A Appendix to "Streaming k-means approximation," N. Ailon, R. Jaiswal, and C. Monteleoni, NIPS 2009.

A.1 Future work

Kanungo *et al.* [KMNP+04] state that a local search heuristic results in a constant factor approximation for k-means, with a polynomial running time. The paper is not self contained with respect to running time analysis, and various key ideas required for completing it appear in [AGKM+04] and [CG99]. Arthur and Vassilvitskii [AV07] report that Kanungo *et al.*'s local search algorithm gives an approximation factor of $O(9+\epsilon)$ in time $O(n^3/\epsilon^d)$, where d is the dimensionality of the data[11]. We do not know what range of ϵ this claim assumes, and what the running time for some fixed ϵ (say, 1) would be. The local search algorithm can be readily plugged into our multi-level algorithm in Section 3.3. Our analysis does highlight, however, the importance of reducing the approximation constants in each invocation of a batch algorithm on memory blocks, because the final approximation constants are exponential in these constants (the power being $\log n/\log M$). Also, it is important to control the polynomial degree of the running time dependence of each invocation. Indeed, assume we can afford a streaming running time of at most $C \times n$ for some constant $C > 0$. If we are using a batch algorithm of running time $C' \times N^p$ on each size-N block for some $C' > 0$, then the maximal block size we can afford will be $\sim (C/C')^{1/p}$. The higher p is, however, the larger the resulting hierarchy depth r, and the worse the final approximation will be. The running time efficiency was, in fact, one of the main motivations for our derivation of k-means# for the purpose of obtaining a constant factor bi-criteria algorithm for k-means. We leave the analysis of plugging in different batch algorithms to our hierarchical solution for streaming k-means to future work.

A.2 Proof of Theorem 3.1

Proof. As mentioned in Section 3.1, the a' approximation of the number of centers is a direct consequence of the algorithm, so it remains to bound the approximation of the k-means objective.

Recall that the k-means cost of a set of centers T, with respect to a point set $S \subset \mathbb{R}^d$, is defined as $cost(T) = \sum_{x\in S} w(x) \cdot D(x,T)^2$, where $w(x)$ denotes the weight associated with the point x.[12] We will denote the optimal clustering by $T^* = \{t_1^*, t_2^*, \ldots, t_k^*\}$. Thus $T^* = \arg\min_{T\subset\mathbb{R}^d \,:\, |T|=k} cost(T)$. For a given set of cluster "centers" T, we will use the notation $t(x)$ to denote the element of T closest to x.

We will make use of the following lemmas, which extend the lemmas in [GMMM+03] (using the exposition of Dasgupta's lecture notes [Das08]), to the case of the k-means objective.

Lemma A.1. *$cost(S, T) \le 2\sum_{i=1}^{\ell} cost(S_i, T_i) + 2\, cost(S_w, T)$*

Proof. We start by rewriting the k-means cost by separating it into the sum over each part (in the partition made by the first step of the algorithm), of the cost of that part.

$$\begin{aligned} cost(S,T) = \sum_{i=1}^{\ell}\sum_{x\in S_i} D(x,T)^2 &\le \sum_{i=1}^{\ell}\sum_{x\in S_i}\left(D(x,t_i(x)) + D(t_i(x),T)\right)^2 \\ &\le 2\sum_{i=1}^{\ell}\sum_{x\in S_i} D(x,t_i(x))^2 + 2\sum_{i=1}^{\ell}\sum_{x\in S_i} D(t_i(x),T)^2 \\ &= 2\sum_{i=1}^{\ell} cost(S_i,T_i) + 2\sum_{i=1}^{\ell}\sum_{j=1}^{|T_i|} |S_{ij}| D(t_{ij},T)^2 \\ &= 2\sum_{i=1}^{\ell} cost(S_i,T_i) + 2\, cost(S_w,T) \end{aligned}$$

[11]The dependence on d could probably be taken care of using dimension reduction techniques, which we will not elaborate on here.

[12]For the unweighted case, we can assume that $w(x) = 1$ for all x.

The first inequality follows from applying the triangle inequality, $D(x,T) \leq D(x,t_i(x)) + D(t_i(x),T)$. The second inequality follows from applying $(a+b)^2 \leq 2a^2 + 2b^2$, to each term in the sum. □

First we will upper bound $\sum_{i=1}^{\ell} cost(S_i, T_i)$.

Lemma A.2. $\sum_{i=1}^{\ell} cost(S_i, T_i) \leq\ b \cdot cost(S, T^*)$

Proof.

$$\sum_{i=1}^{\ell} cost(S_i, T_i) \quad \leq \quad \sum_{i=1}^{\ell} b \cdot \min_{T' \subset \mathbb{R}^d} cost(S_i, T') \leq \sum_{i=1}^{\ell} b \cdot cost(S_i, T^*) \leq\ b \cdot cost(S, T^*)$$

The first inequality is due to T_i being the result of A which provides a b approximation to the optimal cost, for each S_i □

Now we will upper bound $cost(S_w, T)$.

Lemma A.3. $cost(S_w, T) \leq 2b' \cdot (\sum_{i=1}^{\ell} cost(S_t, T_i) +\ cost(S, T^*))$

Proof. First,

$$cost(S_w, T) \leq b' \cdot \min_{T' \subset \mathbb{R}^d} cost(S_w, T') \leq b' \cdot cost(S_w, T^*),$$

where the first inequality is due to T being the result of A' which provides a b' approximation to the optimal cost, for input S_w. The second inequality follows from the optimality of the right hand side for S_w. We can now $cost(S_w, T^*)$ bound as follows.

$$\begin{aligned} cost(S_w, T^*) &= \sum_{i=1}^{\ell} \sum_{j=1}^{|T_i|} |S_{ij}| D(t_{ij}, T^*)^2 \\ &\leq 2 \sum_{i=1}^{\ell} \sum_{j=1}^{|T_i|} \sum_{x \in S_{ij}} D(x, t_{ij})^2 + 2 \sum_{i=1}^{\ell} \sum_{j=1}^{|T_i|} \sum_{x \in S_{ij}} D(x, t^*(x))^2 \\ &= 2 \sum_{i=1}^{\ell} \sum_{x \in S_i} D(x, t_i(x))^2 + 2 \sum_{i=1}^{\ell} \sum_{x \in S_i} D(x, t^*(x))^2 \\ &= 2 \sum_{i=1}^{\ell} cost(S_t, T_i) + 2\, cost(S, T^*) \end{aligned}$$

The first inequality uses the triangle inequality and then $(a+b)^2 \leq 2a^2 + 2b^2$, similar to the proof of Lemma A.1. □

To attain the Theorem, we simply apply substitions from Lemmas A.2 and A.3 to the statement of Lemma A.1. □

A.3 Additional experimental results

The experimental set-up is described in the paper. Here we report standard deviations on the experiments run.

k	BL	DC-1	DC-2
5	$1.3302 \cdot 10^8$	$2.3433 \cdot 10^8$	$4.2539 \cdot 10^8$
10	$2.9615 \cdot 10^8$	$1.5782 \cdot 10^8$	$1.8783 \cdot 10^8$
15	$3.1203 \cdot 10^8$	$8.6772 \cdot 10^7$	$1.3998 \cdot 10^8$
20	$3.6956 \cdot 10^8$	$5.4427 \cdot 10^7$	$1.0200 \cdot 10^8$
25	$2.4563 \cdot 10^8$	$4.7795 \cdot 10^3$	$4.2328 \cdot 10^3$

k	BL	DC-1	DC-2
5	$0.0000 \cdot 10^6$	$1.5902 \cdot 10^6$	$2.5717 \cdot 10^6$
10	$1.2051 \cdot 10^6$	$5.2143 \cdot 10^5$	$5.3538 \cdot 10^5$
15	$0.0736 \cdot 10^6$	$3.0826 \cdot 10^5$	$3.2327 \cdot 10^5$
20	$0.2603 \cdot 10^6$	$1.1590 \cdot 10^5$	$2.2730 \cdot 10^5$
25	$0.5821 \cdot 10^6$	$1.2943 \cdot 10^5$	$1.2939 \cdot 10^5$

k	BL	DC-1	DC-2
5	$1.1687 \cdot 10^7$	$4.5518 \cdot 10^7$	$3.6388 \cdot 10^7$
10	$0.0000 \cdot 10^7$	$8.1261 \cdot 10^6$	$1.1827 \cdot 10^7$
15	$0.0373 \cdot 10^7$	$3.3351 \cdot 10^6$	$3.6615 \cdot 10^6$
20	$0.2398 \cdot 10^7$	$2.3456 \cdot 10^6$	$2.0151 \cdot 10^6$
25	$0.0631 \cdot 10^7$	$1.1220 \cdot 10^6$	$9.8168 \cdot 10^5$

Table 3: Standard deviations of the k-means cost (over 10 random restarts per algorithm): a) norm25 dataset, b) Cloud dataset, c) Spambase dataset.