Noisy, Greedy and Not so Greedy k-Means++

Anup Bhattacharya
Indian Statistical Institute, Kolkata, India

Jan Eube
University of Bonn, Germany

Heiko Röglin
University of Bonn, Germany

Melanie Schmidt
University of Cologne, Germany

Abstract

The k-means++ algorithm due to Arthur and Vassilvitskii [4] has become the most popular seeding method for Lloyd's algorithm. It samples the first center uniformly at random from the data set and the other $k-1$ centers iteratively according to D^2-sampling, i.e., the probability that a data point becomes the next center is proportional to its squared distance to the closest center chosen so far. k-means++ is known to achieve an approximation factor of $\mathcal{O}(\log k)$ in expectation.

Already in the original paper on k-means++, Arthur and Vassilvitskii suggested a variation called *greedy k-means++ algorithm* in which in each iteration multiple possible centers are sampled according to D^2-sampling and only the one that decreases the objective the most is chosen as a center for that iteration. It is stated as an open question whether this also leads to an $\mathcal{O}(\log k)$-approximation (or even better). We show that this is not the case by presenting a family of instances on which greedy k-means++ yields only an $\Omega(\ell \cdot \log k)$-approximation in expectation where ℓ is the number of possible centers that are sampled in each iteration.

Inspired by the negative results, we study a variation of greedy k-means++ which we call *noisy k-means++ algorithm*. In this variation only one center is sampled in every iteration but not exactly by D^2-sampling. Instead in each iteration an adversary is allowed to change the probabilities arising from D^2-sampling individually for each point by a factor between $1-\epsilon_1$ and $1+\epsilon_2$ for parameters $\epsilon_1 \in [0,1)$ and $\epsilon_2 \geq 0$. We prove that noisy k-means++ computes an $\mathcal{O}(\log^2 k)$-approximation in expectation. We use the analysis of noisy k-means++ to design a moderately greedy k-means++ algorithm.

2012 ACM Subject Classification Theory of computation → Approximation algorithms analysis; Theory of computation → Facility location and clustering; Theory of computation → Design and analysis of algorithms

Keywords and phrases k-means++, greedy, adaptive sampling

Digital Object Identifier 10.4230/LIPIcs.ESA.2020.18

Funding *Anup Bhattacharya*: supported by an NPDF Fellowship, sponsored by the Government of India.
Heiko Röglin: supported by DFG grant RO 5439/1–1.
Melanie Schmidt: supported by DFG grant SCHM 2765/1–1.

Acknowledgements We thank the reviewers for their detailed comments.

1 Introduction

Clustering is a very important tool in many machine learning applications. The task is to find structure that is hidden in input data in the form of clusters, and to do this in an unsupervised way. Since clusters come with very different properties depending on the application, a variety of clustering algorithms and measures to judge clusterings have arisen in the last decades. Among those, a hugely popular method is Lloyd's algorithm [19] (also called the k-means algorithm), which for example was voted to be one of the ten most influential data mining algorithms in machine learning at the IEEE International Conference on Data Mining (ICDM) in 2006 [24].

Lloyd's algorithm is an iterative local search heuristic operating on points from Euclidean space $\mathbb{R}^d$. The measure that it implicitly strives to optimize is the k-means cost function: For a point set $X \subset \mathbb{R}^d$ and a center set $C \subset \mathbb{R}^d$, the k-means cost function is defined as

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

the sum of the squared distances of all points to their respective center. The k-means problem asks to minimize the k-means cost over all choices of C with $|C| = k$. In an optimal solution of the k-means problem, the centers are means of their clusters, and Lloyd's algorithm iterates between computing the means of all clusters as the new center set and reassigning all points to their closest centers to form new clusters. The k-means cost function is also called *sum of squared errors* because when the means are viewed as representatives of the clusters, then the k-means cost is the squared error of this representation.

The k-means problem is NP-hard [3, 20], and it is also hard to approximate to arbitrary precision [5, 18]. On the positive side, constant-factor approximations are possible, and the best known factor is 6.357 due to a break-through result by Ahmadian et al. [2, 18]. However, the constant-factor approximation algorithms for k-means are not very practical. On the other hand, Lloyd's method is hugely popular in practice, but can produce solutions that are arbitrarily bad in the worst case.

A major result in clustering thus was the k-means++ algorithm due to Arthur and Vassilvitskii [4] in 2007, which enhances Lloyd's method with a fast and elegant initialization method that provides an $\mathcal{O}(\log k)$-approximation in expectation. The k-means++ algorithm samples k initial centers by adaptive sampling, where in each step, a point's probability of being sampled is proportional to its cost in the current solution (we will refer to this kind of sampling as D^2-sampling in the following). After sampling k centers, the solution is refined by using Lloyd's algorithm. Algorithm 1 contains pseudo code for the k-means++ algorithm.

The beauty of the algorithm is that it has a bounded approximation ratio of $\mathcal{O}(\log k)$ in expectation, and at the same time computes solutions that are good (much better than $\Theta(\log k)$) on practical tests. By feeding the computed centers into Lloyd's method, the solutions are refined to even better quality. Nevertheless, Arthur and Vassilvitskii show that the approximation ratio of k-means++ is tight in the worst case: They give an (albeit artificial) example where the expected approximation ratio is $\Omega(\log k)$, and this has been extended by now to examples where k-means++ outputs a $\Omega(\log k)$-approximate solution with high probability [8], and even in the plane [7].

Due to its beneficial theoretical and practical properties, k-means++ has by now become the de-facto standard for solving the k-means problem in practice. What is less known is that the original paper [4] and the associated PhD thesis [22] actually propose a possible

Algorithm 1 The k-means++ algorithm [4].

1: Sample a point c_1 independently and uniformly at random from X.
2: Let $C = \{c_1\}$.
3: **for** $i = 2$ to k **do**
4: **for** $x \in X$ **do**
5: $p(x) := \frac{\min_{c \in C} ||x-c||^2}{\sum_{y \in X} \min_{c \in C} ||y-c||^2}$
6: **end for**
7: Sample a point c_i from X, where every $x \in X$ has probability $p(x)$.
8: Update $C = C \cup \{c_i\}$.
9: **end for**
10: Run Lloyd's algorithm initialized with center set C and output the result.

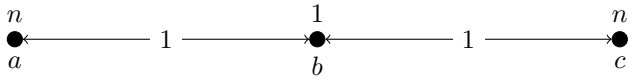

Figure 1 A bad example for the deterministic heuristic that always chooses the current cost minimizer as the next center. An optimal 2-clustering costs less than 1, while a clustering where b is a center costs $\Omega(n)$.

improvement to the k-means++ algorithm: the *greedy k-means++ algorithm*. Here in each of the adaptive sampling steps, not only one center but ℓ possible centers are chosen (independently according to the same probability distribution), and then among these l centers, the one that decreases the k-means cost the most is chosen. This is greedy because a center that reduces the cost in the current step might be a bad center later on (for example if we choose a center that lies between two optimum clusters, thus preventing us from choosing good centers for both on the long run). The original paper [4] says:

> *Also, experiments showed that k-means++ generally performed better if it selected several new centers during each iteration, and then greedily chose the one that decreased Φ (the cost function) as much as possible. Unfortunately, our proofs do not carry over to this scenario. It would be interesting to see a comparable (or better) asymptotic result proven here.*

The intuition is that k-means++ tries to find clusters in the dataset, and with each sample, it tries to find a new cluster that has not been hit by a previously sampled center. This has a failure probability, and the super-constant approximation ratio stems from the probability that some clusters are missed. In this failure event, the algorithm chooses two centers that are close to each other compared to the optimum cost. Greedy k-means++ tries to make this failure event less likely by boosting the probability to find a center from a new cluster that has not been hit previously and greedily choosing the center.

For $\ell = 1$, the greedy k-means++ becomes the k-means++ algorithm, and for very large ℓ it becomes nearly deterministic, a heuristic that always chooses the current minimizer among the whole dataset. It is easy to observe that the latter is not a good algorithm: Consider Figure 1. In the first step, the center that minimizes the overall k-means cost in the next step is b. But if we choose b, then the second greedy center is either a or c, and we end up with a clustering of cost $\Omega(n)$, while the solution $\{a, c\}$ has a cost of 1 (and the optimum solution is even slightly better).

So the crucial question is how to set ℓ, and whether there is an ℓ for which greedy k-means++ outperforms k-means++. It has been shown in [1] that for any optimal clustering of an input data set, k-means++ has in each iteration a constant probability to sample a point from a 'new' optimal cluster, where new means that no point from that cluster has previously been chosen as a center. This leads to a bicriteria approximation, since after $\mathcal{O}(k)$ centers, the algorithm has discovered all optimal clusters in expectation. Following the intuition that stems from this analysis, a natural idea would be to set $\ell = \mathcal{O}(\log k)$: This reduces the probability to pick no point from a new cluster to $\Omega(1/k)$, and by union bound, the failure probability that this event happens in one of the k samples decreases to a constant. This choice is also advertized by Celebi et al. [9], who feature greedy k-means++ in a study of initialization strategies for Lloyd's method. They report that it performs better than k-means++, for a suggested value of $\ell = \log k$. The PhD thesis [22] reports experiments with $\ell = 2$ that outperformed k-means++. It also states that the approximation guarantee of greedy k-means++ is unknown (pp. 62+63).

We initiate the analysis of the greedy k-means++ algorithm. Firstly, we prove that greedy k-means++ is *not* asymptotically better than k-means++. More precisely, we show the following statement.

▶ **Theorem 1.** *For any $k \geq 4$ and any ℓ, there exists a point set $X_{k,\ell}$ such that the expected approximation guarantee of greedy k-means++ is $\Omega(\min\{\ell, k/\log k\} \cdot \log k)$.*

Theorem 1 implies that the worst-case approximation guarantee of greedy k-means++ cannot get better by choosing $\ell > 1$. In particular for $\ell = \log k$, the approximation guarantee worsens to $\Omega(\log^2 k)$.

As indicated in the quote from [4] above, the original proof of k-means++ does not carry over to greedy k-means++, not even if we aim for a higher approximation guarantee like $\mathcal{O}(\ell \log k)$. Roughly speaking, the main problem in the analysis is that while the probability to choose a point as a center can only be increased by a factor of ℓ by the greedy procedure, there is no multiplicative lower bound on how much individual probabilities can be *de*creased. Indeed, if a point $x \in P$ is the worst greedy choice, then its probability to be chosen decreases from some $p(x)$ in the original k-means++ algorithm to $(p(x))^\ell$, which is much smaller than $p(x)$. If this happened to good centers, it could hurt the approximation factor badly.

We proceed to study a different variation of k-means++ which we call the *noisy k-means++* algorithm. This algorithm performs k-means++, but does not sample with exact probabilities. Instead of sampling a point x with probability $p(x)$ as suggested by D^2-sampling, it uses an arbitrary probability $p'(x)$ with $(1-\epsilon_1)p(x) \leq p'(x) \leq (1+\epsilon_2)p(x)$, where $\epsilon_1 \in [0, 1)$ and $\epsilon_2 \geq 0$. If we cast greedy k-means++ as a noisy k-means++ algorithm, we observe that we get a trivial upper bound of $\epsilon_2 = \ell - 1$, however, no trivial *lower* bound on how much the probabilities are skewed.

Noisy k-means++ is also interesting in its own right, since in practice, the probabilities actually computed are prone to rounding errors. Due to the iterative nature of k-means++, it is not at all clear how large the effect of a small rounding can be. We show that the following theorem holds.

▶ **Theorem 2.** *Let T_k denote the set of centers sampled by noisy k-means++ on dataset X and assume that $\frac{k}{\ln k} \geq \max\{18, \frac{24(\epsilon_1+\epsilon_2)(1+\epsilon_2)}{(1-\epsilon_1)^2}\}$. Then,*

$$\mathbb{E}[\Phi(X, T_k)] \leq \mathcal{O}\left(\left(\frac{1+\epsilon_2}{1-\epsilon_1}\right)^3 \cdot \log^2(k) \cdot \mathrm{OPT}_k(X)\right),$$

where $\mathrm{OPT}_k(X)$ denotes the k-means costs of an optimal k-clustering of X. If $\frac{k}{\ln k} \leq \max\{18, \frac{24(\epsilon_1+\epsilon_2)(1+\epsilon_2)}{(1-\epsilon_1)^2}\}$, then $\mathbb{E}[\Phi(X, T_k)] \leq \mathcal{O}\left(\left(\frac{1+\epsilon_2}{1-\epsilon_1}\right)^4 \cdot \log^2\left(\frac{1+\epsilon_2}{1-\epsilon_1}\right) \cdot \mathrm{OPT}_k(X)\right)$.

We use Theorem 2 to analyze a *moderately greedy* variant of k-means++, where the simple idea is that with probability p, we do a normal k-means++ step, and with probability $1-p$, we do a greedy k-means++ step. The idea is that in this variant, a point is never completely disregarded, so we do get a lower bound on the probabilities, yet in many steps, we do still profit from the additional power of greedy k-means++ seen in experiments. For constant p and ℓ, this variant gives an $\mathcal{O}(\log^2 k)$-approximation by Theorem 2.

Techniques

Our lower bound example for greedy k-means++ is close to the original $\Omega(\log k)$ lower bound example in [4] (we contract each cluster to a single location except for one cluster where one point is moved away from the location into the center of the instance, see Section 2).

However, the proof of the lower bound proceeds very differently. Morally, instead of missing clusters (which becomes less likely due to the multiple samples), the failure event is to choose a bad point as a center. This alone is responsible for the $\Omega(\ell \log k)$ lower bound, while the original $\Omega(\log k)$ bound stems from missing clusters.

To analyze noisy k-means++, we build upon an analysis of k-means++ by Dasgupta [12]. Analyzing k-means++ is about 'hitting' clusters. For some fixed optimal solution we call a cluster *covered* if a point is sampled that provides a good enough center for it. An iteration of k-means++ is *wasted* if a point is sampled from an already covered cluster. Dasgupta uses a potential function which accumulates costs over the wasted iterations. To make the connection between k-means++ and this potential function, it is crucial that the expected average cost of the uncovered clusters does not increase over time (in k-means++). For noisy k-means++, this is not true: The cost can increase. We show that the increase can be bounded, roughly by a factor of $\log k$. Then the key difficulty is to analyze the resulting random process which is highly dependent. We analyze an abstract version first and then show how to apply it to the setting of noisy k-means++.

Additional related work

In his master's thesis, Pago [21] shows that for $\ell = \log k$, the example in Figure 1 can be extended such that greedy k-means++ gives an $\Omega(\log k)$-approximation in expectation.

Bachem et al. [6] suggest to speed up k-means++ by replacing the exact sampling according to the probabilities $p(x)$ by a fast approximation based on Markov Chain Monte Carlo sampling. They prove that under certain assumptions on the dataset their algorithm yields the same approximation guarantee in expectation as k-means++, namely $\mathcal{O}(\log k)$. Their algorithm can be viewed as a special case of noisy k-means++. However, their analysis of the approximation factor is based on making the total variation distance between the probability distributions p and p' (in every step) so small that with high probability their algorithm behaves identically to k-means++. In contrast to this, Theorem 2 also applies to choices of ϵ_1 and ϵ_2 for which noisy k-means++ behaves differently from k-means++ with high probability.

Lattanzi and Sohler [17] propose an intermediate improvement step to be executed between the D^2-sampling and Lloyd's algorithm in order to improve the solution quality to a constant factor approximation in expectation. Their algorithm starts with a k-means++ solution and then performs $\mathcal{O}(k \log \log k)$ improvement steps: In each such step, a new center is sampled with D^2-sampling, and if swapping it with an existing center improves the solution, then this swap is performed. While this is a greedy improvement step and thus a bit related to greedy k-means++, their algorithm is closer in spirit to a known local search algorithm by Kanungo et al. [16] which uses center swaps (starting on an arbitrary solution) to obtain a constant-factor approximation, but needs a lot more rounds and is impractical. Very recently, Choo et. al. [10] improved the result by Lattanzi and Sohler and showed that $\mathcal{O}(k)$ swaps are sufficient to achieve a constant factor approximation.

The bicriteria analysis by Aggarwal et al. [1] mentioned above was improved by Wei [23] who showed that for any $\beta > 1$, sampling βk centers with D^2-sampling yields an $\mathcal{O}(1)$-approximation in expectation (with βk centers). Hsu and Telgarsky [15] show that greedy k-means++ for $\ell = \Theta(k)$ leads to a bicriteria O(1)-approximation if $\Theta(k)$ centers are chosen. All above cited works assume that k and d are input parameters; if one of them is a constant, then there exists a PTAS for the problem [11, 13, 14].

In bicriteria results (which, in a sense, also applies to [17] and [10]), the key is to show that a cluster that has not been covered by a good center is found with high probability. For the analysis of greedy k-means++ and noisy k-means++, the main challenge is to bound the expected cost after only k steps.

2 Lower Bound for Greedy k-means++

In this section we construct an instance on which greedy k-means++ yields only an $\Omega(\ell \log k)$-approximation in expectation. More precisely, we analyze Algorithm 2.

Algorithm 2 Greedy k-means++ algorithm [4].

Sample[1] a point c_1 independently and uniformly at random from X.
Let $C = \{c_1\}$.
for $i = 2$ to k **do**
  for $x \in X$ **do**
    $p(x) := \frac{\min_{c \in C} ||x-c||^2}{\sum_{y \in X} \min_{c \in C} ||y-c||^2}$.
  end for
  Sample[1] a set S of ℓ points independently according to this probability distribution.
  Let $c_i = \arg\min_{u \in S} \Phi(X, C \cup \{u\})$.
  Update $C = C \cup \{c_i\}$.
end for
Run Lloyd's algorithm initialized with center set C and output the result.

Note that we only draw one sample in the first step. This is due to the fact that in the first step, k-means++ is guaranteed to discover a new cluster, so there is no reason to draw multiple samples.

The instance is based on a regular $(k-1)$-simplex with side length $\sqrt{2}$. Let the vertices of this simplex be denoted by $v_1, \ldots, v_k$. There are k points each at vertices $v_1, \ldots, v_{k-1}$, $(k-1)$ points at vertex v_k, and there is one point at the center o of the simplex. Let X denote the set of all these points. The simplex can be constructed explicitly in $\mathbb{R}^k$ by letting v_i be the ith canonical unit vector for each i and $o = (1/k, \ldots, 1/k)$. Then it follows that the distance between the center o and any vertex v_i is $\sqrt{(k-1)/k}$.

An optimal clustering $(C_1^\star, \ldots, C_k^\star)$ of this instance is obtained as follows: The clusters $C_1^\star, \ldots, C_{k-1}^\star$ consist of the k points at vertices $v_1, \ldots, v_{k-1}$, respectively, and the cluster $C_k^\star$ consists of the $(k-1)$ points at vertex v_k and the point at the center o. The cost of this clustering is bounded from above by $||o - v_k||^2 = \frac{k-1}{k} = O(1)$.

Consider a k-clustering C obtained by greedy k-means++ that contains the point at o as one of the k centers. The cost of this clustering is at least $(k-1)^2/k = \Omega(k)$ because there exists at least one i such that C has no center at v_i. In the best case this is v_k, which generates the aforementioned cost because the $(k-1)$ points at v_k will be assigned to the center at o. The approximation guarantee of this clustering is $\Omega(k)$. We prove that with sufficiently large probability, greedy k-means++ places one of the centers at o.

Morally, we proceed as follows. We define a failure event F which captures the case that one of the points at v_k is chosen as a center during the execution of greedy k-means++. If this event happens, we cannot show a high lower bound on the approximation guarantee. So we show that F happens at most with constant probability, so with sufficient probability, F does not occur. Then we analyze the probablity that under the condition that F does not occur, o is chosen as a center during the execution of k-means++. This probability increases with every iteration (when the kth center is chosen, there are only o, the points at v_k and the points at one other location v_i left as possible choices). We analyze a simplified random experiment to lower bound the probablity that o is chosen as a center during the iterations $i = 2, \ldots, k$.

[1] In all our algorithms we do sampling with replacement.

▶ **Theorem 1.** *For any $k \geq 4$ and any ℓ, there exists a point set $X_{k,\ell}$ such that the expected approximation guarantee of greedy k-means++ is $\Omega(\min\{\ell, k/\log k\} \cdot \log k)$.*

Proof. Notice that for $\ell = 1$ there is nothing to show since a lower bound of $\Omega(\log k)$ is known for this case. So in the following, we assume that $\ell \geq 2$. Furthermore we consider first only the case that $\ell \leq \frac{k}{20\ln(k-1)}$ and defer the discussion of larger ℓ to the end of the proof.

We consider the point set X constructed above. Consider a k-clustering C obtained by greedy k-means++ that contains the point at o as one of the k centers. The cost of this clustering is at least $(k-1)^2/k = \Omega(k)$ because there exists at least one i such that C has no center at v_i. In the best case this is v_k, which generates the aforementioned cost because the $(k-1)$ points at v_k will be assigned to the center at o. The approximation guarantee of this clustering is $\Omega(k)$. We will prove that with sufficiently large probability, greedy k-means++ places one of the centers at o.

We start the analysis by defining the following events for all $i \in [k]$:

F_i = the center chosen in the ith iteration lies at v_k,

G_i = the center chosen in the ith iteration lies at o,

$H_i = F_i \cup G_i$.

We denote by Φ_i the potential after $i-1$ iterations if in these iterations no point from $C_k^\star$ has been chosen as a center. Since the probability to choose the same v_i more than once is zero, this means that $i-1$ centers from different clusters from $C_1^\star, \ldots, C_{k-1}^\star$ have been chosen. In the remaining $k-i+1$ clusters, k points pay a cost of 2, except for the one point at o which pays $1-1/k$. Thus,

$$\Phi_i = 2((k-i+1)k-1) + 1 - \frac{1}{k}$$

and

$$2((k-i+1)k-1) \leq \Phi_i \leq 2k(k-i+1).$$

We define

$$F = F_1 \cup (F_2 \cap \overline{H_1}) \cup \ldots \cup (F_{k-1} \cap \overline{H_1} \cap \ldots \cap \overline{H_{k-2}})$$

as the event that in one of the first $k-1$ iterations a point at v_k is chosen as a center and that this is the first center chosen from $C_k^\star$. We exclude the last iteration because $\mathbf{Pr}(F_k)$ is significantly higher than $\mathbf{Pr}(F_i)$ for $i \leq k-1$.

We will prove a lower bound for the probability of the event $\overline{F} \cap (G_2 \cup \ldots \cup G_{k-1})$ because if this event happens then the point at o is one of the centers computed by greedy k-means++, i.e., the approximation factor is at least $\Omega(k)$.

If the event F occurs then we cannot prove a lower bound on the approximation guarantee of greedy k-means++. Hence, we will prove an upper bound for the probability of F. Observe that

$$\mathbf{Pr}[F] \leq \sum_{i=1}^{k-1} \mathbf{Pr}[F_i \mid \overline{H_1} \cap \ldots \cap \overline{H_{i-1}}] \cdot \mathbf{Pr}[\overline{H_1} \cap \ldots \cap \overline{H_{i-1}}] \leq \sum_{i=1}^{k-1} \mathbf{Pr}[F_i \mid \overline{H_1} \cap \ldots \cap \overline{H_{i-1}}]$$

and

$$\mathbf{Pr}[F_1] = \frac{k-1}{k^2} \leq \frac{1}{k}.$$

Consider the situation that $1 \leq i-1 \leq k-2$ iterations have already been performed and that in these iterations cluster $C_k^\star$ has not been covered. Then each point from an uncovered cluster $C_j^\star$ with $j < k$ reduces the potential by $2k$. Each point at v_k reduces the potential by $2(k-1)$ and the point at o reduces the potential by

$$(\underbrace{(k-i+1)}_{\geq 2}k-1)(1+1/k)+1-1/k \geq (2k-1)(1+1/k)+1-1/k = 2(k+1-1/k) > 2k.$$

Hence, the points at v_k have the least potential reduction and thus a point at v_k is only selected as new center in iteration i if all ℓ sampled candidates are at v_k. Hence, we obtain

$$\mathbf{Pr}[F_i \mid \overline{H_1} \cap \ldots \cap \overline{H_{i-1}}] = \left(\frac{2(k-1)}{\Phi_i}\right)^\ell.$$

Altogether this implies

$$\mathbf{Pr}[F] \leq \mathbf{Pr}[F_1] + \sum_{i=2}^{k-1} \mathbf{Pr}[F_i \mid \overline{H_1} \cap \ldots \cap \overline{H_{i-1}}] \leq \frac{1}{k} + \sum_{i=2}^{k-1} \left(\frac{2(k-1)}{\Phi_i}\right)^\ell.$$

Together with $\Phi_i \geq 2((k-i+1)k-1)$ this implies

$$\begin{aligned}
\mathbf{Pr}[F] &\leq \frac{1}{k} + \sum_{i=2}^{k-1} \left(\frac{2(k-1)}{2((k-i+1)k-1)}\right)^\ell \\
&= \frac{1}{k} + \sum_{i=2}^{k-1} \left(\frac{k-1}{(k-i+1)k-1}\right)^\ell \\
&\leq \frac{1}{k} + \sum_{i=2}^{k-1} \left(\frac{k}{(k-i+1)k}\right)^\ell \\
&= \frac{1}{k} + \sum_{i=2}^{k-1} \left(\frac{1}{i}\right)^\ell,
\end{aligned}$$

where the inequality in the penultimate line of the calculation follows from $\frac{a}{b} < \frac{a+1}{b+1}$ for $0 < a < b$. Using $\ell \geq 2$ and $k \geq 4$, it follows

$$\mathbf{Pr}[F] \leq \frac{1}{k} + \sum_{i=2}^{k-1} \left(\frac{1}{i}\right)^2 \leq \frac{1}{k} + \sum_{i=2}^{\infty} \left(\frac{1}{i}\right)^2 = \frac{1}{k} + \left(\frac{\pi^2}{6} - 1\right) \leq 0.9.$$

This shows that with constant probability, the failure event F does not occur, i.e., with constant probability none of the points from v_k is chosen as a center in the first $k-1$ iterations.

Now let us consider the probability that the point at o is selected as a center. We have argued above that the potential reduction of the point at o in iteration $2 \leq i \leq k-1$ is larger than $2k$ if cluster $C_k^\star$ has not been covered in the first $i-1$ iterations. We have also seen that any other point reduces the potential by at most $2k$. Hence, in order to select the point at o as center it suffices already if it belongs to the ℓ candidates chosen in iteration i. Denote the event that the jth sample in iteration i is o by G_{ij}. Then for $i \in \{2, \ldots, k-1\}$,

$$\begin{aligned}\mathbf{Pr}[G_i \mid \overline{H_1} \cap \ldots \cap \overline{H_{i-1}}] &= \mathbf{Pr}[\cup_{j=1}^{\ell} G_{ij} \mid \overline{H_1} \cap \ldots \cap \overline{H_{i-1}}] \\ &\geq \sum_{j=1}^{\ell} \mathbf{Pr}[G_{ij} \mid \overline{H_1} \cap \ldots \cap \overline{H_{i-1}}] - \sum_{1 \leq j_1 < j_2 \leq j} \mathbf{Pr}[G_{ij_1} \cap G_{ij_2} \mid \overline{H_1} \cap \ldots \cap \overline{H_{i-1}}] \\ &= \frac{\ell(1-1/k)}{\Phi_i} - \frac{\binom{\ell}{2}(1-1/k)^2}{\Phi_i^2} \\ &\geq \frac{\ell(1-1/k)}{\Phi_i} - \frac{\ell^2(1-1/k)^2}{\Phi_i^2},\end{aligned}$$

where the first inequality follows from Bonferroni inequalities.

Since $\ell \leq k/(20\ln(k-1)) \leq k/2$, we obtain

$$\frac{\ell(1-1/k)}{\Phi_i} \leq \frac{\ell}{\Phi_i} \leq \frac{\ell}{2((k-i+1)k-1)} \leq \frac{\ell}{k} \leq \frac{1}{2}.$$

This is helpful, because for any $a \in \mathbb{R}$ with $0 \leq a \leq 1/2$, it holds that $a - a^2 \geq a/2$. Thus, the previous two inequalities imply

$$\mathbf{Pr}[G_i \mid \overline{H_1} \cap \ldots \cap \overline{H_{i-1}}] \geq \frac{\ell(1-1/k)}{\Phi_i} - \left(\frac{\ell(1-1/k)}{\Phi_i}\right)^2 \geq \frac{\ell(1-1/k)}{2\Phi_i}. \tag{1}$$

Let us now condition on the event $\overline{F}$, which happens with constant probability. Then we can write the probability of the event we care about as

$$\begin{aligned}\mathbf{Pr}[\overline{F} \cap (G_2 \cup \ldots \cup G_{k-1})] &= \mathbf{Pr}[\overline{F}] \cdot \mathbf{Pr}[G_2 \cup \ldots \cup G_{k-1} \mid \overline{F}] = \mathbf{Pr}[\overline{F}] \cdot \sum_{i=2}^{k-1} \mathbf{Pr}[G_i \mid \overline{F}] \\ &\geq \mathbf{Pr}[\overline{F}] \cdot \sum_{i=2}^{k-1} \mathbf{Pr}[G_i \mid \overline{F_1} \cap \ldots \cap \overline{F_{i-1}}],\end{aligned}$$

where we used in the penultimate step that the events G_i are mutually exclusive and in the last step that $\overline{F} \subseteq \overline{F_1} \cap \ldots \cap \overline{F_{i-1}}$. We cannot use (1) directly to bound $\mathbf{Pr}[G_i \mid \overline{F_1} \cap \ldots \cap \overline{F_{i-1}}]$ because the condition is different (in (1) we condition on the event that no point from $C_k^\star$ has been chosen as center in the first $i-1$ iterations while conditioning on $\overline{F_1} \cap \ldots \cap \overline{F_{i-1}}$ only implies that no point at v_k has been chosen as a center).

To prove a lower bound on $\mathbf{Pr}[G_2 \cup \ldots \cup G_{k-1} \mid \overline{F}]$, we consider a different random experiment E. This random experiment consists of $k-2$ iterations numbered from 2 to $k-1$ and each iteration i is successful with probability $\mathbf{Pr}[G_i \mid \overline{H_1} \cap \ldots \cap \overline{H_{i-1}}]$ independent of the other iterations. Then $\mathbf{Pr}[G_2 \cup \ldots \cup G_{k-1} \mid \overline{F}]$ equals the probability that at least one of the iterations of E is successful. Let E' denote the same random experiment as E only with modified success probabilities. In E' iteration i is successful with probability $\frac{\ell(1-1/k)}{2\Phi_i}$. Due to (1) and Bonferroni inequalities and using $k \geq 4$, we obtain

$$\begin{aligned}\mathbf{Pr}[G_2 \cup \ldots \cup G_{k-1} \mid \overline{F}] &= \mathbf{Pr}[\text{at least one success in } E] \\ &\geq \mathbf{Pr}[\text{at least one success in } E'] \\ &\geq \sum_{i=2}^{k-1} \frac{\ell(1-1/k)}{2\Phi_i} - \sum_{2 \leq i < j \leq k-1} \frac{\ell(1-1/k)}{2\Phi_i} \cdot \frac{\ell(1-1/k)}{2\Phi_j} \\ &\geq \sum_{i=2}^{k-1} \frac{\ell(1-1/k)}{4k(k-i+1)} - \sum_{2 \leq i < j \leq k-1} \frac{\ell}{4((k-i+1)k-1)} \cdot \frac{\ell}{4((k-j+1)k-1)}\end{aligned}$$

$$\geq \sum_{i=2}^{k-1} \frac{\ell(1-1/k)}{4k(k-i+1)} - \sum_{2\leq i<j\leq k-1} \frac{\ell}{3k(k-i+1)} \cdot \frac{\ell}{3k(k-j+1)}$$
$$= \frac{\ell(1-1/k)}{4k} \sum_{i=2}^{k-1} \frac{1}{i} - \frac{\ell^2}{9k^2} \sum_{2\leq i<j\leq k-1} \frac{1}{(k-i+1)(k-j+1)}$$
$$\geq \frac{3\ell}{16k} \sum_{i=2}^{k-1} \frac{1}{i} - \frac{\ell^2}{9k^2} \left(\sum_{i=2}^{k-1} \frac{1}{i} \right)^2$$
$$\geq \frac{3\ell}{16k} (\ln(k-1) - 1) - \frac{\ell^2}{9k^2} \ln^2(k-1).$$

For $k \geq 4$, we have $\ln(k-1) - 1 \geq 0.089 \ln(k-1)$. Together with the previous calculation we get

$$\mathbf{Pr}[G_2 \cup \ldots \cup G_{k-1} \mid \overline{F}] \geq 0.0166 \cdot \frac{\ell \cdot \ln(k-1)}{k} - \left(\frac{\ell \cdot \ln(k-1)}{3k} \right)^2$$
$$= \frac{\ell \cdot \ln(k-1)}{k} \cdot \left(0.0166 - \frac{\ell \cdot \ln(k-1)}{9k} \right)$$
$$\geq 0.01 \cdot \frac{\ell \cdot \ln(k-1)}{k},$$

where we used $\ell \leq \frac{0.05 \cdot k}{\ln(k-1)}$ for the last inequality.

Overall we obtain

$$\mathbf{Pr}[\overline{F} \cap (G_2 \cup \ldots \cup G_{k-1})] = \mathbf{Pr}[\overline{F}] \cdot \mathbf{Pr}[G_2 \cup \ldots \cup G_{k-1} \mid \overline{F}]$$
$$\geq 0.1 \cdot 0.01 \cdot \frac{\ell \cdot \ln(k-1)}{k} = \Omega\left(\frac{\ell \cdot \log(k)}{k} \right).$$

If this event happens, then the costs of the clustering are $\Omega(k)$. Hence the expected costs of the clustering computed by greedy k-means++ are $\Omega(\ell \cdot \log(k))$.

Finally let us consider the case $\ell > \frac{k}{20 \ln(k-1)}$. We argue that in this case the approximation guarantee cannot be better than for $\ell = \frac{k}{20 \ln(k-1)}$. To see that this is true, one has to have a closer look at where the upper bound on ℓ has been used in the argument above. It is used twice: once for proving an upper bound on the conditional probability of G_i and once for proving an upper bound on the conditional probability of $G_2 \cup \ldots \cup G_{k-1}$. Both these probabilities increase with ℓ so if ℓ is larger one could simply replace it by $\frac{k}{20 \ln(k-1)}$, leading to a lower bound of $\Omega(k/\log(k) \cdot k) = \Omega(k)$ for the approximation guarantee. ◀

3 Analysis of Noisy k-means++ Seeding

In this section we analyze a noisy seeding procedure, which we call *noisy k-means++* in the following. This procedure iteratively selects k centers from the data set in a similar fashion as k-means++. The only difference is that the probability of sampling a point as the next center is no longer exactly proportional to its squared distance to the closest center chosen so far. The probabilities are only approximately correct. To be more precise, consider an iteration of noisy k-means++. For any point $x \in X$, we denote by p_x the probability that x is chosen by k-means++ as the next center (i.e., p is the uniform distribution in the first iteration and the distribution that results from D^2-sampling in the following iterations). In noisy k-means++ an adversary can choose an arbitrary probability distribution q on X

with $q_x \in [(1-\epsilon_1)p_x, (1+\epsilon_2)p_x]$ for all $x \in X$ where $\epsilon_1 \in [0,1)$ and $\epsilon_2 \geq 0$ are parameters[2]. Then the next center is sampled according to q. This is repeated in every iteration of noisy k-means++ and in every iteration the adversary can decide arbitrarily how to choose q based on the current distribution p that results from D^2-sampling. We analyze the worst-case approximation guarantee provided by noisy k-means++.

The difficulty with noisy k-means++ is that a) it has a high probability to differ from k-means++, and b) the steps are highly dependent on each other, so once the algorithm has deviated, this propagates in the subsequent steps. It may be surprising that such a little change to the algorithm has such a huge effect. After some considerations it is even unclear if noisy k-means++ has any approximation guarantee at all. While we achieve worse guarantees compared to k-means++, we do at least answer this question affirmatively, showing that noisy k-means++ achieves an expected approximation guarantee of $\mathcal{O}(\log^2 k)$. Achieving this requires an intricate analysis of the highly dependent algorithm. We could not make it work with the original proof, so we use an alternative proof by Dasgupta [12] as a starting point. Also in this proof, a crucial step breaks down (the expected average cost of uncovered clusters can now increase, which is not the case for k-means++). This makes the process difficult to analyze and solving this challenge is the main technical contribution of this paper.

Let us first introduce some notation. We denote by $\Phi(X, C)$ the k-means costs of data set X with respect to center set C, i.e.,

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

For $c \in \mathbb{R}^d$ we also write $\Phi(X, c)$ instead of $\Phi(X, \{c\})$ and similarly for $x \in \mathbb{R}^d$ we write $\Phi(x, C)$ instead of $\Phi(\{x\}, C)$. Let $\mathrm{OPT}_k(X)$ denote the optimal k-means costs of dataset X. In the following we assume that a data set X is given and we denote by $(C_1^\star, \dots, C_k^\star)$ an optimal k-clustering of X. For a finite set $X \subset \mathbb{R}^d$, we denote by $\mu(X) = \frac{1}{|X|} \sum_{x \in X} x$ its mean. The following lemma is well-known.

▶ **Lemma 3.** *For any finite $X \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$,*

$$\Phi(C, z) = \Phi(C, \mu(C)) + |C| \cdot ||z - \mu(C)||^2 = \mathrm{OPT}_1(C) + |C| \cdot ||z - \mu(C)||^2.$$

We call an optimal cluster $C_i^\star$ *covered* by (noisy) k-means++ if at least one point from $C_i^\star$ is selected as a center. Arthur and Vassilvitskii [4] observe that covered clusters are well approximated by k-means++ in expectation. In particular, they show that the expected costs of an optimal cluster $C_i^\star$ with respect to the center set computed by k-means++ are at most $2 \cdot \mathrm{OPT}_1(C_i^\star)$ and $8 \cdot \mathrm{OPT}_1(C_i^\star)$ if the cluster is covered in the first or any of the following iterations, respectively. First of all, we carry these observations over to noisy k-means++. The following two lemmata are straightforward adaptations of Lemma 3.2 and Lemma 3.3 in [4].

▶ **Lemma 4.** *Let c_1 denote the first center chosen by noisy k-means++. For each optimal cluster $C_i^\star$,*

$$\mathbb{E}[\Phi(C_i^\star, c_1) \mid c_1 \in C_i^\star] \leq \frac{2(1+\epsilon_2)}{1-\epsilon_1} \cdot \mathrm{OPT}_1(C_i^\star).$$

[2] For better readability, whenever we write $q(x) \leq (1+\epsilon_2)p_x$, we implicitly require $q(x) \leq \min\{1, (1+\epsilon_2)p_x\}$

Proof. In k-means++ the first center is chosen uniformly at random, i.e., each point from X has a probability of $1/|X|$ of being chosen. In noisy k-means++, all points have a probability in $[(1-\epsilon_1)/|X|, (1+\epsilon_2)/|X|]$ of being chosen. Hence, the probability of choosing a point $x \in C_i^\star$ as the first center conditioned on the first center being chosen from $C_i^\star$ is at most $\frac{1+\epsilon_2}{(1-\epsilon_1)|C_i^\star|}$. This implies

$$\begin{aligned}
\mathbb{E}[\Phi(C_i^\star, \{c_1\})] &\leq \sum_{c \in C_i^\star} \frac{1+\epsilon_2}{(1-\epsilon_1)|C_i^\star|}\, \Phi(C_i^\star, c) \\
&= \frac{1+\epsilon_2}{1-\epsilon_1} \cdot \frac{1}{|C_i^\star|} \sum_{c \in C_i^\star} \Phi(C_i^\star, c) \\
&= \frac{1+\epsilon_2}{1-\epsilon_1} \cdot \frac{1}{|C_i^\star|} \sum_{c \in C_i^\star} (\mathrm{OPT}_1(C_i^\star) + |C_i^\star| \cdot ||c - \mu(C_1^\star)||^2) \qquad \text{(Lemma 3)} \\
&= \frac{2(1+\epsilon_2)}{1-\epsilon_1} \cdot \mathrm{OPT}_1(C_i^\star)
\end{aligned}$$

◀

▶ **Lemma 5.** *Consider an iteration of noisy k-means++ after the first one and let $C \neq \emptyset$ denote the current set of centers. We denote by z the center sampled in the considered iteration. Then for any $C \neq \emptyset$ and any optimal cluster $C_i^\star$,*

$$\mathbb{E}[\Phi(C_i^\star, C \cup \{z\}) \mid C, z \in C_i^\star] \leq \frac{8(1+\epsilon_2)}{1-\epsilon_1} \cdot \mathrm{OPT}_1(C_i^\star).$$

Proof. Conditioned on sampling a point from $C_i^\star$, the probability of choosing point $x \in C_i^\star$ as the next center is at most $\frac{1+\epsilon_2}{1-\epsilon_1} \cdot \frac{\Phi(x,C)}{\Phi(C_i^\star, C)}$. If x is chosen as the next center, the costs of any point $p \in C_i^\star$ become $\min\{\Phi(p,C), ||p-x||^2\}$. This implies

$$\begin{aligned}
\mathbb{E}[\Phi(C_i^\star, C \cup \{z\}) \mid C, z \in C_i^\star] &= \sum_{x \in C_i^\star} \mathbf{Pr}[z = x \mid C] \cdot \Phi(C_i^\star, C \cup \{x\}) \\
&\leq \frac{1+\epsilon_2}{1-\epsilon_1} \cdot \sum_{x \in C_i^\star} \frac{\Phi(x,C)}{\Phi(C_i^\star, C)} \sum_{p \in C_i^\star} \min\{\Phi(p,C), ||p-x||^2\}.
\end{aligned} \tag{2}$$

For any two points $x, p \in C_i^\star$, we can write

$$\Phi(x,C) = \Big(\min_{c \in C} ||x-c||\Big)^2 \leq \Big(\min_{c \in C}(||x-p|| + ||p-c||)\Big)^2 \leq 2\Phi(p,C) + 2||x-p||^2.$$

By summing over all p in $C_i^\star$, we get

$$\Phi(x,C) \leq \frac{2}{|C_i^\star|} \sum_{p \in C_i^\star} \Phi(p,C) + \frac{2}{|C_i^\star|} \sum_{p \in C_i^\star} ||x-p||^2.$$

With (2), this implies that $\mathbb{E}[\Phi(C_i^\star, C \cup \{z\}) \mid C, z \in C_i^\star]$ is bounded from above by

$$\begin{aligned}
&\frac{1+\epsilon_2}{1-\epsilon_1} \cdot \sum_{x \in C_i^\star} \frac{\frac{2}{|C_i^\star|}\sum_{p \in C_i^\star} \Phi(p,C) + \frac{2}{|C_i^\star|}\sum_{p \in C_i^\star} ||x-p||^2}{\Phi(C_i^\star, C)} \sum_{p \in C_i^\star} \min\{\Phi(p,C), ||p-x||^2\} \\
&= \frac{1+\epsilon_2}{1-\epsilon_1} \cdot \sum_{z \in C_i^\star} \frac{\frac{2}{|C_i^\star|}\sum_{p \in C_i^\star} \Phi(p,C)}{\sum_{p \in C_i^\star} \Phi(p,C)} \sum_{p \in C_i^\star} \min\{\Phi(p,C), ||p-z||^2\}
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1+\epsilon_2}{1-\epsilon_1} \cdot \sum_{z \in C_i^\star} \frac{\frac{2}{|C_i^\star|} \sum_{p \in C_i^\star} ||p-z||^2}{\sum_{p \in C_i^\star} \Phi(p,C)} \sum_{p \in C_i^\star} \min\{\Phi(p,C), ||p-z||^2\} \\
&\leq \frac{1+\epsilon_2}{1-\epsilon_1} \cdot \sum_{z \in C_i^\star} \frac{2}{|C_i^\star|} \sum_{p \in C_i^\star} ||p-z||^2 + \frac{1+\epsilon_2}{1-\epsilon_1} \cdot \sum_{z \in C_i^\star} \frac{2}{|C_i^\star|} \sum_{p \in C_i^\star} ||p-z||^2 \\
&= \frac{4(1+\epsilon_2)}{1-\epsilon_1} \cdot \sum_{z \in C_i^\star} \frac{1}{|C_i^\star|} \sum_{p \in C_i^\star} ||p-z||^2 \\
&= \frac{4(1+\epsilon_2)}{1-\epsilon_1} \cdot \sum_{z \in C_i^\star} \frac{1}{|C_i^\star|} (\mathrm{OPT}_1(C_i^\star) + |C_i^\star| \cdot ||z - \mu(C_i^\star)||^2) \qquad \text{(Lemma 3)} \\
&= \frac{8(1+\epsilon_2)}{1-\epsilon_1} \cdot \mathrm{OPT}_1(C_i^\star)
\end{aligned}
$$

◀

Consider a run of noisy k-means++. For $t \in [k]$, let H_t and U_t denote the set of all points from X that belong after iteration i to covered and uncovered optimal clusters, respectively. Let u_t denote the number of uncovered clusters after iteration t. Furthermore let T_t denote the set of centers chosen by noisy k-means++ in the first t iterations. We say that iteration t is *wasted* if the center chosen in iteration t comes from H_{t-1}, i.e., if in iteration t no uncovered cluster becomes covered.

▶ **Corollary 6.** *For any $t \in [k]$,*

$$\mathbf{E}\left[\Phi(H_t, T_t)\right] \leq \frac{8(1+\epsilon_2)}{1-\epsilon_1} \cdot \mathrm{OPT}_k(X)$$

Proof. Using Lemma 4 and Lemma 5 we obtain

$$
\begin{aligned}
\mathbf{E}\left[\Phi(H_t, T_t)\right] &= \sum_{i=1}^{k} \mathbf{Pr}[C_i^\star \subseteq H_t] \cdot \mathbf{E}\left[\Phi(C_i^\star, T_t) \mid C_i^\star \subseteq H_t\right] \\
&\leq \sum_{i=1}^{k} \mathbf{Pr}[C_i^\star \subseteq H_t] \cdot \frac{8(1+\epsilon_2)}{1-\epsilon_1} \cdot \mathrm{OPT}_1(C_i^\star) \\
&\leq \frac{8(1+\epsilon_2)}{1-\epsilon_1} \cdot \sum_{i=1}^{k} \mathrm{OPT}_1(C_i^\star) \\
&= \frac{8(1+\epsilon_2)}{1-\epsilon_1} \cdot \mathrm{OPT}_k(X).
\end{aligned}
$$

◀

Corollary 6 implies that the covered clusters contribute in expectation at most $O(\mathrm{OPT}_k(X))$ to the costs of the solution computed by noisy k-means++ (assuming ϵ_1 and ϵ_2 to be constants). The not straightforward part is to prove an upper bound for the costs of the clusters that are not covered by noisy k-means++. For this, we adapt the analysis of k-means++ due to Dasgupta [12]. This analysis is based on a potential function that accumulates costs in every wasted iteration. The potential function has the property that the expected value of the potential function in the end can be bounded and that the total costs accumulated are in expectation at least the costs of the uncovered clusters in the end.

Dasgupta crucially uses that the expected average costs of the uncovered clusters do not increase in k-means++. For noisy k-means++ this is not true anymore in general. Hence, we have to adapt the potential function and the analysis. We define $W_i = 1$ if iteration i is wasted and $W_i = 0$ otherwise. We define the potential function as

$$\Psi_k = \sum_{i=2}^{k} W_i \cdot \frac{\Phi(U_i, T_i)}{u_i}.$$

The easier part is to show that the potential can be bounded from above.

▶ **Lemma 7.** *It holds*

$$\mathbf{E}[\Psi_k] \leq \frac{8(1+\epsilon_2)^2}{(1-\epsilon_1)^2} \cdot (\ln(k)+1) \cdot \mathrm{OPT}_k(X).$$

Proof. Let $i \in \{2,\dots,t\}$. In the following calculation we sum over all realizations $\mathcal{F}_{i-1}$ of the first $i-1$ iterations of noisy k-means++. Any realization $\mathcal{F}_{i-1}$ determines the value of $\Phi(U_{i-1},T_{i-1})$ and u_{i-1}. We use the notation $[\dots]_{\mathcal{F}_{i-1}}$ to express that all terms inside the brackets take the values determined by $\mathcal{F}_{i-1}$. Then

$$\begin{aligned}\mathbf{E}\left[W_i \cdot \frac{\Phi(U_i,T_i)}{u_i}\right] &= \sum_{\mathcal{F}_{i-1}} \mathbf{Pr}[\mathcal{F}_{i-1}] \cdot \mathbf{E}\left[W_i \cdot \frac{\Phi(U_i,T_i)}{u_i} \;\middle|\; \mathcal{F}_{i-1}\right] \\ &= \sum_{\mathcal{F}_{i-1}} \mathbf{Pr}[\mathcal{F}_{i-1}] \cdot \mathbf{Pr}[W_i = 1 \mid \mathcal{F}_{i-1}] \cdot \mathbf{E}\left[\frac{\Phi(U_i,T_i)}{u_i} \;\middle|\; \mathcal{F}_{i-1} \cap (W_i = 1)\right]\end{aligned}$$

Since under the condition that iteration i is wasted the average costs of the uncovered clusters cannot increase, we can upper bound the term above by

$$\begin{aligned}&\sum_{\mathcal{F}_{i-1}} \mathbf{Pr}[\mathcal{F}_{i-1}] \cdot \mathbf{Pr}[W_i = 1 \mid \mathcal{F}_{i-1}] \cdot \left[\frac{\Phi(U_{i-1},T_{i-1})}{u_{i-1}}\right]_{\mathcal{F}_{i-1}} \\ &\leq \sum_{\mathcal{F}_{i-1}} \mathbf{Pr}[\mathcal{F}_{i-1}] \cdot \left[\frac{(1+\epsilon_2)\Phi(H_{i-1},T_{i-1})}{(1-\epsilon_1)\Phi(U_{i-1},T_{i-1})} \cdot \frac{\Phi(U_{i-1},T_{i-1})}{u_{i-1}}\right]_{\mathcal{F}_{i-1}} \\ &= \frac{1+\epsilon_2}{1-\epsilon_1} \cdot \sum_{\mathcal{F}_{i-1}} \mathbf{Pr}[\mathcal{F}_{i-1}] \cdot \left[\frac{\Phi(H_{i-1},T_{i-1})}{u_{i-1}}\right]_{\mathcal{F}_{i-1}} \\ &\leq \frac{1+\epsilon_2}{1-\epsilon_1} \cdot \sum_{\mathcal{F}_{i-1}} \mathbf{Pr}[\mathcal{F}_{i-1}] \cdot \left[\frac{\Phi(H_{i-1},T_{i-1})}{k-i+1}\right]_{\mathcal{F}_{i-1}} \\ &= \frac{1+\epsilon_2}{1-\epsilon_1} \cdot \frac{\mathbf{E}[\Phi(H_{i-1},T_{i-1})]}{k-i+1}.\end{aligned}$$

This implies

$$\begin{aligned}\mathbf{E}[\Psi_k] &= \sum_{i=2}^{k} \mathbf{E}\left[W_i \cdot \frac{\Phi(U_i,T_i)}{u_i}\right] \\ &\leq \frac{1+\epsilon_2}{1-\epsilon_1} \cdot \sum_{i=2}^{k} \frac{\mathbf{E}[\Phi(H_{i-1},T_{i-1})]}{k-i+1}.\end{aligned}$$

With Corollary 6 this yields

$$\mathbf{E}[\Psi_k] \leq \frac{8(1+\epsilon_2)^2}{(1-\epsilon_1)^2} \cdot \mathrm{OPT}_k(X) \sum_{i=2}^{k} \frac{1}{k-i+1} \leq \frac{8(1+\epsilon_2)^2}{(1-\epsilon_1)^2} \cdot (\ln(k)+1) \cdot \mathrm{OPT}_k(X). \quad \blacktriangleleft$$

Now our goal is to prove a lower bound on the potential, namely, we want to prove the following lemma, where we use $D(\epsilon_1,\epsilon_2,\ln k) = \max\{18, \frac{24(\epsilon_1+\epsilon_2)(1+\epsilon_2)}{(1-\epsilon_1)^2}\} \cdot \ln k$.

▶ **Lemma 8.** *If $k \geq D(\epsilon_1,\epsilon_2,\ln k)$, then set $B(\epsilon_1,\epsilon_2,\ln k) = \frac{4(1+\epsilon_2)}{1-\epsilon_1} \cdot \ln(k) + 2$, and otherwise, set $B(\epsilon_1,\epsilon_2,\ln k) = D(\epsilon_1,\epsilon_2,\ln k)$. Then*

$$\mathbf{E}[\Psi_k] \geq \frac{\mathbf{E}[\Phi(U_k,T_k)]}{B(\epsilon_1,\epsilon_2,\ln k)}.$$

The main technical challenge is to bound the increase in the expected average cost of uncovered clusters, namely, the proof of Lemma 8 heavily depends on the following lemma. We state the lemma below and prove it later. We use the notation $[\ldots]_{\mathcal{F}_i}$ to express that all terms inside the brackets take the values determined by $\mathcal{F}_i$.

▶ **Lemma 9.** *Set* $D(\epsilon_1, \epsilon_2, \ln k) = \max\{18, \frac{24(\epsilon_1+\epsilon_2)(1+\epsilon_2)}{(1-\epsilon_1)^2}\} \cdot \ln k$. *If* $k \geq D(\epsilon_1, \epsilon_2, \ln k)$, *then set* $B(\epsilon_1, \epsilon_2, \ln k) = \frac{4(1+\epsilon_2)}{1-\epsilon_1} \cdot \ln(k) + 2$, *and otherwise, set* $B(\epsilon_1, \epsilon_2, \ln k) = D(\epsilon_1, \epsilon_2, \ln k)$. *Then for any* $i \in [k]$ *and any realization* $\mathcal{F}_i$ *of the first* i *iterations*

$$\mathbf{E}\left[\frac{\Phi(U_k, T_k)}{u_k} \,\middle|\, \mathcal{F}_i\right] \leq B(\epsilon_1, \epsilon_2, \ln k) \cdot \left[\frac{\Phi(U_i, T_i)}{u_i}\right]_{\mathcal{F}_i}.$$

We assume Lemma 9 to prove Lemma 8 as follows. Later, we give the proof of Lemma 9.

Proof. For any $i \in \{2, \ldots, k\}$, we obtain using Lemma 9

$$\begin{aligned}
\mathbf{E}\left[W_i \cdot \frac{\Phi(U_i, T_i)}{u_i}\right] &= \sum_{\mathcal{F}_i, [W_i]_{\mathcal{F}_i}=1} \mathbf{Pr}[\mathcal{F}_i] \cdot \left[\frac{\Phi(U_i, T_i)}{u_i}\right]_{\mathcal{F}_i} \\
&\geq \sum_{\mathcal{F}_i, [W_i]_{\mathcal{F}_i}=1} \mathbf{Pr}[\mathcal{F}_i] \cdot \frac{1}{B(\epsilon_1, \epsilon_2, \ln k)} \cdot \mathbf{E}\left[\frac{\Phi(U_k, T_k)}{u_k} \,\middle|\, \mathcal{F}_i\right] \\
&= \frac{1}{B(\epsilon_1, \epsilon_2, \ln k)} \cdot \sum_{\mathcal{F}_i} \mathbf{Pr}[\mathcal{F}_i] \cdot \mathbf{E}\left[W_i \cdot \frac{\Phi(U_k, T_k)}{u_k} \,\middle|\, \mathcal{F}_i\right] \\
&= \frac{1}{B(\epsilon_1, \epsilon_2, \ln k)} \cdot \mathbf{E}\left[W_i \cdot \frac{\Phi(U_k, T_k)}{u_k}\right].
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbf{E}[\Psi_k] &= \sum_{i=2}^{k} \mathbf{E}\left[W_i \cdot \frac{\Phi(U_i, T_i)}{u_i}\right] \\
&\geq \frac{1}{B(\epsilon_1, \epsilon_2, \ln k)} \cdot \sum_{i=2}^{k} \mathbf{E}\left[W_i \cdot \frac{\Phi(U_k, T_k)}{u_k}\right] \\
&= \frac{1}{B(\epsilon_1, \epsilon_2, \ln k)} \cdot \mathbf{E}\left[\left(\sum_{i=2}^{k} W_i\right) \cdot \frac{\Phi(U_k, T_k)}{u_k}\right] \\
&= \frac{1}{B(\epsilon_1, \epsilon_2, \ln k)} \cdot \mathbf{E}\left[u_k \cdot \frac{\Phi(U_k, T_k)}{u_k}\right] \\
&= \frac{\mathbf{E}[\Phi(U_k, T_k)]}{B(\epsilon_1, \epsilon_2, \ln k)}
\end{aligned}$$

◀

With Lemma 7, Lemma 8 and Corollary 6, we prove the main theorem.

▶ **Theorem 2.** *Let* T_k *denote the set of centers sampled by noisy* k*-means++ on dataset* X *and assume that* $\frac{k}{\ln k} \geq \max\{18, \frac{24(\epsilon_1+\epsilon_2)(1+\epsilon_2)}{(1-\epsilon_1)^2}\}$. *Then,*

$$\mathbb{E}[\Phi(X, T_k)] \leq \mathcal{O}\left(\left(\frac{1+\epsilon_2}{1-\epsilon_1}\right)^3 \cdot \log^2(k) \cdot \mathrm{OPT}_k(X)\right),$$

where $\mathrm{OPT}_k(X)$ *denotes the* k*-means costs of an optimal* k*-clustering of* X. *If* $\frac{k}{\ln k} \leq \max\{18, \frac{24(\epsilon_1+\epsilon_2)(1+\epsilon_2)}{(1-\epsilon_1)^2}\}$, *then* $\mathbb{E}[\Phi(X, T_k)] \leq \mathcal{O}\left(\left(\frac{1+\epsilon_2}{1-\epsilon_1}\right)^4 \cdot \log^2\left(\frac{1+\epsilon_2}{1-\epsilon_1}\right) \cdot \mathrm{OPT}_k(X)\right)$.

Proof. For $k \geq \max\{18, \frac{24(\epsilon_1+\epsilon_2)(1+\epsilon_2)}{(1-\epsilon_1)^2}\} \cdot \ln k$, Lemma 7 and Lemma 8 imply

$$\begin{aligned}\mathbf{E}\left[\Phi(U_k, T_k)\right] &\leq B(\epsilon_1, \epsilon_2, \ln k) \cdot \mathbf{E}\left[\Psi_k\right] \\ &\leq B(\epsilon_1, \epsilon_2, \ln k) \cdot \frac{8(1+\epsilon_2)^2}{(1-\epsilon_1)^2} \cdot (\ln(k)+1) \cdot \mathrm{OPT}_k(X) \\ &= O\left(\frac{(1+\epsilon_2)^3}{(1-\epsilon_1)^3} \cdot \log^2(k) \cdot \mathrm{OPT}_k(X)\right).\end{aligned}$$

With Corollary 6 this implies

$$\begin{aligned}\mathbf{E}\left[\Phi(H_k, T_k) + \Phi(U_k, T_k)\right] &\leq O\left(\frac{1+\epsilon_2}{1-\epsilon_1} \cdot \mathrm{OPT}_k(X)\right) + O\left(\frac{(1+\epsilon_2)^3}{(1-\epsilon_1)^3} \cdot \log^2(k) \cdot \mathrm{OPT}_k(X)\right) \\ &\leq O\left(\frac{(1+\epsilon_2)^3}{(1-\epsilon_1)^3} \cdot \log^2(k) \cdot \mathrm{OPT}_k(X)\right)\end{aligned}$$

For $k \leq \max\{18, \frac{24(\epsilon_1+\epsilon_2)(1+\epsilon_2)}{(1-\epsilon_1)^2}\} \cdot \ln k$, we get

$$\begin{aligned}&B(\epsilon_1, \epsilon_2, \ln k) \cdot \frac{8(1+\epsilon_2)^2}{(1-\epsilon_1)^2} \cdot (\ln(k)+1) \cdot \mathrm{OPT}_k(X) \\ &= O\left(\frac{(1+\epsilon_2)^4}{(1-\epsilon_1)^4} \cdot \log^2(k) \cdot \mathrm{OPT}_k(X)\right),\end{aligned}$$

where we use that $\epsilon_1 < 1$, so $\epsilon_1 + \epsilon_2 \leq 1 + \epsilon_2$. This implies

$$\begin{aligned}\mathbf{E}\left[\Phi(H_k, T_k) + \Phi(U_k, T_k)\right] &\leq O\left(\frac{(1+\epsilon_2)^4}{(1-\epsilon_1)^4} \cdot \log^2(k) \cdot \mathrm{OPT}_k(X)\right) \\ &\leq O\left(\frac{(1+\epsilon_2)^4}{(1-\epsilon_1)^4} \cdot \log^2\left(\frac{(1+\epsilon_2)^2}{(1-\epsilon_1)^2}\right) \cdot \mathrm{OPT}_k(X)\right) \\ &= O\left(\left(\frac{1+\epsilon_2}{1-\epsilon_1}\right)^4 \cdot \log^2\left(\frac{1+\epsilon_2}{1-\epsilon_1}\right) \cdot \mathrm{OPT}_k(X)\right),\end{aligned}$$

where we use for the second inequality that either $\sqrt{k} \leq \frac{k}{\ln k} \leq 18$ and then $\ln k \leq O(1)$, or

$$\begin{aligned}\sqrt{k} \leq \frac{k}{\ln k} \leq \frac{24(\epsilon_1+\epsilon_2)(1+\epsilon_2)}{(1-\epsilon_1)^2} &\Rightarrow \ln\sqrt{k} \leq \ln\left(\frac{24(\epsilon_1+\epsilon_2)(1+\epsilon_2)}{(1-\epsilon_1)^2}\right) \\ &\Rightarrow \log k \leq O\left(\log\left(\frac{(1+\epsilon_2)^2}{(1-\epsilon_1)^2}\right)\right).\end{aligned}$$

◀

We conclude this section by discussing Lemma 9. It says that the average potential of uncovered clusters increases by at most a logarithmic multiplicative factor. We first consider the following abstract random experiment whose connection to noisy k-means++ we discuss in the actual proof of Lemma 9 below. Let $a_1, \ldots, a_z \in \mathbb{R}_{\geq 0}$ denote numbers with average value 1. Since there are z numbers with the average equal to one, their sum equals z. We assume that in each step of our experiment with probability $\epsilon \in [0, 1)$ an adversary chooses one of the numbers to be removed and with probability $1 - \epsilon$ a number is removed by proportional sampling (i.e., if number a_i still exists then it is removed with probability a_i/S, where S denotes the sum of the remaining numbers). Note that in this process the number a_i is sampled with probability at least $(1-\epsilon)\frac{a_i}{S}$. Additionally after each step an adversary can arbitrarily lower the value of some numbers. This process is run for ℓ steps and we are interested in an upper bound for the expected average of the numbers remaining after these ℓ steps. We denote this average by A_ℓ.

▶ **Lemma 10.** *Let $\epsilon \in (0,1)$, assume that $\frac{z}{\ln z} \geq \max\{18, \frac{24\epsilon}{(1-\epsilon)^2}\}$, and $\ell \geq z/2$. Then $\mathbf{E}[A_\ell] \leq \frac{4}{1-\epsilon} \cdot \ln(z) + 2$. For $z \leq \max\{18, \frac{24\epsilon}{(1-\epsilon)^2}\} \cdot \ln z$, we observe that $\mathbf{E}[A_\ell] \leq z \leq \max\{18, \frac{24\epsilon}{(1-\epsilon)^2}\} \cdot \ln z$.*

Proof. Let Z denote the number of adversarial steps among the first ℓ steps. Then $\mathbf{E}[Z] = \epsilon\ell$. We denote by $\mathcal{F}_1$ the event that $Z \geq \frac{1+\epsilon}{2} \cdot \ell$. Note that $(1+\epsilon)/2 = \epsilon + (1-\epsilon)/2$ always lies between ϵ and 1.

By Chernoff bound we get

$$\begin{aligned}\mathbf{Pr}[\mathcal{F}_1] = \mathbf{Pr}\left[Z \geq \frac{1+\epsilon}{2}\ell\right] &= \mathbf{Pr}\left[Z \geq \frac{1+\epsilon}{2\epsilon}\epsilon\ell\right] \\ &= \mathbf{Pr}\left[Z \geq \left(1 + \frac{1-\epsilon}{2\epsilon}\right) \cdot \mathbf{E}[Z]\right] \leq \exp\left(-\frac{\min\{\delta, \delta^2\} \cdot \mathbf{E}[Z]}{3}\right)\end{aligned}$$

for $\delta = \frac{1-\epsilon}{2\epsilon}$. We make a case analysis. For $\frac{1-\epsilon}{2\epsilon} \geq 1 \Leftrightarrow \epsilon \leq \frac{1}{3}$, $\min\{\delta, \delta^2\} = \delta$, so we have

$$\mathbf{Pr}[\mathcal{F}_1] \leq \exp\left(-\frac{\delta \cdot \mathbf{E}[Z]}{3}\right) = \exp\left(-\frac{\frac{1-\epsilon}{2\epsilon} \cdot \epsilon\ell}{3}\right) = \exp\left(-\frac{(1-\epsilon)}{6}\ell\right) \leq \exp\left(-\frac{(1-\epsilon)}{12}z\right),$$

where the last inequality follows from $\ell \geq z/2$. We observe that

$$\exp\left(-\frac{(1-\epsilon)}{12}z\right) \leq \frac{1}{z} \Leftrightarrow \frac{z}{\ln z} \geq \frac{12}{1-\epsilon},$$

and by $\epsilon \leq 1/3$ and by our lower bound on $z/\ln z$, we have $\frac{z}{\ln z} \geq 18 \geq \frac{12}{1-\varepsilon}$.

If $\epsilon > 1/3$, we compute similarly that

$$\mathbf{Pr}[\mathcal{F}_1] \leq \exp\left(-\frac{\frac{(1-\epsilon)^2}{(2\epsilon)^2} \cdot \epsilon\ell}{3}\right) \leq \exp\left(-\frac{(1-\epsilon)^2}{12\epsilon}\ell\right) \leq \exp\left(-\frac{(1-\epsilon)^2}{24\epsilon}z\right) \leq \frac{1}{z},$$

where the last inequality follows from $z/(\ln z) \geq \frac{24\epsilon}{(1-\epsilon)^2}$.

If the event $\mathcal{F}_1$ does not happen then in at least $(1 - \frac{1+\epsilon}{2})\ell = \frac{1-\epsilon}{2}\ell \geq \frac{1-\epsilon}{4}z =: z/c'$ steps proportional sampling is used to remove one of the numbers (we set $c' = 4/(1-\epsilon)$). We will show that with high probability after these steps all remaining numbers are at most $2c' \ln z$. Let $\mathcal{F}_2$ denote the event that after z/c' steps of proportional sampling at least one number with final value at least $2c' \ln z$ is remaining. Furthermore, let $\mathcal{E}_i$ denote the event that the ith number a_i remains after z/c' steps of proportional sampling and its final value $\tilde{a}_i$ is at least $2c' \ln z$ (remember that the adversary can decrease numbers during the process but not increase and hence $\tilde{a}_i \leq a_i$). Then $\mathcal{F}_2 = \mathcal{E}_1 \cup \ldots \cup \mathcal{E}_z$. If $\mathcal{E}_i$ occurs then the ith number is in every step at least $\tilde{a}_i \geq 2c' \ln z$. Since the numbers $a_1, \ldots a_z$ have average 1, their sum is z. The sum of the remaining numbers cannot increase during the process. Hence, in every step the probability of taking the ith number is at least $(2c' \ln z)/z$. This implies

$$\mathbf{Pr}[\mathcal{E}_i] \leq \left(1 - \frac{2c' \ln z}{z}\right)^{z/c'} \leq \exp(-2 \ln z) = \frac{1}{z^2}.$$

We use a union bound to obtain

$$\mathbf{Pr}[\mathcal{F}_2] = \mathbf{Pr}[\exists i \in [z] : \mathcal{E}_i] \leq \frac{1}{z}.$$

If neither $\mathcal{F}_1$ nor $\mathcal{F}_2$ occurs then the final value of each remaining number is at most $2c' \ln z$. Hence, in this case, also the average is bounded from above by $2c' \ln z$. Otherwise we only use the trivial upper bound of z for the average of the remaining numbers (observe that initially each a_i is at most z because the average is 1). Altogether we obtain

$$\begin{aligned}\mathbf{E}[A_\ell] &\leq \mathbf{Pr}[\neg\mathcal{F}_1 \wedge \neg\mathcal{F}_2] \cdot 2c' \ln z + \mathbf{Pr}[\mathcal{F}_1 \vee \mathcal{F}_2] \cdot z \\ &\leq 2c' \ln z + (\mathbf{Pr}[\mathcal{F}_1] + \mathbf{Pr}[\mathcal{F}_2]) \cdot z \\ &\leq 2c' \ln z + \left(\frac{1}{z} + \frac{1}{z}\right) \cdot z \\ &= 2c' \ln(z) + 2 = 4/(1-\epsilon) \ln z + 2.\end{aligned}$$

For the second inequality stated in the lemma, we only observe that even if we draw all but one number, the average cannot increase beyond z since the sum of the numbers is z. Thus $\mathbf{E}[A_\ell] \leq z$ is true for any $1 \leq \ell \leq z$. ◀

We prove below that if $\ell < z/2$ then $\mathbf{E}[A_\ell] \leq 2$.

▶ **Lemma 11.** *Let $\ell < z/2$. Then $\mathbf{E}[A_\ell] \leq 2$.*

Proof. In the worst case all steps are adversarial and the ℓ smallest numbers are removed. Then the average of the remaining numbers is at most

$$\frac{z}{z-\ell} < \frac{z}{z-z/2} = 2.$$ ◀

Using Lemma 11, we obtain the following corollary.

▶ **Corollary 12.** *Let $\epsilon \in (0,1)$ and $1 \leq \ell \leq z-1$. Then for $z \geq \max\{18, \frac{24\epsilon}{(1-\epsilon)^2}\} \cdot \ln z$, we get*

$$\mathbf{E}[A_\ell] \leq \frac{4}{1-\epsilon} \cdot \ln z + 2,$$

and for $z \leq \max\{18, \frac{24\epsilon}{(1-\epsilon)^2}\} \cdot \ln z$, we have $\mathbf{E}[A_\ell] \leq \max\{18, \frac{24\epsilon}{(1-\epsilon)^2}\} \cdot \ln z$.

Proof. Follows from Lemma 10 and Lemma 11. ◀

Now we are ready to prove Lemma 9.

Proof of Lemma 9. Given realization F_i, after the first i iterations there are $z = u_i \leq k$ uncovered clusters. Each of them has certain costs with respect to the center set after the first i iterations. The costs of each cluster do not increase in the following iterations anymore because only new centers are added. In any iteration the costs of these clusters may decrease and one uncovered cluster may become covered. If the latter happens, the average costs of the uncovered clusters can increase (if the costs of the uncovered cluster that becomes covered are less than the average costs of the uncovered clusters). Hence, only the non-wasted iterations are of interest.

The costs of the uncovered clusters after the first i iterations correspond to the numbers $a_1, \ldots, a_z$ in the random experiment above. We scaled the instance such that the sum of the a_i is equal to z. This is without loss of generality. In each iteration of noisy k-means++ either a covered cluster is hit again, which can only reduce the numbers a_i, or an uncovered cluster becomes covered, in which case the corresponding number is removed. Conditioned on covering an uncovered cluster, the probability p_i that a_i is removed is at least $\frac{1-\epsilon_1}{1+\epsilon_2} \cdot \frac{a_i}{S}$, where S denotes the sum of the costs of the uncovered clusters (i.e., the sum of the remaining a_i). We can simulate the probability distribution induced by the probabilities p_i

by mixing two distributions: with probability $\frac{1-\epsilon_1}{1+\epsilon_2}$ we do proportional sampling, i.e., we choose a_i with probability $\frac{a_i}{S}$, and with probability $1-\frac{1-\epsilon_1}{1+\epsilon_2}$ we sample according to some other distribution to obtain the right probabilities p_i. In the abstract random experiment analyzed above this second distribution is selected by an adversary. For $\epsilon = \frac{\epsilon_1+\epsilon_2}{1+\epsilon_2} \in (0,1)$ we have

$$1-\epsilon = 1-\frac{\epsilon_1+\epsilon_2}{1+\epsilon_2} = \frac{1+\epsilon_2-(\epsilon_1+\epsilon_2)}{1+\epsilon_2} = \frac{1-\epsilon_1}{1+\epsilon_2}.$$

Hence, Corollary 12 applies to noisy k-means++ with $\epsilon = \frac{\epsilon_1+\epsilon_2}{1+\epsilon_2}$. Observe that then

$$\frac{24\epsilon}{(1-\epsilon)^2} = \frac{24(\epsilon_1+\epsilon_2)(1+\epsilon_2)^2}{(1+\epsilon_2)(1-\epsilon_1)^2} = \frac{24(\epsilon_1+\epsilon_2)(1+\epsilon_2)}{(1-\epsilon_1)^2}.$$ ◀

Bicriteria Approximation

We remark that noisy k-means++ still gives a bicriteria approximation because the probability that an uncovered cluster is hit can only be decreased by a constant factor, and the probability to pick a good center is also still comparably high. The theorem mentioned below follows from [1]. We omit the proof of this theorem in this paper as it easily follows from [1].

▶ **Theorem 13.** *Let S denote a set of $16(\frac{1+\epsilon_2}{1-\epsilon_1})^2(k+\sqrt{k})$ centers sampled using noisy k-means++, then $\Phi(X,S) \leq 20\ \mathrm{OPT}_k(X)$ with probability at least $1-\exp\left(-0.0157\cdot\frac{1-\epsilon_1}{1+\epsilon_2}\right)$.*

Not so greedy k-means++

Consider the following variant of the greedy k-means++ algorithm (Algorithm 3).

Algorithm 3 Moderately greedy k-means++.

Input: Set $X \subseteq \mathbb{R}^d$, integers k, l
Output: $C \subseteq X, |C| = k$
$C = \emptyset$
Sample a point c_1 independently and uniformly at random from X.
Let $C = \{c_1\}$.
for $i = 2$ to k **do**
    With probability p, sample one point c_i with D^2-sampling and set $C = C \cup \{c_i\}$.
    With the remaining probability:
        Sample a set S of ℓ points independently with D^2-sampling from X wrt C.
        Let $c_i = \arg\min_{u\in S} \Phi(X, C\cup\{u\})$.
        Update $C = C \cup \{c_i\}$.
end for
Return C

Let $x \in P$ be any point. Say that $p_i(x)$ is the probability to draw x with one D^2-sample from X based on the center set $c_1,\ldots,c_{i-1}$. Then the probability $q_i(x)$ to sample x in iteration i of the above algorithm satisfies

$$p\cdot p_i(x) \leq q_i(x) \leq [(1-p)\cdot\ell + p]\cdot p_i(x),$$

since with probability p, we do exactly the same as k-means++, and with probability $(1-p)$, we sample ℓ times, which can at most boost the probability by a factor of $(1-p)\cdot\ell$. Assume that p is a constant. Then by Theorem 2, moderately greedy k-means++ has an expected approximation guarantee of $\mathcal{O}(\ell^3\cdot\log^2 k)$ (for large k).

References

1 Ankit Aggarwal, Amit Deshpande, and Ravi Kannan. Adaptive sampling for k-means clustering. In *Proceedings of the 12th and 13th APPROX-RANDOM*, pages 15–28, 2009.

2 Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. Better guarantees for k-means and euclidean k-median by primal-dual algorithms. In *Proceedings of the 58th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pages 61–72, 2017.

3 Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

4 David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the 18th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1027–1035, 2007.

5 Pranjal Awasthi, Moses Charikar, Ravishankar Krishnaswamy, and Ali Kemal Sinop. The hardness of approximation of euclidean k-means. In *Proceedings of the 31st International Symposium on Computational Geometry (SoCG)*, pages 754–767, 2015.

6 Olivier Bachem, Mario Lucic, S. Hamed Hassani, and Andreas Krause. Approximate k-means++ in sublinear time. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 1459–1467, 2016. URL: `http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12147`.

7 Anup Bhattacharya, Ragesh Jaiswal, and Nir Ailon. Tight lower bound instances for k-means++ in two dimensions. *Theoretical Computer Science*, 634:55–66, 2016.

8 Tobias Brunsch and Heiko Röglin. A bad instance for k-means++. *Theoretical Computer Science*, 505:19–26, 2013.

9 M. Emre Celebi, Hassan A. Kingravi, and Patricio A. Vela. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, 40(1):200–210, 2013.

10 Davin Choo, Christoph Grunau, Julian Portmann, and Václav Rozhon. k-means++: few more steps yield constant approximation. *CoRR*, abs/2002.07784, 2020. `arXiv:2002.07784`.

11 Vincent Cohen-Addad, Philip N. Klein, and Claire Mathieu. Local search yields approximation schemes for k-means and k-median in euclidean and minor-free metrics. In *Proceedings of the 57th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pages 353–364, 2016.

12 Sanjoy Dasgupta. Lecture 3 – Algorithms for k-means clustering, 2013. accessed May 8th, 2019. URL: `http://cseweb.ucsd.edu/~dasgupta/291-geom/kmeans.pdf`.

13 Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd ACM Symposium on Theory of Computing (STOC)*, pages 569–578, 2011.

14 Zachary Friggstad, Mohsen Rezapour, and Mohammad R. Salavatipour. Local search yields a PTAS for k-means in doubling metrics. *SIAM Journal on Computing*, 48(2):452–480, 2019.

15 Daniel J. Hsu and Matus Telgarsky. Greedy bi-criteria approximations for k-medians and k-means. *CoRR*, abs/1607.06203, 2016. `arXiv:1607.06203`.

16 Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. A local search approximation algorithm for k-means clustering. *Computational Geometry*, 28(2-3):89–112, 2004.

17 Silvio Lattanzi and Christian Sohler. A better k-means++ algorithm via local search. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 3662–3671, 2019.

18 Euiwoong Lee, Melanie Schmidt, and John Wright. Improved and simplified inapproximability for k-means. *Information Processing Letters*, 120:40–43, 2017.

19 Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. originally published as Bell Laboratories Technical Memorandum in 1957.

20 Meena Mahajan, Prajakta Nimbhorkar, and Kasturi R. Varadarajan. The Planar k-means Problem is NP-Hard. In *Proceedings of the 3rd Workshop on Algorithms and Computation (WALCOM)*, pages 274–285, 2009.

21 Benedikt Pago. Upper and lower bounds for the approximation ratios of incremental and hierarchical clustering algorithms. Master's thesis, University of Bonn, 2018.

22 Sergei Vassilvitskii. *k-means: Algorithms, Analyses, Experiments*. PhD thesis, Stanford University, 2007.

23 Dennis Wei. A constant-factor bi-criteria approximation guarantee for k-means++. In *Proceedings of the Annual Conference on Neural Information Processing Systems 2016 (NIPS)*, pages 604–612, 2016.

24 Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus F. M. Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2008.