Smaller Core-Sets for Balls

Mihai Bădoiu* Kenneth L. Clarkson†

June 14, 2004

Abstract

We prove the existence of small core-sets for solving approximate k-center clustering and related problems. The size of these core-sets is considerably smaller than the previously known bounds, and imply faster algorithms; in particular, we get an algorithm needing $O(dn/\epsilon + (1/\epsilon)^5)$ time to compute an ϵ-approximate minimum enclosing ball (1-center) of n points in d dimensions. We also give a simple gradient-descent algorithm 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 of a given set of points.

1 Introduction

Given a set of points $P \subset R^d$ and value $\epsilon > 0$, a *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 $2/\epsilon$. This is not so far from a lower bound of $1/\epsilon$, which is easily shown by considering a regular simplex in $1/\epsilon$ dimensions. Such a bound is of particular interest for k-center clustering, where the core-set size appears as an exponent of n in the running time.

Our proof is a simple effective construction. 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 compute a 1-center in time $O(dn/\epsilon + (1/\epsilon)^5)$.

In the next section, we prove the core-set bound for 1-centers, and then describe the gradient-descent algorithm. In the conclusion, we state the resulting bound for the general k-center problem.

*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

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 $c_{B(P)}$.*

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 $S_0 = \{p\}$. 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 $c \equiv c_{B(P)}$, $R \equiv r_{B(P)}$, $\lambda_i \equiv r_i/R$, $d_i \equiv ||c - c_i||$ and $K_i \equiv ||c_{i+1} - c_i||$. Since the radius of the minimum enclosing ball is R, there is at least one point $q \in P$ such that $||q - c_i|| \geq R$. If $K_i = 0$ then we are done, since the maximum distance from c_i to any point is at most R. If $K_i > 0$, let H be the hyperplane that contains c_i and is orthogonal to (c_i, c_{i+1}). Let H^+ be the closed half-space bounded by H that does not contain c_{i+1}. By Lemma Lemma 2.1, there must be a point $p \in S_i \bigcap H^+$ such that $||c_i - p|| = r_i = \lambda_i R$, and so $||c_{i+1} - p|| \geq \sqrt{\lambda_i^2 R^2 + K_i^2}$. Therefore,

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

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

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

Using (1) we get that

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

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 $\lambda_i > 1 - \epsilon$, it's enough that $1 + i/2 \geq 1/\epsilon$, or 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 (c, c_i) which contains c. Let H^+ be the closed half-space bounded by H that does not contain c_i and let $H^- = \mathbb{R} \setminus H^+$. 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 $c_{i+1} \in H^+$, by moving c_i towards c we only increase $||c_{i+1} - c||$, and as noted before if $c_i = c$ 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, (c, c_{i+1}) is orthogonal to (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 Conclusions

In this paper we showed the existence of small core-sets for solving 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.

[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.