Optimal Core-Sets for Balls

Mihai Bădoiu[*] Kenneth L. Clarkson[†]

August 23, 2002

Abstract

Given a set of points $P \subset R^d$ and value $\epsilon > 0$, an ϵ-*core-set* $S \subset P$ has the property that the smallest ball containing S is within ϵ of the smallest ball containing P. This paper shows that any point set has an ϵ-core-set of size $\lceil 1/\epsilon \rceil$, and this bound is tight in the worst case. A faster algorithm given here finds an ϵ-core-set of size at most $2/\epsilon$. These results imply the existence of small core-sets for solving approximate k-center clustering and related problems. The sizes of these core-sets are considerably smaller than the previously known bounds, and imply faster algorithms; one such algorithm needs $O(dn/\epsilon + (1/\epsilon)^5)$ time to compute an ϵ-approximate minimum enclosing ball (1-center) of n points in d dimensions. A simple gradient-descent algorithm is also given, for computing the minimum enclosing ball in $O(dn/\epsilon^2)$ time. This algorithm also implies slightly faster algorithms for computing approximately the smallest radius k-flat fitting a set of points.

1 Introduction

Given a set of points $P \subset R^d$ and value $\epsilon > 0$, an ϵ-*core-set* $S \subset P$ has the property that the smallest ball containing S is within ϵ of the smallest ball containing P. That is, if the smallest ball containing S is expanded by $1+\epsilon$, then the expanded ball contains P. It is a surprising fact that for any given ϵ there is a core-set whose size is independent of d, depending only on ϵ. This is was shown by Bădoiu *et al.*[BHI], where applications to clustering were found, and the results have been extended to k-flat clustering.[HV].

While the previous result was that a core-set has size $O(1/\epsilon^2)$, where the constant hidden in the O-notation was at least 64, here we show that there are core-sets of size at most $\lceil 1/\epsilon \rceil$. This matches a lower bound of $\lceil 1/\epsilon \rceil$, as we show simply by considering a regular simplex. Such a bound is of particular interest for k-center clustering, where the core-set size appears as an exponent in the running time. A key lemma in the proof of the upper bound is the fact that the bound for Löwner-John ellipsoid pairs is tight for simplices.

While the existence proof for these optimal core-sets is a relatively slow algorithm, we give a fast construction for a somewhat larger core-set, of size at most $2/\epsilon$. We also give a simple algorithm for computing smallest balls, that looks something like gradient descent; this algorithm serves to prove a core-set bound, and can also be used to prove a somewhat better core-set bound for k-flats. Also, by combining this algorithm with the construction of the core-sets, we can approximate a 1-center in time $O(dn/\epsilon + (1/\epsilon)^5)$.

In the next section, we prove the $2/\epsilon$ core-set bound for 1-centers, and then describe the gradient-descent algorithm. Next we prove a lower bound, and then the matching upper bound. In the conclusion, we state the resulting bound for the general k-center problem.

2 Core-sets for 1-centers

Given a ball B, let c_B and r_B denote its center and radius, respectively. Let $B(P)$ denote the 1-center of P, the smallest ball containing it.

We restate the following lemma, proved in [GIV]:

Lemma 2.1 *If $B(P)$ is the minimum enclosing ball of $P \subset \mathbb{R}^d$, then any closed half-space that contains the center $c_{B(P)}$ also contains a point of P that is at distance $r_{B(P)}$ from $c_{B(P)}$. It follows that for any point q at distance K from $c_{B(P)}$, there is a point q' of P at distance at least $\sqrt{r^2_{B(P)} + K^2}$ from q.*

The last statement follows from the first by considering the halfspace bounded by a hyperplane perpendicular to $\overline{pc_{B(P)}}$, and not containing p.

Theorem 2.2 *There exists a set $S \subseteq P$ of size $2/\epsilon$ such that the distance between $c_{B(S)}$ and any point p of P is at most $(1+\epsilon)r_{B(P)}$.*

Proof: We proceed in the same manner as in [BHI]: we start with an arbitrary point $p \in P$ and set $S_0 = \{p\}$.

[*] MIT Laboratory for Computer Science; 545 Technology Square, NE43-371; Cambridge, Massachusetts 02139-3594; mihai@theory.lcs.mit.edu

[†] Bell Labs; 600 Mountain Avenue; Murray Hill, New Jersey 07974; clarkson@research.bell-labs.com

Let $r_i \equiv r_{B(S_i)}$ and $c_i \equiv c_{B(S_i)}$. Take the point $q \in P$ which is furthest away from c_i and add it to the set: $S_{i+1} \leftarrow S_i \bigcup \{q\}$. Repeat this step $2/\epsilon$ times.

Let $c \equiv c_{B(P)}$, $R \equiv r_{B(P)}$, $\hat{R} \equiv (1+\epsilon)R$, $\lambda_i \equiv r_i/\hat{R}$, $d_i \equiv ||c - c_i||$ and $K_i \equiv ||c_{i+1} - c_i||$.

If all the points are at distance at most $\hat{R}$ from c_i, then we are done. Otherwise, there is at least one point $q \in P$ such that $||q - c_i|| > \hat{R}$. If $K_i = 0$ then we are done, since the maximum distance from c_i to any point is at most R. If $K_i > 0$, then, as mentioned for the lemma above, let H be the hyperplane that contains c_i and is orthogonal to $\overline{c_i c_{i+1}}$. Let H^+ be the closed half-space bounded by H that does not contain c_{i+1}. By Lemma 2.1, there must be a point $p \in S_i \bigcap H^+$ such that $||c_i - p|| = r_i = \lambda_i \hat{R}$, and so $||c_{i+1} - p|| \geq \sqrt{\lambda_i^2 \hat{R}^2 + K_i^2}$. Also, by the triangle inequality the distance from the new center to q is at least $\hat{R} - K_i$, so $\lambda_{i+1}\hat{R} \geq \hat{R} - K_i$. By combining the two inequalities we get

$$\lambda_{i+1}\hat{R} \geq \max(\hat{R} - K_i, \sqrt{\lambda_i^2 \hat{R}^2 + K_i^2}) \qquad (1)$$

We want a lower bound on λ_{i+1} that depends only on λ_i. Observe that the bound on λ_{i+1} is smallest with respect to K_i when

$$\begin{aligned} \hat{R} - K_i &= \sqrt{\lambda_i^2 \hat{R}^2 + K_i^2} \\ \hat{R}^2 - 2K_i\hat{R} + K_i^2 &= \lambda_i^2 \hat{R}^2 + K_i^2 \\ K_i &= \frac{(1-\lambda_i^2)\hat{R}}{2} \end{aligned}$$

Using (1) we get that

$$\lambda_{i+1} \geq \frac{\hat{R} - \frac{(1-\lambda_i^2)\hat{R}}{2}}{\hat{R}} = \frac{1+\lambda_i^2}{2} \qquad (2)$$

Substituting $\gamma_i = \frac{1}{1-\lambda_i}$ in the recurrence (2), we get $\gamma_{i+1} = \frac{\gamma_i}{1-1/(2\gamma_i)} = \gamma_i(1 + \frac{1}{2\gamma_i} + \frac{1}{4\gamma_i^2}\ldots) \geq \gamma_i + 1/2$. Since $\lambda_0 = 0$, we have $\gamma_0 = 1$, so $\gamma_i \geq 1 + i/2$ and $\lambda_i \geq 1 - \frac{1}{1+i/2}$. That is, to get $\lambda_i \geq \frac{1}{1+\epsilon}$, it's enough that $i \geq 2/\epsilon$. ■

3 Simple algorithm for 1-center

The algorithm is the following: start with an arbitrary point $c_1 \in P$. Repeat the following step $1/\epsilon^2$ times: at step i find the point $p \in P$ farthest away from c_i, and move toward p as follows: $c_{i+1} \leftarrow c_i + (p - c_i)\frac{1}{i+1}$.

Claim 3.1 *If $B(P)$ is the 1-center of P with center $c_{B(P)}$ and radius $r_{B(P)}$, then $||c_{B(P)} - c_i|| \leq r_{B(P)}/\sqrt{i}$ for all i.*

Proof: Proof by induction: Let $c \equiv c_{B(P)}$. Since we pick c_1 from P, we have that $||c - c_1|| \leq R \equiv r_{B(P)}$. Assume that $||c - c_i|| \leq R/\sqrt{i}$. If $c = c_i$ then in step i we move away from c by at most $R/(i+1) \leq R/\sqrt{i+1}$, so in that case $||c - c_{i+1}|| \leq R/\sqrt{i+1}$. Otherwise, let H be the hyperplane orthogonal to $\overline{cc_i}$ which contains c. Let H^+ be the closed half-space bounded by H that does not contain c_i and let $H^- \equiv \mathbb{R} \setminus H^+$. Note that the furthest point from c_i in $B(P) \bigcap H^-$ is at distance less than $\sqrt{||c_i - c||^2 + R^2}$ and we can conclude that for every point $q \in P \bigcap H^-$, $||c_i - q|| < \sqrt{||c_i - c||^2 + R^2}$. By Lemma 2.1 there exists a point $q \in P \bigcap H^+$ such that $||c_i - q|| \geq \sqrt{||c_i - c||^2 + R^2}$. This implies that $p \in P \bigcap H^+$. We have two cases to consider:

- If $c_{i+1} \in H^+$, then the distance between c_{i+1} and c is maximized when $c_i = c$. Then, as before, we have $||c_{i+1} - c|| \leq R/(i+1) \leq R/\sqrt{i+1}$. Thus, $||c_{i+1} - c|| \leq R/\sqrt{i+1}$
- if $c_{i+1} \in H^-$, by moving c_i as far away from c and p on the sphere as close as possible to H^-, we only increase $||c_{i+1} - c||$. But in this case, $\overline{cc_{i+1}}$ is orthogonal to $\overline{c_i p}$ and we have $||c_{i+1} - c|| = \frac{R^2/\sqrt{i}}{R\sqrt{1+1/i}} = R/\sqrt{i+1}$.

■

4 A Lower Bound for Core-Sets

Theorem 4.1 *Given $\epsilon > 0$, there is a pointset P such that any ϵ-core-set of P has size at least $\lceil 1/\epsilon \rceil$.*

Proof: We can take P to be a regular simplex with $d+1$ vertices, where $d \equiv \lfloor 1/\epsilon \rfloor$. A convenient representation for such a simplex has vertices that are the natural basis vectors $e_1, e_2, \ldots, e_{d+1}$ of R^{d+1}, where e_i has the i'th coordinate equal to 1, and the remaining coordinates zero. Let core-set S contain all the points of P except one point, say e_1. The circumcenter of the simplex is $(1/(d+1), 1/(d+1), \ldots, 1/(d+1))$, and its circumradius is

$$R \equiv \sqrt{(1 - 1/(d+1))^2 + d/(d+1)^2} = \sqrt{d/(d+1)}.$$

The circumcenter of the remaining points is $(0, 1/d, 1/d, \ldots, 1/d)$, and the distance R' of that circumcenter to e_1 is

$$R' = \sqrt{1 + d/d^2} = \sqrt{1 + 1/d}.$$

Thus

$$R'/R = 1 + 1/d = 1 + 1/\lfloor 1/\epsilon \rfloor \geq 1 + \epsilon,$$

with equality only if $1/\epsilon$ is an integer. The theorem follows. ■

5 Optimal Core-Sets

In this section, we show that there are ϵ-core-sets of size at most $\lceil 1/\epsilon \rceil$. The basic idea is to show that the pointset for the lower bound, the set of vertices of a regular simplex, is the worst case for core-set construction.

We can assume, without loss of generality, that the input set is the set of vertices of a simplex; this follows from the condition that the 1-center of P is determined by a subset $P' \subset P$ of size at most $d+1$: that is, the minimum enclosing ball of P is bounded by the circumscribed sphere of P'. Moreover, the circumcenter of P' is contained in the convex hull of P. That is, the problem of core-set construction for P is reduced to the problem of core-set construction for a simplex $T = \operatorname{conv} P'$, where the minimum enclosing ball $B(T)$ is its circumscribed sphere.

Lemma 5.1 *Let B' be the largest ball contained in a simplex T, such that B' has the same center as the minimum enclosing ball $B(T)$. Then*

$$r_{B'} \le r_{B(T)}/d.$$

Proof: We want an upper bound on the ratio $r_{B'}/r_{B(T)}$; consider a similar problem related to ellipsoids: let $e(T)$ be the maximum volume ellipsoid inside T, and $E(T)$ be the minimum volume ellipsoid containing T. Then plainly

$$\frac{r^d_{B'}}{r^d_{B(T)}} \le \frac{\operatorname{Vol}(e(T))}{\operatorname{Vol}(E(T))},$$

since the volume of a ball B is proportional to r^d_B, and $\operatorname{Vol}(e(T)) \ge \operatorname{Vol}(B')$, while $\operatorname{Vol}(E(T)) \le \operatorname{Vol}(B(T))$. Since affine mappings preserve volume ratios, we can assume that T is a regular simplex when bounding $\operatorname{Vol}(e(T))/\operatorname{Vol}(E(T))$. When T is a regular simplex, the maximum enclosed ellipsoid and minimum enclosing ellipsoid are both balls, and the ratio of the radii of those balls is $1/d$. [H] (In other words, any simplex shows that the well-known bound for Löwner-John ellipsoid pairs is tight.[J]) Thus,

$$\frac{r^d_{B'}}{r^d_{B(T)}} \le \frac{\operatorname{Vol}(e(T))}{\operatorname{Vol}(E(T))} \le \frac{1}{d^d},$$

and so

$$\frac{r_{B'}}{r_{B(T)}} \le \frac{1}{d},$$

as stated. ■

Lemma 5.2 *Any simplex T has a facet F such that $r^2_{B(F)} \ge (1 - 1/d^2) r^2_{B(T)}$.*

Proof: Consider the ball B' of the previous lemma. Let F be a facet of T such that B' touches F. Then that point of contact p is the center of $B(F)$, since p is the intersection of F with the line through $c_{B(T)}$ that is perpendicular to F. Therefore

$$r^2_{B(T)} = r^2_{B'} + r^2_{B(F)},$$

and the result follows using the previous lemma. ■

Next we describe a procedure for constructing a core-set of size $\lceil 1/\epsilon \rceil$.

As noted, we can assume that P is the set of vertices of a simplex T, such that the circumcenter $c_{B(T)}$ is in T. We pick an arbitrary subset P' of P of size $\lceil 1/\epsilon \rceil$. (We might also run the algorithm of S2 until a set of size $\lceil 1/\epsilon \rceil$ has been picked, but such a step would only provide a heuristic speedup.) Let $R \equiv r_{B(P)}$. Repeat the following until done:

- find the point a of P farthest from $c_{B(P')}$;
- if a is no farther than $R(1+\epsilon)$ from $c_{B(P')}$, then return P' as a core-set;
- Let P'' be $P \cup \{a\}$;
- find the facet F of $\operatorname{conv} P''$ with the largest circumscribed ball;
- Let P' be the vertex set of F.

The first step (adding the farthest point a) will give an increased radius to $B(P'')$, while the second step (deleting the point $P'' \setminus \operatorname{vert} F$) makes the set P' more "efficient".

Theorem 5.3 *Any point set $P \subset R^d$ has an ϵ-core-set of size at most $\lceil 1/\epsilon \rceil$.*

Proof: Let r be the radius of $B(P')$ at the beginning of an iteration, and let r' be the radius of $B(P')$ if the iteration completes. We will show that $r' > r$.

Note that if $r \ge R(1-\epsilon^2)$, the iteration will exit successfully: applying Lemma 2.1 to $c_{B(P')}$ and $c_{B(P)}$ (with the latter in the role of "q"), we obtain that there is a point $q' \in P'$ such that

$$R^2 \ge ||c_{B(P)} - q'||^2 \ge r^2 + ||c_{B(P')} - c_{B(P)}||^2,$$

so that

$$\epsilon^2 R^2 \ge R^2 - r^2 \ge ||c_{B(P')} - c_{B(P)}||^2,$$

implying that $c_{B(P')}$ is no farther than ϵR to $c_{B(P)}$, and so $c_{B(P')}$ is no farther than $R(1+\epsilon)$ from any point of

P, by the triangle inquality. We have, if the iteration completes, that

$$
\begin{aligned}
r^2 &< R(1-\epsilon^2) \le \hat{R}^2 \frac{1-\epsilon^2}{(1+\epsilon)^2} \\
&= \hat{R}^2 \frac{1-\epsilon}{1+\epsilon},
\end{aligned} \tag{3}
$$

where $\hat{R} \equiv R(1+\epsilon)$.

By reasoning as for the proof of Theorem 2.2,

$$
r_{B(P'')} \ge \frac{\hat{R} + r^2/\hat{R}}{2}, \tag{4}
$$

and we can use the lower bound of the previous lemma on the size of $B(F)$ to obtain

$$
r' \ge \frac{\hat{R} + r^2/\hat{R}}{2} \sqrt{1 - \frac{1}{\lceil 1/\epsilon \rceil^2}},
$$

and so

$$
\frac{r'}{r} \ge \frac{\hat{R}/r + r/\hat{R}}{2} \sqrt{1-\epsilon^2}.
$$

The right-hand side is decreasing in $r/\hat{R}$, and so, since from (3), $r < \hat{R}\sqrt{(1-\epsilon)/(1+\epsilon)}$, we have

$$
\frac{r'}{r} > \frac{\sqrt{\frac{1-\epsilon}{1+\epsilon}} + \sqrt{\frac{1+\epsilon}{1-\epsilon}}}{2} \sqrt{1-\epsilon^2} = 1.
$$

Therefore $r' > r$ when an iteration completes. Since there are only finitely many possible values for r, we conclude that the algorithm successfully terminates with an ϵ-core-set of size $\lceil 1/\epsilon \rceil$. ■

6 Conclusions

We have proven the existence of small core-sets for k-center clustering. The new bounds are not only asymptotically smaller but also the constant is much smaller that the previous results. These results combined with the techniques from [BHI] and [HV] allow us to get faster algorithms for the k-center problem and j-approximate k-flat respectively. We can solve the k-center problem in $2^{O((k \log k)/\epsilon)} dn$ while the previous bound was $2^{O((k \log k)/\epsilon^2)} dn$. Also, the running time for computing j-approximate k-flat (with or without outliers) is $dn^{O(kj/\epsilon^5)}$, while the previous known bound was $dn^{O(kj/\epsilon^5 \log \frac{1}{\epsilon})}$. By combining the two algorithms above we get an $O(dn/\epsilon + (1/\epsilon)^5)$ time algorithm for computing 1-center which is faster than the previously fastest algorithm, with running time $O(dn/\epsilon^2 + (1/\epsilon)^{10} \log \frac{1}{\epsilon})$.

References

[BHI] Mihai Bădoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. *Proceedings of the 34th Symposium on Theory of Computing*, 2002.

[H] Ralph Howard. The John Ellipsoid Theorem. `http://www.math.sc.edu/ howard/Notes/app-convex-note.pdf`, 1997.

[HV] Sariel Har-Peled, and Kasturi R. Varadarajan. Projective Clustering in High Dimensions using Core-Sets. *Symposium on Computational Geometry*, 2002.

[GIV] Ashish Goel, Piotr Indyk, and Kasturi R. Varadarajan. Reductions among high dimensional proximity problems. *Proceedings of the 12th ACM-SIAM Symposium on Discrete Algorithms*, 2001.

[J] Fritz John. Extremum problems with inequalities as subsidiary conditions. *Studies and Essays Presented to R. Courant on his 60th birthday*, 1948.