞ sense.

Let us therefore consider the case where $\text{cost}(P, F) > \frac{1}{(m\Delta)^c}$ for the optimal $F \in \mathcal{F}'$. In this case, we express the points in P' in terms of an orthogonal basis for the span of V and b, but round the coordinates of each point in P' to the nearest multiple of $\frac{1}{(mn\Delta)^{c_1}}$ where $c_1 > c$ is a sufficiently large integer. We now scale so that the coordinates of points in P' are integers. Note that the magnitude of the largest coordinate is $(mn\Delta)^{O(1)}$.

Now, treating P' as an input to the integer (j, k) projective clustering problem in $(\mathbb{R}^{d'+1}, \cdot)$ we compute a coreset Q using Theorem 4.3. The running time for this step is $n(\log mn)^{O(1)}$.

5.3 Solving the Problem on the Coreset. We need to find a shape F that is contained in V such that $\text{cost}(Q, F) \leq (1+\varepsilon)\text{cost}(Q, F')$ for any shape F' contained in V. Since the size of Q is $(\log(mn))^{O(1)}$, we can afford to use a generic polynomial time algorithm for this. For example, we can consider the discretization for Q similar to that in Theorem 4.2, but this time we actually compute it. We omit the details from this version, and conclude with our main result:

THEOREM 5.1. *Let P be an n-point instance of the integer (j, k)-projective clustering problem $(\mathbb{R}^m, \mathcal{F})$ (the largest magnitude of any coordinate for a point in P is at most $(mn)^{10}$), and $\varepsilon > 0$ be a parameter. There is a randomized algorithm that runs in time $mn(\log(mn))^{O(1)}$ and returns a shape $F \in \mathcal{F}$ such that with constant probability, $\text{cost}(P, F) \leq (1+\varepsilon)\text{cost}(P, F')$ for any $F' \in \mathcal{F}$. Here, j and k are constants but m is not.*

6 Conclusions

We conclude with some remarks on the work described here and directions for future work.

- For the shape fitting problems considered in this article, we have assumed that the distance $\text{dist}(p, F)$ between a point $p \in \mathbb{R}^d$ and shape $F \in \mathcal{F}$ is the minimum Euclidean distance from p to a point in the shape F. The results readily generalize to the case where $\text{dist}(p, F)$ is defined to be the τ-th power of the minimum distance, for $\tau > 0$.
- The results in Theorem 4.3 on small coresets for circle fitting and projective clustering in fixed dimension imply, via techniques that are now standard, that such small coresets can be maintained in an insertion-only streaming setting using small space.
- One direction for future work is on improving the parameters in the connection between L_∞ coresets and sensitivity that is made in Lemma . In this context, we note that to apply the Lemma, it is enough to have an L_∞ δ-coreset for some $0 < \delta < 1$ that is closer to 1 than 0. For some problems, this allows one to get around the difficulty that L_∞ coresets tend to have exponential dependence on the dimension. This fact is illustrated by the j-median example following the Lemma.
- Perhaps the most interesting open problem raised is whether a near-linear algorithm for (j, k) projective clustering is possible in high dimensions, without making the extra assumption that points have integer coordinates that are polynomially bounded. Another question is whether a small L_1 coreset exists for the integer projective clustering problem in high dimensions; note that our work establishes such coresets in constant dimension.

Acknowledgements. We thank the reviewers for insightful feedback.

References

[1] Pankaj K. Agarwal, Sariel Har-peled, Kasturi, and R. Varadarajan. Geometric approximation via coresets. In *Combinatorial and Computational Geometry, MSRI*, pages 1–30. University Press, 2005.

[2] Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. Approximating extent measures of points. *J. ACM*, 51:606–635, July 2004.

[3] Pankaj K. Agarwal, Sariel Har-Peled, and Hai Yu. Robust shape fitting via peeling and grating coresets. *Discrete & Computational Geometry*, 39(1-3):38–58, 2008.

Copyright © SIAM.
Unauthorized reproduction of this article is prohibited.

Downloaded 05/02/16 to 109.230.67.85. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

[4] Pankaj K. Agarwal and Nabil H. Mustafa. k-means projective clustering. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '04, pages 155–165, New York, NY, USA, 2004. ACM.

[5] Pankaj K. Agarwal, Cecilia Magdalena Procopiuc, and Kasturi R. Varadarajan. Approximation algorithms for k-line center. In *ESA*, pages 54–63, 2002.

[6] Charu C. Aggarwal, Joel L. Wolf, Philip S. Yu, Cecilia Procopiuc, and Jong Soo Park. Fast algorithms for projected clustering. In *Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, SIGMOD '99, pages 61–72, New York, NY, USA, 1999. ACM.

[7] Charu C. Aggarwal and Philip S. Yu. Finding generalized projected clusters in high dimensional spaces. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, SIGMOD '00, pages 70–81, New York, NY, USA, 2000. ACM.

[8] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, SIGMOD '98, pages 94–105, New York, NY, USA, 1998. ACM.

[9] Mihai Badoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *STOC*, pages 250–257, 2002.

[10] Saugata Basu, Richard Pollack, and Marie-Françoise Roy. Computing a set of points meeting every cell defined by a family of polynomials on a variety. In *Proceedings of the workshop on Algorithmic foundations of robotics*, pages 537–555, Natick, MA, USA, 1995. A. K. Peters, Ltd.

[11] Ke Chen. On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. *SIAM J. Comput.*, 39:923–947, August 2009.

[12] Chun-Hung Cheng, Ada Waichee Fu, and Yi Zhang. Entropy-based subspace clustering for mining numerical data. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 84–93, New York, NY, USA, 1999. ACM.

[13] Kenneth L. Clarkson. Subgradient and sampling algorithms for ℓ_1 regression. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '05, pages 257–266, Philadelphia, PA, USA, 2005. Society for Industrial and Applied Mathematics.

[14] Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W. Mahoney. Sampling algorithms and coresets for ℓ_p regression. *SIAM J. Comput.*, 38:2060–2078, February 2009.

[15] Amit Deshpande and Kasturi R. Varadarajan. Sampling-based dimension reduction for subspace approximation. In *STOC*, pages 641–650, 2007.

[16] Michael Edwards and Kasturi R. Varadarajan. No coreset, no cry: II. In *FSTTCS*, pages 107–115, 2005.

[17] Dan Feldman, Amos Fiat, and Micha Sharir. Coresets for weighted facilities and their applications. In *FOCS*, pages 315–324, 2006.

[18] Dan Feldman, Amos Fiat, Micha Sharir, and Danny Segev. Bi-criteria linear-time approximations for generalized k-mean/median/center. In *Symposium on Computational Geometry*, pages 19–26, 2007.

[19] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *STOC*, pages 569–578, 2011.

[20] Dan Feldman, Morteza Monemizadeh, and Christian Sohler. A PTAS for k-means clustering based on weak coresets. In *Symposium on Computational Geometry*, pages 11–18, 2007.

[21] Dan Feldman, Morteza Monemizadeh, Christian Sohler, and David P. Woodruff. Coresets and sketches for high dimensional subspace approximation problems. In *SODA*, pages 630–649, 2010.

[22] Sariel Har-Peled. No, coreset, no cry. In *FSTTCS*, pages 324–335, 2004.

[23] Sariel Har-Peled. Coresets for discrete integration and clustering. In *FSTTCS*, pages 33–44, 2006.

[24] Sariel Har-Peled. How to get close to the median shape. *Comput. Geom.*, 36(1):39–51, 2007.

[25] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *STOC*, pages 291–300, 2004.

[26] Sariel Har-Peled and Kasturi R. Varadarajan. High-dimensional shape fitting in linear time. *Discrete & Computational Geometry*, 32(2):269–288, 2004.

[27] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. Linear-time approximation schemes for clustering problems in any dimensions. *J. ACM*, 57(2), 2010.

[28] Michael Langberg and Leonard J. Schulman. Universal epsilon-approximators for integrals. In *SODA*, pages 598–607, 2010.

[29] Cecilia M. Procopiuc, Michael Jones, Pankaj K. Agarwal, and T. M. Murali. A monte carlo algorithm for fast projective clustering. In *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, SIGMOD '02, pages 418–427, New York, NY, USA, 2002. ACM.

[30] Nariankadu D. Shyamalkumar and Kasturi R. Varadarajan. Efficient subspace approximation algorithms. In *SODA*, pages 532–540, 2007.

[31] Antoine Vigneron. Geometric optimization and sums of algebraic functions. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '10, pages 906–917, Philadelphia, PA, USA, 2010. Society for Industrial and Applied Mathematics.

Copyright © SIAM.
Unauthorized reproduction of this article is prohibited.