Optimal Core-Sets for Balls

Mihai Bădoiu* Kenneth L. Clarkson†

May 2006

Abstract

Given a set of points $P \subset \mathbb{R}^d$ 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. Some experimental results are also given, comparing this algorithm with a previous one, and with a more powerful, but slower one.

1 Introduction

Given a set of points $P \subset \mathbb{R}^d$ 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: the center of the smallest ball containing S is within $(1+\epsilon)r_P$ distance to any point of P, where r_P is the radius of the smallest ball containing P. Bădoiu *et al.* showed that for any given ϵ, there is an ϵ-core-set whose size depends only on ϵ, and not on the dimension d [BHI]. That paper also gave applications in approximate k-center and k-flat clustering. (See also [HV] .) Some of these algorithms have a running time that is exponential in the size of an ϵ-core-set, and so it is important to have a tight estimate of that size.

An earlier paper showed that there are core-sets of size at most $2/\epsilon$, but the worst-case lower bound, easily shown by considering regular simplices, is only $\lceil 1/\epsilon \rceil$[BC]. (Another earlier paper independently showed that there are ϵ-core-sets of size $O(1/\epsilon)$, as well as other results related to the minimum enclosing ball problem[KMV]). Here we show that the lower bound is tight: there are always ϵ-core-sets of size $\lceil 1/\epsilon \rceil$. 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.

The existence proof for these optimal core-sets is an algorithm that repeatedly tries to improve an existing core-set by swapping: given $S \subset P$ of size k, it tries to swap a point out of S, and another in from P, to improve the approximation made by S. Our proof shows that a $1/k$-approximate ball can be produced by this procedure. (That is, if the smallest ball containing the output set is expanded by $1 + 1/k$, the resulting ball contains the whole set.) While it is possible to bound the number of iterations of the procedure for a slightly sub-optimal bound, such as $1/(k-1)$, no such bound was found for the optimal case. However, we give experimental evidence that for random pointsets, the algorithm makes no change at all in the core-sets produced by the authors' previous

*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

procedure, whose guaranteed accuracy is only $2/k$. That is, the algorithm given here serves as a fast way of verifying that the approximation ϵ is $1/k$, and not just $2/k$.

We also consider an alternative local improvement procedure, with no performance guarantees, that gives a better approximation accuracy, at the cost of considerably longer running time.

Some notation: given a point set P, let $B(P)$ denote the 1-center of P, that is, the smallest ball containing P. Let c_P denote the center of $B(P)$, and let r_P denote the radius of $B(P)$. (Note that if P is itself a ball, c_P and r_P are the center and radius of that ball.)

The next two sections give the lower and upper bounds, respectively. Section 4 shows that similar bounds hold, using similar constructions, for a slightly different definition of core-sets. Next, the experimental results are given, and then some concluding remarks.

2 A Lower Bound for Core-Sets

Theorem 2.1 *Given $\epsilon > 0$, there exists $d \in \mathbb{N}$ and a point set $P \subset \mathbb{R}^{d+1}$ such that any ϵ-core-set of P has size at least $\lceil 1/\epsilon \rceil$.*

Proof: We can take P to be the set of $d+1$ vertices of a regular d-simplex, 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 $\mathbb{R}^{d+1}$, 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 c_P is $(1/(d+1), 1/(d+1), \ldots, 1/(d+1))$, and its circumradius is

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

The circumcenter c_S is $(0, 1/d, 1/d, \ldots, 1/d)$, and the distance $\|e_1 - c_S\|$ of that circumcenter to e_1 is

$$\|e_1 - c_S\| = \sqrt{1 + d/d^2} = \sqrt{1 + 1/d}.$$

Thus

$$\|e_1 - c_S\|/r_P = 1 + 1/d = 1 + 1/\lfloor 1/\epsilon \rfloor \geq 1 + \epsilon,$$

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

3 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 will need the following lemma, proven in [GIV].

Lemma 3.1 *Any closed half-space that contains the center c_P of the minimal enclosing ball of P also contains a point of P that is at distance r_P from c_P. It follows that for any point q at distance K from c_P, there is a point q' of P at distance at least $\sqrt{r_P^2 + K^2}$ from q.*

Lemma 3.2 *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*

$$r_{B'} \leq r_T/d.$$

Proof: We want an upper bound on the ratio $r_{B'}/r_T$; 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

$$\frac{r_{B'}^d}{r_T^d} \leq \frac{\mathrm{Vol}(e(T))}{\mathrm{Vol}(E(T))},$$

since the volume of a ball B is proportional to r_B^d, and $\mathrm{Vol}(e(T)) \geq \mathrm{Vol}(B')$, while $\mathrm{Vol}(E(T)) \leq \mathrm{Vol}(B(T))$. Since affine mappings preserve volume ratios, we can assume that T is a regular simplex when bounding $\mathrm{Vol}(e(T))/\mathrm{Vol}(E(T))$. 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,

$$\frac{r_{B'}^d}{r_T^d} \leq \frac{\mathrm{Vol}(e(T))}{\mathrm{Vol}(E(T))} \leq \frac{1}{d^d},$$

and so

$$\frac{r_{B'}}{r_T} \leq \frac{1}{d},$$

as stated. ∎

Lemma 3.3 *Any d-simplex T has a facet F such that $r_F^2 \geq (1 - 1/d^2) r_T^2$.*

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_T that is perpendicular to F. Therefore

$$r_B^2 = r_{B'}^2 + r_F^2,$$

and the result follows using the previous lemma. ∎

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

Algorithm. Pick an arbitrary subset $S \subset P$ of size $\lceil 1/\epsilon \rceil$. (We might also run the algorithm of [BC] until a set of size $\lceil 1/\epsilon \rceil$ has been picked, but such a step would only provide a heuristic speedup.) Repeat the following until done:

- Find the point a of P farthest from c_S; let $S_a := S \cup \{a\}$;
- Find the facet F of $\mathrm{conv}\, S_a$ with the largest circumscribed ball, and let $S_{a \setminus b}$ denote the vertex set of F;
- If $r_{S_{a \setminus b}} \leq r_S$, return S as an ϵ-core-set; otherwise set $S := S_{a \setminus b}$, and repeat these steps.

The step yielding S_a generally increases the radius, while the step yielding $S_{a \setminus b}$ makes a set that is more "efficient".

Lemma 3.4 *In the above algorithm, letting $K := \|c_{S_a} - c_S\|$,*

$$r_{S_a} \geq \|a - c_S\| - K$$

and

$$r_{S_a} \geq \sqrt{r_S^2 + K^2},$$

or equivalently, letting D be a value with $D \leq \|a - c_S\|$,

$$\frac{r_{S_a}}{r_S} \geq \max\{\|a - c_S\|/r_S - K/r_S, \sqrt{1 + (K/r_S)^2}\} \geq (D/r_S + r_S/D)/2.$$

Proof: This is a restatement of part of the proof of Theorem 2.2 of an earlier paper[BC]. For completeness, here is the proof. The first inequality follows from $r_{S_a} \geq \|a - c_{S_a}\|$ and the triangle inequality. For the second inequality, by Lemma 3.1, there is a point $q' \in S$ such that

$$r_{S_a} \geq ||c_{S_a} - q'|| \geq \sqrt{r_S^2 + K^2}.$$

The last statement of the lemma follows by picking $K/r_S = (\beta^2 - 1)/2\beta$, where $\beta := D/r_S$; this choice makes the two terms in the maximum equal, minimizing the lower bound expression. The minimum value is then $(\beta + 1/\beta)/2$, as given. ■

Theorem 3.5 *Any point set $P \subset \mathbb{R}^d$ has an ϵ-core-set of size at most $\lceil 1/\epsilon \rceil$.*

Proof: Let $\hat{R} := r_P(1 + \epsilon)$.

We will show that when the $a \in P$ in the first step is farther than $\hat{R}$ from c_S, it must hold that $r_S < r_{S_{a \setminus b}}$, so that the algorithm will not stop at the current iteration.

Suppose

$$||a - c_S|| > \hat{R}. \tag{1}$$

We will first use this assumption to show (2) below.

By the triangle inequality,

$$||c_S - c_P|| \geq ||a - c_S|| - ||a - c_P|| > r_P(1 + \epsilon) - r_P = \epsilon r_P,$$

so

$$||c_S - c_P||^2 > \epsilon^2 r_P^2.$$

Using this bound, and applying Lemma 3.1 to c_S and c_P (with the latter in the role of "q"), we obtain that there is a point $q' \in S$ such that

$$r_P^2 \geq ||c_P - q'||^2 \geq r_S^2 + ||c_S - c_P||^2 > r_S^2 + \epsilon^2 r_P^2,$$

and so

$$r_S^2 < r_P^2(1 - \epsilon^2) = \hat{R}^2 \frac{1 - \epsilon^2}{(1 + \epsilon)^2} = \hat{R}^2 \frac{1 - \epsilon}{1 + \epsilon}.$$

That is, assuming (1),

$$r_S < \hat{R}\sqrt{\frac{1 - \epsilon}{1 + \epsilon}}. \tag{2}$$

The assumption (1) and Lemma 3.4 imply

$$\frac{r_{S_a}}{r_S} \geq \frac{\hat{R}/r_S + r_S/\hat{R}}{2}, \tag{3}$$

by picking $D = \hat{R}$.

Using (3) and the lower bound of Lemma 3.3 on the size of $B(F)$, we obtain

$$\frac{r_{S_{a\setminus b}}}{r_S} \geq \frac{r_{S_a}}{r_S}\sqrt{1 - \frac{1}{\lceil 1/\epsilon \rceil^2}} \geq \frac{\hat{R}/r_S + r_S/\hat{R}}{2}\sqrt{1-\epsilon^2}.$$

The last expression is decreasing in $r_S/\hat{R}$, and so from (2), we have

$$\frac{r_{S_{a\setminus b}}}{r_S} > \frac{\sqrt{\frac{1-\epsilon}{1+\epsilon}} + \sqrt{\frac{1+\epsilon}{1-\epsilon}}}{2}\sqrt{1-\epsilon^2} = 1.$$

Therefore $r_{S_{a\setminus b}} > r_S$ when $||a - c_S|| > \hat{R}$, and so termination of the algorithm implies $||a - c_S|| \leq \hat{R} := r_P(1+\epsilon)$, for all $a \in P$. Since there are only finitely many possible values for r_S, we conclude that the algorithm successfully terminates with an ϵ-core-set of size $\lceil 1/\epsilon \rceil$. ■

4 Alternate Definition

An alternate definition for an ϵ-core-set S bases the size of the ball to contain P not on r_P, but rather on r_S. The following result shows that the above algorithm gives a core-set, in this alternate sense, whose size is best possible for the worst case, as provided by the lower-bound example above.

Theorem 4.1 *Any set P of points has a subset S of size at most $\lceil 1/\epsilon \rceil$ such that every point of P is within $r_S/(1-\epsilon)$ of c_S. There are sets P for which no smaller subset S has this property.*

Proof: The example of Section 2 implies that there are $\epsilon > 0$ so that ϵ-core-sets in this sense must have size at least $1/\epsilon$.

The algorithm yielding the upper bound also yields a core-set in this sense, which can be seen as follows. In the proof of the upper bound, Theorem 3.5, the condition that S should satisfy for this alternate definition is that every point $a \in P$ is within distance $r_{S_*}/(1-\epsilon)$ of c_{S_*}. So (1) is replaced by the assumption that

$$\|a - c_S\| > r_S/(1-\epsilon).$$

Lemma 3.4 and this assumption imply

$$\frac{r_{S_a}}{r_S} \geq (1/(1-\epsilon) + (1-\epsilon))/2 = 1 + \epsilon^2/2(1-\epsilon),$$

where D in the lemma takes the value $r_S/(1-\epsilon)$. Thus

$$\begin{aligned}
\frac{r^2_{S_{a\setminus b}}}{r^2_S} &> \left(\frac{r_{S_a}}{r_S}\right)^2 \sqrt{1-\epsilon^2}^2 \\
&\geq \left(1 + \frac{\epsilon^2}{2(1-\epsilon)}\right)^2 (1-\epsilon^2) \\
&= (1-\epsilon^2)(1 + \epsilon^2/(1-\epsilon) + \epsilon^4/4(1-\epsilon)^2) \\
&= 1 - \epsilon^2 + \epsilon^2(1+\epsilon) + (1-\epsilon^2)\epsilon^4/4(1-\epsilon)^2 \\
&= 1 + \epsilon^3 + \epsilon^4(1+\epsilon)/4(1-\epsilon) \\
&> 1 + \epsilon^3
\end{aligned}$$

and so the assumption $||a - c_S|| > r_S/(1-\epsilon)$ also implies that the algorithm will not exit after this iteration. Therefore, when the algorithm exits, $\|a - c_S\| \leq r_S/(1-\epsilon)$, and as above, the r_S values increase from iteration to iteration, and have only finitely many possible values. ■

The proof implies that the value of r_S increases by a factor of at least $\sqrt{1+\epsilon^3}$ at each iteration. Since we can assume that $r_S \geq r_P/2$ initially, the algorithm requires at most $2\log_{1+\epsilon^3} 2 = O(1/\epsilon^3)$ iterations to obtain an ϵ-core-set in this alternate sense.

5 Experimental Results

Some experimental results on the approximation ratios are shown in Figures 1 through 8, each for different dimensional random data and distributions. The ordinates are the sizes of the core-sets considered, and the abscissas are the percentage increase in radius needed to enclose the whole set, relative to the smallest enclosing sphere.

In the plots,

- (hot start) a plain line shows results for the algorithm given here, starting from the output of the previous algorithm guaranteeing a $2/k$-core-set;
- (old) a dashed line is for the previous algorithm guaranteeing a $2/k$-core-set;
- (random start) a bullet ($\bullet$) is for the algorithm given here, starting from a random subset;
- (1-swap) a dot (.) is for an algorithm that is like the one given here, but that works a little harder: it attempts local improvement by swapping a point into the core-set, and another point out of the core-set. The possible points considered for swapping in are the three farthest from the circumcenter of the current core-set, while the points considered for swapping out are those three whose individual deletion leaves the circumradius as large as possible.

Figures 9 through 16 show the number of iterations needed for the algorithms, using the same graphing scheme.

Note that the random-start algorithm often does as well or better as hot-start algorithm, although a small but non-trivial number of iterations are required, while often the hot-start algorithm needs few or no iterations: the optimal algorithm serves as a confirmation that the "old" algorithm returns a better result than guaranteed.

We also performed tests of the gradient-descent method described in [BC]. The algorithm is quite simple: start with an arbitrary point $c_1 \in p$. Repeat the following step K times: at step i find the point $p \in P$ farthest away from the current center c_i and move towards p as follows: $c_{i+1} \leftarrow c_i + (p - c_i)\frac{1}{i+1}$. For $K = 1/\epsilon^2$, this algorithm produces a point which is at distance at most ϵ away from the true center. For this requirement, it can be shown that this algorithm is tight on the worst case for the case of a simplex. However, if we require that the farthest away point from the point produced is at distance at most $(1+\epsilon)r_P$, it is not clear if the analysis of the algorithm is tight. In fact, to our surprise, in our experiments the distance between the point produced and the farthest away point is 99.999% of the time under $(1+1/K)r_P$ and always under $(1+1.1/K)r_P$. We tested the algorithm under normal and uniform distributions. An empiric argument to try to explain this unexpected behaviour is the following: it has been noted that the algorithm picks most (but not all) of the points from a small subset in a repetitive way, i.e., for example one point can appear every $5-10$ iterations. Now, if you only pick 2 points A and B in an alternate way (A,

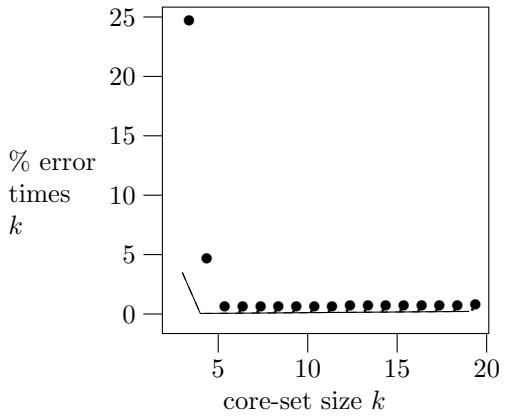

Figure 1: $d = 3$, normal

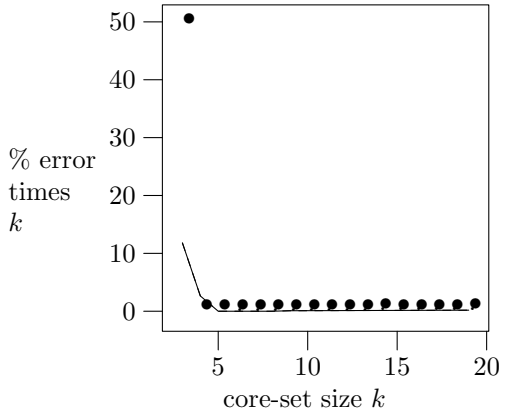

Figure 2: $d = 3$, uniform

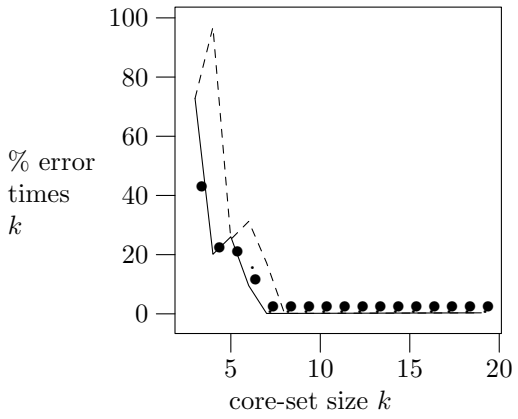

Figure 3: $d = 10$, normal

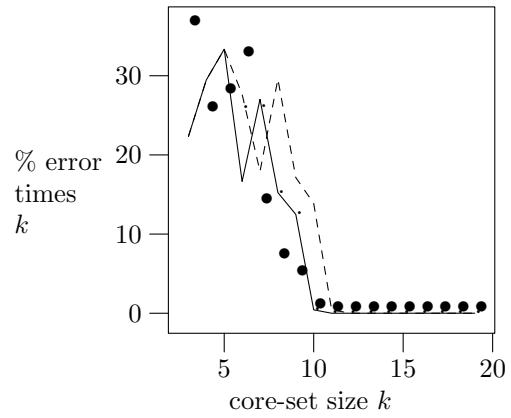

Figure 4: $d = 10$, uniform

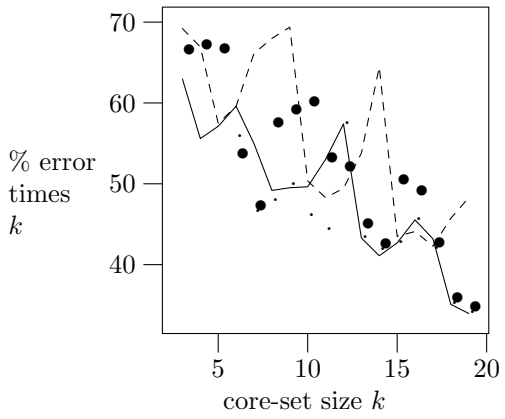

Figure 5: $d = 100$, normal

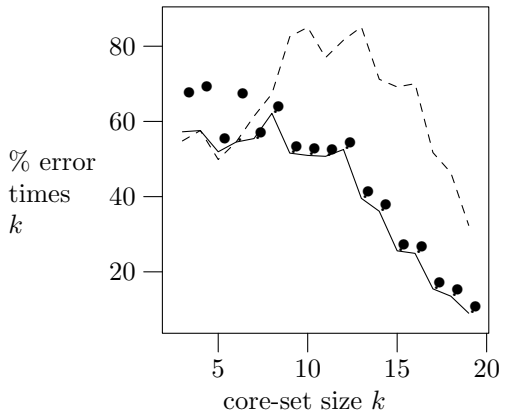

Figure 6: $d = 100$, uniform

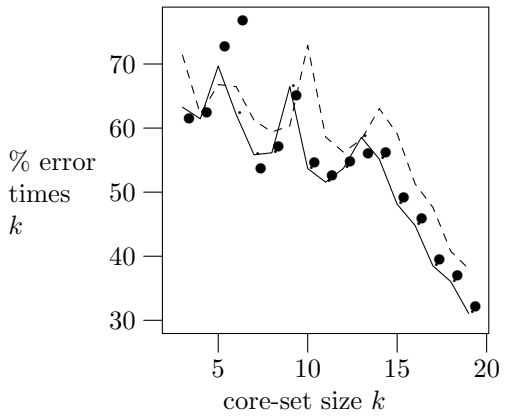

Figure 7: $d = 200$, normal

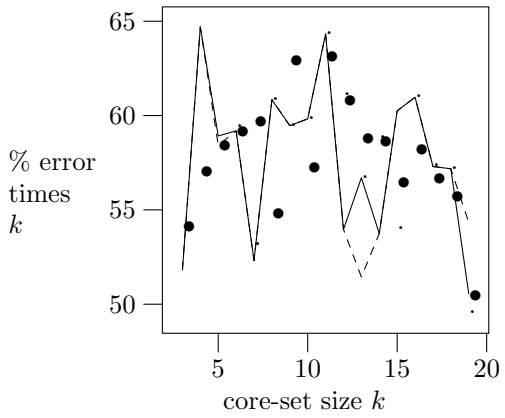

Figure 8: $d = 200$, uniform

B, A, B, ...), (i.e., subcase of the case when the solution is given by 2 points), the solution will converge quickly to the subspace spanned by A and B and it's easy to see that the error within the subspace will be at most $1/K$ after K steps. This empiric argument seems to give some intuition on why the algorithm give so much better error in practice. It may also be possible to prove this algorithm converges much faster theoretically.

Figures 17 through 19 show convergence results for the "gradient descent" algorithm. They show the percentage overestimate of the radius of the minimum enclosing ball, as a function of the number of iterations i. The first two figures show results for $d = 2, 3, 10, 100$, and 200, and the final figure shows the results for point distributed in an annulus with $d = 10$. Note that the error is often less than $1/i$ and never more than a small multiple of it.

6 Conclusions

In this paper we have proven the existence of optimal-sized core-sets for k-center clustering. We have also performed experimental tests and observed that in practice the error is much lower than the error that is guaranteed for a variety of core-set construction algorithms and the gradient-descent algorithm explained in [BC].

References

[BC] Mihai Bădoiu and Kenneth L. Clarkson. Smaller Core-Sets for Balls. *Proc. Symp. on Discrete Algorithms*, 2003.

[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.

[KMV] Piyush Kumar, Joseph S. B. Mitchell, and E. Alper Yıldırım. Approximate minimum enclosing balls in high dimensions using core-sets *Journal of Experimental Algorithmics, 2003.*

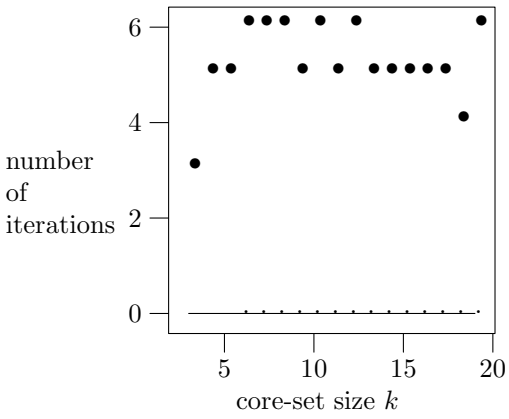

Figure 9: $d = 3$, normal

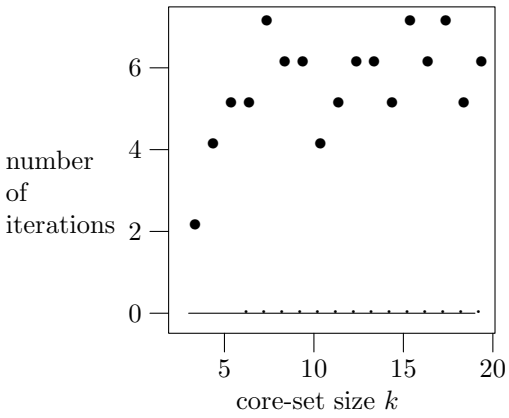

Figure 10: $d = 3$, uniform

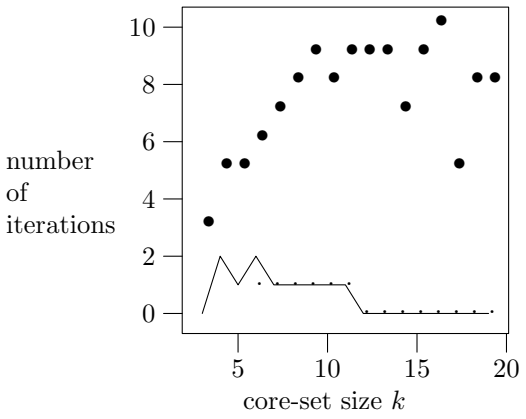

Figure 11: $d = 10$, normal

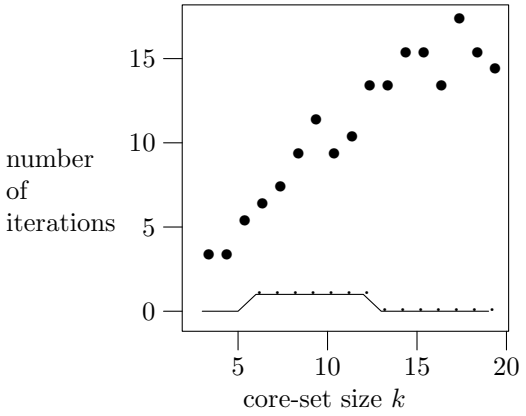

Figure 12: $d = 10$, uniform

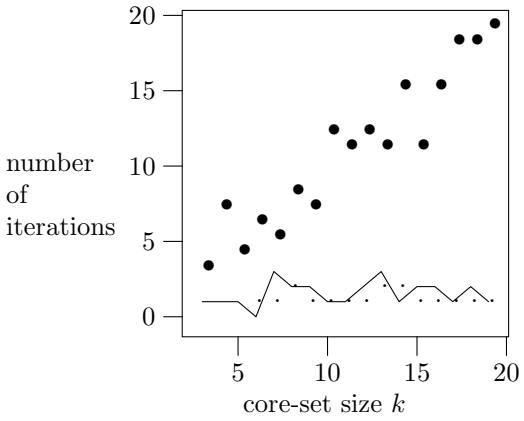

Figure 13: $d = 100$, normal

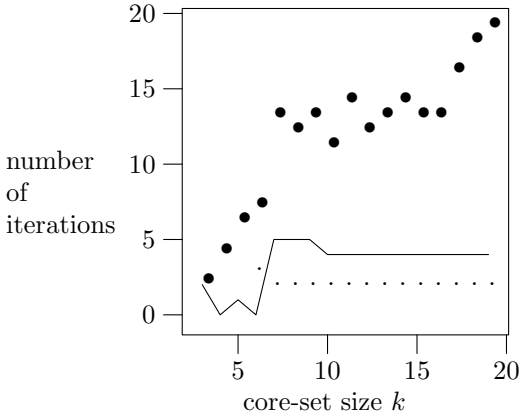

Figure 14: $d = 100$, uniform

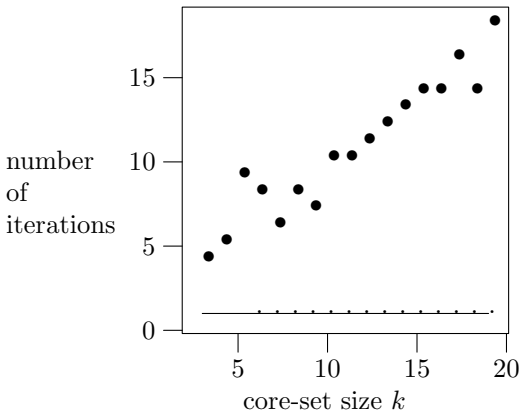

Figure 15: $d = 200$, normal

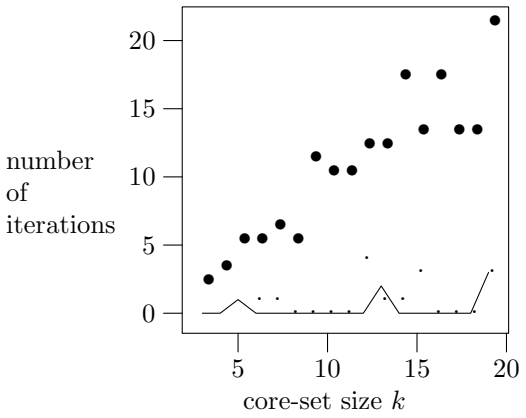

Figure 16: $d = 200$, uniform

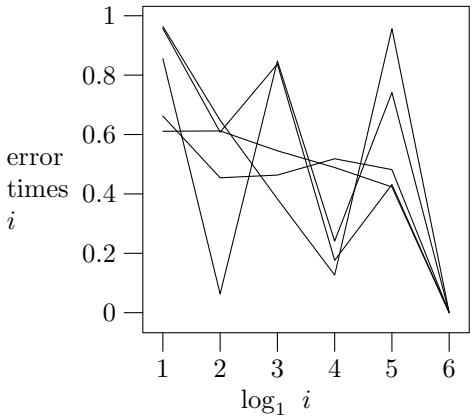

Figure 17: normal

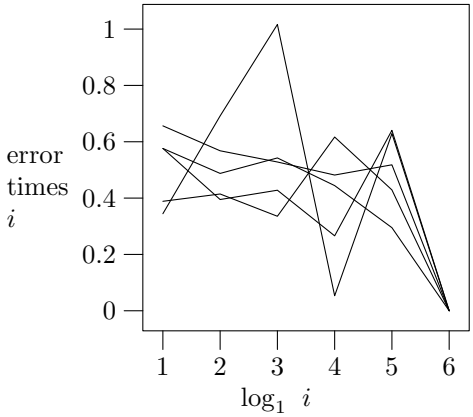

Figure 18: uniform

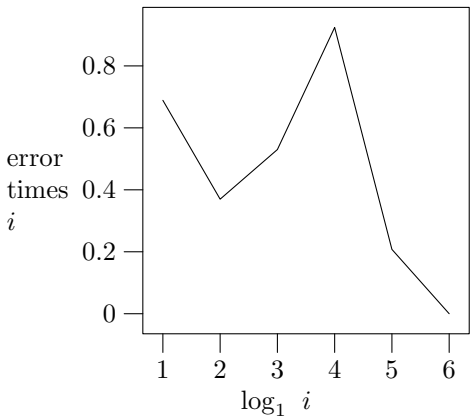

Figure 19: annulus