On Lloyd's algorithm: new theoretical insights for clustering in practice

Cheng Tang & Claire Monteleoni

George Washington University

tangch@gwu.edu & cmontel@gwu.edu

A paradox for "k-means clustering"

k-means objective ϕ of $C = \{c_i, i \in [k]\}$ on a dataset X:

$$\phi_X(C) = \sum_{x \in X} \|x - C(x)\|^2, \text{ where } C(x) = \arg\min_{c \in C} \|x - c\|$$

Even though approximation algorithms exist, they are rarely used for applications. Instead, a few **heuristics**, most notably Lloyd's algorithm, are preferred and often successful in practice.

Lloyd's algorithm (a.k.a. the "k-means" algorithm)

Input: dataset X, $|X| = n$); k; samples size m, $m > k$.

1. Seeding: select an initial set of k centroids C_0
2. Repeat Lloyd's update until convergence or a **stopping criterion** is met.

   $S_t \leftarrow \{V(\nu_r) \cap X, \nu_r \in C_{t-1}, r \in [k]\}$

   $C_t \leftarrow \{m(S_r), S_r \in S_t, r \in [k]\}$

Output: C

Known results on Lloyd's algorithm

- No worst-case global performance guarantee (w.r.t. the k-means objective), and its running time can be exponential [6] on bad instances.
- Practically successful, and continues to be used and adapted

  Stochastic (mini-batch) k-means for large-scale clustering [2, 5].

  Spherical k-means for training single-layer NN (dictionary learning) [3].

Possible explanation for the paradox?

- Maybe datasets encountered in practice are not worst-case instances.
- Good solutions for applications may not need to optimize the k-means objective.

Our goal

Analyze Lloyd's algorithm and and its variants under **data clusterability** assumptions, beyond the scope of k-means clustering.

Main idea

To analyze the convergence property of Lloyd's algorithm, we need to first find a sufficient condition for which it indeed converges.

Characterizing "basin of attraction" for Lloyd's algorithm For any clustering T_* with centroids C_*, if it is an attractor for Lloyd's algorithm, then the following holds: if at some t, $\Delta^t < \delta$ implies $\Delta^{t^+} < \delta, \forall t^+ \geq t$, where $\Delta^t = \max_{r \in [k]} \|c_t^r - c_*^r\|$.

What conditions guarantee being an approximate attractor?

1. Δ^t is sufficiently small
2. C_* is a **well-clusterable** solution

We can also show Δ^t being small alone is not sufficient to guarantee convergence!

Our clusterability assumption

Definition 1 (($d_{rs}^*(f)$-center separability)**.** *A dataset-solution pair (X, T_*) satisfies $d_{rs}^*(f)$-center separability if we redefine $d_{rs}^*(f)$ above as $d_{rs}^*(f) := f\sqrt{\phi_*}(\frac{1}{\sqrt{n_r}} + \frac{1}{\sqrt{n_s}})$, where ϕ_* is the k-means cost of T_*.*

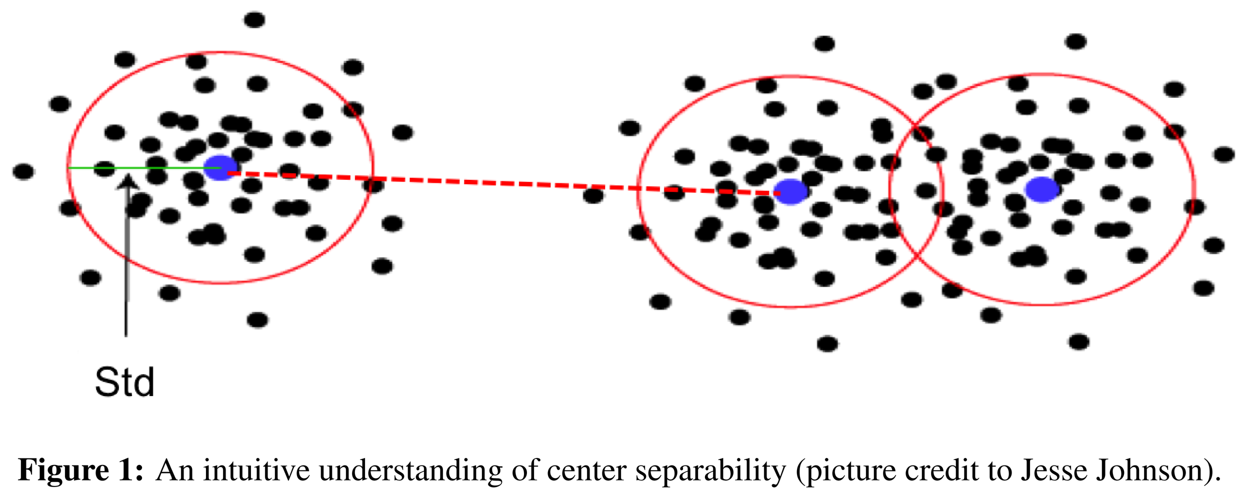

Figure 1: An intuitive understanding of center separability (picture credit to Jesse Johnson).

Interpretation: $C_T(T_*)$ is a good solution if for any of pair of its cluster r, s,

$$\|m(T_r) - m(T_s)\| > f\sqrt{\phi_*}(\frac{1}{\sqrt{n_r}} + \frac{1}{\sqrt{n_s}}) \geq f(std(r) + std(s))$$

Results

*Our results build on and generalizes the previous line of work [4, 1].

Global convergence of Lloyd's algorithm

Assume there is a dataset-solution pair (X, T_*) satisfying $d_{rs}^*(f)$-center separability, with $f > 32$.

Theorem 1 (Convergence rate)**.** *If at iteration t, $\forall r \in [k]$, $\|c_t^r - c_*^r\| < \beta_t \frac{\sqrt{\phi_*}}{\sqrt{n_r}}$ with $\beta_t < \max\{\gamma\frac{f}{8}, \frac{128}{9f}\}$ with $\gamma < 1$, then $\forall r \in [k]$, $\|c_{t+1}^r - c_*^r\| < \beta_{t+1}\frac{\sqrt{\phi_*}}{\sqrt{n_r}}$, with $\beta_{t+1} < \max\{\frac{\gamma}{2}\frac{f}{8}, \frac{128}{9f}\}$.*

Theorem 2 (Performance guarantee)**.** *If we cluster X using Algorithm **??**, where we choose a g-approximate k-means algorithm with $g < \frac{f^2}{128} - 1$ for the seeding, and execute Lloyd's update until convergence, then all but $\frac{81}{8f^2}$ fraction of the points will be correctly classified with respect to T_*.*

Interpretation: Any $O(k)$-approximation seeding + Lloyd's update works, and Lloyd's algorithm has linear convergence before reaching plateu.

Stochastic k-means for large-scale clustering (ongoing work)

Three challenges in analyzing stochastic k-means algorithm.

A scalable seeding algorithm The initialization of centers should not depend on the data size. We showed **running single-linkage on a uniform random sample of data**, an example of *Buckshot algorithm*, is a good seeding w.h.p.

Per-iteration convergence analysis We need to adapt the batch analysis of Lloyd's algorithm to the stochastic setting.

Lemma 1 (The stochastic gradient lemma)**.** *If $\|c_t^r - c_*^r\|^2 \leq \beta_t^2\frac{\phi_*}{n_r}$, then $E\{\|c_{t+1}^r - c_*^r\|^2 | F_t, \eta_{t+1}^r = \eta\} \leq (1-\eta)\beta_t^2\frac{\phi_*}{n_r} + \eta\beta_{t+1}\beta_t\frac{\phi_*}{n_r} + \tilde{E}\eta^2\|m(\hat{S}_r) - \nu_r\|^2$.*

A practical stopping criterion The stopping criterion needs to be practically verifiable, and locally measured.

- Our first proposed criterion $\frac{\delta^t}{\|c_{t-1}^r - c_{t-1}^s\|} < thres$, where $\delta^t := \max_{r \in [k]} \|c_t^r - c_{t-1}^r\|$. This is inspired by stopping criterion in practice.

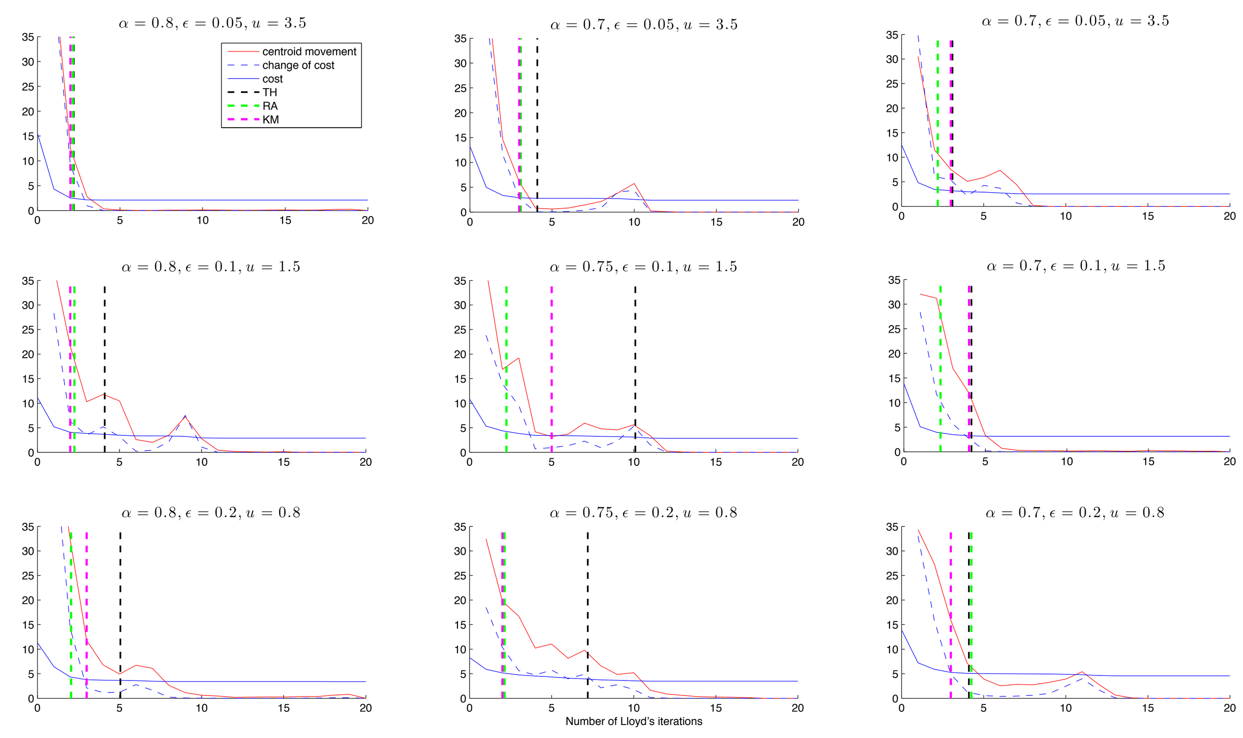

Figure 2: In each subfigure, Lloyd's algorithm is initialized on a solution with some degree of clusterability. We plot δ^t, $\Delta\phi_t$, and ϕ_t (scaled differently for convenient display) versus t; the vertical bars marks the stopped iteration according to different stopping criteria. The subfigures vary by clusterability of the dataset, parameterized by α, ϵ, u; clusterability decreases from top to bottom and from left to right.

- Our second proposed stopping criterion is based on a random quantity; given an i.i.d. sample M from X, we calculate a random quantity $\hat{\delta}(M) := \max_{x \in M} d(x, \{c_t^r, r \in [k]\})$. The second stopping criterion is $\frac{\hat{\delta}(M)}{\min \|c_t^r - c_t^s\|} < thres$. We can show if there exists a well-clusterable solution, C_*, then the second criterion is satisfied if and only if C_t is close to C_* w.h.p.

Questions we want to address in the future

- How to adapt the current analysis framework to other variants of Lloyd's algorithm?
- What kind of data yields what kind of clusterability? I.e., how does clusterability relate to other structural assumption of data? For example,
- How does clusterability change with respect to preprocessing?

References

[1] Pranjal Awasthi and Or Sheffet. Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 15th International Workshop, APPROX 2012, and 16th International Workshop, RANDOM 2012, Cambridge, MA, USA, August 15-17, 2012. Proceedings*, pages 37–49, 2012.

[2] Léon Bottou and Yoshua Bengio. Convergence properties of the k-means algorithms. In *Advances in Neural Information Processing Systems 7, [NIPS Conference, Denver, Colorado, USA, 1994]*, pages 585–592, 1994.

[3] Adam Coates, Andrew Y. Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, pages 215–223, 2011.

[4] Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pages 299–308, 2010.

[5] D. Sculley. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 1177–1178, 2010.

[6] Andrea Vattani. k-means requires exponentially many iterations even in the plane. *Discrete & Computational Geometry*, 45(4):596–616, 2011.