Bi-criteria Linear-time Approximations for Generalized k-Mean/Median/Center

Dan Feldman[*]
School of Computer Science
Tel Aviv University
Tel Aviv 69978, Israel
dannyf@post.tau.ac.il

Amos Fiat[*]
School of Computer Science
Tel Aviv University
Tel Aviv 69978, Israel
fiat@post.tau.ac.il

Danny Segev
School of Mathematical Sciences
Tel Aviv University
Tel Aviv 69978, Israel
segevd@post.tau.ac.il

Micha Sharir[†]
School of Computer Science
Tel Aviv University
Tel Aviv 69978, Israel
michas@post.tau.ac.il

ABSTRACT

We consider the problem of approximating a set P of n points in $\mathbb{R}^d$ by a collection of j-dimensional flats, and extensions thereof, under the standard median / mean / center measures, in which we wish to minimize, respectively, the sum of the distances from each point of P to its nearest flat, the sum of the squares of these distances, or the maximal such distance. Such problems cannot be approximated unless P=NP but do allow bi-criteria approximations where one allows some leeway in both the number of flats and the quality of the objective function. We give a very simple bi-criteria approximation algorithm, which produces at most $\alpha(k, j, n) = \log n \cdot (jk \log\log n)^{O(j)}$ flats, which exceeds the optimal objective value for any k j-dimensional flats by a factor of no more than $\beta(j) = 2^{O(j)}$. Given this bi-criteria approximation, we can use it to reduce the approximation factor arbitrarily, at the cost of increasing the number of flats. Our algorithm has many advantages over previous work, in that it is much more widely applicable (wider set of objective functions and classes of clusters) and much more efficient — reducing the running time bound from $O(n^{\mathrm{poly}(k,j)})$ to $dn \cdot (jk)^{O(j)}$. Our algorithm is randomized and successful with probability 1/2 (easily boosted to probabilities arbitrarily close to 1).

Categories and Subject Descriptors

F.2.2 [**Theory of Computation**]: Analysis of Algorithms and Problem Complexity—*Nonnumerical Algorithms and Problems*

General Terms

Algorithms, Theory

Keywords

Geometric optimization, k-mean, k-median, k-center, approximation, bi-criteria approximation

1. INTRODUCTION

Clustering is one of the central problems in computer science. It is relevant to issues of unsupervised learning, classification, databases, spatial range-searching, data-mining, etc. Input points to be clustered are often in very high dimensional space, (*e.g.*, documents represented as a bag of English words in 600,000-dimensional space or gene expression data for 10,000 genes).

Let $P \subset \mathbb{R}^d$ be a set of n points in d-dimensional space. A reasonable goal is to "approximate" P by a small collection, F, of "shapes" in $\mathbb{R}^d$. Depending on the problem, elements of F may be restricted to be single points, lines, j-dimensional subspaces ($j < d$), affine spaces, or other, nonlinear shapes in $\mathbb{R}^d$. For a point $p \in P$, let $c(p) \in F$ be the element $x \in F$ closest to p (ties broken arbitrarily). Every $c \in F$ represents a "cluster", and a point $p \in P$ is said to belong to cluster $c(p)$. The set F is called a *projective clustering* of P.

Typically, the projective clustering problem pre-specifies the class of allowable cluster shapes in F and their number, k. The *value* of a projective clustering F is some function of the distances between points $p \in P$ and their associated clusters $c(p)$, $p \in P$. A good projective clustering is one of small value. Common objectives are to minimize $\sum_{p\in P} \mathrm{dist}(p, c(p))$, the corresponding sum of squared distances, or the maximal such distance. A projective clustering F that minimizes one of these three main objective

[*] Supported by the German-Israel Foundation for Scientific Research and Development.

[†] Supported by NSF Grant CCF-05-14079, by a grant from the U.S.-Israeli Binational Science Foundation, by grant 155/05 from the Israel Science Fund, and by the Hermann Minkowski–MINERVA Center for Geometry at Tel Aviv University.

SCG'07, June 6–8, 2007, Gyeongju, South Korea.

functions, is referred to as the *k-median*, *k-mean*, or *k-center*, respectively. For example, the 2-flat k-median is a set, F, of k 2-flats in $\mathbb{R}^d$, that minimizes the sum of distances $\sum_{p \in P} \operatorname{dist}(p, c(p))$. See Tables 1 and 2 for a variety of results concerning projective clustering problems. Additional applications, heuristics, and implementations of projective clustering includes PROCLUS [5], ORCLUS [6], DOC [22], and CLIQUE [7]. Heuristics for projective clustering that are based on heuristics for k-means can be found in [2], with more references therein.

When the number of objects, k, or the dimension, d, are part of the input, almost all such projective clustering problems are *NP*-hard [19]. It is therefore natural to seek approximation algorithms for projective clustering problems. A *c-approximation* algorithm for a k-projective clustering problem should produce a k-projective clustering, F, with value not greater than c times the smallest value of any k-projective clustering.

Unfortunately, even for planar point sets $P \subset \mathbb{R}^2$, it is *NP*-complete to determine if there exist k lines (1-flats) whose union covers P [19], when k is part of the input. If the k lines indeed cover the points of P, then the sum of distances, sum of squared distances, and maximal distance, are all zero. Hence, any finite approximation to the k-line median, mean, or center problems is *NP*-hard, even for point sets P in the plane. In Table 1 we summarize recent work on approximate projective clustering. (Note that all algorithms for an arbitrary k in the table are (at least) exponential in k.)

Given that approximate k-projective clustering is intractable for non-constant k, it is natural to try to find a *bi-criteria approximation.* For points $P \subset \mathbb{R}^d$, an (α, β) bi-criteria approximation for k-projective clustering by j-dimensional flats is a set F of α j-dimensional flats whose value is within a factor of β from the minimal value of any k j-dimensional flats. The parameters α and β may depend on k, j, d, and n, where the dependence on n should be small (say, polylogarithmic), or — even better — independent of n. In Table 2 we summarize the current state of affairs regarding such bi-criteria approximations for projective clustering. Our results appear in rows marked $\star$ in Table 2.

Our main result is an algorithm that produces such an (α, β) bi-criteria approximation for k-projective clustering, for point sets in any dimension $d \geq 1$, by lines or flats of any dimension $j < d$. Our algorithm is motivated by and related to prior work on bi-criteria approximations for other problems, in particular [13], [15], [17], etc. We achieve a bi-criteria approximation with $\alpha(k, j, n) = \log n \cdot (jk \log n \log n)^{O(j)}$ and $\beta(j) = 2^{O(j)}$, in time $dn \cdot (jk)^{O(j)}$. Furthermore, this bi-criteria approximation holds simultaneously for all three objective functions: median, mean, and center. It is noteworthy that the running time has only linear dependence on both the dimension d, and the number of input points n. We also observe that one can refine the solution so as to decrease the approximation factor β to an arbitrarily small value, at the cost of increasing the number α of flats. The last row of Table 2 gives such a variant. This can be done by computing a set of flats using large β, and then surround each of them by a "grid" of additional parallel j-flats; see [4, 11].

As Table 2 states, prior work on such approximations has only dealt with very limited projective clustering problems, and only for k-center clustering problems.

1.1 Some implications of bi-criteria approximation

As mentioned above, Table 1 includes projective clustering approximation results from a related companion paper [11]. Rows marked $\star\star$ describe an FPTAS for the mean and median objective functions for any number k of line clusters or for a single j-flat cluster, $j \geq 2$. The FPTAS of [11] uses as a starting point (and as black box) a bi-criteria approximation — the subject of the current paper.

We remark that many other results follow from our bi-criteria approximation. For example, using a bi-criteria approximation, one can derive an FPTAS for the k-line center clustering problem that takes $O(n)$ time, improving upon the $O(n \log n)$ bound of [4]. One can also derive explicit and efficient constructions for related *coresets* (see [1, 11]), previously unknown, such as coresets for a single j-flat or for k-lines (center/mean/median). Some of these developments are given in the companion paper [11] while others constitute work in progress.

2. RESULTS

2.1 Informal overview

We seek a small set F of α j-dimensional flats so that the value of the objective function (median, mean, or center) is not much larger than that of the optimal k j-dimensional flats. One can view our algorithm as an instance of the following "meta algorithm" for a bi-criteria projective clustering for input point sets $P \subset \mathbb{R}^d$:

- Choose a set $P' \subset P$ of size $\geq |P|/2$, and a set F' of k' j-dimensional flats (for some parameter k'), such that the value of the objective function (or, rather, of all three objective functions) for F' on P' is no more than c times the value of the optimal k j-dimensional flats (for P) on P', for some constant factor c.

- Set $P = P \setminus P'$ and repeat until P is very small, in which case take F' to be the set of all j-dimensional flats spanned by P.

As $|P|$ keeps shrinking by factors of 2, this process can be repeated at most $\log |P|$ times. By taking the union of the sets F' , we get a set F of $k' \log |P|$ j-dimensional flats, for which the value of the objective function, over the entire P, is off by no more than a factor of c.

In fact, our real algorithm, given below, is very similar to the meta algorithm above, with the following (minor and technical) variations:

- The set F' is simply a set of j-dimensional flats determined by a small set of randomly chosen points from P.

- It follows from the fact that the points were chosen at random that, with some non-trivial probability, some large set of "good" points, $P' \subseteq P$, has the property that the optimal solution, computed only over P', has approximately the same value as the one for F'.

- In fact, this set P' consists of a large fraction of the $|P|/2$ points of P that are closest to the flats of F'. The intuition comes from the argument that *many* of the points near the flats of F' are not much farther

Flat dim. j	$k = \#$ Flats	Objective Function	Approx	Ref.	Time
1	$k \geq 1$	median/mean	FPTAS	$\star\star$	$nd \cdot k^{O(1)} + (\epsilon^{-d} \log n)^{O(dk^2)}$
$j \geq 1$	1	median	FPTAS	$\star\star$	$nd \cdot (jk)^{O(j^2)} + (\epsilon^{-1} \operatorname{polylog} n)^{O(d^2 j^2)}$
$j \geq 1$	1	mean	Exact	SVD [21]	$\min \left\{ O(nd^2), O(n^2 d) \right\}$
$j \geq 1$	1	mean	PTAS	[9, 14, 23]	$nd \, \mathrm{poly}(j, 1/\epsilon)$
$j \geq 1$	$k \geq 1$	mean	PTAS*	[8]	$d(n/\epsilon)^{O(jk^3/\epsilon)}$
$j \geq 1$	1	median	PTAS	[24]	$nd \cdot 2^{O(j/\epsilon \log^2(1/\epsilon))}$
$j = 1, d = 2$	1	median	Exact	[10]	$O(n^{4/3} \log^2 n)$
$j \geq 1$	$k \geq 1$	median	PTAS*	[24]	$d(n/\epsilon)^{\mathrm{poly}(j,k,1/\epsilon))}$
$j \geq 1$	1	center	PTAS	[16]	$dn^{O\left(j/\epsilon^5 \log(1/\epsilon)\right)}$
$j \geq 1$	1	center	PTAS	[20]	$dn \cdot \exp\left(\frac{2^{O(j^2)}}{\epsilon^2} \log \frac{1}{\epsilon}\right)$
$j \geq 1$	$k \geq 1$	center	PTAS	[16]	$dn^{O\left(jk/\epsilon^5 \log(1/\epsilon)\right)}$
1	$k \geq 1$	center	FPTAS	[4]	$n \log n \cdot \epsilon^{O(-d-k)} k^{O(k)} + \log n \cdot (k/\epsilon)^{O(d^2 k^2)}$
$j = 1, d = 2$	2	center	Exact	[18]	$O(n^2 \log^2 n)$
$j = 1, \, d = 2$	2	center	3-approx	[3]	$O(n \log n)$

Table 1: Approximate projective clustering (not bi-criteria). The input is a set $P \subset \mathbb{R}^d$, $|P| = n$, the goal is to find a good approximation for P using k j-dimensional flats. Unless P=NP, all such approximations must be superpolynomial in k. The first two rows above, marked $\star\star$, give results from a companion paper [11].

*This requires prior knowledge of the value of the optimal solution. However, using results from this paper we can avoid this requirement (see [12]).

$P \subset \mathbb{R}^d$	Flat dim. j	$k = \#$ Flats	Objective Function	α	β	Ref.	Time
$d = 2$	$j = 1$	$k \geq 1$	center	$O(k \log k)$	6	[3]	$O(nk^2 \log^4 n)$
$d = 2$	$j = 1$	$1 \leq k \leq n^{1/6}$	center	$O(k \log k)$	1	[13]	$O(n \log k)$
$d = 3$	$j = 2$	$k \geq 1$	center	$O(k \log k)$	24	[3]	$O(n^{3/2} k^{11/4} \log^{O(1)}(n))$
$d \geq 1$	$j = 1$	$k \geq 1$	center	$O(dk \log k)$	8	[3]	$O(dnk^3 \log^4 n)$
$d \geq 1$	$j \geq 1$	$k \geq 1$	center mean median	$\log n \cdot (jk \log \log n)^{O(j)}$	$2^{O(j)}$	$\star$	$dn \cdot (jk)^{O(j)}$
$d \geq 1$	$j \geq 1$	$k \geq 1$	center mean median	$(2^d jk \log n)^{O(j)}$	1/2	$\star$	$dn(jk)^{O(j)} + (2^d jk \log n)^{O(j)}$

Table 2: Results on bi-criteria approximate projective clustering. The input is a set $P \subset \mathbb{R}^d$, $|P| = n$, the goal is to find an approximation for P using α j-dimensional flats to within a β factor off the optimal such approximation by k j-dimensional flats. The last two entries are the contribution of this paper. Our bi-criteria approximation holds simultaneously for all three main objective functions.

from F' than they are to some other (arbitrary) set of k flats.

- Unfortunately, not all points "close" to F' have the property that F' is a good approximation to the optimal set of flats; these are "bad" points.

- Fortunately, we can amortize the high contribution to the objective function by these "bad" points against the next round of points to be chosen. The contribution to the objective function, appropriately scaled, of the good points of the next round will dominate that of the current "bad" points.

2.2 The Algorithm

We first briefly review some notation.

***j*-Flats.** For $d \geq 1$ and $0 \leq j \leq d-1$, a j-flat in $\mathbb{R}^d$ is a shorthand notation for a j-dimensional flat in $\mathbb{R}^d$. For example, a 0-flat is a point, a 1-flat is a line, and a $(d-1)$-flat is a hyperplane in $\mathbb{R}^d$. For a multi-set X of $j+1$ points in $\mathbb{R}^d$, we denote by $\mathrm{flat}(X)$ a j-flat that passes through all the points of X. If there is more than one such flat, we choose one of them arbitrarily. For $k \geq 1$ an integer, we denote by $\mathbb{F}(k,j,d)$ the collection of sets that contain at most k flats in $\mathbb{R}^d$, each of dimension of at most j.

Euclidean Distance. For a j-flat f and a point p in $\mathbb{R}^d$, we denote by $\mathrm{dist}(p,f)$ the minimum *Euclidean distance* from p to f. For a set of flats F, we denote by $\mathrm{dist}(p,F) = \min_{f\in F}\mathrm{dist}(p,f)$ the distance of p to its nearest flat in F. The pseudo-code of our bi-criteria algorithm is given in Figure 1.

THEOREM 2.1. *Let P be a set of n points in $\mathbb{R}^d$, and k,j integers, such that $k \geq 1$ and $0 \leq j \leq d-1$. Then the procedure* APPROX-K-J-FLATS(P,k,j)*, given in Figure 1, returns a set F of $\log n \cdot (jk \log\log n)^{O(j)}$ j-flats, such that, with probability at least $1/2$, we have*

$$\begin{aligned}
\sum_{p\in P}\mathrm{dist}(p,F) &\leq 2^{j+2} \min_{F^*\in\mathbb{F}(k,j,d)} \sum_{p\in P}\mathrm{dist}(p,F^*),\\
\sum_{p\in P}\big(\mathrm{dist}(p,F)\big)^2 &\leq 2^{2j+3} \min_{F^*\in\mathbb{F}(k,j,d)} \sum_{p\in P}\big(\mathrm{dist}(p,F^*)\big)^2,\\
\max_{p\in P}\mathrm{dist}(p,F) &\leq 2^{j+1} \min_{F^*\in\mathbb{F}(k,j,d)} \max_{p\in P}\mathrm{dist}(p,F^*).
\end{aligned}$$

The running time of this procedure is $dn\cdot(jk)^{O(j)}$.

PROOF. Let F^* be an arbitrary set of flats in $\mathbb{F}(k,j,d)$. The proof relies on the following theorem which is the main technical contribution of this paper.

THEOREM 2.2. *Let P be a set of n points in $\mathbb{R}^d$, and k,j integers, such that $k\geq 1$ and $0\leq j\leq d-1$. Let F be the set of flats that is returned by the bi-criteria approximation algorithm* APPROX-K-J-FLATS(P,k,j) *(see Fig. 1). For an arbitrary set of flats $F^*\in\mathbb{F}(k,j,d)$, define $P_{\mathrm{bad}} = \{b\in P \mid \mathrm{dist}(b,F) > 2^{j+1}\mathrm{dist}(b,F^*)\}$. Then, with probability at least $1/2$, we can map each point $b\in P_{\mathrm{bad}}$ to a distinct point $p\in P\setminus P_{\mathrm{bad}}$, such that $\mathrm{dist}(b,F)\leq 2^{j+1}\mathrm{dist}(p,F^*)$.*

Using Theorem 2.2, we prove the inequalities in Theorem 2.1 as follows. Assuming that the property of Thorem 2.2 does hold (which happens with probability at least 1/2), we have

$$\begin{aligned}
\sum_{p\in P}\mathrm{dist}(p,F) &= \sum_{p\in P\setminus P_{\mathrm{bad}}}\mathrm{dist}(p,F) + \sum_{b\in P_{\mathrm{bad}}}\mathrm{dist}(b,F)\\
&\leq \sum_{p\in P\setminus P_{\mathrm{bad}}}\Big(\mathrm{dist}(p,F) + 2^{j+1}\mathrm{dist}(p,F^*)\Big)\\
&\leq 2^{j+2}\sum_{p\in P}\mathrm{dist}(p,F^*),
\end{aligned}$$

where the first inequality follows from the matching of points in P_{bad} to points in $P\setminus P_{\mathrm{bad}}$, and the second inequality follows from the definition of $P\setminus P_{\mathrm{bad}}$. The same arguments imply the other two inequalities. That is,

$$\begin{aligned}
\sum_{p\in P}\big(\mathrm{dist}(p,F)\big)^2 &= \sum_{p\in P\setminus P_{\mathrm{bad}}}\big(\mathrm{dist}(p,F)\big)^2\\
&\quad + \sum_{b\in P_{\mathrm{bad}}}\big(\mathrm{dist}(b,F)\big)^2\\
&\leq \sum_{p\in P\setminus P_{\mathrm{bad}}}\Big(\big(\mathrm{dist}(p,F)\big)^2\\
&\quad +2^{2j+2}\big(\mathrm{dist}(p,F^*)\big)^2\Big)\\
&\leq 2^{2j+3}\sum_{p\in P}\big(\mathrm{dist}(p,F^*)\big)^2.
\end{aligned}$$

$$\begin{aligned}
\max_{p\in P}\mathrm{dist}(p,F) &= \max\Big\{\max_{p\in P\setminus P_{\mathrm{bad}}}\mathrm{dist}(p,F),\\
&\qquad \max_{b\in P_{\mathrm{bad}}}\mathrm{dist}(b,F)\Big\}\\
&\leq \max_{p\in P\setminus P_{\mathrm{bad}}}\Big\{\mathrm{dist}(p,F),\\
&\qquad 2^{j+1}\mathrm{dist}(p,F^*)\Big\}\\
&\leq 2^{j+1}\max_{p\in P}\mathrm{dist}(p,F^*).
\end{aligned}$$

We next analyze the size of F and the time for its construction. Since the size of Q is reduced by at least half in each iteration, we have $t_{\max}-1\leq\log n$ iterations. In line 10, at most $\big(32k(j+1)\big)^{j+1}$ flats are added to F (by the bound in line 2), so the overall size of the output set of flats is

$$\begin{aligned}
&\sum_{t=1}^{t_{\max}-1}\lceil 32k(j+1)\big(2+\log(j+1)+\log k\\
&\qquad +\min\{t,\log\log n\}\big)\rceil^{j+1} + \big(32k(j+1)\big)^{j+1}\\
&= \sum_{t=1}^{t_{\max}-1}\big(O(jk)\cdot O(jk+\log\log n)\big)^{j+1} + \big(O(jk)\big)^{j+1}\\
&= \log n\cdot(jk\log\log n)^{O(j)}.
\end{aligned}$$

The running time of the t^{th} iteration is dominated by the running time of Line 6, which is (using brute force)

$$\begin{aligned}
O(d\,|Q|\cdot|F'|) = O(dn/2^t)\cdot\big(32k(j+1)(2+\log(j+1)\\
+\log k+t)\big)^{j+1}.
\end{aligned}$$

Summing this over t, we get a sum of the form

$$O\left(dn\cdot(32jk)^{j+1}\sum_{t\geq 1}\frac{\big(2+\log(jk)+t\big)^{j+1}}{2^t}\right) = dn\cdot f(j,k),$$

```
Algorithm Approx-k-j-Flats($P, k, j$)
Input. A set of $n$ points $P \subset \mathbb{R}^d$, and two integers $k \geq 1$, $0 \leq j \leq d-1$.
Output. A set of $j$-flats $F$ that satisfies Theorem 2.2.
$t \leftarrow 1$, $Q \leftarrow P$, $F \leftarrow \emptyset$
while $|Q| \geq 32k(j+1)$
    for $i \leftarrow 0$ to $j$
        Pick a random sample $S_i$ of $\lceil 32k(j+1)\big(2 + \log(j+1) + \log k + \min\{t, \log\log n\}\big)\rceil$ points
        from $Q$, each chosen uniformly at random and independently.
    $F' \leftarrow \{\mathrm{flat}(X) \mid X \in S_0 \times S_1 \times \cdots \times S_j\}$.
    Compute a set $R_t \subseteq Q$ of the closest $\lceil |Q|/2 \rceil$ points to $F'$, where ties are broken arbitrarily.
    $F \leftarrow F \cup F'$
    $Q \leftarrow Q \setminus R_t$
    $t \leftarrow t+1$
$F \leftarrow F \cup \{\mathrm{flat}(X) \mid X \in Q^{j+1}\}$
$t_{\max} \leftarrow t$, $R_{t_{\max}} \leftarrow Q$ (used only for analysis)
```

Figure 1: The bi-criteria algorithm Approx-k-j-Flats**.**

where

$$\begin{aligned} f(j,k) &= (jk)^{O(j)} \sum_{t \geq 1} \frac{\big(2 + t + \log(jk)\big)^{j+1}}{2^t} \\ &= (jk)^{O(j)} \sum_{t=1}^{\lfloor \log(jk) \rfloor} \frac{\big(2\log(jk)\big)^{j+1}}{2^t} \\ &\quad + (jk)^{O(j)} \sum_{t \geq \lfloor \log(jk) \rfloor + 1} \frac{t^{j+1}}{2^t} \\ &= (jk)^{O(j)} \Big[\big(\log(jk)\big)^{j+1} + j^{j+1}\Big] = (jk)^{O(j)}. \end{aligned}$$

□

The probability that the resulting set F of Approx-k-j-Flats satisfies the inequalities of Theorem 2.1 can be made arbitrarily close to 1, by running Approx-k-j-Flats repeatedly x times with independent random choices each time. Then we take the three sets which minimize the three expressions in Theorem 2.1. The union of these sets will satisfy all three inequalities, with probability at least $1 - 1/2^x$.

3. PROOF OF THEOREM 2.2

We first provide a brief overview of the proof. It begins with Lemma 3.1, which is a simple probabilistic lemma, giving a bound on the size of a random sample from a set Q that guarantees, with high probability, that it hits each of k given subsets of Q of some given size.

Lemma 3.2 says that if we choose an arbitrary line ℓ through the origin, and a line $\mathrm{sp}(b)$ connecting some arbitrary point b to the origin, then for all points whose angle with ℓ is greater than the angle between $\mathrm{sp}(b)$ and ℓ, the distance to $\mathrm{sp}(b)$ is at most a constant factor times the distance to ℓ. This observation is later generalized to higher-dimensional flats in Lemma 3.3.

Lemma 3.4 deals with one iteration of the algorithm. It uses the preceding lemmas argue that the set of flats F' chosen by the algorithm has the property that the set of bad points (points close to F' that are much closer to F^*) is small.

Finally, the proof amortizes the contribution of the (few) bad points against the contribution of other good points, concluding the proof of the theorem.

Lemma 3.1. *Let Q be a set of m points, $k \geq 1$ an integer, and $c > k$ a parameter. Let $Q_1, Q_2, \ldots, Q_k$ be any k subsets of Q, each containing β points. Assume that we pick at least $(m/\beta)\ln c$ random independent samples from Q (with or without repetitions). Then the probability that at least one of the subsets does not contain any sample point is at most k/c.*

Proof. The probability that the first sampled point is not contained in Q_1 is $1 - \beta/m$. Therefore, the probability that none of the sampled points are in Q_1 is at most

$$\left(1 - \frac{\beta}{m}\right)^{(m/\beta)\ln c} \leq e^{-\ln c} = \frac{1}{c}.$$

Clearly, similar calculation hold for any Q_i, $1 \leq i \leq k$. Hence, the probability that at least one of these sets does not contain any sample point is at most k/c. □

In the following analysis, we use the notation $\mathrm{sp}(X)$ for the linear span of a set X; when X is a singleton b, the shorthand notation $\mathrm{sp}(b)$ thus denotes the line through b and the origin.

Lemma 3.2. *Let ℓ be a line in $\mathbb{R}^d$ that passes through the origin. Let Q be a set of points in $\mathbb{R}^d$. Then, for any natural number $\beta \leq |Q|$ there is a set $B \subseteq Q$ of β points, such that for all $b \in B$ and $q \in Q \setminus B$*

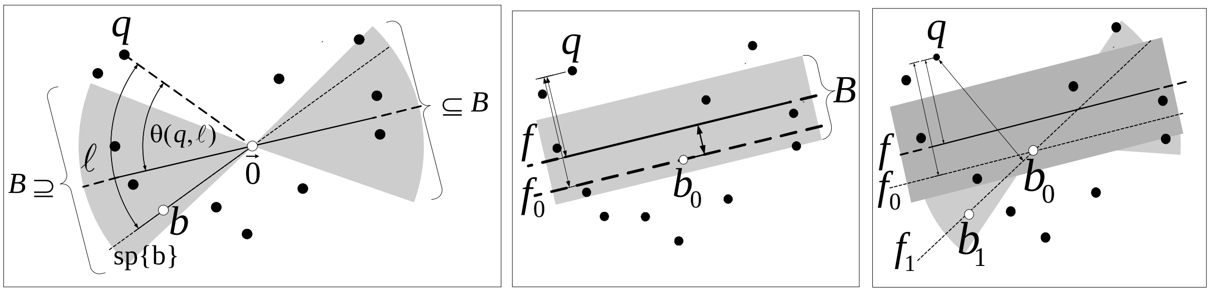
$$\mathrm{dist}(q, \mathrm{sp}(b)) \leq 2\,\mathrm{dist}(q, \ell).$$

Proof. For a point $q \in Q$, denote by $\theta(q, \ell)$ the acute angle formed by the lines $\mathrm{sp}(q)$ and ℓ; see Figure 2(a) for the planar case. Let $B \subseteq Q$ be the set consisting of the β points q with the smallest values of $\theta(q, \ell)$, and let $b \in B$. For $q \in Q \setminus B$ we thus have $\theta(b, \ell) \leq \theta(q, \ell)$, and therefore

$$\theta(q, \mathrm{sp}(b)) \leq \theta(q, \ell) + \theta(b, \ell) \leq 2\theta(q, \ell)$$

or, $\theta(q, \mathrm{sp}(b))/2 \leq \theta(q, \ell)$, which implies that

$$\begin{aligned} \sin\theta(q, \mathrm{sp}(b)) &= 2\sin\frac{\theta(q, \mathrm{sp}(b))}{2}\cos\frac{\theta(q, \mathrm{sp}(ab))}{2} \\ &\leq 2\sin\frac{\theta(q, \mathrm{sp}(b))}{2} \leq 2\sin\theta(q, \ell). \end{aligned}$$



(a) $\theta(q, \text{sp}(b)) \le 2\theta(q, \ell)$ (b) $\text{dist}(q, f_0) \le 2\,\text{dist}(q, f)$ (c) $\text{dist}(q, f_1) \le 4\,\text{dist}(q, f)$

Figure 2: The case of one line in the plane ($d = 2$, $j = 1$). (a) The set B contains the β points in the gray areas. (b) The set B consists of the β points of P closest to f. (c) $\text{dist}(q, f_1) \le 2\,\text{dist}(q, f_0) \le 4\,\text{dist}(q, f)$ for every q outside the gray area ($q \in Q \setminus Q_{\text{bad}}$).

The distance from q to $\text{sp}(b)$ can then be bounded by

$$\begin{aligned}\text{dist}(q, \text{sp}(b)) &= \|q\| \sin\theta(q, \text{sp}(b)) \\ &\le \|q\| \cdot 2\sin\theta(q, \ell) = 2\,\text{dist}(q, \ell).\end{aligned}$$

□

Lemma 3.3. *Let Q be a set of points in $\mathbb{R}^d$, and $f = \text{sp}(v_1, \ldots, v_{j-1}, v_j)$ be a j-dimensional subspace of $\mathbb{R}^d$, for some given tuple of j mutually orthogonal unit vectors $v_1, \ldots, v_j$. Then, for any natural number $\beta \le n$, there exists a subset $B \subseteq Q$ of β points, such that for every point $b \in B$, and the corresponding subspace $f(b) = \text{sp}(b, v_2, v_3, \ldots, v_j)$, we have*

$$\text{dist}\left(q, f(b)\right) \le 2\,\text{dist}(q, f),$$

for all $q \in Q \setminus B$.

Proof. Let $\{v_{j+1}, \ldots, v_d\}$ be a set of vectors that span the subspace orthogonal to f. For a point $x \in \mathbb{R}^d$ we denote by x' the projection of x onto the subspace $M = \text{sp}(v_1, v_{j+1}, v_{j+2}, \ldots, v_d)$. For a set $X \subseteq \mathbb{R}^d$, we define $X' = \{x' \mid x \in X\}$.

By substituting $P = Q'$, a as the origin, and $\ell = \text{sp}(v_1')$ in Lemma 3.2, we conclude that for any natural number $\beta \le n$ there exists a set $B' \subseteq Q'$ of β points, such that for every $b' \in B'$, the corresponding line $\text{sp}(b')$ satisfies

$$\text{dist}(q', \text{sp}(b')) \le 2\,\text{dist}(q', \text{sp}(v_1')) = 2\,\text{dist}(q', \text{sp}(v_1)), \quad (3.1)$$

for all $q' \in Q' \setminus B'$ (by construction, $v_1' = v_1$). We define B to be the set of those $b \in Q$ such that $b' \in B'$. We claim that B satisfies the property asserted in the lemma, that is, for each point $b \in B$, its corresponding subspace $f(b) = \text{sp}(b, v_2, v_3, \ldots, v_j)$ satisfies $\text{dist}\left(q, f(b)\right) \le 2\,\text{dist}(q, f)$, for all $q \in Q \setminus B$.

Indeed, let q be any point in $Q \setminus B$. By definition,

$$\text{dist}\left(q, f(b)\right) = \min_{u \in f(b)} \|q - u\| = \min_{u \in f(b)} \left\|q' + (q - q' - u)\right\|.$$

Since q' is the projection of q onto $M = \text{sp}(v_1, v_{j+1}, v_{j+2}, \ldots, v_d)$, we conclude that $q - q'$ is in $\text{sp}(v_2, v_3, \ldots, v_j) \subseteq f(b)$. Hence, by the previous equation,

$$\begin{aligned}\text{dist}\left(q, f(b)\right) &= \min_{u \in f(b)} \left\|q' + (q - q' - u)\right\| \\ &= \min_{w \in f(b)} \left\|q' + w\right\| = \text{dist}\left(q', f(b)\right).\end{aligned} \quad (3.2)$$

Since $b - b' \in \text{sp}(v_2, v_3, \ldots, v_j)$, we also have

$$f(b) = \text{sp}(b, v_2, v_3, \ldots, v_j) = \text{sp}(b', v_2, \ldots, v_j),$$

and the length of the projection of q' onto $f(b)$ is therefore

$$\sqrt{\sum_{i=2}^{j} (q' \cdot v_i)^2 + \frac{(q' \cdot b')^2}{\|b'\|^2}} = q' \cdot b' / \|b'\|.$$

By the Pythagorean Theorem we then get

$$\text{dist}\left(q', f(b)\right) = \sqrt{\|q'\|^2 - (q' \cdot b' / \|b'\|)^2} = \text{dist}(q', \text{sp}(b')). \quad (3.3)$$

Substituting this in (3.2) gives

$$\text{dist}\left(q, f(b)\right) = \text{dist}\left(q', f(b)\right) = \text{dist}(q', \text{sp}(b')).$$

Similarly, by replacing $f(b)$ and b with f and v_1, respectively, in the last equations, we get

$$\text{dist}(q, f) = \text{dist}(q', f) = \text{dist}(q', \text{sp}(v_1)).$$

Combining the last two equations in (3.1) gives us

$$\begin{aligned}\text{dist}\left(q, f(b)\right) &= \text{dist}(q', \text{sp}(b')) \\ &\le 2\,\text{dist}(q', \text{sp}(v_1)) = 2\,\text{dist}(q, f),\end{aligned}$$

which completes the proof of the lemma. □

Lemma 3.4. *Let F^* be a set of k arbitrary j-flats in $\mathbb{R}^d$, where $k \ge 1$ and $0 \le j \le d - 1$. Consider the sets Q and F' at the time line 7 is executed in the t^{th} iteration of* Approx-k-j-Flats(P, k, j), *and define $Q_{\text{bad}} = \{q \in Q \mid \text{dist}(q, F') > 2^{j+1}\,\text{dist}(q, F^*)\}$. Then $|Q_{\text{bad}}| \le |Q|/16$ with probability at least $1 - 2^{-2-\min\{t, \log\log n\}}$.*

Proof. For a j-flat $f \in F^*$, let $B \subset Q$ be the set of the $\beta = \lfloor |Q| / (16k(j+1)) \rfloor$ points of Q, closest to f, where ties are broken arbitrarily; see Fig. 2(b). Fix a point $b_0 \in B$, and let f_0 be the j-flat that is parallel to f and passes through b_0. Note that for every point $q \in Q \setminus B$ we have $\text{dist}(b_0, f) \le \text{dist}(q, f)$ by definition of B. Thus,

$$\text{dist}(q, f_0) \le \text{dist}(q, f) + \text{dist}(b_0, f) \le 2\,\text{dist}(q, f). \quad (3.4)$$

Without loss of generality, we assume that the point b_0 is the origin (i.e., $b_0 = \vec{0}$), and $f_0 = \text{sp}(v_1, v_2, \ldots, v_{j-1}, v_j)$, for some given tuple of j mutually orthogonal unit vectors $v_1, \ldots, v_j$.. By Lemma 3.3, there exists a set $B(b_0) \subseteq Q$ of β points, such that for every $b_1 \in B(b_0)$, and the corresponding j-flat $f_1 = \text{sp}(b_1, v_2, \ldots, v_j)$, we have

$$\text{dist}(q, f_1) \le 2\,\text{dist}(q, f_0) \quad (3.5)$$

for all $q \in Q \setminus B(b_0)$; see Fig.2(c).

Fix a point $b_1 \in B(b_0)$. By substituting $f = f_1$ in Lemma 3.3, we conclude that there is a set $B(b_0, b_1) \subseteq Q$ of β points, such that for every $b_2 \in B(b_0, b_1)$, and the corresponding j-flat $f_2 = \mathrm{sp}(b_1, b_2, v_3, v_4, \ldots, v_j)$, we have

$$\mathrm{dist}(q, f_2) \le 2 \,\mathrm{dist}(q, f_1),$$

for all $q \in Q \setminus B(b_0, b_1)$. Combining (3.4) and (3.5) with the last equation yields

$$\mathrm{dist}(q, f_2) \le 2 \,\mathrm{dist}(q, f_1) \le 4 \,\mathrm{dist}(q, f_0) \le 8 \,\mathrm{dist}(q, f),$$

for all $q \in Q \setminus \big(B \cup B(b_0) \cup B(b_0, b_1)\big)$.

Similarly, by induction, for every j-flat $f \in F^*$, and $0 \le i \le j$, there is a set $B_f(b_0^f, b_1^f, \ldots, b_{i-1}^f) \subseteq Q$ of β points (for $i = 0$, we denote the set simply as B_f), such that for every $b_i^f \in B_f(b_0^f, b_1^f, \ldots, b_{i-1}^f)$, and the corresponding j-flat $f_i = b_0^f + \mathrm{sp}(b_1^f, b_2^f, \ldots, b_i^f, v_{i+1}^f, \ldots, v_j^f)$, we have

$$\mathrm{dist}(q, f_i) \le 2^{i+1} \,\mathrm{dist}(q, f), \tag{3.6}$$

for all $q \in Q \setminus \bigcup_{0 \le i \le j} B_f(b_0^f, \ldots, b_{i-1}^f)$.

We claim that with probability at least $1 - 2^{-2-\min\{t, \log\log n\}}$, for each $f \in F^*$ and $0 \le i \le j$, the set S_i contains a point $b_i^f \in B_f(b_0^f, b_1^f, \ldots, b_{i-1}^f)$. Indeed, we have k sets B_f, of size β each, for $f \in F^*$. Lemma 3.1 shows that if we sample at least $\dfrac{|Q|}{\beta} \ln c$ points from Q, the probability that at least one of the sets B_f will not contain any sample point is at most k/c. Let $c = 2^{2+\log(j+1)+\log k+\min\{t,\log\log n\}}$, and note that, by Line 2 of Approx-k-j-Flats, we have $|Q| / \big(32k(j+1)\big) \ge 1$, so

$$\beta = \lfloor |Q| / \big(16k(j+1)\big) \rfloor \ge |Q| / \big(32k(j+1)\big).$$

Hence

$$\begin{aligned}&(|Q|/\beta) \ln c \le \\ &\quad \lceil 32k(j+1)\big(2 + \log(j+1) + \log k + \min\{t, \log\log n\}\big)\rceil \\ &\quad = |S_0|,\end{aligned}$$

and thus the probability that S_0 misses at least one of the sets B_f is at most

$$\begin{aligned}k/c &= k/2^{2+\log(j+1)+\log k+\min\{t,\log\log n\}} \\ &\le 2^{-2-\log(j+1)-\min\{t,\log\log n\}}.\end{aligned}$$

Assume that this event does not arise (which happen with probability at least $1 - 2^{-2-\log(j+1)-\min\{t,\log\log n\}}$). Pick a point $b_0^f \in B_f \cap S_0$ for each $f \in F^*$, and consider the k sets $B_f(b_0^f)$, $f \in F^*$. As in the case for S_0, it can be shown that S_1 misses at least one of the sets $B_f(b_0^f)$ with probability at most $k/c \le 2^{-2-\log(j+1)-\min\{t,\log\log n\}}$.

By repeating this process, we conclude that, for every $f \in F^*$, the set $S_0 \times S_1 \times \ldots \times S_j$ contains a $(j+1)$-tuple $b_0^f, b_1^f, \ldots, b_j^f$ such that $b_i^f \in B_f(b_0^f, \ldots, b_{i-1}^f)$ for each $0 \le i \le j$, with probability at least $1 - (j+1)k/c \ge 1 - (j+1) \cdot 2^{-2-\log(j+1)-\min\{t,\log\log n\}} \ge 1 - 2^{-2-\min\{t,\log\log n\}}$. This implies that, with the same probability, F' contains a j-flat f_j that passes through $b_0^f, b_1^f, \ldots, b_j^f$ for every $f \in F^*$. Refer to this event as E, and assume that it occurs. In this case, by (3.6), $\mathrm{dist}(q, f_j) \le 2^{j+1} \,\mathrm{dist}(q, f)$ for all $q \in Q \setminus \bigcup_{0 \le i \le j} B_f(b_0^f, \ldots, b_{i-1}^f)$, where b_i^f is one of the points in $S_i \cap B_f(b_0^f, \ldots, b_{i-1}^f)$ which, since we assume that E occurs, is nonempty. Hence, $Q_{\mathrm{bad}} \subseteq \bigcup_{f \in F^*} \bigcup_{0 \le i \le j} B_f(b_0^f, \ldots, b_{i-1}^f)$. Since, by construction, each of the sets in the union is of size β, we get

$$\begin{aligned}|Q_{\mathrm{bad}}| &\le \left| \bigcup_{f \in F^*} \bigcup_{0 \le i \le j} B_f(b_0^f, \ldots, b_{i-1}^f) \right| \\ &\le (j+1)k\beta \le |Q| / 16\end{aligned} \tag{3.7}$$

with probability at least $1 - 2^{-2-\min\{t,\log\log n\}}$. This completes the proof of Lemma 3.4. □

Now we are ready to prove Theorem 2.2.

Proof. Note that $R_1, R_2, \ldots, R_{t_{\max}}$ is a partition of P, and for every $p \in R_{t_{\max}}$ we have $\mathrm{dist}(p, F) = 0$, by Line 10 (*i.e.*, $P_{\mathrm{bad}} \cap R_{t_{\max}} = \emptyset$). Thus,

$$P_{\mathrm{bad}} = \bigcup_{1 \le t \le t_{\max}-1} P_{\mathrm{bad}} \cap R_t. \tag{3.8}$$

Consider the sets Q and F' at the time line 7 is executed, in some t^{th} iteration $(1 \le t \le t_{\max} - 1)$ of Approx-k-j-Flats, and define

$$Q_{\mathrm{bad}} = \{b \in Q \mid \mathrm{dist}(b, F') > 2^{j+1} \,\mathrm{dist}(b, F^*)\}.$$

We first prove that, with probability at least $1 - 2^{-2-\min\{t,\log\log n\}}$, we have

$$|Q_{\mathrm{bad}} \cap R_t| \le |R_{t+1} \setminus Q_{\mathrm{bad}}|. \tag{3.9}$$

Indeed, In Lemma 3.4 we prove that, with probability at least $1 - 2^{-2-\min\{t,\log\log n\}}$, we have $|Q_{\mathrm{bad}}| \le |Q|/16$. By Line 2 $|Q| \ge 20$, so, by definition of R_{t+1} we have $|Q|/5 \le \lfloor |Q|/4 \rfloor \le |R_{t+1}|$. Hence,

$$\begin{aligned}|Q_{\mathrm{bad}}| &\le |Q|/16 < |Q|/5 - |Q|/16 \\ &\le |R_{t+1}| - |Q_{\mathrm{bad}}| \le |R_{t+1} \setminus Q_{\mathrm{bad}}|,\end{aligned} \tag{3.10}$$

with probability at least $1 - 2^{-2-\min\{t,\log\log n\}}$.

Since $F \subseteq F'$, and every point in R_t is closer to F' than any point in R_{t+1}, we have by (3.9) that we can map each point $b \in Q_{\mathrm{bad}} \cap R_t$ to a different point $p \in R_{t+1} \setminus Q_{\mathrm{bad}}$, such that

$$\mathrm{dist}(b, F) \le \mathrm{dist}(b, F') \le \mathrm{dist}(p, F') \le 2^{j+1} \,\mathrm{dist}(p, F^*).$$

Because $P_{\mathrm{bad}} \cap R_t \subseteq Q_{\mathrm{bad}} \cap R_t$, and $R_{t+1} \setminus Q_{\mathrm{bad}} \subseteq R_{t+1} \setminus P_{\mathrm{bad}}$, we conclude that we can map each point $b \in P_{\mathrm{bad}} \cap R_t$ to a different point $p \in R_{t+1} \setminus P_{\mathrm{bad}}$ such that $\mathrm{dist}(b, L) \le 2^{j+1} \,\mathrm{dist}(p, L^*)$, with probability at least $1 - 2^{-2-\min\{t,\log\log n\}}$. Thus, the probability that this holds for all the $t_{\max} - 1 \le \log n$ iterations is at least

$$\begin{aligned}&1 - \sum_{t=1}^{t_{\max}} 2^{-2-\min\{t,\log\log n\}} \\ &= 1 - \sum_{t=1}^{\lfloor \log\log n \rfloor} 2^{-2-t} - \sum_{t=\lfloor \log\log n \rfloor+1}^{t_{\max}} 2^{-2-\log\log n} \\ &\ge 1 - \frac{1}{4} - \frac{\log n}{2^{2+\log\log n}} = \frac{1}{2}.\end{aligned}$$

Using (3.8), this concludes the proof of Theorem 2.2. □

Acknowledgment

We would like to thank Sariel Har-Peled for very helpful discussions and comments concerning the topics of this paper.

4. REFERENCES

[1] P.K. Agarwal, S. Har-Peled and K. R. Varadarajan, Geometric approximation via coresets, *in Combinatorial and Computational Geometry*, Cambridge University Press, New York, Vol. 52, pp. 1–30, 2005.

[2] P. K. Agarwal and N. H. Mustafa, k-Means projective clustering, *Proc. 23rd Annu. ACM Symposium on Principles of Database Systems*, 2004, pp.155–165.

[3] P. K. Agarwal and C. M. Procopiuc, Approximation algorithms for projective clustering, *Proc. 11th Annu. ACM-SIAM Symposium on Discrete Algorithms*, 2000, pp. 538–547.

[4] P. K. Agarwal, C. M. Procopiuc and K. R. Vaeadarajan, Approximation algorithms for k-line center, *Proc. 10th Annu. European Symposium on Algorithms*, Vol. 2461, *Lecture Notes Comput. Sci.*, 2002, pp. 54–63.

[5] C. C. Aggarwal, C. M. Procopiuc, J. L. Wolf, and P. S. Yu, Fast algorithms for projected clustering, *Proc. ACM-SIGMOD Intl. Conf. Managment of Data*, 1999, pp. 61–72.

[6] C. C. Aggarwal and P. S. Yu, Finding generalized projected clusters in high dimensional spaces, *Proc. ACM-SIGMOD Intl. Conf. on Management of Data*, 2000, pp. 544–555.

[7] R. Agrawal, J. Gehrke, D. Gunopulos and P. Raghavan, Automatic subspace clustering of high-dimensional data for data mining applications, *Proc. ACM-SIGMOD Intl. Conf. on Management of Data*, 1998, pp. 94–105.

[8] A. Deshpande, L. Rademacher, S. Vempala and G. Wang, Matrix approximation and projective clustering via volume sampling, *Proc. 17th Annu. ACM-SIAM Symposium on Discrete Algorithms*, 2006, pp. 1117–1126.

[9] A. Deshpande and S. Vempala, Adaptive Sampling and Fast Low-Rank Matrix Approximation, *Proc. 10th International Workshop on Randomization and Computation, RANDOM 2006, pp. 292–303.*

[10] T. K. Dey, Improved bounds for planar k-sets and related problems, *Discrete Comput. Geom.* 19 (1998), 373–382.

[11] D. Feldman, A. Fiat and M. Sharir, Coresets for weighted facilities and their applications, *Proc. 47th Annu. IEEE Symposium on Foundations of Computer Science*, 2006, pp. 315–324.

[12] D. Feldman, M. Monemizadeh, and C. Sohler, Coresets with negative weightes and their applications for high dimensional clustering, manuscript, 2007.

[13] S. Har-Peled, Clustering motion, *Proc. 42nd Annu. IEEE Symposium on Foundations of Computer Science*, 2001, pp. 84–93.

[14] S. Har-Peled, Low rank matrix approximation in linear time, *Manuscript*, 2006.

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

[16] S. Har-Peled and K. Varadarajan, Projective clustering in high dimensions using coresets, *Proc. 18th Annu. ACM Symposium on Computational Geometry*, 2002, pp. 312–318.

[17] P, Indyk, Sublinear time algorithms for metric space problems, *Proc. 31st Annu. ACM Symposium on Theory of Computing*, 1999, pp. 428–434.

[18] J. W. Jaromczyk and M. Kowaluk, The two-line center problem from a polar view: A new algorithm and data structure, *Lecture Notes in Computer Science*, Vol. 955, Springer Verlag, 1995, pp. 13–18.

[19] N. Megiddo and A. Tamir, Finding least distance lines, *SIAM J. Algebraic Discrete Methods* 4 (1983), 207–211.

[20] R. Panigrahy, Minimum enclosing polytope in high dimensions, Computing Research Repository, cs.CG/0407020, 2004.

[21] K. Pearson, On lines and planes of closest fit to systems of points in space. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, 1901, pp. 559–572.

[22] C. M. Procopiuc, M. Jones, P. K. Agarwal and T. M. Murali, A Monte Carlo algorithm for fast projective clustering, *Proc. ACM SIGMOD Intl. Conf. on Management of Data*, 2002, pp. 418–427.

[23] T. Sarlós, Improved approximation algorithms for large matrices via random projections, *Proc. 47th Annu. IEEE Symposium on Foundations of Computer Science*, 2006, pp. 143–152.

[24] N. D. Shyamalkumar and K. R. Varadarajan, Efficient subspace approximation algorithm, *Proc. 18th Annu. ACM-SIAM Symposium on Discrete Algorithms*, 2007.