Comprehension Report-1-1

Base Paper:

Title: **Simple & Sharp Analysis of k-means||** Author(s): Va'clav, 2020, ICML

Report Objectives:

Fundamental understanding of approximation algorithms w.r.t k-means problem statement (Lloyd's algorithm, kmeans++, kmeans||) along with understanding of their approximation guarantees to the extent of practical application.

Introduction:

k-means|| [2], is a **distributed** variant of the k-means++ [3] algorithm which is an approximation algorithm for the k-means problem. The most popular solution to k-means is more formally known as Lloyd's algorithm (Which is also an approximation algorithm). This paper provides an improvement of the approximation guarantee of k-means||'s over-seeding step and proves that this bound is tight. Additionally, the Author of this paper also provides a refined simple analysis of the initial over-seeding procedure, followed by a sharp analysis of lower and upper bounds of the new approximation guarantee.

Preliminary Notation:

Let $X \subset \mathbb{R}^d$, we call $x \in X$ a point and $\mathbb{R}^d$ denotes a euclidean space of d-dimensions .

For $Y \subset X$ we denote its mean vector as follows, $\mu_Y = \frac{1}{|Y|}\sum_{y \in Y} y$, where μ_Y is called the **centroid** of Y. The distance between two points $x_i, x_j \in X$ is denoted as, $||x_i - xj|| = \sqrt{< x_i - x_j, x_i - x_j >}$. The distance between a point x and a set of points $Y$is, $d(x, Y) = \min_{y \in Y} ||x - y||$. We denote the cost function between $X, C \subset R^d$ as, $\varphi_X(C) = \sum_{x \in X} \min_{c \in C} ||x - c||^2$.

k-means problem formulation:

Let Σ be the power-set of $\mathbb{R}^d$, according to $\sigma-$ algebra. Lets call $\Gamma^k \subset \Sigma$, where $\forall \gamma \in \Gamma^k, |\gamma| = k \in \mathbb{Z}$. Given a $X \subset \mathbb{R}^d, k \in \mathbb{Z}$. Find a set $C \in \Gamma^k$ such that we get $\min_{C \in \Gamma^k} \varphi_X(C)$. We call C^* the optimal set of centers and φ^* the optimal cost.*This problem has shown to be NP-Hard.* [4]

k-means++ (Brief Overview):

The goal of k-means++ is to provide an approximation guarantee that is within constant bounds of the optimal solution of the k-means problem **for all instances** (meaning worst case guarantee).

It has been proven to be $O(log(k))$-competitive to the optimal solution of k-means [3]. The analysis of k-means++ relies on the proof of 3 lemmas, which are as follows: (Note: 1, 2 are proved in notes)

1. Let $z \in X \subset \mathbb{R}^d$ then, $\sum \varphi_X(z) - \sum \varphi_X^* = |X| ||\mu_X - z||^2$ [3]
2. Let $A \subset \mathbb{R}^d$ and $p \in A$ be a point that is sampled at random according to the uniform distribution. Then, $E[\varphi_A(p)] \leq 2\varphi_A^*$. [1, 3]
3. Let $A \subset \mathbb{R}^d$ and $p \in A$ be a point that is sampled at random according to the D^2- distribution .i.e. $\frac{\varphi_p(C)}{\varphi_A(C)}$. Let $C \subset \mathbb{R}^d$ (a random set of centers). Then, $E[\varphi_A(C \cup p)] \leq 8\varphi_A^*$. [1, 3]

k-means++ initialization algorithm: [2, 3]

function k-means-initialization(X, k):

1. $C \leftarrow \{ x \in X \}$, where x is a point sampled randomly according to the uniform distribution of X
2. while $|C| < k$ do
3. $\quad C \leftarrow C \cup x$, where $x \in X$ ~ D^2
4. end while
5. Lloyd's algorithm is then run on C as the initialization of the set of centers for X, k.

Reasoning behind D^2-distribution:

Additional note on step 5: At each iteration of t-loop we're picking multiple points in X based on the probability of picking that point (e.g if x in X has 0.9 probability it is very likely to be added to C' but there is a 0.1 chance that it is NOT added to C'). D^2 will give higher weights to those points in X that are further away from the current C.

k-means|| Algorithm: [1,2]

function k-means|| (X, k):

1. $\ell \leftarrow \Omega(k)$
2. $C \leftarrow \emptyset$
3. $C \leftarrow C \cup x$, where $x \in X \sim U$
4. $\psi \leftarrow \varphi_X(C)$
5. for $O(log\psi)$ times do
6. $C' \leftarrow \emptyset$
7. $C' \leftarrow C' \cup \{ x | \forall x \in X, x$ is sampled according to $p_x = \min(1, \frac{\ell\varphi_x(C)}{\varphi_X(C)}) \}$
8. $C \leftarrow C \cup C'$
9. end for
10. $W_C \leftarrow \{ w_c | \forall c \in C, w_c := \sum_{x \in X} 1\{ \varphi_x(c) = \min_{c \in C} ||x - c||^2 \} \}$
11. Re-cluster C into k-clusters using weights W_C in any weighted clustering algorithm (e.g. k-means++)
12. return C

Warmup Simple Analysis:

Theorem 1:

Suppose $t = O(log\frac{\varphi_X^*}{\varphi^*})$ & $\ell \geq k$, then over-seeding gives C s.t. $E[\varphi_X(C)] = O(\varphi^*)$.

Note: $\exists \varphi_X^* \neq \varphi^*$. But, $\varphi_X(\mu_X) = \varphi_X^*$.

Definition 1: (Settled Clusters)

Let $A \subset \Sigma_X$ s.t. $\mu_A \in C^* \in \Gamma^k \subset \Sigma_{\mathbb{R}^d}$ s.t. $|C^*| = k$.

A is settled w.r.t $C \in \Gamma^k \iff \varphi_A(C) \leq 10\varphi_A^*$. Otherwise, A is unsettled. (10?)

Lemma 4:

Let $C \in \Gamma$ be current set of centers, during over-seeding of k-means||.

Probability of A being unsettled at next over-seeding sampling step is $exp(-\frac{\ell\varphi_A(C)}{5\varphi_X(C)})$.

Proof of Lemma 4:

This proof is based on the following, Lemma 3 (L-3), Markov's Inequality (M.I.), Definition 1(D-1) & $1 + x \leq e^x$. Let Y be a +r.v. & $a > 0$

Let C be current centers during over-seeding, $p \in X$ currently sampled point (Acc. to 7.).

$C' = C \cup p$ (de-notion)

M.I. states, $P(Y \geq a) = \frac{E[Y]}{a}$.

$Y = \varphi_A(C')$ & $a = 10\varphi_A^*$. (According to D-1).

$\therefore P(\varphi_A(C') \geq 10\varphi_A^*) \leq \frac{E[\varphi_A(C')]}{10\varphi_A^*} \leq \frac{8}{10}$

$\implies P(\varphi_A(C') < 10\varphi_A^*) \geq \frac{1}{5}$.

$\exists A' \subset A$ such that A becomes "settled".

.i.e. $\frac{1}{5} \leq \frac{\varphi_{A'}(C')}{\varphi_A(C')}$, where

$\varphi_{A'}(C') = \sum\{\forall \varphi_{a'}(C') | a' \in A', \varphi_{a'}(C') \leq 10\varphi_A(C')\}$

(Note: best to read as probability of picking A' from A s.t. A is settled is at-least).

We know from step-7 in k-means|| algorithm that each point $x \in X$ is sampled with $p_x = \min(1, \frac{\ell\varphi_x(C')}{\varphi_X(C')})$.

$\therefore 1 \leq \frac{\ell\varphi_x(C')}{\varphi_X(C')} \implies x$ is sampled.

Which means **if**, $\frac{\varphi_X(C')}{\ell} \leq \varphi_x(C')$ **and** $x \in A'$ we sample x and say that A is settled.

Else,

$P(Y \geq a) \leq \prod_{x \in A'}(1 - \frac{\ell\varphi_x(C')}{\varphi_X(C')})$

$\leq \exp(-\sum_{x \in A'} \frac{\ell\varphi_x(C')}{\varphi_X(C')})$

$\leq \exp(-\frac{\ell\varphi_A(C')}{5\varphi_X(C')})$ (due to $1 + x \leq e^x$) $\square$

Proof of Theorem 1: (Uses Assumption)

Preliminary Notation: From here on φ_Y^t means "cost of point set Y after t sampling rounds. φ_U^t denotes total cost of unsettled clusters after t iterations.

Proposition 1: (Used as assumption, a.k.a ~Theorem 2 from [3])

$E[\varphi_U^{t+1}] \leq (1 - \frac{1}{50})\varphi_U^t.$

From Lemma 2, we know that after step 3 in k-means||. $E[\varphi_X(c)] \leq 2\varphi_X(\mu_X)$.

Note: Proof of Lemma 2 in notes is for $=$ not for $\leq$.

$\therefore E[\varphi_U^{t+1}] \leq \frac{49}{50}\varphi_U^t + 20\varphi^*$

$\implies$ for T times (total iterations) we will get,

$E[\varphi_U^T] \leq 2(\frac{49}{50})^T\varphi_X(\mu_X) + 20\varphi^* \sum_{t=0}^{T-1}(\frac{49}{50})^t$

$\leq 2(\frac{49}{50})^T\varphi_X(\mu_X) + 1000\varphi^*.$

$\therefore T = O(log\frac{\varphi_X(\mu_X)}{\varphi^*})$

$\because \varphi^T \leq \varphi_U^T + 10\varphi^*$ yields desired claim. $\square$

References:

1. "Simple & Sharp Analysis of k-means||"- Va'clav, 2020
2. "k-means++: The Advantages of Careful Seeding"-Authur & Vassilvitskii, 2007
3. "Scalable K-Means++" -Bahmani, 2012
4. NP-hardness of Euclidean sum-of-squares clustering - (Aloise, 2009; Mahajan, 2009)