A Survey on Information Theoretic Bounds for the K-Means Algorithm

Gurpreet Singh
150259

Ritesh Baldva
rb3447

Abstract

Lloyd's algorithm [12], commonly known as the k-means algorithm, provides an intuitive sub-optimal algorithmic solution to the data clustering problem. The k-means algorithm has a number of variants and applications in different domains and is the most popular clustering algorithm in Machine Learning. Despite its simplicity and popularity, there is lack of information theoretic bounds on the clustering solutions. It is affected by initialization strategies, along with the contuinity and unboundedness of the loss function. In this survey, we peruse some recent strategies which try to find theoretic bounds for the k-means solution based on assumptions on the underlying data distribution or modified initialization strategies.

1 Introduction

The data clustering problem that we deal with can be defined as minimizing the cost which is defined for a set of n points in $\mathbb{R}^d$ as the sum of squared l2 norm distance between each point and the closest center out of k chosen centers. This, in general, is NP-hard. [12] proposed a local search solution which begins with k arbitrary centers and iteratively improves the clustering to greedily minimize the cost. The simplicity and the ability to parallelize the algorithm [19] makes it one of the most popular clustering algorithms. However, due to the arbitrary initialization procedure and the poperties of the loss function, the algorithm suffers from unbounded cost estimates.

In section 3, we explore some uniform deviation bounds derived for the k-means algorithm. The strategies discussed assume different conditions for the underlying probability distributions.

The k-means++ [3] algorithm affords a well designed initialization strategy which helps us bound the expected cost within $\mathcal{O}(\log k)$ factor of the optimal cost. Despite the bounded results on expectation, the k-means++ strategy has some caveats that are highlighed in Section 4 along with some recent advances which try to tackle these issues. We also state the theoretical guarantees achieved by some of the recent seeding based strategies for k-means initialization without proof.

2 Uniform Deviation Bounds

If we use the Lloyd's algorithm for identifying the k-centres by approximating the k-means cost function, one problem that needs to be addressed is how well do the selected centers fit the underlying the data distribution. This problem refers to calculating to bounding the following quantity,

$$\left| \frac{1}{m} \sum_{j=1}^{m} \min_i \|x_j - p_i\|_2^2 - \int \min_i \|x - p_i\|_2^2 d\rho(x) \right| \tag{1}$$

In the above equation 1, ρ refers to the probability distribution of the data and $\{p_i\}_{i=1}^{k}$ refer to the set of selected centers. The paper establishes the above deviation to follow $\mathcal{O}(m^{-1/2})$, where m is the number of samples from the distribution data.

After running the algorithm, we obtain the cost on the samples, but evaluating the cost according to the probability measure can still make it unbounded. The idea is that the generally the samples will exist from areas of high density and the term $\int \| x - p \|_2^2 \, d\rho(x)$ will be negligibly small for points which are far away from the centers, because of low density at those points. Thus, we can create balls around such centers and be able to move around within the balls to chose a new center and the effect of that new center from points far away will still be of the same order as previously. This can be seen easily with the help of the triangle inequality. Thus we have two tasks, bounding the deviations within the balls and the deviation for points which lie outside the balls. Let's introduce the notation to follow the lemmas for finally bounding the deviation rates, which is the same as defined in the paper.

1. A convex differentiable function f, which for the hard k-means problem is, $f = \| x \|_2^2$. Let's also assume that the function is α-strongly convex and β-smooth. Let's also define the related Bregman divergence with respect to f, as $B_f(x, y)$.
2. Order-p moment bound M on ρ, i.e $\mathop{\mathbb{E}}_{\rho} \left[\left\| X - \mathop{\mathbb{E}}_{\rho} [\, X \,] \right\|^l \right] \leq M \;\forall l \in [1, p]$
3. P represents the set of centers and the corresponding cost of a single point x, with respect to P and f, as $\phi_f(x; P) = \min_{p \in P} B_f(x, p)$. Then, k-means cost of a set of points with respect to a probability measure ν is $\mathop{\mathbb{E}}_{\nu} \left[\, \phi_f(X; P) \, \right]$.

We first start with bounding the deviations between the k-th moments of the true measure and the empirical measure. This is done by first establishing a ball in the euclidean space which satisfies the following [Lemma A.6] in [18], where τ is a moment map,

$$B = \left\{ x \in \mathcal{X} : \| \tau(X) \| \leq (M/\epsilon)^{1/(p-k)} \right\} \tag{2}$$

This results in the following bound on the moment on the true measure ρ,

$$\int_B \| \tau(x) \|^k \, d\rho(x) \leq \epsilon \tag{3}$$

The deviations between the k-th moments is derived through the Minkowski's inequality [Lemma A.7] in [18] and a concentration bound established using bounded moments of order p [Lemma A.3][18],

$$\left| \int_B \| \tau(x) \|^k \, d\hat{\rho}(x) - \int_B \| \tau(x) \|^k \, d\rho(x) \right| \leq \sqrt{\frac{M' e p'}{2m}} \left(\frac{2}{\delta} \right)^{1/p'} \tag{4}$$

which holds with probability at least $1-\delta$ over draw of the sample size $m \geq p'/(M'e)$. In the above equation 4, $p' \geq 1$ and $M' = 2^{p'}\epsilon$ and the relation holds for the ball with the maximum radius defined among all possible moment orders less than p.

The main result for bounding the deviations in the k-means cost is defined int Theorem 3.2 of [18]. The actual idea lies behind breaking the deviation into different terms using triangle inequality. From the complete space, we first reduce down to space (ball) where a decent amount of probability mass is present. Then we concentrate on the centers that are in and around this ball. From here we focus on a particular possible center of a disk that covers this ball. We try to bound the k-mean costs with respect to this ball locally through moment methods and use bracketing to bound the costs when reducing from the whole space to this actual ball.

$$\begin{aligned}\left|\int \phi_f(x;P)d\rho(x) - \int \phi_f(x;P)d\rho(x)\right| \leq & \left|\int \phi_f(x;P)d\rho(x) - \int_B \phi_f(x;P\cap C)d\hat{\rho}(x)\right| \\ & + \left|\int_B \phi_f(x;P\cap C)d\rho(x) - \int_B \phi_f(x;Q)d\rho(x)\right| \\ & + \left|\int_B \phi_f(x;Q)d\rho(x) - \int_B \phi_f(x;Q)d\hat{\rho}(x)\right| \\ & + \left|\int_B \phi_f(x;Q)d\hat{\rho}(x) - \int_B \phi_f(x;P\cap C)d\hat{\rho}(x)\right| \\ & + \left|\int_B \phi_f(x;P\cap C)d\hat{\rho}(x) - \int \phi_f(x;P)d\hat{\rho}(x)\right|\end{aligned}$$

In the above breakdown the ball B represents the ball we concentrate on and C represents the region around this ball, which is considered viable to get centers from P. Finding the bracketing functions is easy, for the lower bound, it's trivially 0 and for the upper bound we can split using the triangle inequality around the expected mass.

$$\begin{aligned} l(x) &= 0 \\ u(x) &= 4(\beta/2)\left\| x - \mathop{\mathbb{E}}_{\rho}[\,X\,]\right\|_2^2\end{aligned}$$

Notice that $u(x)$ can be used as moment function for the equation 4 and with triangle inequality, thus used to bound the fifth term accordingly. We bound the first term with ϵ itself. For the third term, since Q is a cover of the region surrounding the ball B, we use the following fact that,

$$\sup_{x\in B}\left|\min_{p\in P} B_f(x,p) - \min_{q\in Q} B_f(x,q)\right| \leq \epsilon \tag{5}$$

which can be shown using properties of Bregman divergence only to have that,

$$\left|\int_B \phi_f(x;Q)d\rho(x) - \int_B \phi_f(x;Q)d\hat{\rho}(x)\right| \leq 4(\beta/2)R_C^2\sqrt{\frac{1}{2m}\log\left(\frac{2\,|\mathcal{N}|^k}{\delta}\right)} \tag{6}$$

Plugging all the above bounds back, we get the bound for the deviation, where the value of ϵ is then chosen carefully to give the required bound for k-means problem. This results in a bound of the order $O(m^{-1/2+\min(1/4,2/p)})$, which results in $O(m^{-1/2})$ when $p \to \infty$.

In a more recent approach, Bachem et al. [5] show a consistent bound of $\mathcal{O}\left(m^{-1/2}\right)$. This is achieved by

reformulation of the deviation problem from 1, to a functional form like below,

$$\left| \frac{1}{m} \sum_{j=1}^{m} \min_i \|x_j - p_i\|_2^2 - \int \min_i \|x - p_i\|_2^2 d\rho(x) \right| \leq \frac{\epsilon}{2} \mathop{\mathbb{E}}_{\rho} \left[\phi_f(X; P) \right] + \frac{\epsilon}{2} \mathop{\mathbb{E}}_{\rho} \left[\phi_f(X; \mathop{\mathbb{E}}_{\rho} [X]) \right] \tag{7}$$

The authors claim in the above paper that along with bounding the deviation, the above formulation also allows for scale-invariance and unbounded solution space. The second term helps in controlling the scale-invariance through the variance term and the first term helps in generalising the bound uniformly for all sets of k-centers Q. The bound on the sample size, in the paper is obtained using in terms of the kurtosis bound on the distribution and a generalization of VC dimension, psuedo-dimension.

3 Seeding for K-Means

The K-means algorithm is extremely sensitive to initialization. Almost all of the techniques which try to improve the convergence of the K-means strategy to a better optima revolve around finding better initialization means. This is known as seeding. In this section, we dicuss some of the earlier seeding strategies as well as the state-of-the-art popular strategies along with any information theoretic bounds they provide.

In the native k-means algorithm, instead of finding a seeding, the data points are randomly assigned to a cluster and the seeding is chosen by computing the means of each data cluster. An improvement to this and one of the oldes methods is the Forgy Approach [8] where k datapoints are chosen uniformly at random and used the initial seeds. A variant of this is to run the k-means algorithm R times each with a random initialization of the cluster means using the Forgy Approach. This is one of the most common approaches of seeding. This strategy helps in selecting a better seeding and somewhat reduces the sensitivity of the initialization, however, does not provide any definite results, nor help us derive a concentration bound on the convergence of the algorithm to a better optima.

Gonzalez [9] proposed an almost deterministic strategy (Maxmin) to choose the seeds where only the first seed is chosed randomly. The MaxMin strategy is as follows; choose the first point uniformly at random from the dataset. At any timestep, choose a point from the dataset whose minimum distance is the maximum from any of the centroids (hence the name). The MaxMin strategy was shown to perform better than the original seeding strategy, the argument being that this strategy is based on inter- cluster distance optimization. [10] proposed a completely deterministic variant of the MaxMin strategy which selects the seeds based on datapoints which improve the loss by the most amount.

There are many researches which conduct an empirical comparison of the aforementioned approaches citepena, celebi, helan, douglas. [17] provides a nice review on the different surveys and comparasons on some seeding strategies for k-means algorithms. None of the above methods, however, afford any theoretical guarantees on the learnt clustering when using Lloyd's Algorithm.

Due to the arbitrary initialization of the cluster means, it is not possible to bound the loss even in expectation. In order to circumvent this, [3] proposed an initialization strategy, called the k-means++ strategy, which allows us to bound the loss of the k-means algorithm in expectation. The k-means++ initialization strategy is presented in algorithm 1. This is one of the most common seeding strategies used for k-means. Most of the newly proposed strategies are based on k-means++. Before discussing these attempts, we dig a little

deeper into the k-means++ strategy in the following text.

3.1 K-Means++ Seeding

The idea of the k-means++ strategy is to sample the means as far away from each other as possible. This maximizes the coverage of the data points and, hence, bounds the cost of the clustering.

Algorithm 1: K-Means++

Input: Data samples $Q = \{\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_n\} \sim P^n$
Output: Cluster means $C = \{\mathbf{c}_1, \mathbf{c}_2 \dots \mathbf{c}_k\} \sim \mathbb{R}^k$

Procedure:

1. Uniformly at random pick $\mathbf{c}_1$ from $\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_n$.
2. Let $C = \{\mathbf{c}_1\}$.
3. For $i = 2 \dots k$, do
   (a) Sample a new center $\mathbf{c}_i$ using D^2 weighting from the dataset. D^2 weighting assigns a probability to each sample as follows
   $$\mathbb{P}\left[C_i = \mathbf{x} \,\middle|\, C \right] = \begin{cases} \frac{1}{Z} d^2(\mathbf{x}, C) & \text{if } \exists j \text{ s.t. } \mathbf{x}_j = \mathbf{x} \\ 0 & \text{else} \end{cases}$$
   where Z is the normalization constant.
   (b) Set $C = C \cup \{\mathbf{c}_j\}$
4. Return C as the set of required k centers.

The bound on the expectation as presented in the original paper is given in Theorem 3.1

Theorem 3.1. [3]. *For any set of data points $Q = \{\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_n\}$ in $\mathbb{R}^d$, we can find a clusering C using the k-means++ strategy such that in expecation, the cost of the clustering is bounded within $\mathcal{O}(\log k)$ factor of the optimal cost. More specifically, we have*

$$\mathop{\mathbb{E}}_{C}[\,\phi_C\,] \;\leq\; 8\,(\log k + 2)\,phi_{OPT}$$

Note that since k-means always only reduces the cost, if the initial clusters afford a bound on the expected cost, then the cost for the final clusters found with the k-means algorithm will also satisfy the same bound.

Without going into the details of the proof, we present the main lemmas used to derive the bound in expecation and some interpretations behind them.

Lemma 3.1. *Let A be a cluster from optimal clustering $\mathcal{C}_{OPT}$ of the samples Q, and let $\mathcal{C}$ be a clustering with one center chosen uniformly at random from A. Then, we have*

$$\mathbb{E}[\,\phi_{\mathcal{C}}(A)\,] \;=\; 2\,\phi_{OPT}(A)$$

This lemma entails that for any set of points, a point chosen at random will always have bounded optimal cost in expctation. This establishes the base case of the k-means++ algorithm for single point clusters.

Lemma 3.2. *Let A be an arbirary cluster from the optimal clustering $\mathcal{C}_{OPT}$ and $\mathcal{C}$ be some arbitrary clustering. If we add a random center to $\mathcal{C}$ from A sampled using D^2 weighting, then we have*

$$\mathbb{E}\left[\,\phi(A)\,\right] \;\leq\; 8\;\phi_{OPT}(A)$$

The above two lemmas combined tell us that if we pick a single point from a cluster either using D^2 weighting or uniformly at random, the expected loss is bounded by a factor of 8 from the optimal cost for that cluster. This essentially means that if we can pick one point out of each cluster, our expected cost will be only a constant times the optimal cost. However, this is obviously not always possible because we do not know the actual clustering and therefore ensuring each cluster has a representative in the initial means is not possible. The third lemma allows us to bridge this gap and provides a way for us to prove Theorem 3.1.

Lemma 3.3. *Let $\mathcal{C}$ be any arbitrary clustering we have chosen. Choose $u > 0$ clusters from $\mathcal{C}_{OPT}$ that do not have any representative in the clustering $\mathcal{C}$. Let Q_u represent the set of points in these u clusters. Also let $Q_c = Q \backslash Q_u$. Now suppose we add $t \leq u$ random centers to $\mathcal{C}$ sampled from Q using D^2-weighting. Let $\mathcal{C}'$ denote the resulting clustering and let ϕ' denote the corresponding potential. Then $\mathbb{E}\left[\,\phi'\,\right]$ is bounded as follows*

$$\mathbb{E}\left[\,\phi'\,\right] \leq (\phi(\mathcal{X}_c) + 8\,\phi_{OPT}(\mathcal{X}_u)) \cdot (1 + H_t) + \frac{u-t}{u}\phi(\mathcal{X}_u)$$

where H_t denotes the harmonic sum, i.e. $H_t = \sum_{i=1}^{t} \frac{1}{i}$.

The intuition behind the above lemma is as follows. As discussed earlier, we want to choose one point for each cluster. The probability of choosing an uncovered cluster A in the above setting would be $\phi_{\mathcal{C}} A / \phi_{\mathcal{C}}$. If the cluster A is not covered, we would expect it's cost to be high, and therefore the probability of choosing a point from A would be high as well. Moreover, if there is a point chosen from A, then the cost would be low and therefore there is lower probability of another point being chosen from A with D^2 weighting.

This means that by using D^2 weighting, we are promoting cluster diversity and, therefore, trying to cover as many clusters we can. This is also the idea behind the proof the above lemma, and subsequently contributes to the intuition and the proof of the K-means++ strategy as a whole.

Ailon et al. [2] further extend the k-means++ strategy to build sampling strategies which reduce the original dataset to size $\mathcal{O}\left(k \log k\right)$ and affords $\mathcal{O}\left(1\right)$ approximation guarantees with a constant probability. Aggarwal et al. [1] show that it is possible with an arbitrarily high probability a constant approximation guarantee based on the same D^2 sampling strategy. Both the approaches build on the D^2 sampling by using a divide and conquer approach to build bi-criteria approximation for the k-means problem.

Ailon et al. [2] develop an alternate strategy called k-means# which samples each centroid $3 \cdot \log k$ times and feeds these samples to the k-means++ algorithm to generate the final clusters. This divide and conquer strategy is a single pass streaming algorithm which affords the same order guarantees as

k-means++.

Even though k-means++ has been one of the most popular techniques for seeding, it is not devoid of problems. Sieranoja [17] suggests that k-means++ generally works well but is more suitable for small to mid-sized datasets. One of the reasons to not choose this strategy for large datasets is that each iteration of the seeding strategy requires $\mathcal{O}(nd)$ time. This means that we have to make k passes of $\mathcal{O}(nd)$ time.

One might wonder that each pass of the Lloyd's algortihm also takes $\mathcal{O}(nd)$ time, however, it is possible to parallelize the K-means algorithm [19]. K-Means++, on the other hand, cannot be parallelized as each iteration depends on the sample from the previous iteration. There have been several attempts to remedy this [7; 5; 6; 4] which we explore in the next subsection.

3.2 Fast and Scalable Seeding with K-Means++

As pointed out in the previous subsection, k-means++ faces the issue of parallelization. Bahmani et al. [7] propose a sampling solution which resembles the approach suggested by Ailon et al. [2]. This sampling strategy proposed provides an additive approximation error along with an $\mathcal{O}(\log k)$ approximation factor in expectation.

The idea behind the algorithm (given in algorithm box 2) is to sample k points in each iteration of the D^2 sampling strategy and do this for $\log n$ iterations. Over these generated samples, k-means++ is run and the cluster centers that are output are used as the initial seeds for the k-means algorithm.

Algorithm 2: K-Means$\|$

Input: Data samples $Q = \{\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_n\} \sim P^n$

Output: Cluster means $C = \{\mathbf{c}_1, \mathbf{c}_2 \dots \mathbf{c}_k\} \sim \mathbb{R}^k$

Procedure:

1. Uniformly at random, sample $\mathbf{c}$ from Q
2. Initialize $C = \{\mathbf{c}\}$ and corresponding clustering $\mathcal{C}$
3. Let $\psi = \phi_{\mathcal{C}}(Q)$
4. Do $\mathcal{O}(\log \psi)$ times
   - Sample $\mathbf{c}$ from Q independently with probability
   $$\mathbb{P}\left[\mathbf{c} = \mathbf{x} \,\middle|\, C\right] = \begin{cases} \frac{1}{Z} d^2(\mathbf{x}, C) & \text{if } \exists j \text{ s.t. } \mathbf{x}_j = \mathbf{c} \\ 0 & \text{else} \end{cases}$$
   for so weighting factor l
   - $C = C \cup C'$
5. For $\mathbf{c} \in C$, set the weight of $\mathbf{c}$ be the number of points in Q closer to $\mathbf{c}$ than any other point in C
6. Recluster the weighted points in C in to k clusters using k-means++ and return the obtained centers.

Theorem 3.2. [7]. *Let* $\alpha = exp\left(-\left(1 - e^{-l/(2k)}\right)\right)$. *If* C *is the set of centers at the beginning of an iteration of algorithm hyperlinkalgo:22 and* C' *is the set of centers added, then*

$$\mathbb{E}\left[\,\phi_{C\cup C'}\,\right] \;\leq\; 8\phi_{OPT} + \frac{1+\alpha}{2}\phi_C$$

The above theorem is the main result shown by Bahmani et al. [7]. This algorithm allowed the runtime of the initialization strategy go down simply because of parallelization trading off with the computation cost of the process. The authors also note that the algorithm works well, empirically, even if the inner loop is run only a constant number of times. Moreover, the convergence of the actual Lloyd's algorithm is shown to improve with k-means‖ seeding [7].

Bachem et al. [6] argued that the bound in Theorem citethm:scalable does not scale with l as the scaling factor can be at least $1/2$. In their work, Bachem et al. [6] improve the theoretical bound for k-means| which depends on the variance of the underlying probability distribution. This bound is stated in Theorem 3.3.

Theorem 3.3. [6]. *Suppose if the points generated by the seeding strategy in algorithm 2 is represented as* C. *A clustering of* Q *based on the centroids obtained from clustering* C, *say* $\mathcal{C}$ *is at most a constant factor from the optimal potential of the clustering of the original data points* Q *along with an additiver term that depends on the variance of the underlying probability distribution. More specifically, we have*

$$\mathbb{E}\left[\,\phi_{\mathcal{C}}(Q)\,\right] \;\leq\; 2\left(\frac{k}{el}\right)^{t} Var\,(Q) + 26\phi_{OPT}$$

According to Theorem 3.3, using k-means++ on the subsampled points gives a $\log k$ factor approximation of the optimal potential with an additive error as showing in the equation.

An alternate work [5] tried to replace D^2 sampling with an MCMC approximation (K-MC^2) which allows the runs to be parallel and requires fewer iterations than k-means++. The proposal distribution for the said MCMC is a uniform distribution over the sample points, therefore assigning $1/n$ to each data point. The acceptance probability for the chain is given as follows

$$\mathbb{P}\left[\,\text{accept}\,(x)\;\middle|\;x_{j-1}\,\right] \;=\; \min\left\{\frac{d^2(x,C)}{d^2(x_{j-1},C)}, 1\right\}$$

Bachem et al. [5] argue that the total variation distance between the MCMC sample distribution and the actual distribution decreases exponentially with the chain length, say m. With this argument, the authors prove a convergence bound as shown in Theorem 3.4.

Theorem 3.4. [5]. *Let* $k > 0$ *and* $0 < \epsilon < 1$. *Let* $p++(C)$ *be the probability of sampling a seeding* C *using k-means++ and* $p_{meme}(C)$ *be the probability of seeding* C *using K-MC*^{2}. *Then, we have*

$$\|\,p_{meme} - p_{++}\,\|_{TV} \;\leq\; \epsilon$$

for a chain length $\mathcal{O}\left(\gamma' \log \frac{k}{\epsilon}\right)$ *where*

$$\gamma' \;=\; \max_{C\subset Q, \|\,C\,\| \,\leq\, k} \max_{\mathbf{x}\in Q} \frac{d^2(x,C)}{\sum_{\mathbf{x}'\in Q} d^2(\mathbf{x}', C')}$$

The above theorem proves that the samples generated using the K-MC^2 are, in expectation, are close to the samples generated using k-means++. Therefore, we can achieve results close to k-means++

without each iteration spending a complete pass over the whole dataset.

Bachem et al. [4] extend these results to result some of the implicit assumptions made by K-MC^2. The authors argue that K-MC^2 does not work well with heavy tailed distributions. Moreover, they refine the theoretical analysis to rid of theses assumptions and allow the algorithm to afford the same results up to a constant factor.

There are countless other approaches which all offer strategies which afford certain kinds of convergence bounds. There hasn't been a successful research on provably converging to the optimal point up to a constant factor. We try to show a simple bound based on convergence inequalities for depenedent random variables for the k-means++ strategy in the next and concluding subsection.

3.3 Deriving a Trivial Bound for K-Means++

Our objective is to derive a concentration bound on the initial cost of the following form.

Suppose the cost of a clustering C achieved by the seeding strategy given in k-means++ is represented as ϕ_C. We want to show that with probability greater than $1 - \delta$, we have

$$\phi_C \;\leq\; \mathbb{E}\left[\,\phi_C\,\right] + \epsilon$$

We have seen such concentration bounds often in information theory. However, most concentration inequalities assume the random variables over which we are computing the expectation to be independent. This is not satisfied in our case, since the sampling of the centers is done sequentially and the probability distribution depends on each of the centers sampled previously.

In order to overcome this, we plan to use concentration bounds which do not assume independence between random variables. There have been a series of forays into deriving such concentration bounds [13; 14; 16; 11; 15] all along similar lines. A general result derived [11] claims the following concentration bound.

Theorem 3.5. *Let $X = \{X_i\}_{i=1}^n$ be a set of dependent random variables all taking values in a countable space $\mathcal{S}$ such that the set of random variables is a coordinate projection over the probability space $(\mathcal{S}^n, \mathcal{F}, \mathbb{P})$ where $\mathcal{F}$ is the powerset of $\mathcal{S}^n$ and $\mathbb{P}$ is a probability measure on $(\mathcal{S}^n, \mathcal{F})$. For an l-Lipschitz function $\varphi : \mathcal{S}^n \to \mathbb{R}$ (with respect to hamming distance), then for any $\epsilon > 0$, we have*

$$\mathbb{P}\{\,|\,\varphi - \mathbb{E}\left[\,\varphi\,\right]\,|\,\geq \epsilon\} \;\leq\; 2\,\exp\left(-\frac{\epsilon^2}{2nl^2\,\|\,\Delta_n\,\|_\infty^2}\right)$$

where Δ is a matrix whose i, j element is defined as follows

$$(\Delta_n)_{i,j} \;=\; \begin{cases} 1 & \text{if } i = j \\ \eta_{i,j} & \text{if } i < j \\ 0 & \text{else} \end{cases}$$

$\eta_{i,j}$ is defined for all $1 \leq i < j \leq n$ as follows

$$\eta_{i,j} \;=\; \max_{\substack{\mathbf{x}_{1:i-1}, w, \bar{w} \\ \mathbb{P}\left[\,X_{1:i} = \mathbf{y}_{1:i-1} w\,\right] > 0 \\ \mathbb{P}\left[\,X_{1:i} = \mathbf{y}_{1:i-1} \bar{w}\,\right] > 0}} \left\|\, \mathbb{P}\left[\, X_{j:n} \,\middle|\, \mathbf{x}_{1:i-1} w \,\right] - \mathbb{P}\left[\, X_{j:n} \,\middle|\, \mathbf{x}_{1:i-1} \bar{w} \,\right] \right\|_{TV}$$

where $\|\cdot\|_{TV}$ signifies the total variation between two probability distributions.

Suppose instead of generating k centers using the k-means++ strategy, we generate m centers sequentially considering only the previous $k-1$ centers for D^2 -weighting. Let these set of centers be denoted as $\{\mu_i\}_{i=1}^{m}$. Then, define the average cost $\overline{\phi}$ as follows

$$\overline{\phi} \;=\; \frac{1}{m-k+1} \sum_{i=1}^{m-k+1} \phi_{\mu_{i:i+k-1}} \tag{8}$$

Note that if we choose the minimum over each quantity in the RHS of equation 8, then that would be a lower bound on $\overline{\phi}$. Therefore, bounding $\overline{\phi}$ helps us bound the cost for one set of centers.

The interesting thing about this is that in expectaion, the previous bound for k-means++ still holds. For the sake of brevity, we exclude the proof of this from this report and simply state it as a lemma.

Lemma 3.4. *If we sample m centers using the k-means++ strategy, then we have*

$$\mathbb{E}\left[\,\overline{\phi}\,\right] \;\leq\; 8\left(\log k + 2\right)\phi_{OPT}$$

where $\overline{\phi}$ is as defined in equation 8.

Note that the given series is a Markov chain with a memory of $k-1$ *i.e.* each timestep depends on the previous $k-1$ timesteps. Example 2.15 from [15] discusses a concentration bound for exactly such an example and derives a bound as follows.

Supppose we have a function f that satisfies the bounded differences property with constant c and a set of random variables $\{X_i\}_{i=1}^{m}$ which follow a markov chain with memory k, then

$$\mathbb{P}\left[\, f(\mathbf{X}) - \mathbb{E}\left[\, f(\mathbf{X})\,\right] > \epsilon\,\right] \;\leq\; \exp\left(-\frac{\epsilon^2}{4kmc^2}\right)$$

Now the function $\overline{\phi}$ satisfies the bounded differences property with $c = n \cdot D^2$ where D is the diameter of the data points *i.e.* the maximum euclidian distance between any two data points. Using this, we can write the following theorem.

Theorem 3.6. *If we sample m centers using the k-means++ strategy with each sample using the previous $k-1$ centers for D^2 -weighting, we can say that with probability at least $1-\delta$,*

$$\overline{\phi} \;\leq\; \mathbb{E}\left[\,\overline{\phi}\,\right] + 2\sqrt{(k-1)\log\frac{1}{\delta}}\,\frac{n \cdot D^2}{\sqrt{m-k+1}}$$

where $\overline{\phi}$ is defined as in equation 8.

This bound, however, is trivial. The reason is that if $m = \Omega\left(k\right)$, then the error term is of the order of $n \cdot D^2$, which is the upper bound of the cost and is, therefore, trivial. If, however, we choose $m = \Omega\left(n^2\right)$, then we remove the dependency on n. This, however, means that the sampling will be of order $\mathcal{O}\left(kdn^3\right)$ which is large for even a moderately sized dataset. Even if somehow can sample this many centers, the bound could still be trivial if the diameter of each cluster is small but the clusters themselves are far apart. Therefore, such a bound does not help us bound the cost.

References

[1] Ankit Aggarwal, Amit Deshpande, and Ravindran Kannan. Adaptive sampling for k-means clustering. In *APPROX-RANDOM*, pages 15–28, 01 2009. doi: 10.1007/978-3-642-03685-9_2.

[2] Nir Ailon, Ragesh Jaiswal, and Claire Monteleoni. Streaming k-means approximation. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 10–18. Curran Associates, Inc., 2009. URL http://papers.nips.cc/paper/3812-streaming-k-means-approximation.pdf.

[3] David Arthur and Sergei Vassilvitskii. K-means++: The Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, page 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics. ISBN 9780898716245.

[4] Olivier Bachem, Mario Lucic, Hamed Hassani, and Andreas Krause. Fast and provably good seedings for k-means. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 55–63. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6478-fast-and-provably-good-seedings-for-k-means.pdf.

[5] Olivier Bachem, Mario Lucic, S. Hamed Hassani, and Andreas Krause. Approximate k-means++ in sublinear time. In *AAAI*, 2016.

[6] Olivier Bachem, Mario Lucic, and Andreas Krause. Distributed and provably good seedings for k-means in constant rounds. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 292–300. JMLR.org, 2017.

[7] Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar, and Sergei Vassilvitskii. Scalable k-means++. *Proc. VLDB Endow.*, 5(7):622–633, March 2012. ISSN 2150-8097. doi: 10.14778/2180912.2180915. URL https://doi.org/10.14778/2180912.2180915.

[8] E. W. Forgy. Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *Biometrics*, 21:768–769, 1965. URL https://ci.nii.ac.jp/naid/10009668881/en/.

[9] Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293 – 306, 1985. ISSN 0304-3975. doi: https://doi.org/10.1016/0304-3975(85)90224-5. URL http://www.sciencedirect.com/science/article/pii/0304397585902245.

[10] Leonard Kaufman and Peter Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. O'Reilly, 09 2009. ISBN 9780470317488.

[11] Leonid Kontorovich and Kavita Ramanan. Concentration Inequalities for Dependent Random Variables via the Martingale Method. *The Annals of Probability*, 2008.

[12] S. Lloyd. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 1982.

[13] K Marton. A Measure Concentration Inequality for Contracting Markov Chains. *Geometric and Functional Analysis*, 1996.

[14] K Marton. K. Measure Concentration for a Class of Random Processes. *Probability Theory and Related Fields*, 1998.

[15] Daniel Paulin. Concentration inequalities for markov chains by marton couplings and spectral methods. *Electron. J. Probab.*, 20:32 pp., 2015. doi: 10.1214/EJP.v20-4039. URL https://doi.org/10.1214/EJP.v20-4039.

[16] Paul-Marie Samson. Concentration of Measure Inequalities for Markov Chains and ϕ-Mixing Processes. *The Annals of Probability*, 2000.

[17] Sami Sieranoja. How much k-means can be improved by using better initialization and repeats? *Pattern Recognition*, 93, 04 2019. doi: 10.1016/j.patcog.2019.04.014.

[18] Matus Telgarsky and Sanjoy Dasgupta. Moment-based uniform deviation bounds for k-means and friends. *CoRR*, abs/1311.1903, 2013. URL `http://arxiv.org/abs/1311.1903`.

[19] Weizhong Zhao, Huifang Ma, and Qing He. Parallel k-means clustering based on mapreduce. In Martin Gilje Jaatun, Gansen Zhao, and Chunming Rong, editors, *Cloud Computing*, pages 674–679, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-10665-1.