Adaptive Sampling for k-Means Clustering

Ankit Aggarwal[1], Amit Deshpande[2], and Ravi Kannan[2]

[1] IIT Delhi
zenithankit@gmail.com
[2] Microsoft Research India
{amitdesh,kannan}@microsoft.edu

Abstract. We show that *adaptively* sampled $O(k)$ centers give a constant factor bi-criteria approximation for the k-means problem, with a constant probability. Moreover, these $O(k)$ centers contain a subset of k centers which give a constant factor approximation, and can be found using LP-based techniques of Jain and Vazirani [JV01] and Charikar et al. [CGTS02]. Both these algorithms run in effectively $O(nkd)$ time and extend the $O(\log k)$-approximation achieved by the k-means++ algorithm of Arthur and Vassilvitskii [AV07].

1 Introduction

k-means is a popular objective function used for clustering problems in computer vision, machine learning and computational geometry. The k-means clustering problem on given n data points asks for a set of k centers that minimizes the sum of squared distances between each point and its nearest center. To write it formally, the k-means problem asks: Given a set $X \subseteq \mathbb{R}^d$ of n data points and an integer $k > 0$, find a set $C \subseteq \mathbb{R}^d$ of k centers that minimizes the following potential function.

$$\phi(C) = \sum_{x \in X} \min_{c \in C} \|x - c\|^2$$

We denote by $\phi_A(C) = \sum_{x \in A} \min_{c \in C} \|x - c\|^2$ the contribution of points in a subset $A \subseteq X$. Let C_{OPT} be the set of optimal k centers. In the optimal solution, each point of X is assigned to its nearest center in C_{OPT}. This induces a natural partition on X as $A_1 \cup A_2 \cup \cdots \cup A_k$ into disjoint subsets.

There is a variant of the k-means problem known as the *discrete k-means problem* where the centers have to be points from X itself. Note that the optima of the k-means problem and its discrete variant are within constant factors of each other. There are other variants where the objective is to minimize the sum of p-th powers of distances instead of squares (for $p \geq 1$), or to be more precise,

$$\left(\sum_{x \in X} \min_{c \in C} \|x - c\|^p \right)^{1/p}.$$

The $p = 1$ case is known as the *k-median problem* and the $p = \infty$ case is known as the *k-center problem*. Moreover, one can also ask the discrete k-means problem over arbitrary metric spaces instead of $\mathbb{R}^d$.

I. Dinur et al. (Eds.): APPROX and RANDOM 2009, LNCS 5687, pp. 15–28, 2009.

1.1 Previous Work

It is NP-hard to solve the k-means problem exactly, even for $k = 2$ [ADHP09], [Das08, KNV08] and even in the plane [MNV09]. Constant factor approximation algorithms are known based on linear programming techniques used for facility location problems but their running time is super-linear in n [JV01]. Kanugo et al. [KMN+04] give a $(9 + \epsilon)$-approximation via local search but in running time $O(n^3\epsilon^{-d})$ that has exponential dependence on d. There are polynomial time approximation schemes with running time linear in n and d but exponential or worse in k [dlVKKR03, HPM04, KSS04, Mat00, Che09]. Such a dependence on k may well be unavoidable, as shown in the case of the discrete k-median problem [GI03].

On the other hand, the most popular algorithm for the k-means problem is a simple iterative-refinement heuristic due to Lloyd [Llo82]: start with k arbitrary (or random) centers, compute the clusters defined by them, define the means of these clusters as the new centers, re-compute clusters and repeat. Lloyd's method is fast in practice but is guaranteed to converge only to a local optimum. In theory, the worst-case running time of Lloyd's heuristic is exponential even in the plane [Vat09]; however, a plausible explanation for its popularity could be its polynomial smoothed complexity [AMR09].

In attempts to bridge this gap between theory and practice, several randomized algorithms have been proposed based on the idea of sampling a subset of points as centers to get a constant factor approximation in time effectively $O(nkd)$. These centers could then be used to initialize the Lloyd's method. Mettu and Plaxton [MP02] and Ostrovsky et al. [ORSS06] give constant factor approximations but their results do not work unconditionally for all data sets.

The most relevant to our paper is a randomized algorithm called k-means++ due to Arthur and Vassilvitskii [AV07]. They propose a simple *adaptive* sampling scheme (they call it as D^2 sampling): in each step, pick a point with probability proportional to its current cost (i.e, its squared distance to the nearest center picked so far) and add it as a new center. This is similar to a greedy 2-approximation algorithm for the k-center problem that picks a point with the maximum cost in each step [Gon85]. Arthur and Vassilvitskii show that *adaptively* sampled k centers give, in expectation, an $O(\log k)$-approximation for the k-means problem. This also means, by Markov inequality, that we get an $O(\log k)$-approximation with a constant probability.

Similar sampling schemes have appeared in the literature on clustering of data streams [GMM+03, COP03] and online facility location [Mey01]. However, these sampling schemes are not as simple and their analysis is quite different.

Arthur and Vassilvitskii's analysis of their $O(\log k)$-approximation relies heavily on a non-trivial induction argument (Lemma 3.3 of [AV07]). Reverse engineering the same argument, they show a lower bound example where adaptively sampled k centers give $\Omega(\log k)$-approximation, in expectation. However, their lower bound is misleading in the sense that even though the expected error for adaptive sampling on this example is high, it gives an $O(1)$-approximation with high probability. The starting point for our work was the following question: Do

adaptively sampled k centers *always* give a *constant* factor approximation, with a *constant* probability?

1.2 Our Results

In Section 2, we extend the results of Arthur and Vassilvitskii to show that adaptively sampled $O(k)$ centers give a constant factor bi-criteria approximation for the k-means problem, with a constant probability. This probability of success can be boosted to arbitrary $(1-\delta)$ by repeating the algorithm $O(\log(1/\delta))$ times and taking the best solution.

In Section 3, we show that our adaptively picked $O(k)$ centers contain a subset of k centers that gives a constant factor approximation for the k-means problem, and this k-subset can be found by solving a weighted k-means problem on $O(k)$ points using the LP-based techniques of Jain and Vazirani [JV01] and Charikar et al. [CGTS02]. This gives us a randomized $O(1)$-approximation for the k-means problem with running time effectively $O(nkd)$.

Our proof techniques bypass the inductive argument of [AV07] and are general enough so as to be applicable in a wide range of other problems, such as facility location, where *adaptive* sampling could be useful.

In Appendix 4, we give a simpler proof of Arthur and Vassilvitskii's $\Omega(\log k)$ lower bound on the expected error of adaptively picked k centers to explain why their lower bound is misleading.

2 Bi-criteria Approximation by Adaptive Sampling

For a given set of centers, the current cost that each point pays in the k-means objective is its squared distance to the nearest center. In each step of *adaptive* sampling, we pick a point with probability proportional to its current cost and make it a new center. In this section, we show that adaptively sampling $O(k)$ points from the given data set itself gives a constant factor approximation for the k-means problem, with a constant probability.

Bi-criteria approximation by adaptive sampling
Input: a set $X \subseteq \mathbb{R}^d$ of n points and $k > 0$.
Output: a set $S \subseteq X$ of size $t = \lceil 16(k+\sqrt{k}) \rceil$.
Initialize $S_0 = \emptyset$.
For $i = 1$ to t do:

1. Pick a point x from the following distribution:
   $\Pr(\text{picking } x) \propto \phi_{\{x\}}(S_{i-1}) = \min_{c \in S_{i-1}} \|x - c\|^2$.
   (Note: For $i = 1$ step, the distribution is uniform.)
2. $S_i \leftarrow S_{i-1} \cup \{x\}$.
3. $i \leftarrow i + 1$.

Return $S \leftarrow S_t$.

Theorem 1. *Let $S \subseteq X$ be the subset of $t = \lceil 16(k+\sqrt{k})\rceil = O(k)$ points picked by the sampling algorithm given above. Then*

$$\phi(S) \leq 20\phi(C_{OPT}),$$

with probability at least 0.03. *(This probability could be boosted to $1-\delta$ by repeating the algorithm $\log(1/\delta)$ times and picking the best of the subsets.) The running time of our algorithm is $O(nkd)$.*

To prove correctness of our algorithm, we first analyze one step. Let S_{i-1} be the set of points obtained after the $(i-1)$-th step of our algorithm. In step i, we define

$$\begin{aligned}\mathsf{Good}_i &= \{A_j \ : \ \phi_{A_j}(S_{i-1}) \leq 10\phi_{A_j}(C_{OPT})\}\\ \mathsf{Bad}_i &= \{A_1, A_2, \ldots, A_k\} \setminus \mathsf{Good}_i\end{aligned}$$

Observe that at each step we pick a point with probability proportional to its cost at the current step. We first show that at each step, either we are already within a small constant factor of the optimum or we pick a point from Bad_i with high probability.

Lemma 1. *In the i-th step of our algorithm, either $\phi(S_{i-1}) \leq 20\phi(C_{OPT})$ or else the probability of picking a point from some cluster in Bad_i is $\geq 1/2$.*

Proof. Suppose $\phi(S_{i-1}) > 20\phi(C_{OPT})$. Then the probability of picking x from some cluster in Bad_i is equal to

$$\begin{aligned}\mathsf{Pr}\left(x \in A_j \text{ from some } A_j \in \mathsf{Bad}_i\right) &= \frac{\sum_{A_j \in \mathsf{Bad}_i} \phi_{A_j}(S_{i-1})}{\phi(S_{i-1})}\\ &= 1 - \frac{\sum_{A_j \in \mathsf{Good}_i} \phi_{A_j}(S_{i-1})}{\phi(S_{i-1})}\\ &\geq 1 - \frac{10\sum_{A_j \in \mathsf{Good}_i} \phi_{A_j}(C_{OPT})}{20\phi(C_{OPT})}\\ &\geq 1 - 1/2\\ &= 1/2.\end{aligned}$$

Note that once a cluster becomes good at some stage then it continues to remain good, i.e. $\mathsf{Good}_i \subseteq \mathsf{Good}_{i+1}$. Good clusters are those clusters that are being covered well enough by the centers we have chosen so far. We analyze a bad cluster and show how the algorithm makes it good.

Here is an important fact about the mean of a point set that we will use throughout the analysis. It can be thought of as an analog of the parallel axis theorem about moment of inertia from elementary physics.

Proposition 1. *Let μ be the mean of a set of points $A \subseteq \mathbb{R}^d$ and let $y \in \mathbb{R}^d$ be any point. Then*

$$\sum_{x \in A} \|x-y\|^2 = \sum_{x \in A} \|x-\mu\|^2 + |A|\,\|y-\mu\|^2.$$

Proof. Folklore. See Lemma 2.1.

Consider a cluster $A \in \mathsf{Bad}_i$. Let μ be the center of A in C_{OPT} and let $|A| = m$. (We drop the subscript j in A_j for the sake of simplicity.) Define $r = \sqrt{\phi_A(C_{OPT})/m}$, the root-mean-square optimal cost for points in A. Furthermore, let y be the point closest to μ in S_{i-1} and $d = \|\mu - y\|$. Observe that since $A \in \mathsf{Bad}_i$,

$$\begin{aligned}
10\phi_A(C_{OPT}) &\leq \phi_A(S_{i-1}) && \text{because } A \in \mathsf{Bad}_i \\
&= \sum_{x \in A} \min_{c \in S_{i-1}} \|x - c\|^2 \\
&\leq \sum_{x \in A} \|x - y\|^2 \\
&= \phi_A(C_{OPT}) + m\|\mu - y\|^2 && \text{by Proposition 1} \\
&= \phi_A(C_{OPT}) + md^2
\end{aligned}$$

Therefore,

$$d \geq \sqrt{\frac{9\phi_A(C_{OPT})}{m}} = 3r.$$

Define $B(\alpha) = \{x \in A \ : \ \|x - \mu\| \leq \alpha r\}$, where $0 \leq \alpha \leq 3 \leq d/r$. This is the set of points from A which are close to the center. The set $B(\alpha)$ is a good set to sample points from because any point $b \in B(\alpha)$ makes A a good cluster as shown below.

Lemma 2. *Let A be any cluster defined by C_{OPT} and let $b \in B(\alpha)$, for $0 \leq \alpha \leq 3$. Then*

$$\phi_A(S_{i-1} \cup \{b\}) \leq 10\phi_A(C_{OPT}).$$

Proof

$$\begin{aligned}
\phi_A(S_{i-1} \cup \{b\}) &= \sum_{x \in A} \min_{c \in S_{i-1} \cup \{b\}} \|x - c\|^2 \\
&\leq \sum_{x \in A} \|x - b\|^2 \\
&= \phi_A(C_{OPT}) + m\|\mu - b\|^2 && \text{by Proposition 1} \\
&\leq \phi_A(C_{OPT}) + m(\alpha r)^2 \\
&= (1 + \alpha^2)\phi_A(C_{OPT}) \\
&\leq 10\phi_A(C_{OPT}) && \text{since } \alpha \leq 3.
\end{aligned}$$

Now we show that $B(\alpha)$ contains a large fraction of points in A.

Lemma 3

$$|B(\alpha)| \geq m\left(1 - \frac{1}{\alpha^2}\right), \qquad \textit{for } 1 \leq \alpha \leq 3.$$

Proof

$$\begin{aligned}
\phi_A(C_{OPT}) &\geq \phi_{A\setminus B(\alpha)}(C_{OPT}) \\
&= \sum_{x\in A\setminus B(\alpha)} \min_{c\in C_{OPT}} \|x-c\|^2 \\
&= \sum_{x\in A\setminus B(\alpha)} \|x-\mu\|^2 \\
&\geq |A\setminus B(\alpha)|\,(\alpha r)^2 \\
&= \left(1-\frac{|B(\alpha)|}{m}\right) m(\alpha r)^2 \\
&= \left(1-\frac{|B(\alpha)|}{m}\right) \alpha^2 \phi_A(C_{OPT}),
\end{aligned}$$

which implies that

$$|B(\alpha)| \geq m\left(1-\frac{1}{\alpha^2}\right).$$

The following lemma states that the cost of $B(\alpha)$ is a substantial fraction of the cost of A with respect to the current S_{i-1} and thus also lower bounds the probability of the next point being chosen from $B(\alpha)$ given that it belongs to A.

Lemma 4

$$\mathsf{Pr}\left(x\in B(\alpha) \mid x\in A \text{ and } A\in \mathsf{Bad}_i\right) = \frac{\phi_{B(\alpha)}(S_{i-1})}{\phi_A(S_{i-1})} \geq \frac{(3-\alpha)^2}{10}\left(1-\frac{1}{\alpha^2}\right).$$

Proof. To prove the above lemma, we obtain an upper bound on $\phi_A(S_{i-1})$ and a lower bound on $\phi_{B(\alpha)}(S_{i-1})$ as follows.

$$\begin{aligned}
\phi_A(S_{i-1}) &= \sum_{x\in A} \min_{c\in S_{i-1}} \|x-c\|^2 \\
&\leq \sum_{x\in A} \|x-y\|^2 \\
&= \phi_A(C_{OPT}) + m\,\|\mu-y\|^2 \qquad \text{by Proposition 1} \\
&= m(r^2+d^2).
\end{aligned}$$

Observe that $\alpha r \leq d$ and $d = \|\mu-y\| \min_{c\in S_{i-1}} \|\mu-c\|$. For any $b\in B(\alpha)$ and any $c\in S_{i-1}$, we have

$$\|b-c\| \geq \|\mu-c\| - \|b-\mu\| \geq d - r\alpha \qquad \text{by triangle inequality.}$$

Thus, $\min_{c\in S_{i-1}} \|b-c\| \geq d-r\alpha$. Using this, we lower bound $\phi_{B(\alpha)}(S_{i-1})$ as follows.

$$\begin{aligned}\phi_{B(\alpha)}(S_{i-1}) &= \sum_{b\in B(\alpha)} \min_{c\in S_{i-1}} \|b-c\|^2 \\ &\geq |B(\alpha)|\,(d-\alpha r)^2 \\ &\geq m\left(1-\frac{1}{\alpha^2}\right)(d-\alpha r)^2 \qquad \text{from Lemma 3.}\end{aligned}$$

Putting these together we get

$$\mathsf{Pr}\left(x\in B(\alpha) \mid x\in A \text{ and } A\in \mathsf{Bad}_i\right) = \frac{\phi_{B(\alpha)}(S_{i-1})}{\phi_A(S_{i-1})} \geq \frac{(1-1/\alpha^2)(d-\alpha r)^2}{r^2+d^2}.$$

Observe that $(d-\alpha r)^2/(r^2+d^2)$ is an increasing function of d for $d\geq 3r\geq \alpha r$. Therefore,

$$\mathsf{Pr}\left(x\in B(\alpha) \mid x\in A \text{ and } A\in \mathsf{Bad}_i\right) \geq \left(1-\frac{1}{\alpha^2}\right)\frac{(3-\alpha)^2}{10}.$$

Lemma 5. *Suppose the point x picked by our algorithm in the i-th step is from $A\in\mathsf{Bad}_i$ and $S_i = S_{i-1}\cup\{x\}$. Then*

$$\mathsf{Pr}\left(\phi_A(S_i)\leq 10\phi_A(C_{OPT}) \mid x\in A \text{ and } A\in\mathsf{Bad}_i\right)\geq 0.126.$$

Proof. Immediately follows from Lemma 2 and Lemma 4 using $\alpha = 1.44225$ (by numerically maximizing the expression in α).

We want to show that in each step, with high probability, we pick a bad cluster A and make it good. Our proof uses the following well known facts about super-martingales.

Definition 1. *A sequence of real valued random variables $J_0, J_1, \ldots, J_t$ is called a* super-martingale *if for every $i>1$, $\mathsf{E}\left[J_i \mid J_0,\ldots,J_{i-1}\right]\leq J_{i-1}$.*

Super-martingales have the following concentration bound.

Theorem 2. (Azuma-Hoeffding inequality) *If $J_0, J_1, \ldots, J_t$ is a super-martingale with $J_{i+1}-J_i\leq 1$, then $\mathsf{Pr}\left(J_t\geq J_0+\delta\right)\leq \exp(-\delta^2/2t)$.*

Proof. (Proof of Theorem 1) By Lemma 1 and Lemma 5, we have

$$\begin{aligned}&\mathsf{Pr}\left(|\mathsf{Bad}_{i+1}|<|\mathsf{Bad}_i|\right)\\ &=\mathsf{Pr}\left(x\in A \text{ for some } A\in\mathsf{Bad}_i\right)\mathsf{Pr}\left(\phi_A(S_i)\leq 10\phi_A(C_{OPT})\,|\,x\in A \text{ and } A\in\mathsf{Bad}_i\right)\\ &\geq \frac{1}{2}\cdot 0.126\\ &= 0.063.\end{aligned}$$

For each step define an indicator variable X_i as follows.

$$X_i = \begin{cases}1 & \text{if } |\mathsf{Bad}_{i+1}| = |\mathsf{Bad}_i|\\ 0 & \text{otherwise.}\end{cases}$$

Thus, $\mathsf{Pr}(X_i = 0) \geq p = 0.063$ and $\mathsf{E}[X_i] \leq 1 - p$. Further, we define

$$J_i = \sum_{1 \leq j \leq i} (X_j - (1-p)).$$

Then $J_{i+1} - J_i \leq 1$ and

$$\begin{aligned}\mathsf{E}[J_i \mid J_0, \ldots, J_{i-1}] &= \mathsf{E}[J_{i-1} + X_i - (1-p) \mid J_0, \ldots, J_{i-1}] \\ &= J_{i-1} + \mathsf{E}[X_i \mid J_0, \ldots, J_{i-1}] - (1-p) \\ &\leq J_{i-1},\end{aligned}$$

which means that $J_1, J_2, \ldots, J_t$ is a super-martingale. So using Theorem 2 we get the following bound.

$$\mathsf{Pr}(J_t \geq J_0 + \delta) \leq \exp(-\delta^2/2t),$$

which means

$$\mathsf{Pr}\left(\sum_{i=1}^{t}(1 - X_i) \geq pt - \delta\right) \geq 1 - \exp(-\delta^2/2t).$$

Choosing $t = (k + \sqrt{k})/p \leq 16(k + \sqrt{k})$ and $\delta = \sqrt{k}$, we obtain

$$\begin{aligned}\mathsf{Pr}\left(\sum_{i=1}^{(k+\sqrt{k})/p}(1 - X_i) \geq k\right) &\geq 1 - \exp\left(\frac{-pk}{2(k+\sqrt{k})}\right) \\ &\geq 1 - \exp(-p/4).\end{aligned}$$

Therefore,

$$\mathsf{Pr}\Big(\text{there are no bad clusters after } (k+\sqrt{k})/p \text{ steps}\Big) \geq 1 - \exp(-p/4) \geq 0.03,$$

or equivalently

$$\mathsf{Pr}(\phi(S) \leq 10\phi(C_{OPT})) \geq 0.03.$$

There is nothing special about the approximation factor 20 in the proof above. One could start with any factor more than 4 and repeat the same proof. The higher the approximation factor, the better are the bounds on the probability and the number of centers picked. We get the following result as a straightforward generalization.

Theorem 3. *Our bi-criteria algorithm, when run for $t = O(k/\epsilon \cdot \log(1/\epsilon))$ steps, gives a $(4+\epsilon)$-approximation for the k-means problem, with a constant probability.*

3 Picking a k-Subset of S

If we use our bi-criteria solution S to cluster X, then every $x \in X$ is assigned to its closest point in S. This induces a natural partition of $X = X_1 \cup X_2 \cup \cdots \cup X_t$ into t disjoint subsets. Let $|X_i| = n_i$ and μ_i be the mean of points in X_i. Then for all i,

$$\phi_{X_i}(\{\mu_i\}) \leq \phi_{X_i}(S)$$

Weighted k-means clustering: Given a set $X \subseteq \mathbb{R}^d$ and weights w_i for each point $x_i \in X$, find a set $C \subseteq \mathbb{R}^d$ of k centers that minimizes the following potential function.

$$\phi'(C) = \sum_{x_i \in X} \min_{c \in C} w_i \left\| x_i - c \right\|^2.$$

We denote by $\phi'_A(C) = \sum_{x_i \in A} \min_{c \in C} w_i \left\| x_i - c \right\|^2$ the contribution of points in a subset $A \subseteq X$.

Using the bi-criteria solution S, we define a weighted k-means problem with points $X' = \{\mu_i \ : \ 1 \leq i \leq t\}$ and weights n_i assigned to point μ_i, respectively. Let C'_{OPT} denote the optimal solution for this weighted k-means problem.

Lemma 6

$$\phi'(C'_{OPT}) \leq 2\phi(C_{OPT}) + 2\phi(S).$$

Proof. By triangle inequality, for any $x \in X$ we have

$$\min_{c \in C_{OPT}} \left\| \mu_i - c \right\| \leq \left\| \mu_i - x \right\| + \min_{c \in C_{OPT}} \left\| x - c \right\|.$$

Therefore,

$$\min_{c \in C_{OPT}} \left\| \mu_i - c \right\|^2 \leq 2 \left\| \mu_i - x \right\|^2 + 2 \min_{c \in C_{OPT}} \left\| x - c \right\|^2$$

Summing over all $x \in X_i$,

$$\begin{aligned} \min_{c \in C_{OPT}} n_i \left\| \mu_i - c \right\|^2 &\leq \sum_{x \in X_i} 2 \left\| \mu_i - x \right\|^2 + 2 \min_{c \in C_{OPT}} \left\| x - c \right\|^2) \\ &\leq 2\phi_{X_i}(S) + 2\phi_{X_i}(C_{OPT}). \end{aligned}$$

Thus,

$$\begin{aligned} \phi'(C'_{OPT}) &\leq \phi'(C_{OPT}) \\ &= \sum_{1 \leq i \leq t} \min_{c \in C_{OPT}} n_i \left\| \mu_i - c \right\|^2 \\ &\leq \sum_{1 \leq i \leq t} 2\phi_{X_i}(S) + 2\phi_{X_i}(C_{OPT}) \\ &= 2\phi(S) + 2\phi(C_{OPT}) \end{aligned}$$

Theorem 4. *Let C be an β-approximation to the weighted k-means problem, i.e., $\phi'(C) \leq \beta\phi'(C'_{OPT})$. Then,*

$$\phi(C) \leq (2\beta + 1)\phi(S) + 2\beta\phi(C_{OPT}).$$

Proof. In the solution C, let μ_i be assigned to the center $c_j \in C$.

$$\begin{aligned}\sum_{x \in X_i} \min_{c \in C} \|x - c\|^2 &\leq \sum_{x \in X_i} \|x - c_j\|^2 \\ &= \sum_{x \in X_i} \|x - \mu_i\|^2 + n_i \|\mu_i - c_j\|^2 \quad \text{by Proposition 1} \\ &\leq \phi_{X_i}(S) + n_i \min_{c \in C} \|\mu_i - c\|^2 .\end{aligned}$$

Therefore,

$$\begin{aligned}\phi(C) &= \sum_{x \in X} \min_{c \in C} \|x - c\|^2 \\ &= \sum_{1 \leq i \leq t} \sum_{x \in X_i} \min_{c \in C} \|x - c\|^2 \\ &\leq \sum_{1 \leq i \leq t} \phi_{X_i}(S) + n_i \min_{c \in C} \|\mu_i - c\|^2 \\ &= \phi(S) + \phi'(C) \\ &\leq \phi(S) + \beta\phi'(C'_{OPT}) \\ &\leq (2\beta + 1)\phi(S) + 2\beta\phi(C_{OPT}).\end{aligned}$$

Note that Theorem 4 implies that a constant factor approximation to the weighted k-means problem constructed from our bi-criteria solution S is also a constant factor approximation to our original k-means problem. The advantage is that the weighted k-means problem is defined only on $O(k)$ points instead of n points. Interestingly, previous works on k-means clustering and a closely related problem of k-median clustering ([JV01],[CGTS02]) generalize to weighted k-means problem as well. This is because [CGTS02] solves the weighted k-median problem and the solution generalizes to distances where even a weak triangle inequality is satisfied. In case of squared Euclidean distance, for example,

$$\|x - z\|^2 \leq 2\left(\|x - y\|^2 + \|y - z\|^2\right).$$

We omit the details as the proofs are essentially the same as in [CGTS02]. These are LP-based algorithms and since the number of variables in our weighted k-means instance is $O(k)$ the overall running time of our sampling coupled with the LP-based algorithm for the resulting weighted k-means problem has running time $O(nkd + \text{poly}(k, \log n))$, which is effectively $O(nkd)$.

4 Simplified Lower Bound

Arthur and Vassilvitskii [AV06] prove that adaptive sampling for the k-means clustering gives an $O(\log k)$ approximation, in expectation. They also show an

example where adaptive sampling gives expected error at least $\Omega(\log k)$ times the optimum. Both these proofs are based on a tricky inductive argument.

In this note, we give a simplified proof of their lower bound. The example for lower bound is the same. Consider n points where they are grouped into k sets $S_1, S_2, \ldots, S_k$ of size n/k each. The points in each S_i form vertices of a regular simplex and the centers of these simplices $S_1, S_2, \ldots, S_k$ form vertices of a larger regular simplex. The smaller simplices live in different dimensions so that

$$\|x - y\| = \begin{cases} \delta & \text{if } x, y \in S_i \text{ for the same } i \\ \Delta & \text{if } x \in S_i \text{ and } y \in S_j \text{ for } i \neq j \end{cases}$$

The optimal k-means clustering uses centers of these regular simplices $S_1, S_2, \ldots, S_k$ and has error

$$\mathrm{OPT} = \frac{n-k}{2}\delta^2.$$

The probability that adaptive sampling picks all k centers from different S_i's is

$$\begin{aligned}
&\mathsf{Pr}\,(\text{adaptive sampling covers all } S_1, S_2, \ldots, S_k) \\
&= \prod_{i=1}^{k-1} \left(1 - \frac{i\left(\frac{n}{k} - 1\right)\delta^2}{\frac{n}{k}(k-i)\Delta^2 + i\left(\frac{n}{k} - 1\right)\delta^2}\right) \\
&\geq \prod_{i=1}^{k-1} \left(1 - \frac{i(n-k)\delta^2}{n(k-i)\Delta^2}\right) \\
&\geq 1 - \sum_{i=1}^{k-1} \frac{i(n-k)\delta^2}{n(k-i)\Delta^2} \qquad \text{by Weierstrass product inequality} \\
&= 1 - \frac{n-k}{n}\frac{\delta^2}{\Delta^2}\sum_{i=1}^{k-1}\frac{i}{k-i} \\
&\geq 1 - \frac{\delta^2}{\Delta^2}\sum_{i=1}^{k-1}\frac{k-i}{i} \\
&\geq 1 - \frac{\delta^2}{\Delta^2}k\left(\sum_{i=1}^{k-1}\frac{1}{i} - 1\right) \\
&\geq 1 - \frac{\delta^2}{\Delta^2}k\log k.
\end{aligned}$$

In fact, we will fix n, k, δ and use $\Delta \gg n, k, \delta$.

$$\begin{aligned}
&\mathsf{Pr}\,(\text{adaptive sampling covers all } S_1, S_2, \ldots, S_k) \\
&= \prod_{i=1}^{k-1} \left(1 - \frac{i\left(\frac{n}{k} - 1\right)\delta^2}{\frac{n}{k}(k-i)\Delta^2 + i\left(\frac{n}{k} - 1\right)\delta^2}\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \prod_{i=1}^{k-1} \left(1 - \frac{i(n-k)\delta^2}{2n(k-i)\Delta^2}\right) \\
&\leq 1 - \frac{1}{2}\sum_{i=1}^{k-1} \frac{i(n-k)\delta^2}{2n(k-i)\Delta^2} && \text{for } \Delta \gg k\delta \\
&= 1 - \frac{n-k}{2n}\frac{\delta^2}{\Delta^2}\sum_{i=1}^{k-1}\frac{i}{k-i} \\
&\leq 1 - \frac{\delta^2}{4\Delta^2}\sum_{i=1}^{k-1}\frac{k-i}{i} && \text{for } n \gg k \\
&\leq 1 - \frac{\delta^2}{8\Delta^2}k\left(\sum_{i=1}^{k-1}\frac{1}{i} - 1\right) \\
&= 1 - \frac{\delta^2}{8\Delta^2}k\log k.
\end{aligned}
$$

Thus

$$\mathsf{Pr}\left(\text{adaptive sampling covers all } S_1, S_2, \ldots, S_k\right) = 1 - \Theta\left(\frac{\delta^2}{\Delta^2}k\log k\right).$$

If our adaptive sampling covers all $S_1, S_2, \ldots, S_k$ then it's error is

$$\mathrm{Err}_{\text{no miss}} = (n-k)\delta^2,$$

whereas even if we miss (i.e., do not cover) one of the S_i's the error is at least

$$\mathrm{Err}_{\text{some miss}} \geq \frac{n}{k}\Delta^2.$$

So the expected error for adaptive sampling is given by

$$
\begin{aligned}
\mathsf{E}\left[\mathrm{Err}\right] &\geq \left(1 - \Theta\left(\frac{\delta^2}{\Delta^2}k\log k\right)\right)\mathrm{Err}_{\text{no miss}} + \Theta\left(\frac{\delta^2}{\Delta^2}k\log k\right)\mathrm{Err}_{\text{some miss}} \\
&\geq \left(1 - \Theta\left(\frac{\delta^2}{\Delta^2}k\log k\right)\right)(n-k)\delta^2 + \Theta\left(\frac{\delta^2}{\Delta^2}k\log k\right)\frac{n}{k}\Delta^2 \\
&\geq (n-k)\delta^2 + \frac{1}{\Delta^2}\cdot \text{some term} + \Theta(\log k)n\delta^2 \\
&= \Omega(\log k)\frac{n-k}{2}\delta^2 && \text{using } n \gg k \text{ and } \Delta \to \infty \\
&= \Omega(\log k)\mathrm{OPT}.
\end{aligned}
$$

Notice that even though the expected error is $\Omega(\log k)\mathrm{OPT}$, we get a constant factor approximation when the adaptive sampling covers all $S_1, S_2, \ldots, S_k$, which happens with a high probability.

5 Conclusion

We present a *simple* bi-criteria constant factor approximation algorithm for the k-means problem using *adaptive* sampling. Our proof techniques can be generalized to prove similar results for other variants of the k-means problem such as the k-median problem, or more generally, the ℓ_p version where we want to minimize the sum of p-th powers of distances rather than squares. This follows because of the weak triangle inequalities satisfied by the p-th powers of Euclidean distances, which gives us a weak form of the parallel axis theorem (i.e., Proposition 1). For the ℓ_p version, we get a similar bi-criteria algorithm where the number of centers picked by the algorithm is $O(k)$, where the constant depends exponentially on p.

Arthur and Vassilvitskii [AV07] show that adaptively sampled k centers give an $O(\log k)$-approximation for the k-means problem, in expectation (and hence also with a constant probability, by Markov inequality). In this paper, we show that adaptively sampled $O(k)$ centers give an $O(1)$-approximation for the k-means problem, with a constant probability. Looking at the lower bound example example (see Appendix 4) it is tempting to conjecture that adaptively sampled k centers give an $O(1)$-approximation for the k-means problem, with a constant probability. It would be nice to settle this conjecture.

Acknowledgements. The second author would like to thank Kasturi Varadarajan for several helpful discussions and Jaikumar Radhakrishnan for suggesting the analogy of Proposition 1 with the parallel axis theorem in elementary physics.

References

[ADHP09] Aloise, D., Deshpande, A., Hansen, P., Popat, P.: NP-hardness of Euclidean sum-of-squares clustering. Machine Learning 75(2), 245–248 (2009)

[AMR09] Arthur, D., Manthey, B., Röglin, H.: k-means has polynomial smoothed complexity (2009), `http://arxiv.org/abs/0904.1113`

[AV06] Arthur, D., Vassilvitskii, S.: How slow is the k-means method?. In: Annual Symposium on Computational Geometry (SOCG) (2006)

[AV07] Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. In: ACM-SIAM Symposium on Discrete Algorithms (SODA) (2007)

[CGTS02] Charikar, M., Guha, S., Tardos, M., Shmoys, D.: A constant factor approximation for the k-median problem. Journal of Computer and System Sciences (2002)

[Che09] Chen, K.: On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. Submitted to SIAM Journal on Computing (SICOMP) (2009)

[COP03] Charikar, M., O'Callaghan, L., Panigrahy, R.: Better streaming algorithms for clustering problems. In: ACM Symposium on Theory of Computing (STOC), pp. 30–39 (2003)

[Das08] Dasgupta, S.: The hardness of k-means clustering, Tech. Report CS2008-0916, UC San Diego (2008)

[dlVKKR03] de la Vega, F., Karpinski, M., Kenyon, C., Rabani, Y.: Approximation schemes for clustering problems. In: ACM Symposium on Theory of Computing (STOC), pp. 50–58. ACM Press, New York (2003)

[GI03] Guruswami, V., Indyk, P.: Embeddings and non-approximability of geometric problems. In: ACM-SIAM Symposium on Discrete Algorithms (SODA) (2003)

[GMM+03] Guha, S., Meyerson, A., Mishra, N., Motwani, R., O'Callaghan, L.: Clustering data streams: Theory and practice. IEEE Transactions on Knowledge and Data Engineering 15(3), 515–528 (2003)

[Gon85] Gonzalez, T.: Clustering to minimize the maximum intercluster distance. Theoretical Computer Science 38, 293–306 (1985)

[HPM04] Har-Peled, S., Mazumdar, S.: On core-sets for k-means and k-median clustering. In: ACM Symposium on Theory of Computing (STOC), pp. 291–300 (2004)

[JV01] Jain, K., Vazirani, V.: Approximation algorithms for metric facility loca- tion and k-median problems using the primal-dual schema and Lagrangian relaxation. Journal of ACM 48, 274–296 (2001)

[KMN+04] Kanugo, T., Mount, D., Netanyahu, N., Piatko, C., Silverman, R., Wu, A.: A local search approximation algorithm for k-means clustering. Computational Geometry 28(2-3), 89–112 (2004)

[KNV08] Kanade, G., Nimbhorkar, P., Varadarajan, K.: On the NP-hardness of the 2-means problem (unpublished manuscript) (2008)

[KSS04] Kumar, A., Sabharwal, Y., Sen, S.: A simple linear time $(1 + \epsilon)$- approximation algorithm for k-means clustering in any dimensions. In: IEEE Symposium on Foundations of Computer Science (FOCS), pp. 454–462 (2004)

[Llo82] Lloyd, S.: Least squares quantization in pcm. IEEE Transactions on Information Theory 28(2), 129–136 (1982)

[Mat00] Matoušek, J.: On approximate geometric k-clustering. Discrete and Computational Geometry 24(1), 61–84 (2000)

[Mey01] Meyerson, A.: Online facility location. In: IEEE Symposium on Foundations of Computer Science (FOCS) (2001)

[MNV09] Mahajan, M., Nimbhorkar, P., Varadarajan, K.: The planar k-means problem is NP-hard. In: Das, S., Uehara, R. (eds.) WALCOM 2009. LNCS, vol. 5431, pp. 274–285. Springer, Heidelberg (2009)

[MP02] Mettu, R., Plaxton, C.: Optimal time bounds for approximate clustering. Machine Learning, 344–351 (2002)

[ORSS06] Ostrovsky, R., Rabani, Y., Schulman, L., Swamy, C.: The effectiveness of Lloyd-type methods for the k-means problem. In: IEEE Symposium on Foundations of Computer Science (FOCS), pp. 165–176 (2006)

[Vat09] Vattani, A.: k-means requires exponentially many iterations even in the plane. In: Annual Symposium on Computational Geometry (SOCG) (2009)