Analyzing D^α seeding for k-means

Etienne Bamas [1] **Sai Ganesh Nagarajan** [2] **Ola Svensson** [3]

Abstract

One of the most popular clustering algorithms is the celebrated D^α seeding algorithm (also know as k-means++ when $\alpha = 2$) by Arthur and Vassilvitskii (2007), who showed that it guarantees in expectation an $O(2^{2\alpha} \cdot \log k)$-approximate solution to the (k,α)-clustering cost (where distances are raised to the power α) for any $\alpha \geq 1$. More recently, Balcan, Dick, and White (2018) observed experimentally that using D^α seeding with $\alpha > 2$ can lead to a better solution with respect to the standard k-means objective (i.e. the $(k, 2)$-clustering cost). In this paper, we provide a rigorous understanding of this phenomenon. For any $\alpha > 2$, we show that D^α seeding guarantees in expectation an approximation factor of

$$O_\alpha\left(\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{2-4/\alpha} \cdot (g_\alpha \cdot \min\{\ell, \log k\})^{2/\alpha}\right)$$

with respect to the standard k-means cost of any underlying clustering; where g_α is a parameter capturing the concentration of the points in each cluster, $\sigma_{\max}$ and $\sigma_{\min}$ are the maximum and minimum standard deviation of the clusters around their center, and ℓ is the number of distinct mixing weights in the underlying clustering (after rounding them to the nearest power of 2). For instance, if the underlying clustering is defined by a mixture of k Gaussian distributions with equal cluster variance (up to a constant-factor), then our result implies that: (1) if there are a constant number of mixing weights, any constant $\alpha > 2$ yields a constant-factor approximation; (2) if the mixing weights are arbitrary, any constant $\alpha > 2$ yields an $O\left(\log^{2/\alpha} k\right)$-approximation, and $\alpha = \Theta(\log \log k)$ yields an $O(\log \log k)^3$-approximation. We complement these results by some lower bounds showing that the dependency on g_α and $\sigma_{\max}/\sigma_{\min}$ is tight. Finally, we provide an experimental validation of the effects of the aforementioned parameters when using D^α seeding.

1. Introduction

Clustering is a quintessential machine learning problem with numerous practical applications in medicine (Alashwal et al., 2019), image segmentation (Shi & Malik, 2000; Burney & Tariq, 2014), market analysis (Chiu et al., 2009) and anomaly detection (Münz et al., 2007), to name a few. One of the most popular formulations is the k-means problem that requires us to pick k centers such that the sum of the squared distance from each data point to its closest center is minimized. The k-means problem is NP-hard even in 2 dimensions (Mahajan et al., 2009) and most research is therefore focused on heuristics and approximation algorithms. For a long time, a heavily used heuristic for this problem has been the Lloyd's algorithm, with Expectation Maximization (EM) style updates for the centers after an initial set of k centers are chosen uniformly at random from the data. While this method finds a local optimum, it is known not to have any approximation guarantees and it could have an exponential run time in the worst case (Arthur & Vassilvitskii, 2006).

The k-means++ method. (Arthur & Vassilvitskii, 2007) came up with the elegant k-means++ method that carefully selects the initial centers (also called D^2 seeding) such that the next center is a data point that is chosen with probability that is proportional to its squared distance to its closest center, selected thus far (see (Ostrovsky et al., 2013) for a concurrent work on a similar algorithm as well). This intuitively makes sense as this initialization is more likely to discover new clusters (that are far away) than simply selecting centers uniformly at random. Indeed, they proved that the initial choice of centers already forms a $O(\log k)$ approximation (in expectation) for this problem. They complemented this upper bound with a family of instances where the expected cost of the D^2 seeding is a factor $\Omega(\log k)$ times the optimum cost, showing that their analysis is tight.

[1]ETH AI Center, Zurich, Switzerland. [2]Zuse Institute Berlin [3]EPFL. Correspondence to: Etienne Bamas <etienne.bamas@inf.ethz.ch>, Sai Ganesh Nagarajan <nagarajan@zib.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Limitations on clusterable instances. While the D^2 seeding method provides clear improvements over uniformly at random initialization in an elegant and efficient manner, the family of instances that show the tightness of the analysis indicates some of its limitations. The family of instances presented in (Arthur & Vassilvitskii, 2007) is indeed highly clusterable: k regular simplices of radius one (each of n/k points) and the pairwise distance between the centers of two simplices are Δ. As Δ tends to ∞ (i.e., the instance becomes more and more clusterable), the expected approximation guarantee of the D^2 seeding method tends to $\Theta(\log k)$.

For such clusterable instances, the issue is that the D^2 seeding method does not put enough probability mass on discovering new clusters. This phenomenon was already observed in the original paper by (Arthur & Vassilvitskii, 2007), and they proposed a greedy variant that takes several samples at each iteration (increasing the probability that at least one hits a new yet undiscovered cluster) and makes a greedy choice among them. This greedy variant has worse guarantees in the worst case (see (Grunau et al., 2023b) for a recent nearly tight analysis). However, the greedy variant shows better experimental performance (after all, we usually look for k clusters when the data is clusterable), and is currently the method implemented in the popular Scikit-learn library (Pedregosa et al., 2011). Specifically, at each iteration $2 + \log(k)$ points are sampled, and, among them, the point that decreases the objective the most is greedily chosen.

Data-driven approach. More recently, Balcan et al. (Balcan et al., 2018) proposed a data-driven approach in order to address the aforementioned limitation of the D^2 seeding method on clusterable data. Instead of always using D^2 seeding for the initial centers, they proposed to use D^α seeding where α is now a parameter of the algorithm. In D^α seeding, a point is selected as the next center with probability proportional to its α-powered distance to its closest center, selected thus far[1]. One can observe that a large choice of $\alpha > 2$ increases the probability that a sampled center will discover a new yet undiscovered cluster which is advantageous. At the same time, a large α makes the algorithm more sensitive to outliers (which is also the reason why the greedy variant of k-means++ is worse in the worst case). Hence the selection of α should depend on the kind of instances one wants to solve, which motivates a data-driven approach. One of the main results in (Balcan et al., 2018) is that this is feasible. They showed that the parameter α is learnable in the sense that if we assume the instance is drawn from some unknown distribution $\mathcal{D}$, then with only polynomially many (in the instance size and other relevant parameters) samples and a polynomial running time, one can compute a parameter $\tilde{\alpha}$ for the sampling that is almost optimum for the given distribution $\mathcal{D}$. This is especially interesting as it shows that setting α to a good value on a given distribution is in principle a task that is manageable. Additionally, Balcan et al. (Balcan et al., 2018) complemented their theoretical results with an experimental analysis that shows that setting α equal to 2 is not always the best choice. For instance, on the MNIST dataset, they find that setting α close to 4 is a significantly better choice than $\alpha = 2$. This is even more striking in the case where $\mathcal{D}$ is a mixture of Gaussians, in which case setting α close to 20 seems the best choice. This highlights the fact that in practice, one can outperform the popular k-means++ algorithm of (Arthur & Vassilvitskii, 2007) by tweaking the parameter α. Yet, they do not provide any quantitative understanding of this phenomenon nor provide any approximation guarantees on these instances with different α. This is the main focus of this paper.

[1] We remark that D^α seeding was already considered in (Arthur & Vassilvitskii, 2007) but they studied it on a cost that was proportional to the distances raised to the power α (i.e. the (k, α)-clustering cost) instead of the standard k-means objective. They showed that D^α seeding was $O(2^{2\alpha} \log k)$-approximate for this cost function. The use of D^α on other objectives, including the standard squared distance objective was first introduced in (Balcan et al., 2018) in a data-driven approach.

Our contributions. Our main contribution is a theoretical analysis of the advantage of D^α seeding, proving that it leads to constant-factor approximation guarantees for a large class of instances, including a balanced mixture of k Gaussians, where the standard k-means++ algorithm is already no better than $\Omega(\log k)$ (see Section 3). We remark that a beyond worst-case analysis is essential as it is easy to see that $\alpha = 2$ is an optimal choice in the worst-case (just as the greedy variant of k-means++ is worse in the worst-case).

In our beyond worst-case analysis, we identify natural data-dependent parameters that measure (i) how concentrated points are in clusters (the parameter g_α), (ii) the ratio of the maximum and minimum standard deviation of the optimal clusters around their mean ($\frac{\sigma_{\max}}{\sigma_{\min}}$), and (iii) how balanced clusters are in terms of number of points (the parameter ℓ). Using these parameters, we show that D^α seeding guarantees for any $\alpha > 2$ an

$$O_\alpha\left(\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{2-4/\alpha} \cdot (g_\alpha \cdot \min\{\ell, \log k\})^{2/\alpha}\right)$$

approximation with respect to the standard k-means objective (the formal statement can be found in Section 2). We further show that the dependence on the first two parameters is necessary and tight (formal statement in Section 2), and this dependence gives a theoretical explanation of the importance of selecting α as a function of the data (Section 2.2). We leave it as an interesting open problem to understand the necessity of the third parameter ℓ.

Finally, a more open-ended direction following our work is

to give a beyond-worst-case analysis of the greedy variant of k−means++. We take a first step in this direction by proving a negative result: we give a family of instances where the natural parameters (i)-(iii) are all constant (and thus D^α seeding yields a constant-factor approximation guarantee for any constant $\alpha > 2$) but greedy k−means++ as implemented in the Scikit-learn library (with $\Theta(\log k)$ samples per iteration) has a super constant approximation guarantee (see Section 2).

1.1. Further related works

Several variants have been studied since the original publication of the k−means++ method in (Arthur & Vassilvitskii, 2007; Ostrovsky et al., 2013). Aggarwal et al. (Aggarwal et al., 2009) obtained an $O(1)$-approximation with constant probability by selecting $O(k)$ centers, which was improved later by (Wei, 2016; Makarychev et al., 2020). Bahmani et al. (Bahmani et al., 2012) provided a scalable version k-means++ that is even more practical, and (Bachem et al., 2017; Cohen-Addad et al., 2020) provided faster ways for randomly selecting the centers. Recently, Lattanzi and Sohler (Lattanzi & Sohler, 2019) obtained a $O(1)$-approximation with additional $O(k \log\log k)$ steps of local search after using k-means++ to choose the initial centers and this was improved by Choo et al. (Choo et al., 2020) to obtain a 10^{30}-approximation with ϵk steps of local search. Our constant factor guarantees are arguably smaller without any additional steps of local search but are applicable to the appropriate family of instances, whilst their method is applicable in the worst case across all instances. Moreover, their local search methods can be augmented on top of our guarantees of D^α-seeding, which may offer significant improvements.

Another variant of k−means++ particularly relevant to our setting is the *noisy* k−means++ algorithm, in which points are sampled according to the standard D^2 seeding, but an adversary is allowed to perturb the sampling probabilities by some multiplicative factor of $(1 \pm \varepsilon)$. For this case, (Bhattacharya et al., 2020) showed an $O(\log^2 k)$ upper-bound, which was then improved to the classic $O(\log k)$ by (Grunau et al., 2023a). In our case, the sampling probability of a point can be completely different between D^2 seeding and D^α seeding. Since the analysis of noisy k−means++ is substantially more difficult than the analysis of the classic k−means++ algorithm, it might come as a surprise that it is still possible to obtain non-trivial guarantees in our case.

Tangentially, one could study algorithms for learning the cluster centers when the data is instantiated from a mixture of Gaussians (Dasgupta, 1999; Arora & Kannan, 2005). Furthermore, specific clustering algorithms are created under specific assumptions on the instances with various clusterability notions (Ackerman & Ben-David, 2009). However, these clusterability notions are often (computationally) hard to check and the algorithms are not as efficient and simple as the seeding-based algorithms, which also work well in practice (without any assumptions). Finally, the main inspiration of this paper is from the idea of data-driven clustering by Balcan et al. (Balcan et al., 2018).

1.2. Preliminaries and notations

To formally introduce the k-means problem and the seeding algorithms, we will need to work with a metric space $(\mathbb{R}^d, \|\cdot\|_2)$. Since we always work with the Euclidean norm, we will drop the subscript in the notation and write the Euclidean norm of a vector x to be $\|x\|$. If we are given some data points $\mathcal{X} \subset \mathbb{R}^d$, the cost of our data $\mathcal{X}$ associated with a given set of t centers Z_t can be defined as follows:

$$\text{cost}^{(2)}(\mathcal{X}, Z_t) := \sum_{x \in \mathcal{X}} \min_{c \in Z_t} \|x - c\|^2. \tag{1}$$

Now note that the centers Z_t define a natural partition of the data, in that: $C_j = \{x \in \mathcal{X} : c_j = \arg\min_{c \in Z_t} \|x - c\|^2\}$. Then one can write the cost equivalently in the following useful way:

$$\text{cost}^{(2)}(\mathcal{X}, Z_t) = \sum_{j=1}^{t} \sum_{x \in C_j} \|x - c_j\|^2. \tag{2}$$

Furthermore, if one has a candidate clustering $\mathcal{C}$, then each corresponding center can be computed as $c_j = \frac{1}{|C_j|} \sum_{x \in C_j} x$, which are the centroids of the corresponding clusters. We will denote the optimal centroids by $\{\mu_1, \mu_2, \ldots, \mu_k\}$. By a slight abuse of notations, we might drop the subscript to identify a cluster $C \in \mathcal{C}$, and μ_C will refer to the mean of that cluster. Let $\mathcal{C}_{\texttt{OPT}}$ be the optimal clustering whose corresponding centers are $Z_{\texttt{OPT}}$. By definition,

$$\text{cost}^{(2)}(\mathcal{X}, Z_{\texttt{OPT}}) = \min_{Z \subset \mathbb{R}^d : |Z| = k} \text{cost}^{(2)}(\mathcal{X}, Z). \tag{3}$$

The k−means++ algorithm. We will now describe the class of parameterized seeding algorithms as in (Arthur & Vassilvitskii, 2007; Balcan et al., 2018) for the k-means objective. For any $\alpha \in [0, \infty]$, the general D^α seeding procedure chooses k centers as follows:

1. The first center $z_1 \in \mathcal{X}$ is chosen uniformly at random from the data points.
2. Let Z_t be the set of t centers chosen so far, such that $t < k$. The next center, $z_{t+1} \in \mathcal{X}$ is chosen with probability given by:

$$p^{(\alpha)}_{\mathcal{X}}(z) := \mathbb{P}(z \text{ is sampled} | Z_t) = \frac{\min_{c \in Z_t} \|z - c\|^\alpha}{\sum_{z \in \mathcal{X}} \min_{c \in Z_t} \|z - c\|^\alpha} \tag{4}$$

The classic $k-\texttt{means++}$ algorithm is the special case of D^2 seeding. For any $\alpha \geq 1$, Arthur and Vassilvitskii (Arthur & Vassilvitskii, 2007) show that D^α-seeding procedure is an $O(2^{2\alpha} \log k)$ approximation in expectation if the cost function is given by:

$$\text{cost}^{(\alpha)}(\mathcal{X}, Z_k) := \sum_{x \in \mathcal{X}} \min_{c \in Z_k} \|x - c\|^\alpha \,. \tag{5}$$

Our focus in this paper is to provide guarantees on the D^α seeding algorithm with $\alpha > 2$ for the standard k-means objective (i.e. $\alpha = 2$).

The greedy $k-$means++ algorithm. Although this is not the main focus of the paper, it is helpful to define briefly here the greedy variant of $k-\texttt{means++}$. The greedy variant with m samples works as follows.

1. The first center $z_1 \in \mathcal{X}$ is chosen uniformly at random from the data points.

2. Let Z_t be the set of t centers chosen so far, such that $t < k$. We select a set of m candidate centers $z_1, z_2, \ldots, z_m$ where each candidate is sampled according to the probability distribution

$$p_{\mathcal{X}}(z) := \mathbb{P}(z \text{ is sampled} | Z_t) = \frac{\min_{c \in Z_t} \|z - c\|^2}{\sum_{z \in \mathcal{X}} \min_{c \in Z_t} \|z - c\|^2} \tag{6}$$

   The next added center z_{t+1} is selected to be the one which decreases the cost the most, among all the candidate centers $z_1, z_2, \ldots, z_m$.

Usually m is selected to mildly increase with the input size. For instance the standard scikit-learn library implements the greedy version of $k-\texttt{means++}$ using $m = \Theta(\log k)$ candidates at each step (see (Pedregosa et al., 2011)).

2. Our Results

In this section, we define formally the natural parameters that D^α seeding depends on and state formally our results. Moreover, we will provide a short discussion on the necessity of the dependence that will clarify our claims. The last part of this section focuses on using our results to provide recommendations on choosing α in different scenarios.

Before we move on to stating our results on the D^α seeding, we need to define the following quantities with respect to the optimal[2] clustering $\mathcal{C}_{\texttt{OPT}} = \{C_1, C_2, \ldots, C_k\}$.

[2]Although we state our definitions and results with respect to the optimal clusters, our results hold for any reference clustering that satisfies the aforementioned properties.

1. We define σ_C as the standard deviation of the points inside cluster $C \in \mathcal{C}_{\texttt{OPT}}$. More precisely,

$$\sigma_C := \sqrt{\frac{\sum_{x \in C} \|x - \mu_C\|^2}{|C|}} \,. \tag{7}$$

   Following this, $\sigma_{\max}$ is defined as the maximum standard deviation of points inside a given cluster, i.e. $\sigma_{\max} := \max_{C \in \mathcal{C}_{\texttt{OPT}}} \sigma_C$, and similarly $\sigma_{\min} := \min_{C \in \mathcal{C}_{\texttt{OPT}}} \sigma_C$.

2. We need a parameter g_α that measures the concentration of the distances of the points to the centroid μ_C in a cluster C:

$$g_\alpha := \max_{C \in \mathcal{C}_{\texttt{OPT}}} \frac{(1/|C|) \cdot \text{cost}^{(\alpha)}(C, \mu_C)}{(\text{cost}^{(2)}(C, \mu_C)/|C|)^{\alpha/2}} \tag{8}$$

$$= \max_{C \in \mathcal{C}_{\texttt{OPT}}} \frac{(1/|C|) \cdot \sum_{x \in C} d^\alpha(x, \mu_C)}{\left((1/|C|) \cdot \sum_{x \in C} d^2(x, \mu_C)\right)^{\alpha/2}} \,. \tag{9}$$

   One can see that g_α is equal to the α^{th} absolute standardized moment of one cluster C (see Chapter 4 in (Kenney, 1939) for a reference).

3. Finally, we need a parameter ℓ to control the number of distinct weights of clusters (where the weight of a cluster C is simply equal to the number of points $|C|$). Formally, for any integer $i \geq 0$ we let k_i to be the number of clusters of $\mathcal{C}_{\texttt{OPT}}$ whose weight lies in the interval $[2^i, 2^{i+1})$. Then we define the following key parameter:

$$\ell := |\{i \in \mathbb{N} \mid k_i > 0\}| \,, \tag{10}$$

   which is the number of intervals of the form $[2^i, 2^{i+1})$ containing the weight of at least one cluster. For instance, if all clusters have the same weight then $\ell = 1$. If all the clusters have weights in some interval $[m, M]$ then $\ell \leq \log_2(M/m)$.

Remark 2.1. Note that we can express $\texttt{OPT}(C)$ for some optimal cluster C, in terms of its standard deviation by $\texttt{OPT}(C) = |C|\sigma_C^2$, and the total cost of the optimal clustering is $\texttt{OPT} = \sum_{C \in \mathcal{C}_{\texttt{OPT}}} |C|\sigma_C^2$.

Given the aforementioned definitions, the main result that we show in this paper is the following theorem, whose formal proof appears in Section 3.

Theorem 2.2. *For any clustering $(C_1, C_2, \ldots, C_k)$ of cost $\texttt{OPT}$, and any $\alpha > 2$, the D^α seeding procedure returns a clustering of expected cost at most $\texttt{OPT}$ times*

$$O\left(f(\alpha) \cdot \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{2-4/\alpha} \cdot (g_\alpha \cdot \min\{\ell, \log k\})^{2/\alpha}\right) \,,$$

where $f(\alpha) := \frac{\alpha^2}{(\alpha/2-1)^{2/\alpha} \cdot (1 - 2^{2/\alpha - 1})}$. In particular, $f(\alpha) = O(\alpha/\varepsilon)^2$ if $\alpha \in [2 + \varepsilon, \infty)$ for some small $\varepsilon > 0$.

An immediate consequence of Theorem 2.2 is that D^α seeding with $\alpha > 2$ fixed yields a constant-factor approximation guarantee for instances consisting of k regular simplices (i.e. all points in a cluster are arranged in a regular simplex) of the same radius (which implies $\sigma_{\max}/\sigma_{\min} = 1$) and same size (which implies $\ell = 1$); a case that includes the described $\Omega(\log k)$ lower bound instances for the standard D^2 seeding from (Arthur & Vassilvitskii, 2007). Indeed, one can check that $g_\alpha = 1$ in that case. We remark that the above stated guarantees do not require the clusters to be separated (as they are e.g. in the described lower bound instances of D^2 seeding), which is a common assumption is several other works (see e.g. (Ostrovsky et al., 2013; Ackerman & Ben-David, 2009)).

We complement our results with the following lower bounds, which are fairly intuitive to prove.

Theorem 2.3. *There exists an instance with a clustering of cost* $\texttt{OPT}$ *such that* $\ell = 1$, $\sigma_{\max}/\sigma_{\min} = 1$, *and the* D^α *seeding procedure returns a clustering of expected cost at least*

$$\Omega\left(g_\alpha\right)^{2/\alpha} \cdot \texttt{OPT}\,,$$

and another instance instance with a clustering of cost $\texttt{OPT}$ *such that* $\ell = 1$, $(g_\alpha)^{2/\alpha} = O(1)$, *and the* D^α *seeding procedure returns a clustering of expected cost at least*

$$\Omega\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{2-4/\alpha} \cdot \texttt{OPT}\,.$$

The formal proof of Theorem 2.3 can be found in Appendix D.1 and Appendix D.2. Moreover, one can use the lower bound instance in Appendix D.2 to obtain a single instance with $k = 2$ and such that D^α seeding is no better than an $\omega(1)$-approximation for all $\alpha \geq 2 + \varepsilon$ for some small fixed $\epsilon > 0$[3]. Finally, as mentioned in introduction, we also prove a lower bound on the greedy variant of $\texttt{k-means++}$.

Theorem 2.4. *There exists an instance with* k *clusters for which* D^α *seeding guarantees a constant factor approximation in expectation for any fixed* $\alpha > 2$, *and such that the greedy* k-$\texttt{means++}$ *algorithm with* $f(k)$ *samples is not better than an* $\Omega(\log\log f(k))^2$ *approximation in expectation.*

This highlights that D^α seeding can be superior in theory to the greedy variant. The proof of this last theorem can be found in Appendix D.3.

2.1. Discussion on the parameters

As we see, the guarantee of D^α seeding as stated in Theorem 2.2 has a dependence on g_α, $\sigma_{\max}/\sigma_{\min}$, and $\min\{\ell, \log k\}$. Here we discuss these dependencies.

[3]However, it would be interesting to see if one can obtain a single instance which is $\Omega(\log k)$ for all values of $\alpha \geq 2$ simultaneously.

The parameter g_α. The moment condition can be seen as a characterization of the concentration of a cluster of points. For instance, one way that g_α could be non-constant is when the cluster has many outliers that are still part of the cluster. On the contrary, if our clusters are generated by a Gaussian mixture, then g_α (for any constant $\alpha \geq 2$) is a constant (in fact, it is not difficult to compute and see that $g_\alpha \leq \alpha^\alpha$ for Gaussian distributions). If we are in the infinite number of samples limit, where each cluster becomes defined by a density function f on some domain $\mathcal{D}$, then g_α is equal to

$$\frac{\int_{\mathcal{D}} \|x-\mu\|^\alpha f(x)dx}{\left(\int_{\mathcal{D}} \|x-\mu\|^2 f(x)dx\right)^{\alpha/2}}\,,$$

To obtain a better understanding of our guarantees, we give below the value of $(g_\alpha)^{2/\alpha}$ for a few common distributions. W.l.o.g. we re-normalize to assume unit variance of each distribution (i.e. the denominator is equal to 1 in the definition of g_α).

1. Perhaps the most classic distribution is the Gaussian distribution with unit variance. In this case, the standardized moment is equal to $O((\alpha/2)^{\alpha/2})$ thus $(g_\alpha)^{2/\alpha}$ is $O(\alpha)$. Furthermore, for multivariate Gaussians, $g_\alpha = O(\alpha^\alpha d^\alpha/2^{d/2})$. Meaning, in higher dimensions g_α decreases rapidly and this is well-supported by our understanding that a Gaussian distribution tends to behave like "balls" in higher dimensions (see Section 3.3.3 in (Vershynin, 2020)). In Remark 2.5, we detail how this g_α can be used to obtain the guarantees for mixture of Gaussians (as claimed in the abstract).

2. For the exponential distribution, which has slightly heavier tails than Gaussian, $\texttt{Exp}(\lambda)$, the α^{th} moment is $O(\alpha!)$ and thus $(g_\alpha)^{2/\alpha}$ is $O(\alpha^2)$.

3. Now consider for instance, a univariate student-t distribution with degree of freedom $\nu > 0$, has its density function given by, $\frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-((\nu+1)/2)}$ (WikipediaT). It is well known that α^{th} moment exists only if $\alpha < \nu$. Thus lower the degree of freedom, the heavier the tail gets, and g_α is bounded only when $\alpha < \nu$, in which case it is roughly $O(\nu^\alpha)$ and thus $(g_\alpha)^{2/\alpha} = O(\nu^2)$.

Remark 2.5. The claims made in abstract for instances drawn from a mixture of Gaussian distributions are now easy to see. Suppose the mixture of k Gaussians $\mathcal{X} \sim \sum_{i=1}^k w_i \mathcal{N}(\mu_i, \Sigma_i)$, satisfies $\max_{i,j\in[k]} \text{tr}(\Sigma_i)/\text{tr}(\Sigma_j) = O(1)$ so that $\sigma_{\max}/\sigma_{\min} = O(1)$. Therefore, in that case we obtain an approximation guarantee of $O(\alpha^2 g_\alpha^{2/\alpha} \cdot \min\{\ell, \log k\}^{2/\alpha}) = O(\alpha^3 \cdot \min\{\ell, \log k\}^{2/\alpha})$. If the mixing weights w_i are all equal, then $\ell = 1$ and we obtain an

approximation ratio of $O(\alpha^3)$, which is constant for any constant $\alpha > 2$. If the mixing weights are arbitrary, our approximation ratio is at most $O(\alpha^3 \min\{\ell, \log k\}^{2/\alpha})$, which is $O(\log\log k)^3$ for $\alpha = \Theta(\log\log k)$. Finally we show that we can achieve this approximation even in the case of a finite number of samples taken from an arbitrary mixture of Gaussians (see Appendix C.2).

We show in Appendix D.2 that the dependency on g_α in Theorem 2.2 is tight. As a simple example that highlights the intuition, consider the instance given in Figure 1. The red cluster is drawn from a standard 2-dimensional Gaussian law. The blue cluster consists of many points highly concentrated at distance δ from the mean of the red cluster, and one single point at distance $\delta + \Delta$ from the mean of the red cluster. For this blue cluster, g_α will be unbounded when Δ tends to ∞ for any $\alpha > 2$. Note that we can choose the parameters in this instance so that (i) both clusters have the same variance and (ii) both clusters have the same number of points so the other parameters do not play a role here. In this situation, there is still $1/2$ probability that the first center is selected in the red cluster (the first center is always chosen uniformly at random), and conditioned on that fact, the D^α seeding (for $\alpha > 2$) will give way too much probability to the isolated point in the blue cluster, which is a serious issue. Our lower bound construction in Appendix D.2 is a simple formalization of this intuition.

The parameter $\sigma_{\max}/\sigma_{\min}$. The dependence on $\sigma_{\max}/\sigma_{\min}$ is necessary and in fact tight (see Appendix D.1). The main issue in sampling with large α is that the algorithm might sample repeatedly from a cluster with large standard deviation more often, and thus it might fail to discover some other clusters.

The parameter ℓ. The dependence on $\min\{\ell, \log k\}$ remains an intriguing open problem, and it is unclear to us if any dependency on ℓ or k is needed when $\alpha > 2$. Note that for α going to infinity, the parameter ℓ should matter less and less since D^α seeding becomes equivalent to picking the furthest point. This behavior is accurately reflected in our bound.

2.2. On choosing α

Our theorem states that there is a trade-off in choosing α. We already know that $\alpha = 2$ may not be the best choice and this is due to the well-clusterable instance of simplices of equal sidelength that are sufficiently far apart. However, Theorem 2.2 implies that any $\alpha > 2$ is a constant factor approximation in this case, and this is because D^α is more aggressive in discovering new clusters. But is it in our best interest to set $\alpha \to \infty$? Interestingly, Balcan et al. (Balcan et al., 2018) show that the best α is learnable, hence selecting the best α is a task that is manageable when there is a training set. Moreover, our theorem predicts a new phenomenon that is not present in the experiments in (Balcan et al., 2018). For a mixture of balanced Gaussians, (Balcan et al., 2018) obtain experimental results whose pattern roughly matches the one shown in left-hand side of Figure 2. This experiment corresponds to a Gaussian mixture with the *same* covariance matrix (namely identity). However, our theory indicates that there is a dependence on the variances that is necessary and it appears in the right-hand side of Figure 2, where one of the Gaussians has much larger variance. Note that our approximation factor has dependence $(\sigma_{\max}/\sigma_{\min})^{2-4/\alpha}$, and to mitigate this effect one can choose some α that is not too large but still greater than 2. In fact, our result suggests a simple strategy. Using the training set, one can obtain an estimate of the key parameters $g_\alpha, \sigma_{\max}/\sigma_{\min}, \ell$, and use these estimates to sufficiently narrow down the search for the best α, hence speeding-up the data-driven approach proposed by (Balcan et al., 2018).

Note that the experiments mentioned here do not use additional steps of Lloyd's algorithm. Since it is quite common to use this algorithm after the seeding, we run additional experiments in Section B. Interestingly, we observe that the general pattern does not change much even after running Lloyd's algorithm until convergence. This means that choosing $\alpha > 2$ has some significant benefits over $\alpha = 2$, even if we run Lloyd's algorithm after the seeding.

As a final note, we mention that it is a common wisdom among practitioners that k-means is a good objective, except when the clusters might have varying sizes and density, or when there are many outliers (Google; scikit). In this context, "varying sizes and density" can be interpreted as the parameters ℓ and $\sigma_{\max}/\sigma_{\min}$, while outliers correspond to the parameter g_α. If one believes this common wisdom, then our result essentially implies that whenever k-means is a good clustering objective, then choosing $\alpha > 2$ should be almost always better than $\alpha = 2$.

3. Proof Sketch of Theorem 2.2

The proof of Theorem 2.2 is inspired by a very clean potential function analysis of the D^2 seeding algorithm by (Dasgupta, 2013). In a similar fashion, it is useful to bound the potential increase at each step. However, as we will see later, the potential function that is used for the D^2 seeding analysis does not seem to work for D^α seeding, and some additional ideas are required. While there are other examples of potential-based analysis of seeding algorithms for k-means in (Aggarwal et al., 2009), to the best of our knowledge, the potential that we use is the first to be able to obtain interesting guarantees when the sampling of a point can be arbitrarily far from being proportional to its cost. This is generally non-trivial as evidenced by recent results in Grunau et al. (2023a); Bhattacharya et al. (2020). We de-

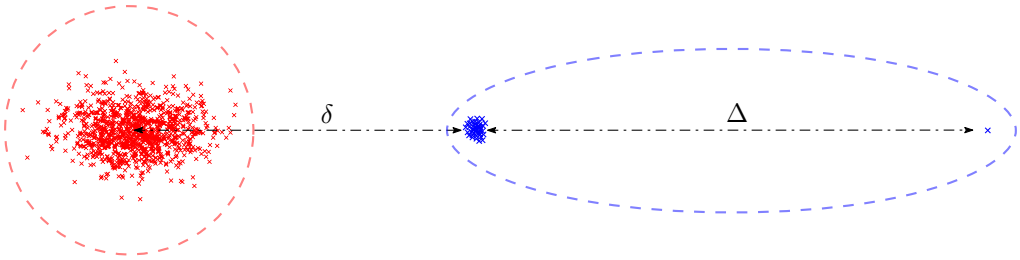

Figure 1. An instance with $k = 2$.

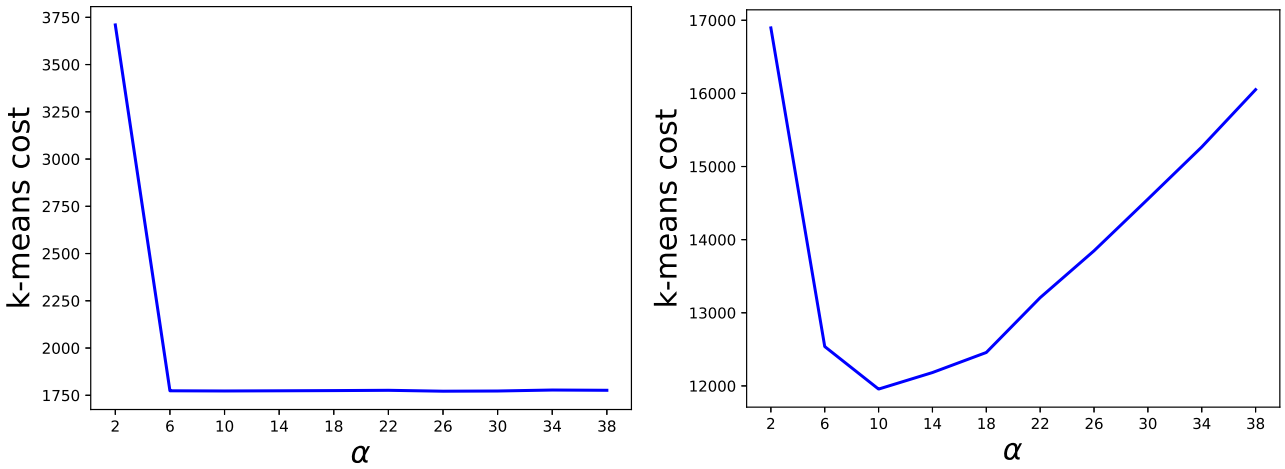

Figure 2. Performance of D^α seeding for two instances $\mathcal{D}_1$ (on the left) and $\mathcal{D}_2$ (on the right), from a balanced mixture of k Gaussians with $k = 4$ and $d = 2$. The centers/means of the Gaussians for both instances are placed on the vertices of a square of side length 100. However, the covariance matrices for $\mathcal{D}_1$ is $\{I, I, I, I\}$, whilst the covariance matrices for $\mathcal{D}_2$ is $\{800I, I, I, I\}$, where I is the identity matrix in two dimensions.

fer a more detailed discussion of the novelty of our potential function to Appendix C.1.

As in (Dasgupta, 2013), at every iteration t, it is useful to keep track of the set of optimal clusters from which a center has already been chosen (i.e. the *hit* clusters) and the complement of this set which is the set of *undiscovered* clusters. Formally, we define H_t to be the set of *hit* clusters after selecting a set of t centers denoted by Z_t, i.e.

$$H_t := \{C \in \mathcal{C}_{\texttt{OPT}} : C \cap Z_t \neq \emptyset\} ,$$

where we recall that $\mathcal{C}_{\texttt{OPT}}$ is the set of clusters in the optimum solution. U_t is defined to be the set of remaining *undiscovered* clusters, i.e, $U_t := \mathcal{C}_{\texttt{OPT}} \setminus H_t$. Furthermore, we define $\text{cost}_t^{(2)}(C)$ as a shorthand to denote the cost induced by the points in the cluster C, after the set Z_t of t centers are chosen. More formally,

$$\text{cost}_t^{(2)}(C) := \text{cost}^{(2)}(C, Z_t) = \sum_{x \in C} \min_{z \in Z_t} \|x - z\|^2 .$$

Since we analyze the D^α seeding, we also need to work with the α-cost:

$$\text{cost}_t^{(\alpha)}(C) := \text{cost}_t^{(\alpha)}(C, Z_t) = \sum_{x \in C} \min_{z \in Z_t} \|x - z\|^\alpha .$$

For any set S of clusters, we define $\text{cost}_t^{(2)}(S) := \sum_{C \in S} \text{cost}_t^{(2)}(C)$, and $\text{cost}_t^{(\alpha)}(S) := \sum_{C \in S} \text{cost}_t^{(\alpha)}(C)$. Moreover we can talk about the cost of a single point at iteration t as $\text{cost}_t^{(2)}(x) := \min_{z \in Z_t} \|x - z\|^2$. The main challenge of the proof is to upperbound the expected cost of undiscovered clusters after selecting our k centers. The cost of hit clusters can be upperbounded fairly easily, in a similar manner as in the analysis of the D^2 seeding in (Arthur & Vassilvitskii, 2007).

The potential function. Now we proceed to define our potential function which will be used to upperbound the cost of undiscovered clusters. For each $i \geq 0$, we define S_i to be the set of clusters in $C \in \mathcal{C}_{\texttt{OPT}}$ such that $|C|$ lies in the interval $[2^i, 2^{i+1})$ (recall that we also defined $k_i := |S_i|$). For each $i \geq 0, t \geq 0$, we define an integer $\tau_i(t) \geq 0$ which will be a local counter, relevant only for the clusters in S_i at

iteration t. z_t is defined to be the center selected at iteration t, and $U_t^{(i)} := U_t \cap S_i$ the set of undiscovered clusters in S_i. For each $i \geq 0$ and time t, we will define an integer $w_i(t) \geq 0$ corresponding to the number of iterations that are considered *wasted* by the clusters in S_i at time t. We use the word *wasted* to follow the intuition given in (Dasgupta, 2013) where an iteration t is wasted when the selected center z_t belongs to an already discovered center (in particular this new center does not discover a new cluster).

Some intuition. In our potential function, the counter τ_i will intuitively count how many iterations were relevant to the set of clusters S_i. Once τ_i reaches the value k_i, we will consider that the set S_i was given enough tries to cover its clusters. During the sampling, it might be that some cluster is hit twice, which will increase the wasted counter w_j for all $j \geq 0$ (and τ_j also increases for all $j \geq 0$). It might be counter intuitive that when a cluster in S_i is hit twice then this still counts as a try in other sets S_j even if $j \neq i$. Indeed, it might be much more natural to simply count this try only for the set of clusters S_i. However, proceeding in this more intuitive way would create a serious issue that some set S_j might get less than k_j tries, because some other set $S_{j'}$ gets more tries than needed.

Formally, we initialize $t = 0$, $w_i(0) = 0$ for all $i \in \mathbb{N}$, and $\tau_i(0) = 0$ for all $i \in \mathbb{N}$. Then, we maintain these quantities as follows. At time $t \geq 0$, let C be the cluster the next center z_t is chosen from. Let $i \geq 0$ be the integer such that $C \in S_i$. Then, there are two cases.

1. If $C \in U_t^{(i)}$, then we set (for all $j \geq 0$)

$$\tau_j(t+1) = \begin{cases} \tau_j(t) + 1 \text{ if } j = i \text{ and } \tau_j(t) < k_j \\ \tau_j(t) \text{ otherwise.} \end{cases}$$

   and $w_j(t+1) = w_j(t)$ for all $j \geq 0$.

2. Otherwise if $C \in H_t$, then we set (for all $j \geq 0$)

$$\tau_j(t+1) = \begin{cases} \tau_j(t) + 1 \text{ if } \tau_j(t) < k_j \\ \tau_j(t) \text{ if } \tau_j(t) \geq k_j \,, \end{cases}$$

   and

$$w_j(t+1) = \begin{cases} w_j(t) + 1 \text{ if } \tau_j(t) < k_j \\ w_j(t) \text{ if } \tau_j(t) \geq k_j \,. \end{cases}$$

Based on these quantities, we define the potential function as follows. First, we define

$$\phi_i(t) := \frac{w_i(t)}{|U_t^{(i)}|} \cdot (2^i)^{1-2/\alpha} \cdot \sum_{C \in U_t^{(i)}} (\mathrm{cost}_t^{(\alpha)}(C))^{2/\alpha} \,. \quad (11)$$

The final potential function can now be defined as

$$\phi(t) := \sum_{i \geq 0} \phi_i(t) \,. \quad (12)$$

Remark 3.1. Note that we introduce the quantity

$$(2^i)^{1-2/\alpha} \cdot \sum_{C \in U_t^{(i)}} (\mathrm{cost}_t^{(\alpha)}(C))^{2/\alpha} \,.$$

It is not obvious that this is a good quantity to control, since we are interested in upper-bounding the quantity

$$\sum_{C \in U_t^{(i)}} \mathrm{cost}_t^{(2)}(C) \,.$$

However, by a simple application of a standard convexity inequality, one will notice that

$$(2^{i+1})^{1-2/\alpha} \cdot \sum_{C \in U_t^{(i)}} (\mathrm{cost}_t^{(\alpha)}(C))^{2/\alpha} \geq \sum_{C \in U_t^{(i)}} \mathrm{cost}_t^{(2)}(C) \,.$$

Using this potential, we are ready to proceed with the main proof. Akin to Dasgupta's analysis (Dasgupta, 2013), we split the proof in three main parts. In Section A.1, we show that $\phi(k)$ is indeed an upper bound on the final cost of undiscovered clusters. In Section A.2, we upper bound the cost of hit clusters using g_α. In Section A.3 we upper-bound the increase of the potential function. Finally, we complete the proof in Section A.4.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

This work was supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number MB22.00054. Part of this work had been carried out when Sai Ganesh Nagarajan was at EPFL.

References

Ackerman, M. and Ben-David, S. Clusterability: A theoretical study. In *Artificial intelligence and statistics*, pp. 1–8. PMLR, 2009.

Aggarwal, A., Deshpande, A., and Kannan, R. Adaptive sampling for k-means clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques: 12th International Workshop, APPROX*

2009, and 13th International Workshop, RANDOM 2009, Berkeley, CA, USA, August 21-23, 2009. Proceedings, pp. 15–28. Springer, 2009.

Alashwal, H., El Halaby, M., Crouse, J. J., Abdalla, A., and Moustafa, A. A. The application of unsupervised clustering methods to alzheimer's disease. *Frontiers in computational neuroscience*, 13:31, 2019.

Arora, S. and Kannan, R. Learning mixtures of separated nonspherical gaussians. *The Annals of Applied Probability*, 15(1A):69–92, 2005.

Arthur, D. and Vassilvitskii, S. How slow is the k-means method? In *Proceedings of the twenty-second annual symposium on Computational geometry*, pp. 144–153, 2006.

Arthur, D. and Vassilvitskii, S. k-means++ the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035, 2007.

Artin, E. *The gamma function*. Courier Dover Publications, 2015.

Bachem, O., Lucic, M., and Krause, A. Distributed and provably good seedings for k-means in constant rounds. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 292–300. PMLR, 2017.

Bahmani, B., Moseley, B., Vattani, A., Kumar, R., and Vassilvitskii, S. Scalable k-means++. *arXiv preprint arXiv:1203.6402*, 2012.

Balcan, M.-F. F., Dick, T., and White, C. Data-driven clustering via parameterized lloyd's families. *Advances in Neural Information Processing Systems*, 31, 2018.

Bhattacharya, A., Eube, J., Röglin, H., and Schmidt, M. Noisy, greedy and not so greedy k-means++. In *28th Annual European Symposium on Algorithms (ESA 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

Burney, S. A. and Tariq, H. K-means cluster analysis for image segmentation. *International Journal of Computer Applications*, 96(4), 2014.

Chiu, C.-Y., Chen, Y.-F., Kuo, I.-T., and Ku, H. C. An intelligent market segmentation system using k-means and particle swarm optimization. *Expert systems with applications*, 36(3):4558–4565, 2009.

Choo, D., Grunau, C., Portmann, J., and Rozhon, V. k-means++: few more steps yield constant approximation. In *International Conference on Machine Learning*, pp. 1909–1917. PMLR, 2020.

Cohen-Addad, V., Lattanzi, S., Norouzi-Fard, A., Sohler, C., and Svensson, O. Fast and accurate k-means++ via rejection sampling. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

D^α Seeding. Link to code. https://github.com/saiganesh223/Alpha-Seeding-for-k-means, 2024.

Dasgupta, S. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pp. 634–644. IEEE, 1999.

Dasgupta, S. Algorithms for k-means clustering. https://cseweb.ucsd.edu/~dasgupta/291-geom/kmeans.pdf, 2013.

Google. k-means advantages and disadvantages. https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages?hl=en. Accessed: 2023-10-06.

Grunau, C., Özüdoğru, A. A., and Rozhoň, V. Noisy k-means++ revisited. In *31st Annual European Symposium on Algorithms (ESA 2023)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2023a.

Grunau, C., Özüdogru, A. A., Rozhon, V., and Tetek, J. A nearly tight analysis of greedy k-means++. In *Proceedings of the 2023 ACM-SIAM Symposium on Discrete Algorithms, SODA 2023, Florence, Italy, January 22-25, 2023*, pp. 1012–1070. SIAM, 2023b.

Hardy, G. H., Littlewood, J. E., and Pólya, G. *Inequalities*. Cambridge university press, 1952.

Kenney, J. F. *Mathematics of statistics*. D. Van Nostrand, 1939.

Lattanzi, S. and Sohler, C. A better k-means++ algorithm via local search. In *International Conference on Machine Learning*, pp. 3662–3671. PMLR, 2019.

Mahajan, M., Nimbhorkar, P., and Varadarajan, K. The planar k-means problem is np-hard. In *International workshop on algorithms and computation*, pp. 274–285. Springer, 2009.

Makarychev, K., Reddy, A., and Shan, L. Improved guarantees for k-means++ and k-means++ parallel. *Advances in Neural Information Processing Systems*, 33:16142–16152, 2020.

Münz, G., Li, S., and Carle, G. Traffic anomaly detection using k-means clustering. In *Gi/itg workshop mmbnet*, volume 7, 2007.

Ostrovsky, R., Rabani, Y., Schulman, L. J., and Swamy, C. The effectiveness of lloyd-type methods for the k-means problem. *Journal of the ACM (JACM)*, 59(6):1–22, 2013.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

scikit. k-means clustering. https://scikit-learn.org/stable/modules/clustering.html. Accessed: 2024-01-18.

Shi, J. and Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

van der Vaart, A. and Wellner, J. A. Weak convergence and empirical processes. 1996.

Vershynin, R. High-dimensional probability. *University of California, Irvine*, 2020.

Wei, D. A constant-factor bi-criteria approximation guarantee for k-means++. *Advances in neural information processing systems*, 29, 2016.

WikipediaT. Student-t distribtuion. https://en.wikipedia.org/wiki/Student%27s_t-distribution. Accessed: 2023-10-07.

A. Proof of Theorem 2.2

In the whole proof of the main result, we will work with a slightly modified parameter g_α. We define

$$\hat{g}_\alpha := \max_{C \in \mathcal{C}_{\text{OPT}}} \frac{(1/|C|^2) \cdot \sum_{z \in C} \text{cost}^{(\alpha)}(C, z)}{(\text{cost}^{(2)}(C, \mu_C)/|C|)^{\alpha/2}} \,. \tag{13}$$

To see how this relates to the g_α parameter we can write

$$\begin{aligned}
(1/|C|^2) \cdot \sum_{z \in C} \text{cost}^{(\alpha)}(C, z) &= \frac{1}{|C|^2} \sum_{z \in C} \sum_{x \in C} d^\alpha(x, z) \\
&\leq \frac{1}{|C|^2} \sum_{z \in C} \sum_{x \in C} 2^\alpha (d^\alpha(x, \mu_C) + d^\alpha(z, \mu_C)) \\
&= \frac{2^\alpha}{|C|^2} \left(|C| \sum_{x \in C} d^\alpha(x, \mu_C) + |C| \sum_{y \in C} d^\alpha(y, \mu_C) \right) = \frac{2^{\alpha+1}}{|C|} \cdot \sum_{x \in C} d^\alpha(x, \mu_C) \,,
\end{aligned}$$

where the second line uses the triangle inequality and the standard inequality $(x + y)^\alpha \leq 2^\alpha (x^\alpha + y^\alpha)$ for any $x, y \geq 0$.

Hence we clearly have

$$\hat{g}_\alpha \leq 2^{\alpha+1} \cdot g_\alpha \,,$$

and since all our guarantees will be involve the quantity $(\hat{g}_\alpha)^{2/\alpha}$, we can simply hide the factor $(2^{\alpha+1})^{2/\alpha}$ in the big O and replace $\hat{g}_\alpha$ by g_α.

Now we are ready to use the potential function defined in Section 3 for our proof.

A.1. Relating the potential function and the cost of undiscovered clusters

This part is fairly straightforward, using a few lemmas.

Lemma A.1. *For all $i \in \mathbb{N}$, we have that $w_i(k) \geq |U_k^{(i)}|$.*

Proof. Consider the quantity $\delta_i(t) := |U_t^{(i)}| - w_i(t)$. Clearly, we have that $\delta_i(0) = k_i$. Next, notice that for every time-step $t < k$ such that $\tau_i(t+1) = \tau_i(t) + 1$, the quantity $\delta_i(t)$ decreases by 1. Indeed, either the algorithm discovers a new cluster in $|U_t^{(i)}|$ (in which case $|U_{t+1}^{(i)}| = |U_t^{(i)}| - 1$), or the algorithm wastes an iteration in some cluster, in which case $w_i(t+1) = w_i(t) + 1$.

Finally, we claim that there are at least k_i such iterations. Note that if $\tau_i(t) < k_i$, then the only way we have that τ_i does not increase is if the algorithm discovers a new cluster in $|U_t^{(j)}|$ for some $j \neq i$. This can happen at most $\sum_{j \neq i} k_j$ times. Hence there must be at least $k - \sum_{j \neq i} k_j = k_i$ iterations where the counter τ_i increases. If we denote by t_i the first time at which $\tau_i(t_i) = k_i$, this implies $t_i \leq k$ and that that $\delta_i(t_i) = 0$ hence $w_i(t_i) = |U_{t_i}^{(i)}|$. Finally, note that for $t > t_i$, $w_i(t)$ does not change anymore, and $|U_t^{(i)}|$ can only decrease; which concludes the proof. □

Lemma A.2. *We have that $\phi(k) \geq \sum_{i \in \mathbb{N}} \text{cost}^{(2)}(U_k^{(i)}, Z_k)/2 = \text{cost}_k^{(2)}(U_k)/2$.*

Proof. Using Lemma A.1, we have

$$\phi(k) = \sum_{i \in \mathbb{N}} \phi_i(k) \geq (1/2) \cdot (2^{i+1})^{1-2/\alpha} \cdot \sum_{C \in U_k^{(i)}} (\text{cost}^{(\alpha)}(C, Z_k))^{2/\alpha} \,.$$

Using Jensen's inequality (see Appendix C.3 for a reference) and the fact that $|C| \leq 2^{i+1}$ for all $C \in U_k^{(i)}$, we have that

$$(2^{i+1})^{1-2/\alpha} (\text{cost}^{(\alpha)}(C, Z_k))^{2/\alpha} \geq \text{cost}_k^{(2)}(C)$$

for all $C \in U_k^{(i)}$, which concludes the proof. □

A.2. The cost of hit clusters

In this section, we give upper bounds on the expected cost of clusters that were hit during the seeding process. These proofs are similar to the ones found in (Dasgupta, 2013; Arthur & Vassilvitskii, 2007). In fact, the first two lemmas are taken directly from (Arthur & Vassilvitskii, 2007).

Lemma A.3 (From (Arthur & Vassilvitskii, 2007))**.** *Assume some arbitrary set T of centers have already been selected, and $z \in C$ is added next using D^α-sampling. Then,*

$$\mathbb{E}\left[\mathrm{cost}^{(\alpha)}(C, T \cup \{z\}) \mid z \in C\right] \leq 2^{2\alpha} \cdot \mathrm{cost}^{(\alpha)}(C, \mu_C) \; .$$

Lemma A.4 (From (Arthur & Vassilvitskii, 2007))**.** *Assume some point z is selected uniformly at random among the points belonging to some cluster C. Then, for any $\alpha \geq 2$,*

$$\mathbb{E}\left[\mathrm{cost}^{(\alpha)}(C, z)\right] \leq 2^{\alpha} \cdot \mathrm{cost}^{(\alpha)}(C, \mu_C) \; .$$

The next lemma deals with the expected squared cost of the hit clusters during the seeding process.

Lemma A.5. *Assume some arbitrary set T of centers have already been selected, and $z \in C$ is added next using D^α seeding. Then,*

$$\mathbb{E}\left[\mathrm{cost}^{(2)}(C, T \cup \{z\}) \mid z \in C\right] \leq (4e + (\alpha+1)^2 \cdot (\hat{g}_\alpha)^{2/\alpha}) \cdot \mathrm{cost}^{(2)}(C, \mu_C) \; .$$

Proof. We start by noting that if $\mathrm{cost}^{(2)}(C, T) \leq (\alpha+1)^2 \cdot (\hat{g_\alpha})^{2/\alpha} \mathrm{cost}^{(2)}(C, \mu_C)$ then the lemma already holds since adding an additional center can only decrease the cost of C. Therefore, we assume this is not the case in the rest of the proof. We can write

$$\mathbb{E}\left[\mathrm{cost}^{(2)}(C, T \cup \{z\}) \mid z \in C\right] = \sum_{z \in C} \frac{\mathrm{cost}^{(\alpha)}(z, T)}{\mathrm{cost}^{(\alpha)}(C, T)} \cdot \sum_{x \in C} \min\{\mathrm{cost}^{(2)}(x, T), \|x - z\|^2\} \; .$$

Let us upperbound the quantity $\mathrm{cost}^{(\alpha)}(z, T)$. For this, let us fix any $x \in C$. Then, if we denote by t_x the point in T which is closest to $x \in C$, we have that

$$\mathrm{cost}^{(\alpha)}(z, T) \leq \|z - t_x\|^\alpha \leq (\|z - x\| + \|x - t_x\|)^\alpha \leq (\alpha+1)^\alpha \cdot \|z - x\|^\alpha + (1 + 1/\alpha)^\alpha \cdot \|x - t_x\|^\alpha \; ,$$

using the triangle inequality and a case distinction whether $\|z - x\| > (1/\alpha)\|x - t_x\|$ or not. Averaging this upper bound over all $x \in C$, we obtain that

$$\begin{aligned} \mathrm{cost}^{(\alpha)}(z, T) &\leq \frac{(\alpha+1)^\alpha \cdot \mathrm{cost}^{(\alpha)}(C, z)}{|C|} + \frac{(1 + 1/\alpha)^\alpha \cdot \mathrm{cost}^{(\alpha)}(C, T)}{|C|} \\ &\leq \frac{(\alpha+1)^\alpha \cdot \mathrm{cost}^{(\alpha)}(C, z)}{|C|} + \frac{e \cdot \mathrm{cost}^{(\alpha)}(C, T)}{|C|} \; . \end{aligned}$$

Hence we can rewrite

$$\begin{aligned} &\mathbb{E}\left[\mathrm{cost}^{(2)}(C, T \cup \{z\}) \mid z \in C\right] \\ &\leq \sum_{z \in C} \frac{\frac{(\alpha+1)^\alpha \cdot \mathrm{cost}^{(\alpha)}(C, z)}{|C|} + \frac{e \cdot \mathrm{cost}^{(\alpha)}(C, T)}{|C|}}{\mathrm{cost}^{(\alpha)}(C, T)} \cdot \sum_{x \in C} \min\{\mathrm{cost}^{(2)}(x, T), ||x - z||^2\} \\ &\leq \frac{(\alpha+1)^\alpha}{|C|} \sum_{z \in C} \frac{\mathrm{cost}^{(\alpha)}(C, z)}{\mathrm{cost}^{(\alpha)}(C, T)} \cdot \mathrm{cost}^{(2)}(C, T) + \frac{e}{|C|} \sum_{z \in C} \mathrm{cost}^{(2)}(C, z) \; . \end{aligned}$$

To finish the argument, note that the second term in the last line corresponds to e times the expected cost of C if we pick once center $z \in C$, uniformly at random. By Lemma A.4, this is at most $(4e) \cdot \mathrm{cost}^{(2)}(C, \mu_C)$. For the second term, we use Equation (13) to write

$$\frac{(\alpha+1)^\alpha}{|C|} \sum_{z \in C} \frac{\mathrm{cost}^{(\alpha)}(C, z)}{\mathrm{cost}^{(\alpha)}(C, T)} \cdot \mathrm{cost}^{(2)}(C, T) \leq ((\alpha+1)^\alpha \hat{g_\alpha}) \cdot (\mathrm{cost}^{(2)}(C, \mu_C))^{\alpha/2} \cdot \frac{\mathrm{cost}^{(2)}(C, T)}{|C|^{\alpha/2-1} \mathrm{cost}^{(\alpha)}(C, T)} \; .$$

Using Jensen's inequality, we obtain that $|C|^{\alpha/2-1}\text{cost}^{(\alpha)}(C,T) \geq (\text{cost}^{(2)}(C,T))^{\alpha/2}$. Hence we finally get that

$$\frac{(\alpha+1)^\alpha}{|C|}\sum_{z\in C}\frac{\text{cost}^{(\alpha)}(C,z)}{\text{cost}^{(\alpha)}(C,T)}\cdot\text{cost}^{(2)}(C,T) \leq ((\alpha+1)^\alpha \hat{g_\alpha})\cdot\frac{(\text{cost}^{(2)}(C,\mu_C))^{\alpha/2}}{(\text{cost}^{(2)}(C,T))^{\alpha/2-1}}\ .$$

Using our assumption that $\text{cost}^{(2)}(C,T) \geq (\alpha+1)^2\cdot(\hat{g_\alpha})^{2/\alpha}\text{cost}^{(2)}(C,\mu_C)$, we clearly get

$$\frac{(\alpha+1)^\alpha}{|C|}\sum_{z\in C}\frac{\text{cost}^{(\alpha)}(C,z)}{\text{cost}^{(\alpha)}(C,T)}\cdot\text{cost}^{(2)}(C,T) \leq ((\alpha+1)^2\cdot(\hat{g_\alpha})^{2/\alpha})\cdot\text{cost}^{(2)}(C,\mu_C)\ ,$$

which finishes the proof. □

The last lemma relates the squared cost and the α-powered cost of any cluster.

Lemma A.6. *For any cluster C, we have*

$$\text{cost}^{(\alpha)}(C,\mu_C) \leq \hat{g_\alpha}\cdot|C|\cdot(\sigma_C)^\alpha\ .$$

Proof. We note that $\text{cost}^{(\alpha)}(C,\mu_C) \leq \frac{1}{|C|}\sum_{z\in C}\text{cost}^{(\alpha)}(C,z)$ (using Jensen's inequality, and the convexity of the function $y \mapsto \sum_{z\in C}||z-y||^\alpha$). Hence, we obtain

$$\begin{aligned}\text{cost}^{(\alpha)}(C,\mu_C) &\leq \left(\frac{1}{|C|}\sum_{z\in C}\text{cost}^{(\alpha)}(C,z)\right)\\ &\leq |C|^{1-\alpha/2}\cdot\hat{g_\alpha}\cdot(\text{cost}^{(2)}(C,\mu_C))^{\alpha/2} = \hat{g_\alpha}\cdot|C|\cdot(\sigma_C)^\alpha\ ,\end{aligned}$$

where the second inequality uses our definition of $\hat{g}_\alpha$. □

A.3. The increase of potential

In this section, we bound the final potential $\phi(k)$. First, we analyze the increase of local potential ϕ_i individually, then we use these results to bound the final expected potential $\mathbb{E}[\phi(k)]$.

A.3.1. The increase in a weight class

Lemma A.7. *Let B_t be the event that the t-th center is selected from H_t. Then, for any $t>0$, $i\geq 0$, and any past choice of centers Z_{t-1}, we have that*

$$\mathbb{E}\Big[\phi_i(t)-\phi_i(t-1)\mid\{Z_{t-1},B_t\}\Big] \leq (\tau_i(t)-\tau_i(t-1))\cdot\frac{(2^i)^{1-2/\alpha}}{|U_{t-1}^{(i)}|}\cdot\sum_{C\in U_{t-1}^{(i)}}(\text{cost}_{t-1}^{(\alpha)}(C))^{2/\alpha}\ .$$

Proof. Fix some $i\geq 0$. If $\tau_i(t-1)=\tau_i(t)$, we have by definition $w_i(t)=w_i(t-1)$ and clearly the potential ϕ_i cannot increase. Otherwise, we simply note that $w_i(t)=w_i(t-1)+1$, and the result clearly follows. Note that in both cases, we use the fact that $(\text{cost}_t^{(\alpha)}(C))^{2/\alpha} \leq (\text{cost}_{t-1}^{(\alpha)}(C))^{2/\alpha}$. □

Lemma A.8. *Let $A_t^{(i)}$ be the event that the t-th center is selected from and undiscovered cluster belonging to the weight class i. Then, for any $t>0$, $i,j\geq 0$, any past choice of centers Z_{t-1}, we have that*

$$\mathbb{E}\Big[\phi_j(t)-\phi_j(t-1)\mid\{Z_{t-1},A_t^{(i)}\}\Big] \leq 0.$$

Proof. We first note that since $z_t\notin H_t$, by definition we have $w_j(t)=w_j(t-1)$ for all $j\neq i$, hence $\phi_j(t)\leq\phi_j(t-1)$ for all $j\neq i$. To bound the change of ϕ_i we notice that $w_i(t)=w_i(t-1)$, and that $|U_t^{(i)}|=|U_{t-1}^{(i)}|-1$. Let C_z be the cluster which is selected in $U_{t-1}^{(i)}$. We claim that

$$\mathbb{E}\Big[(\text{cost}_{t-1}^{(\alpha)}(C_z))^{2/\alpha}\mid\{Z_{t-1},A_t^{(i)}\}\Big] \geq \frac{1}{|U_{t-1}^{(i)}|}\cdot\sum_{C\in U_{t-1}^{(i)}}(\text{cost}_{t-1}^{(\alpha)}(C))^{2/\alpha}\ . \tag{14}$$

Indeed, note that

$$\begin{aligned}
\mathbb{E}\left[(\mathrm{cost}_{t-1}^{(\alpha)}(C_z))^{2/\alpha} \mid \{Z_{t-1}, A_t^{(i)}\}\right] &= \sum_{C\in U_{t-1}^{(i)}} \frac{\mathrm{cost}_{t-1}^{(\alpha)}(C)}{\sum_{C\in U_{t-1}^{(i)}} \mathrm{cost}_{t-1}^{(\alpha)}(C)} \cdot \left(\mathrm{cost}_{t-1}^{(\alpha)}(C)\right)^{2/\alpha} \\
&= \frac{|U_{t-1}^{(i)}|}{\sum_{C\in U_{t-1}^{(i)}} \mathrm{cost}_{t-1}^{(\alpha)}(C)} \cdot \sum_{C\in U_{t-1}^{(i)}} \frac{\mathrm{cost}_{t-1}^{(\alpha)}(C) \cdot \left(\mathrm{cost}_{t-1}^{(\alpha)}(C)\right)^{2/\alpha}}{|U_{t-1}^{(i)}|} \\
&\overset{(1)}{\geq} \frac{|U_{t-1}^{(i)}|}{\sum_{C\in U_{t-1}^{(i)}} \mathrm{cost}_{t-1}^{(\alpha)}(C)} \cdot \sum_{C\in U_{t-1}^{(i)}} \frac{\mathrm{cost}_{t-1}^{(\alpha)}(C)}{|U_{t-1}^{(i)}|} \cdot \sum_{C\in U_{t-1}^{(i)}} \frac{(\mathrm{cost}_{t-1}^{(\alpha)}(C))^{2/\alpha}}{|U_{t-1}^{(i)}|} \\
&= \sum_{C\in U_{t-1}^{(i)}} \frac{(\mathrm{cost}_{t-1}^{(\alpha)}(C))^{2/\alpha}}{|U_{t-1}^{(i)}|} .
\end{aligned}$$

The inequality (1) can be obtained using Lemma C.2 (Chebyshev's sum inequality) by considering the ordered sequence $\left(\mathrm{cost}_{t-1}^{(\alpha)}(C)\right)_{C\in U_{t-1}^{(i)}}$ and $\left((\mathrm{cost}_{t-1}^{(\alpha)}(C))^{2/\alpha}\right)_{C\in U_{t-1}^{(i)}}$. Thus we have:

$$\begin{aligned}
&\mathbb{E}\left[\phi_i(t) \mid \{Z_{t-1}, A_t^{(i)}\}\right] \\
&\leq \frac{w_i(t)}{|U_{t-1}^{(i)}| - 1} \cdot (2^i)^{1-2/\alpha} \cdot \left(\sum_{C\in U_{t-1}^{(i)}} (\mathrm{cost}_t^{(\alpha)}(C))^{2/\alpha} - \mathbb{E}\left[(\mathrm{cost}_{t-1}^{(\alpha)}(C_z))^{2/\alpha} \mid \{Z_{t-1}, A_t^{(i)}\}\right] \right) \\
&\leq \frac{w_i(t)}{|U_{t-1}^{(i)}| - 1} \cdot (2^i)^{1-2/\alpha} \cdot \left(\frac{|U_{t-1}^{(i)}| - 1}{|U_{t-1}^{(i)}|} \cdot \sum_{C\in U_{t-1}^{(i)}} (\mathrm{cost}_{t-1}^{(\alpha)}(C))^{2/\alpha} \right) \\
&= \phi_i(t-1).
\end{aligned}$$

□

The previous two lemmas upper-bound the increase of ϕ_j conditioned on events B_t or $A_t^{(i)}$. Using this two bounds, the next lemma upper-bounds the expected increase of the potential ϕ_j, removing the conditioning on the events B_t, $A_t^{(i)}$.

Lemma A.9. *For every $i \geq 0, \alpha > 2, t > 0$, we have*

$$\mathbb{E}\left[\phi_i(t) - \phi_i(t-1) \mid \{\tau_i(t) = \tau_i(t-1) + 1\}\right] \leq h(\alpha) \cdot \left(2^i\right)^{1-2/\alpha} \cdot (k_i - \tau_i(t-1))^{-2/\alpha} \cdot \left((2^{2\alpha} \hat{g_\alpha}) \sum_{C\in\mathcal{C}_{OPT}} |C| (\sigma_C)^\alpha \right)^{2/\alpha},$$

where $h(\alpha) = \frac{(\alpha/2-1)^{1-2/\alpha}}{\alpha/2}$, and

$$\mathbb{E}\left[\phi_i(t) - \phi_i(t-1) \mid \{\tau_i(t) = \tau_i(t-1)\}\right] \leq 0 .$$

Proof. If $\tau_i(t) = \tau_i(t-1)$, then either we can apply Lemma A.7 or Lemma A.8, and in both cases the expected potential ϕ_i can only decrease.

If $\tau_i(t) = \tau_i(t-1) + 1$ then either z_t belongs to $U_{t-1}^{(i)}$, in which case the potential ϕ_i can only decrease in expectation (by Lemma A.8 again). Therefore there remains only the case that $z_t \in H_{t-1}$ (we denote this event by B_t). In this case, let us denote by Z_{t-1} the current set of selected centers. Using Lemma A.7 and the previous cases we have that

$$\begin{aligned}
&\mathbb{E}\left[\phi_i(t)-\phi_i(t-1) \mid \{\tau_i(t)=\tau_i(t-1)+1, Z_{t-1}\}\right] \\
&\quad\le (2^i)^{1-2/\alpha}\cdot\frac{\sum_{C\in U_{t-1}^{(i)}}(\mathrm{cost}_{t-1}^{(\alpha)}(C))^{2/\alpha}}{|U_{t-1}^{(i)}|}\cdot\mathbb{P}[B_t \mid \{\tau_i(t)=\tau_i(t-1)+1, Z_{t-1}\}] \\
&\quad\le (2^i)^{1-2/\alpha}\cdot\frac{\sum_{C\in U_{t-1}^{(i)}}(\mathrm{cost}_{t-1}^{(\alpha)}(C))^{2/\alpha}}{|U_{t-1}^{(i)}|}\cdot\frac{\mathrm{cost}_{t-1}^{(\alpha)}(H_{t-1})}{\mathrm{cost}_{t-1}^{(\alpha)}(H_{t-1})+\mathrm{cost}_{t-1}^{(\alpha)}(U_{t-1}^{(i)})} \\
&\quad\le (2^i)^{1-2/\alpha}\cdot\frac{\sum_{C\in U_{t-1}^{(i)}}(\mathrm{cost}_{t-1}^{(\alpha)}(C))^{2/\alpha}}{|U_{t-1}^{(i)}|}\cdot\frac{\mathrm{cost}_{t-1}^{(\alpha)}(H_{t-1})}{\mathrm{cost}_{t-1}^{(\alpha)}(H_{t-1})+|U_{t-1}^{(i)}|\cdot\left(\frac{\sum_{C\in U_{t-1}^{(i)}}(\mathrm{cost}_{t-1}^{(\alpha)}(C))^{2/\alpha}}{|U_{t-1}^{(i)}|}\right)^{\alpha/2}},
\end{aligned}$$

where the last inequality uses Jensen's inequality. If we consider the last expression as a function of $X := \frac{\sum_{C\in U_{t-1}^{(i)}}(\mathrm{cost}_{t-1}^{(\alpha)}(C))^{2/\alpha}}{|U_{t-1}^{(i)}|}$ (the other quantities being fixed), one can see that this expression attains a maximum value for

$$X=\left(\frac{\mathrm{cost}_{t-1}^{(\alpha)}(H_{t-1})}{|U_{t-1}^{(i)}|\cdot(\alpha/2-1)}\right)^{2/\alpha}.$$

Plugging in this value, we obtain that

$$\begin{aligned}
&\mathbb{E}\left[\phi_i(t)-\phi_i(t-1) \mid \{\tau_i(t)=\tau_i(t-1)+1, Z_{t-1}\}\right] \\
&\quad\le\frac{(\alpha/2-1)^{1-2/\alpha}}{\alpha/2}\cdot(2^i)^{1-2/\alpha}\cdot\frac{(\mathrm{cost}_{t-1}^{(\alpha)}(H_{t-1}))^{2/\alpha}}{|U_{t-1}^{(i)}|^{2/\alpha}} \\
&\quad\le\frac{(\alpha/2-1)^{1-2/\alpha}}{\alpha/2}\cdot(2^i)^{1-2/\alpha}\cdot(k_i-\tau_i(t-1))^{-2/\alpha}\cdot(\mathrm{cost}_{t-1}^{(\alpha)}(H_t))^{2/\alpha},
\end{aligned}$$

where the second inequality uses the fact that $|U_{t-1}^{(i)}|\ge k_i-\tau_i(t-1)$. By the law of total expectation, we have that

$$\begin{aligned}
&\mathbb{E}\left[\phi_i(t)-\phi_i(t-1) \mid \{\tau_i(t)=\tau_i(t-1)+1\}\right]=\mathbb{E}_{Z_{t-1}}\left[\mathbb{E}\left[\phi_i(t)-\phi_i(t-1) \mid \{\tau_i(t)=\tau_i(t-1)+1, Z_{t-1}\}\right]\right] \\
&\quad\le\frac{(\alpha/2-1)^{1-2/\alpha}}{\alpha/2}\cdot(2^i)^{1-2/\alpha}\cdot(k_i-\tau_i(t-1))^{-2/\alpha}\cdot\mathbb{E}_{Z_{t-1}}\left[\left(\mathrm{cost}_{t-1}^{(\alpha)}(H_{t-1})\right)^{2/\alpha}\right] \\
&\quad\le\frac{(\alpha/2-1)^{1-2/\alpha}}{\alpha/2}\cdot(2^i)^{1-2/\alpha}\cdot(k_i-\tau_i(t-1))^{-2/\alpha}\cdot\left(\mathbb{E}_{Z_{t-1}}\left[\mathrm{cost}_{t-1}^{(\alpha)}(H_{t-1})\right]\right)^{2/\alpha} \\
&\quad\le\frac{(\alpha/2-1)^{1-2/\alpha}}{\alpha/2}\cdot(2^i)^{1-2/\alpha}\cdot(k_i-\tau_i(t-1))^{-2/\alpha}\cdot\left((2^{2\alpha}\hat{g_\alpha})\sum_{C\in\mathcal{C}_{\mathrm{OPT}}}|C|(\sigma_C)^\alpha\right)^{2/\alpha},
\end{aligned}$$

where the second inequality uses Jensen's inequality, and the last inequality uses Lemmas A.6, A.4, and Lemma A.3. Indeed, we clearly have the expected α-powered cost of a hit cluster is at most its expected cost after the first time it is hit, which, using Lemmas A.3 and A.4 is at most $2^{2\alpha}$ times $\mathrm{cost}^{(\alpha)}(C,\mu_C)$. □

A.3.2. The global increase

In this part, we are ready to bound the final expected value of $\phi(k)$ with the following lemma.

Lemma A.10. *For any $\alpha>2$, we have that*

$$\frac{\mathbb{E}\left[\phi(k)\right]}{OPT}\le f(\alpha)\cdot(\hat{g_\alpha})^{2/\alpha}\cdot\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{2-4/\alpha}\cdot\min\{\ell,\log(k)\}^{2/\alpha},$$

where $f(\alpha)=\frac{16(\alpha/2-1)^{1-2/\alpha}}{\alpha/2-1}\cdot\frac{2-2^{2/\alpha-1}}{1-2^{2/\alpha-1}}$.

Proof. Using Lemma A.9, we obtain that

$$\begin{aligned}
\mathbb{E}\left[\phi_i(k)\right] &\leq h(\alpha)\cdot\left(2^i\right)^{1-2/\alpha}\cdot\left((2^{2\alpha}\hat{g_\alpha})\sum_{C\in\mathcal{C}_{\texttt{OPT}}}|C|(\sigma_C)^\alpha\right)^{2/\alpha}\cdot\sum_{t=0}^{k_i-1}(k_i-t)^{-2/\alpha}\\
&\leq h(\alpha)\cdot\left(2^i\right)^{1-2/\alpha}\cdot\left((2^{2\alpha}\hat{g_\alpha})\sum_{C\in\mathcal{C}_{\texttt{OPT}}}|C|(\sigma_C)^\alpha\right)^{2/\alpha}\cdot\int_0^{k_i}(k_i-u)^{-2/\alpha}du\\
&\leq h(\alpha)\cdot\left(2^i\right)^{1-2/\alpha}\cdot\left((2^{2\alpha}\hat{g_\alpha})\sum_{C\in\mathcal{C}_{\texttt{OPT}}}|C|(\sigma_C)^\alpha\right)^{2/\alpha}\cdot\frac{(k_i)^{1-2/\alpha}}{1-2/\alpha}\,.
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
\mathbb{E}[\phi(k)]=\sum_{i\geq 0}\mathbb{E}[\phi_i(k)] &\leq \frac{h(\alpha)}{1-2/\alpha}\cdot\left((2^{2\alpha}\hat{g_\alpha})\sum_{C\in\mathcal{C}_{\texttt{OPT}}}|C|(\sigma_C)^\alpha\right)^{2/\alpha}\cdot\sum_{i\geq 0}(2^ik_i)^{1-2/\alpha}\\
&=\frac{8h(\alpha)(\hat{g_\alpha})^{2/\alpha}}{1-2/\alpha}\cdot\left(\sum_{C\in\mathcal{C}_{\texttt{OPT}}}|C|(\sigma_C)^\alpha\right)^{2/\alpha}\cdot\sum_{i\geq 0}(2^ik_i)^{1-2/\alpha}\,.
\end{aligned}$$

Comparing this with optimum clustering, we get using basic algebraic manipulations

$$\begin{aligned}
\frac{\mathbb{E}[\phi(k)]}{\texttt{OPT}} &= \frac{8h(\alpha)(\hat{g_\alpha})^{2/\alpha}}{1-2/\alpha}\cdot\frac{\left(\sum_C|C|(\sigma_C)^\alpha\right)^{2/\alpha}}{\sum_C|C|(\sigma_C)^2}\cdot\sum_{i\geq 0}(2^ik_i)^{1-2/\alpha}\\
&=\frac{8h(\alpha)(\hat{g_\alpha})^{2/\alpha}}{1-2/\alpha}\cdot\frac{\left(\sum_C|C|\left(\frac{\sigma_C}{\sigma_{\max}}\right)^\alpha\sigma_{\max}^\alpha\right)^{2/\alpha}}{\sum_C|C|\left(\frac{\sigma_C}{\sigma_{\min}}\right)^2\cdot\sigma_{\min}^2}\cdot\sum_{i\geq 0}(2^ik_i)^{1-2/\alpha}\\
&\leq\frac{8h(\alpha)(\hat{g_\alpha})^{2/\alpha}}{1-2/\alpha}\cdot\frac{\left(\sum_C|C|\left(\frac{\sigma_C}{\sigma_{\max}}\right)^2\right)^{2/\alpha}\cdot\sigma_{\max}^2}{\sum_C|C|\left(\frac{\sigma_C}{\sigma_{\min}}\right)^2\cdot\sigma_{\min}^2}\cdot\sum_{i\geq 0}(2^ik_i)^{1-2/\alpha}\\
&\leq\frac{8h(\alpha)(\hat{g_\alpha})^{2/\alpha}}{1-2/\alpha}\cdot\frac{\left(\sum_C|C|\left(\sigma_C\right)^2\right)^{2/\alpha}\cdot\sigma_{\max}^{2-4/\alpha}}{\sum_C|C|\left(\frac{\sigma_C}{\sigma_{\min}}\right)^2\cdot\sigma_{\min}^2}\cdot\sum_{i\geq 0}(2^ik_i)^{1-2/\alpha}\\
&\leq\frac{8h(\alpha)(\hat{g_\alpha})^{2/\alpha}}{1-2/\alpha}\cdot\frac{\left(\sum_C|C|\left(\frac{\sigma_C}{\sigma_{\min}}\right)^2\right)^{2/\alpha}\cdot(\sigma_{\min})^{4/\alpha}\cdot\sigma_{\max}^{2-4/\alpha}}{\sum_C|C|\left(\frac{\sigma_C}{\sigma_{\min}}\right)^2\cdot\sigma_{\min}^2}\cdot\sum_{i\geq 0}(2^ik_i)^{1-2/\alpha}\\
&\leq\frac{8h(\alpha)(\hat{g_\alpha})^{2/\alpha}}{1-2/\alpha}\cdot\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{2-4/\alpha}\cdot\frac{\sum_{i\geq 0}(2^ik_i)^{1-2/\alpha}}{\left(\sum_C|C|\right)^{1-2/\alpha}}\\
&\leq\frac{16h(\alpha)(\hat{g_\alpha})^{2/\alpha}}{1-2/\alpha}\cdot\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{2-4/\alpha}\cdot\frac{\sum_{i\geq 0}(2^ik_i)^{1-2/\alpha}}{\left(\sum_{i\geq 0}2^ik_i\right)^{1-2/\alpha}}\,,
\end{aligned}$$

where the fourth inequality is obtained using

$$\left(\sum_C|C|\left(\frac{\sigma_C}{\sigma_{\min}}\right)^2\right)^{2/\alpha-1}\leq\left(\sum_C|C|\right)^{2/\alpha-1}\,.$$

To conclude the proof, we need to upper bound the ratio

$$\frac{\sum_{i\geq 0}(2^i k_i)^{1-2/\alpha}}{\left(\sum_{i\geq 0} 2^i k_i\right)^{1-2/\alpha}}$$

for any integer sequence $(k_i)_{i\geq 0}$ satisfying the constraint $\sum_{i\geq 0} k_i = k$. Using Jensen's inequality (note that $x \mapsto x^{1-2/\alpha}$ is concave), we immediately obtain

$$\frac{\sum_{i\geq 0}(2^i k_i)^{1-2/\alpha}}{\left(\sum_{i\geq 0} 2^i k_i\right)^{1-2/\alpha}} \leq \ell^{2/\alpha} \,. \tag{15}$$

For the second upperbound, note that to have equality in Jensen's inequality, we must have that $2^i k_i = 2^j k_j$ for all $i, j \geq 0$ such that $k_i \neq 0, k_j \neq 0$. Intuitively, this can only happen when $\ell = O(\log k)$. Formally, let us denote by L the maximum $i \geq 0$ such that $k_L \neq 0$. We then build a decreasing sequence of indices as follows. We define i_1 to be equal to L. Then, assuming we defined the indices $i_1, i_2, \dots i_x$, we define i_{x+1} to be the highest index $0 \leq i < i_x$ such that $k_{i_{x+1}} \geq 2k_{i_x}$. We stop until it is not possible to find an index i_{x+1} fitting those conditions anymore. We denote by $\mathcal{I}$ the set of indices that were selected, and $\mathcal{J} := \mathbb{N} \setminus \mathcal{I}$ its complement.

We notice that, by construction, $|\mathcal{I}| \leq \log k$ since $\sum_{i\in\mathcal{I}} k_i \leq \sum_{i\geq 0} k_i = k$. Then clearly we have that

$$\frac{\sum_{i\geq 0}(2^i k_i)^{1-2/\alpha}}{\left(\sum_{i\geq 0} 2^i k_i\right)^{1-2/\alpha}} \leq \frac{\sum_{i\in\mathcal{I}}(2^i k_i)^{1-2/\alpha}}{\left(\sum_{i\in\mathcal{I}} 2^i k_i\right)^{1-2/\alpha}} + \frac{\sum_{i\in\mathcal{J}}(2^i k_i)^{1-2/\alpha}}{\left(\sum_{i\in\mathcal{I}} 2^i k_i\right)^{1-2/\alpha}} \,. \tag{16}$$

The first term on the right-hand side can be upperbounded by $|\mathcal{I}|^{2/\alpha} \leq (\log k)^{2/\alpha}$, using Jensen's inequality again. As for the second term, let us denote by $i_1, i_2, \dots, i_{|\mathcal{I}|}$ the set of indices selected to be in $\mathcal{I}$. Then, we write

$$\begin{aligned}
\sum_{i\in\mathcal{J}}(2^i k_i)^{1-2/\alpha} &= \sum_{x=1}^{|\mathcal{I}|} \sum_{j\in(i_{x+1},i_x)} (2^j k_j)^{1-2/\alpha} \\
&\leq \sum_{x=1}^{|\mathcal{I}|} \sum_{j\in(i_{x+1},i_x)} (2^j 2k_{i_x})^{1-2/\alpha} \\
&\leq \sum_{x=1}^{|\mathcal{I}|} \sum_{j\in(i_{x+1},i_x)} (2^{i_x} 2^{j-i_x} 2k_{i_x})^{1-2/\alpha} \\
&\leq \sum_{i\in\mathcal{I}}(2^i k_i)^{1-2/\alpha} \cdot \left(\sum_{j=0}^{+\infty} \left(2^{-i}\right)^{1-2/\alpha}\right) \\
&= \frac{\sum_{i\in\mathcal{I}}(2^i k_i)^{1-2/\alpha}}{1-2^{2/\alpha-1}} \,.
\end{aligned}$$

Therefore the second right-hand side term in Equation (16) can be upper bounded by

$$\frac{(\log k)^{2/\alpha}}{1-2^{2/\alpha-1}} \,,$$

which concludes the proof. □

A.4. Wrapping it up

Using Lemma A.10 in combination with Lemma A.2, we obtain that the final expected cost of undiscovered clusters is at most

$$h(\alpha) \cdot (\hat{g_\alpha})^{2/\alpha} \cdot \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{2-4/\alpha} \cdot \min\{\ell, \log(k)\}^{2/\alpha} \cdot \texttt{OPT} \,,$$

where $h(\alpha) := \frac{16(\alpha/2-1)^{1-2/\alpha}}{\alpha/2-1} \cdot \frac{2-2^{2/\alpha-1}}{1-2^{2/\alpha-1}}$. Regarding the cost of hit clusters, we use Lemma A.5 and Lemma A.4 to see that the expected cost of hit clusters is at most

$$(4e + (\alpha+1)^2 \cdot (\hat{g_\alpha})^{2/\alpha}) \cdot \texttt{OPT} \, .$$

Therefore, if we define $f(\alpha) := (4e + (\alpha+1)^2) \cdot \frac{16(\alpha/2-1)^{1-2/\alpha}}{\alpha/2-1} \cdot \frac{2-2^{2/\alpha-1}}{1-2^{2/\alpha-1}} = O\left(\frac{\alpha^2}{(\alpha/2-1)^{2/\alpha} \cdot (1-2^{2/\alpha-1})}\right)$, we obtain precisely the result of Theorem 2.2.

B. Additional Experiments

In this section, we provide additional experiments to further validate our claims. The code can be found in (D^α Seeding). Particularly, we run the D^α-seeding on the following instances:

1. 1) $\mathcal{D}_1$: A mixture of 4 Gaussians with the centers/means of the Gaussians placed on the corners of a square with edge length of $2\Delta = 100$ (default). All the covariances are identity matrices (in $d = 2$).
2. 2) $\mathcal{D}_2$: A mixture of 4 Gaussians with the centers/means of the Gaussians placed on the corners of a square with edge length of $2\Delta = 100$ (default). All the covariances are identity matrices (in $d = 2$), except one that has a covariance $\sigma^2 I$, with $\sigma^2 = 400$.
3. 3) $\mathcal{D}_3$: A mixture of 8 Gaussians with the centers/means of the Gaussians placed on the corners of a cube with edge length of $2\Delta = 100$ (default). All the covariances are identity matrices (in $d = 3$).
4. 4) $\mathcal{D}_4$: A mixture of 8 Gaussians with the centers/means of the Gaussians placed on the corners of a cube with edge length of $2\Delta = 100$ (default). All the covariances are identity matrices (in $d = 3$), except one that has a covariance $\sigma^2 I$, with $\sigma^2 = 800$.
5. 5) $\mathcal{D}_5$: A mixture of 4 bivariate student-t distributions (see Figure 5), with different degrees of freedom, $\nu = \{1.6, 2, 5, 10\}$, where the location parameter is centered on the corners of a square with edge length of $2\Delta = 100$ (default).

We generate these instance with $n = 1000$ samples from the appropriate mixture distribution. Then we consider $\alpha = \{2, 6, 10, \ldots, 38\}$, i.e, starting from 2 and increments of 4 for the Gaussian instances, i.e, $\mathcal{D}_1$ through $\mathcal{D}_4$.

In all experiments, we show the average performance over $N = 5000$ trial runs for each value of α. In some cases we also run Lloyd's algorithm until convergence after the seeding.

Figures 4 and 3, confirm our main claims that for Gaussians of the same/similar covariances, it is always best to choose α that is much larger than 2. This can be seen in both the $\mathcal{D}_1$ and $\mathcal{D}_3$. Although, the costs seem to remain a constant for $\alpha \geq 6$. However, the story is different when there is one Gaussian with a sufficiently different covariance, then one can see from the figures of $\mathcal{D}_2$ and $\mathcal{D}_4$ that there is a trade-off. Moreover, we can see that even if we run Lloyd's algorithm until convergence the general pattern does not change much. A small exception here is Figure 3, on the instance $\mathcal{D}_2$, we observe that our theoretical approximation of the seeding step with different variances must point to good values of α being smaller in order to minimize the effect of $(\frac{\sigma_{\max}}{\sigma_{\min}})^{2-4/\alpha}$. However, the seeds planted by larger values of α are able hit more of the optimum clusters, only that the chosen point could be sub-optimal and hence only a few steps of Lloyds is able to fix it. This might explain why Lloyds might be able to fix the issue in some cases where α is big. In contrast, when $\alpha = 2$ the seeding likely misses some optimum clusters and additional Lloyds steps do not offer much help in that case because it gets stuck in a local optimum.

Finally, with the experiments on student-t distribution, we intend to show the effect of g_α on the performance of D^α seeding. Here when the degrees of freedom (df) are $1 < df \leq 2$, the variance is ∞ (see (WikipediaT)). As seen in Figure 6, when $df = 1.6$, large values of α show a significant degradation in their performance. But when $df = 2$, $\alpha \geq 4$ performs better than $\alpha = 2$. In general, one might argue that if such a pattern is observed where $\alpha = 2$ performs better than $\alpha > 2$, it may be that the distributions could have many outliers. In this case, it might be that the k-means objective is probably not appropriate and one should consider more robust objectives such as k-medians. Now with $df = 5, 10$, where the distributions become more "Gaussian" like, we see a similar pattern as the one observed in Gaussian mixtures, where a large α is preferred. Crucially, we also notice that in this case, the number of Lloyd's steps required for convergence it much lower for $\alpha > 2$ in comparison to the number needed when $\alpha = 2$.

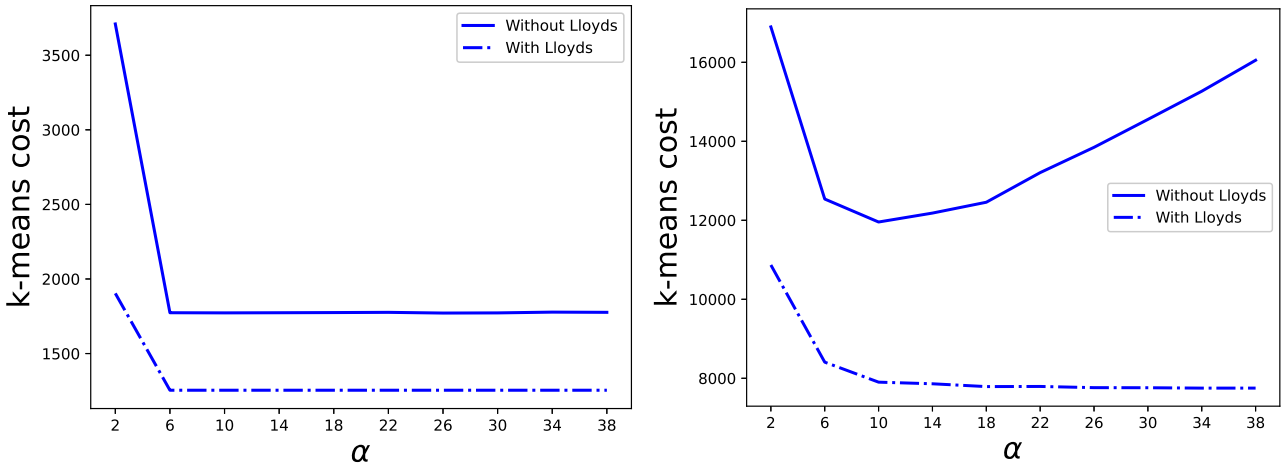

Figure 3. Performance of D^α seeding for two instances $\mathcal{D}_1$ (on the left) and $\mathcal{D}_2$ (on the right).

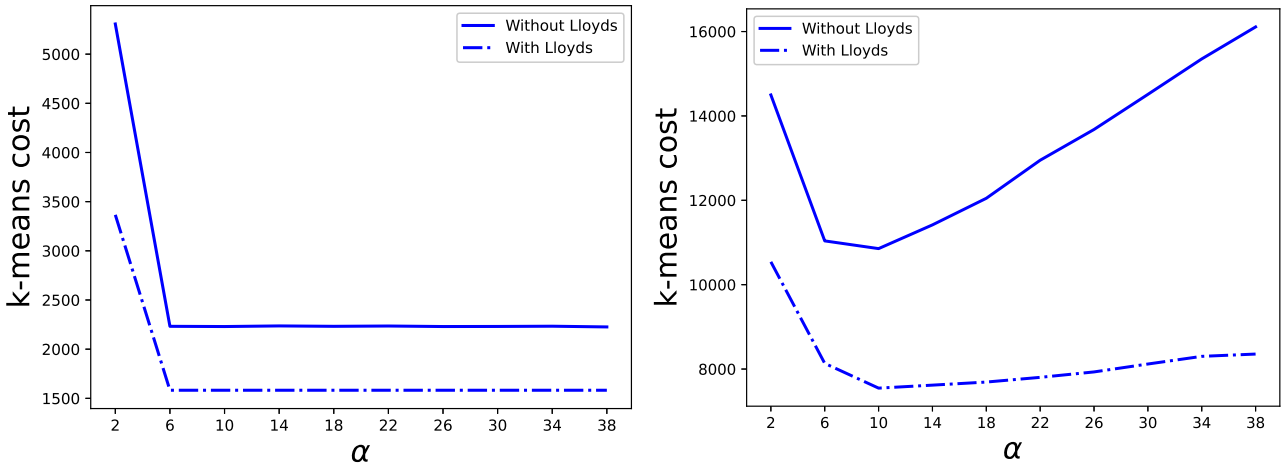

Figure 4. Performance of D^α seeding for two instances $\mathcal{D}_3$ (on the left) and $\mathcal{D}_4$ (on the right). The green curve indicates the cost right after the seeding step and the blue curve indicates the cost if we additionally run Lloyd's steps until convergence.

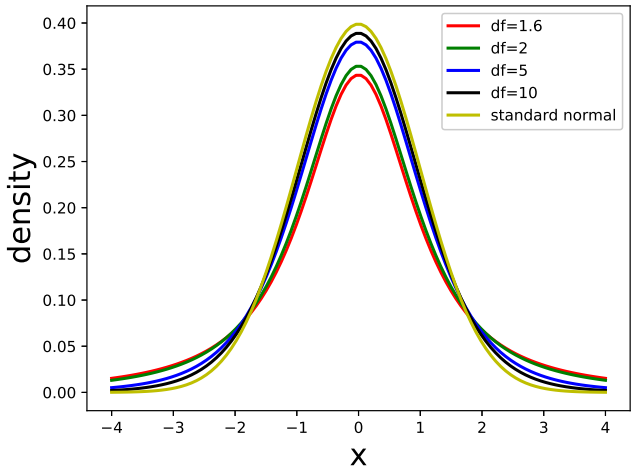

Figure 5. The pdf of univariate student-t distributions with different degrees of freedom.

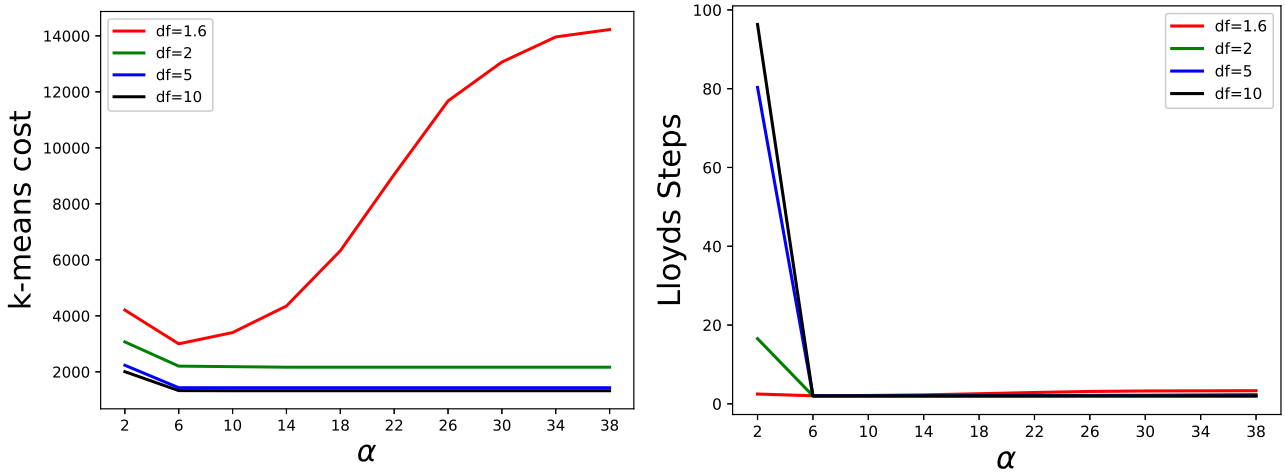

Figure 6. Performance of D^α seeding for the instance $\mathcal{D}_5$. The figure on the left shows the performance, and on the right is the number of Lloyd's steps needed until convergence.

C. Additional Discussion and Useful Inequalities

C.1. On potential functions of previous works

Here, we discuss more in details why a new potential function is needed when dealing with D^α seeding. Many of the previous works on `k-means++` rely on a more natural potential function. For instance in (Dasgupta, 2013), the potential function is defined as

$$\psi(t) = \frac{W_t}{|U_t|} \cdot \text{cost}^{(2)}(U_t, Z_t) \ ,$$

where U_t is the number of undiscovered clusters, and W_t is the number of iterations where a center was selected in an already discovered cluster. We notice two main differences with our potential function. The first one is that we partitioned the clusters by weight classes, and the second is that replaced the expression

$$\text{cost}^{(2)}(U_t^{(i)}, Z_t)$$

by a more involved

$$(2^i)^{1-2/\alpha} \cdot \sum_{C \in U_t^{(i)}} (\mathrm{cost}^{(\alpha)}(C, Z_t))^{2/\alpha} .$$

Let us discuss the second difference first, and for the purpose of simplicity let us assume there is a unique weight class. Then it is very tempting to use the same potential function as (Dasgupta, 2013). Unfortunately, we run into issues when the next center is selected in an undiscovered cluster.

Indeed, in that case it is quite crucial in the proof that the potential should decrease in expectation, and an increase in potential here seems like a quite fatal flaw. To show that the potential decreases in the case of D^2 seeding, one simply notices that if we had picked one of the undiscovered clusters uniformly at random and removed it from $\mathrm{cost}^{(2)}(U_t, Z_t)$, then the potential would already decrease. Now, D^2 seeding can only do better than that, since heavy clusters have even more chance of being sampled. But what about D^α seeding? Unfortunately, this is not true anymore, as it can be that there are two clusters C_1, C_2 such that $\mathrm{cost}^{(2)}(C_1, Z_t) > \mathrm{cost}^{(2)}(C_2, Z_t)$, but $\mathrm{cost}^{(\alpha)}(C_1, Z_t) < \mathrm{cost}^{(\alpha)}(C_2, Z_t)$, i.e. the orders are reversed. And this can happen even if g_α is fairly small.

Therefore, one needs a potential where the undiscovered clusters are ranked in the same order of importance as D^α seeding weights them. Of course, this potential should also scale as a square of the euclidean distance. This is where our potential comes in.

Now about the first difference, note that because of scaling issues we had to add an extra $n^{1-2/\alpha}$ if all the clusters are of weight n. If the clusters have different weight, the idea of partitioning them using geometric grouping comes naturally.

C.2. Guarantees for finite sample mixture of Gaussians

By assumption we have $\mathcal{X} \sim \sum_{i=1}^k w_i \mathcal{N}(\mu_i, \Sigma_i)$, with $\max_{i,j \in [k]} \mathrm{tr}(\Sigma_i)/\mathrm{tr}(\Sigma_j) = O(1)$ and $w_i > 0$ for all $i \in [k]$ (w.l.o.g.). In the infinite sample case (where the sums are replaced by expectations) we have that $C_i = \{x \in \mathcal{X} : x \sim \mathcal{N}(\mu_i, \Sigma_i)\}$ is the optimal clustering (with infinite samples), i.e, the true centers of the Gaussians are the parameters that maximize the (population) likelihood. Below we show for a mixture of gaussian distributions that $g_\alpha = O(\alpha^{\alpha/2})$ which implies that in the case of arbitrary weights the worst-case approximation is $O(\alpha^2 (g_\alpha)^{2/\alpha} \log^{2/\alpha} k)$, thus, when we set $\alpha = \Theta(\log\log(k))$, we have an $O(\log\log k)^3$-approximation. When $\ell = O(1)$, we get a $O(\alpha^3)$ approximation which is a constant for α being a constant. Hence in the infinite sample limit the result holds. Now we use the quantities that are computed above (when considering infinite samples) and we show that these quantities do not change much with high probability when sufficiently many samples are drawn from the Gaussian mixture.

Consider the reference clustering $C_i = \{x \in \mathcal{X} : x \sim \mathcal{N}(\mu_i, \Sigma_i)\}$, which is optimal in the case of infinite samples. We let $|\mathcal{X}| = N$ and we consider first the case when all the covariances Σ_i are full rank. Now we have $\mathbb{E}[|C_i|] = N w_i$ and we let $w_{\min} := \min_{i \in [k]} w_i$.

We can define the "sampled" version of centroids and the standard deviation of each cluster. That is, $\hat{\mu}_i := \frac{1}{|C_i|} \sum_{x \in C_i} x$, and

$$\hat{r_i} := \sqrt{\frac{\sum_{x \in C_i} \|x - \hat{\mu}_i\|^2}{|C_i|}} .$$

To ensure that the above definition makes sense, we will require sufficiently many samples w.r.t $w_{\min}$. Also, we will effectively replace $\hat{\mu_i}$ by μ_i in the above equation. Thus, we will condition on the following events being true.

1. $E_1 = \{$Each reference cluster i has $\Theta(N w_i)$ points$\}$.
2. $E_2 = \{\|\Sigma_i^{-1/2} (\hat{\mu}_i - \mu_i)\| < \varepsilon, \text{ for some arbitrary } \varepsilon > 0, \text{ for each } i \in [k]\}$.

We remark that if $w_i = \Theta(1/k)$ for all i, then we have as result of E_1, that $\ell = O(1)$ w.h.p, in which case we get $O(\alpha^3)$ approximation.

Continuing our proof, since we have a multivariate Gaussian distribution, it holds that, $\mathbb{E}\left[\|X_i - \mu_i\|^2\right] = \mathrm{tr}(\Sigma_i)$, where $\mathrm{tr}(.)$ represents the trace of a matrix.

$$\mathrm{tr}(\Sigma_i) = \sum_{j=1}^{d} \lambda_{ij},$$

such that for each $i \in [k]$, we have, $\lambda_{i1} \geq \lambda_{i2} \geq \ldots \geq \lambda_{id} \geq 0$. Without loss of generality let λ_{i1} be the largest eigenvalue for Σ_i. Note that the covariance matrix is always positive semi-definite. Furthermore, when the covariance matrix is full-rank, the eigenvalues are positive. Now, we have to characterize g_α. Using the definition in (8), in the expected sense, one has to compute:

$$g^{(i)}(\alpha) := \frac{\mathbb{E}_{X_i,X_j}\left[\|X_i - X_j\|^\alpha\right]}{\left(\mathbb{E}\left[\|X_i - \mu_i\|^2\right]\right)^{\alpha/2}}.$$

We also note that,

$$2\mathbb{E}\left[\|X_i - \mu_i\|^\alpha\right] \leq \mathbb{E}_{X_i,X_j}\left[\|X_i - X_j\|^\alpha\right] \leq 2^{\alpha+1}\mathbb{E}\left[\|X_i - \mu_i\|^\alpha\right] \tag{17}$$

Thus instead we can re-define,

$$\tilde{g}^{(i)}(\alpha) := \frac{\mathbb{E}\left[\|X_i - \mu_i\|^\alpha\right]}{\left(\mathbb{E}\left[\|X_i - \mu_i\|^2\right]\right)^{\alpha/2}}.$$

Here, the denominator is $\left(\mathbb{E}\left[\|X_i - \mu_i\|^2\right]\right)^{\alpha/2} = (\text{tr}(\Sigma_i))^{\alpha/2}$. Now consider the transformation, $Y_i := \Sigma_i^{-1/2}(X_i - \mu_i)$. Now Y_i is a standard multivariate Gaussian, i.e, $Y_i \sim \mathcal{N}(0, I_d)$, where is the I_d is the identity matrix is d dimensions. Then, $\mathbb{E}_{\mathcal{N}(\mu_i,\Sigma_i)}\left[\|X_i - \mu_i\|^\alpha\right] = \mathbb{E}_{\mathcal{N}(0,I_d)}\left[\|\Sigma_i^{1/2}Y_i\|^\alpha\right] \leq \|\Sigma_i^{1/2}\|^\alpha \mathbb{E}_{\mathcal{N}(0,I_d)}\left[\|Y_i\|^\alpha\right]$. Here, we have:

$$\begin{aligned}
\mathbb{E}_{\mathcal{N}(0,I_d)}\left[\|Y_i\|^\alpha\right] &= \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\|y\|^\alpha \exp\left(-\frac{\|y\|^2}{2}\right)dy \\
&= \frac{\Gamma(\frac{(\alpha+d)}{2})}{2^{d/2}\Gamma(\frac{d}{2})}. \\
&\leq (\alpha/2)^{\alpha/2}2^{\alpha/2}\frac{(d/2)^{\alpha/2}}{2^{d/2}} \\
&\leq \frac{\alpha^{\alpha/2}}{2^{\alpha/2}}\,.
\end{aligned}$$

From this we can compute,

$$\begin{aligned}
\tilde{g}^{(i)}(\alpha) &\leq \frac{\|\Sigma_i^{1/2}\|^\alpha \alpha^{\alpha/2}}{4(\text{tr}(\Sigma_i))^{\alpha/2}} \\
&= \frac{(\lambda_{i1})^{\alpha/2}}{(\text{tr}(\Sigma_i))^{\alpha/2}}\frac{\alpha^{\alpha/2}}{4} \\
&\leq \frac{\alpha^{\alpha/2}}{2^{\alpha/2}} \\
&= O((\alpha/2)^{\alpha/2})\,.
\end{aligned}$$

Thus, the quantity g_α as defined in (8) in its expected form is at most $O((2\alpha)^{\alpha/2})$, by using the relation in (17). Now, we move on to the concentration bounds for the following value:

$$\hat{r_i} = \sqrt{\frac{\sum_{x\in C_i}\|x - \mu_i\|^2}{|C_i|}}\,.$$

First, we have i.i.d samples from the mixture distribution. Consider the random variable $\|x - \mu_i\|^2$, with $x \in C_i$, the expectation of this random variable is $\mathbb{E}\left[\|X_i - \mu_i\|^2\right]$. Then using Chebyshev's inequality, we have

$$\Pr\left(|\hat{r_i}^2 - \mathbb{E}\left[\|X_i - \mu_i\|^2\right]| \geq \varepsilon\right) \leq \frac{\text{Var}\left(\|X_i - \mu_i\|^2\right)}{|C_i|\varepsilon^2}\,. \tag{18}$$

Now, we can set $\varepsilon = \frac{\mathbb{E}\left[\|X_i - \mu_i\|^2\right]}{2}$, then the concentration bound becomes,

$$\begin{aligned}
\Pr\left(\left|\hat{r_i}^2 - \mathbb{E}\left[\|X_i - \mu_i\|^2\right]\right| \geq \frac{\mathbb{E}\left[\|X_i - \mu_i\|^2\right]}{2}\right) &\leq \frac{4\tilde{g}_i(4)(\mathbb{E}\left[\|X_i - \mu_i\|^2\right])^2 - 4\mathbb{E}\left[\|X_i - \mu_i\|^2\right])^2}{|C_i|(\mathbb{E}\left[\|X_i - \mu_i\|^2\right])^2} \\
&\leq \frac{O(1)}{|C_i|}\,.
\end{aligned}$$

Thus, we have w.h.p , $\hat{r_i^2} = \Theta(r_i^2) = \Theta(\text{tr}(\Sigma_i))$, for all $i \in [k]$. This holds when $N = \Omega(k^2/w_{\min}^2)$, then the success probability is $1 - O(1/\sqrt{N})$. Now, we can condition on this event, to get a bound on g_α. Particularly, we must look at the α moment, i.e, we must use the value of $\tilde{g}^{(i)}(\alpha)$. Now the plan is to show that $\tilde{g}_\alpha$ as defined is a constant w.h.p. For this, it suffices to show that the following quantity

$$\hat{r_i}^{(\alpha)} := \frac{\sum_{x \in C_i} \|x - \mu_i\|^\alpha}{|C_i|},$$

is $\Theta(\mathbb{E}\left[\|X_i - \mu_i\|^\alpha\right])$ w.h.p. Assuming this is true, then we have $g_\alpha \leq 2^\alpha \hat{r_i}^{(\alpha)} \leq \Theta(\mathbb{E}\left[\|X_i - \mu_i\|^\alpha\right]) = O((2\alpha)^{\alpha/2})$. This is also due to the fact that $\hat{r_i^2} = \Theta(r_i^2)$. Now, to see how the concentration holds, we again use Chebyshev's inequality, when the random variable is $\|x - \mu_i\|^\alpha$ for $x \in C_i$, again with the expectation being $\mathbb{E}\left[\|x - \mu_i\|^\alpha\right]$. Repeating, the arguments, from the previous case, i.e, using Equation (18) for the aforementioned random variable,

$$\begin{aligned}\Pr\left(\left|\hat{r_i}^{(\alpha)} - \mathbb{E}\left[\|X_i - \mu_i\|^\alpha\right]\right| \geq \frac{\mathbb{E}\left[\|X_i - \mu_i\|^\alpha\right]}{2}\right) &\leq \frac{4\tilde{g}^{(i)}(2\alpha)(\mathbb{E}\left[\|X_i - \mu_i\|^\alpha\right])^2}{|C_i|(\mathbb{E}\left[\|X_i - \mu_i\|^\alpha\right])^2} \\ &\quad - \frac{4\mathbb{E}\left[\|X_i - \mu_i\|^\alpha\right])^2}{|C_i|(\mathbb{E}\left[\|X_i - \mu_i\|^\alpha\right])^2} \\ &\leq \frac{O_\alpha(1)}{|C_i|}.\end{aligned}$$

Thus, we have that $g_\alpha = O_\alpha(1)$, w.h.p, when $N = \Theta_\alpha(k^2/w_{\min}^2)$.

Events E_1 and E_2 occur with high probability: To end this proof, we show that events E_1 and E_2 occur with high probability as well. When we have a mixture of distributions, we have the number of points sampled from each distribution $(|C_1|, |C_2|, \ldots, |C_k|)$ is distributed as multinomial distribution with parameters $(N, \{w_i\})$. When we have k balanced distributions, $p_i = 1/k$. Now, we can use the following concentration bound called the Bretagnolle-Huber-Carol inequality (van der Vaart & Wellner, 1996) (Proposition A.6.6), which states that:

$$\Pr(\sum_{i=1}^{k} ||C_i| - Nw_i|) \geq 2\sqrt{N}\delta) \leq 2^k \exp(-2\delta^2), \quad \delta > 0 \tag{19}$$

By setting $\delta = \sqrt{\frac{2N}{Nk+1}}$, we have that for $N = \Omega(k)$, event E_1 holds w.h.p.

Now, coming to E_2, let us focus on each Gaussian separately and let us condition on event E_1. For Gaussian i, we have that with probability $1 - \delta$, $\|\Sigma_i^{-1/2}(\hat{\mu_i} - \mu_i)\| < \varepsilon/k$ with $N = \Omega(\frac{k^3 d}{\varepsilon^2}\log^2(1/\delta))$ samples. Hence, by choosing a small enough (δ, ε), we can guarantee that with sufficiently many samples $N = \tilde{\Omega}(\text{poly}(k, d, \frac{1}{\min_{i\in[k]}\lambda_{id}}))$ [4], E_1, holds with high probability. Now E_2, also implies that the "reference" centers/clusters that we picked in the beginning are in fact close to the optimal centers with high probability.

Degenerate Covariance: If the covariance matrix is not full rank, we can still define the Gaussian distribution in the appropriate affine subspace $\mathbb{R}^{d'}$, where $d' \leq d$, and $\text{rank}(\Sigma_i) = d'$. We can repeat the previous analysis, as our analysis is done for each cluster separately and obtain the same conclusion.

When we have $N = \tilde{\Omega}(\text{poly}(k, d, \frac{1}{\min_{i\in[k]}\lambda_{id}}, \frac{1}{w_{\min}}))$, we get the desired finite sample guarantee.

C.3. Useful inequalities

In our proof, we use several times the following inequalities, which we give here for completeness.

Lemma C.1 (Jensen's inequality (Hardy et al., 1952))**.** *For any set of non-negative real numbers $\{y_i\}_{i=1}^n$ it holds that:*

$$\sum_{i=1}^{n} y_i^\gamma \geq n^{1-\gamma}\left(\sum_{i=1}^{n} y_i\right)^\gamma, \tag{20}$$

if $\gamma \geq 1$. If $0 \leq \gamma < 1$, the reversed inequality holds.

[4] Here, $\tilde{\Omega}$ hides log factors

Lemma C.2 (Chebyshev's sum inequality (Hardy et al., 1952)). *For any sequence of real numbers $\{y_i\}_{i=1}^n$ and $\{w_i\}_{i=1}^n$, such that $y_1 \geq y_2 \geq \ldots \geq y_n$ and $w_1 \geq w_2 \geq \ldots \geq w_n$, it holds that:*

$$\frac{1}{n}\sum_{i=1}^{n} y_i w_i \geq \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) \cdot \left(\frac{1}{n}\sum_{i=1}^{n} w_i\right). \tag{21}$$

D. Lower bounds

D.1. The dependence on $\sigma_{\max}/\sigma_{\min}$ is tight

We consider an instance with one cluster of n points arranged on a regular simplex of side-length R, while the other $k-1$ clusters contain n points arranged on a regular simplex of side-length 1. Note that for a cluster made of a regular simplex, we have that $g_\alpha \leq 2^\alpha$, hence $(g_\alpha)^{2/\alpha} = O(1)$, and the concentration of clusters will not play any role in this construction.

The clusters of radius 1 are pairwise separated by a distance δ, while the cluster of radius R is at infinite distance from all the other clusters. We choose R and δ so that $R^\alpha = 10 \cdot k\delta^\alpha$. Then, at each iteration $t \geq 3$, the probability of sampling from the cluster of radius R is at least

$$\frac{nR^\alpha}{nk\delta^\alpha + nR^\alpha} \geq 1/2.$$

Hence, the expected number of missed clusters is at least $k/3$, which implies that the expected cost of the returned solution is at least $(nk/3)\cdot\delta^2$ while OPT costs at most $nR^2 + nk$. We select δ so that $R^2 = k$ which is satisfied for $\delta = k^{1/2-1/\alpha}$. In this case, the expected competitive ratio is at least $\Omega(k^{1-2/\alpha})$, and one can check that $\sigma_{\max} = R = k^{1/2}$, and $\sigma_{\min} = 1$.

D.2. The dependence on g_α is tight

Consider the following instance. There are two clusters of n points each. The first clusters C_1 consists of n points on a regular simplex. All points in C_1 are at distance 1 from each other, and at distance 1 from the centroid of C_1. The second cluster has 2 groups of points. There are $n-1$ points at the origin, and 1 point at coordinates $(\Delta, 0, \ldots, 0)$. We place the cluster C_1 so that its centroid lies at coordinates $(-\delta, 0, \ldots, 0)$, with $\delta = \Delta/n^{1/\alpha}$.

Second, we select Δ so that $\sigma_{\max}/\sigma_{\min} = 1$. For this, we would simply need to set Δ so that

$$(\Delta/n)^2 \cdot (n-1) + \Delta^2(1-1/n)^2 = n$$

which gives $\Delta = \Theta(\sqrt{n})$. The cost of the given clustering is equal to $2n$.

However, we can see that g_α for the second cluster is not a constant. Indeed we can compute that

$$g_\alpha = \frac{(1/n)\cdot\sum_{z\in C_2}\mathrm{cost}^{(\alpha)}(C_1, z)}{n^{1-\alpha/2}\cdot(\mathrm{cost}^{(2)}(C_2, \mu_{C_2}))^{\alpha/2}} \tag{22}$$

$$= \frac{(1/n)\cdot(2(n-1)\Delta^\alpha)}{n} \tag{23}$$

$$= \Theta\left(n^{\alpha/2-1}\right). \tag{24}$$

To analyze the D^α seeding procedure on this instance, let us denote by $\mathcal{B}_1$ the event that the first sampled center belongs to C_1, and by $\mathcal{B}_2$ the event that the second sample belongs to C_2 and lies at coordinate $(\Delta, 0, \ldots, 0)$.

Clearly we have that

$$\mathbb{P}[\mathcal{B}_1] = 1/2\,,$$

and

$$\mathbb{P}[\mathcal{B}_2 \mid \mathcal{B}_1] \geq \frac{\Delta^\alpha}{n + (n-1)\cdot\delta^\alpha + \Delta^\alpha} \geq 1/3\,,$$

by our choice of δ. This implies that $\mathbb{P}[\mathcal{B}_1 \cap \mathcal{B}_2] \geq 1/6$ and the expected cost of the output clustering is at least

$$\frac{1}{6}\cdot(n-1)\cdot\delta^2 = \Theta(n^{1+1-2/\alpha}) = \Omega(n^{1-2/\alpha}\cdot\texttt{OPT}) = \Omega((g_\alpha)^{2/\alpha})\cdot\texttt{OPT}\,.$$

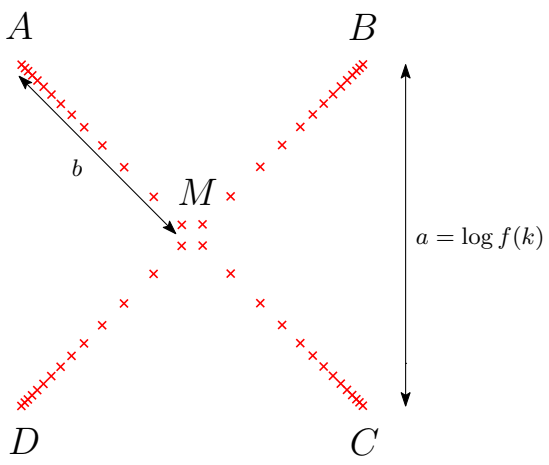

Figure 7. A schematic view of a group of 4 clusters. The densest parts are at the vertices A, B, C, D. All the points in a cluster lie on the segment from a vertex to the center of mass of the square.

D.3. A lower bound for greedy `k-means++`

In this section, we prove Theorem 2.4. In particular, it implies a super constant lower bound for the greedy variant which uses $f(k)$ samples, as long as $f(k)$ is super constant. We emphasize that here we did not try to optimize the lower bound. The main purpose of this section is to show that the greedy algorithm with a superconstant number of samples cannot guarantee a constant factor in expectation. Consider the following construction against greedy k-means++ with $f(k)$ samples. We will place our points in the euclidean space. For ease of construction, we will take the distributional point of view in this instance, i.e. we assume that each cluster C_i contains infinitely many points placed in $\mathbb{R}^d$ according to some distribution f_i. Moreover, we will assume that these clusters are balanced (i.e. they have the same weight) which means here that the global distribution of all points in all the clusters has probability density $f(x) = \frac{1}{k} \cdot \sum_{i \in [k]} f_i(x)$ for all $x \in \mathbb{R}^2$.

In our instance, we will create $k/4$ groups of 4 clusters as follows. First, we describe how to construct one group. We create an square $ABCD$ of side length $a = \log(f(k))$ and we denote by M the centroid of this square, and by $b = a/\sqrt{2}$ the length of half a diagonal in the square. In the cluster C_1, the probability density will be a one-dimensional density on the segment $[AM]$ defined as follows

$$f_1(x) = \begin{cases} \frac{e^{-||x-A||_2}}{1-e^{-b}} \text{ if } x \in [AM] \text{ and} \\ 0 \text{ otherwise.} \end{cases}$$

We recall here that $AM = b$ so the above function is indeed a probability density. The centroid of the first cluster will be equal to the point $\mu_1 \in [AM]$ such that

$$A\mu_1 = \frac{\int_0^b x e^{-x} dx}{1 - e^{-b}} = 1 - \frac{b}{e^b - 1} \, .$$

We repeat the construction similarly for the clusters C_2, C_3, and C_4, replacing A by B, C, and D respectively. Note that we needed only 2 dimensions to construct our first group. We refer the reader to Figure 7 for an illustation of the construction.

Then, we add $k/4 - 1$ extra dimensions and we fix the point M to be the origin of $\mathbb{R}^{k/4+1}$. Finally, we make $k/4 - 1$ additional copies $T_2, T_3, \ldots, T_{k/4}$ of our first square and we arrange their centroids $M_1 = (0)_{k/4+1}, M_2, \ldots, M_{k/4}$ in a regular simplex of side-length $\Delta = 100 \cdot (f(k))^3 \cdot k$. One should think of Δ as a much bigger distance than a (essentially, we have $k/4$ squares that are infinitely far from each other). Each group of 4 cluster will be denoted G_j for $j \in [k/4]$.

We prove the following simple statements.

Claim D.1. For any cluster C_i, we have that

$$\lim_{a \to \infty} \text{cost}^{(2)}(C_i) = 1.$$

Proof. We can write

$$\begin{aligned}\text{cost}^{(2)}(C_i) &= \frac{\int_0^b (x - A\mu_i)^2 e^{-x} dx}{1 - e^{-b}} \\ &= \frac{e^b + e^{-b} - b^2 - 2}{(e^b - 1) \cdot (1 - e^{-b})} ,\end{aligned}$$

which clearly tends to 1 as b goes to infinity. □

Claim D.2. For any cluster C_i and any fixed $\alpha \geq 2$ and $b > 10$, we have that

$$g_\alpha = \frac{\int_0^b \int_0^b |x - y|^\alpha f_i(x) f_i(y) dx dy}{\left(\int_0^b (x - A\mu_i)^2 f_i(x) dx\right)^{\alpha/2}} \leq 4 \cdot \Gamma(\alpha + 1) , \tag{25}$$

where Γ is the gamma function (Artin, 2015).

Proof. From Claim D.1, we already have that the denominator tends to 1 as a goes to infinity. Hence, we only need to upper bound the numerator. We proceed with the simple change of variable $u = (x + y)$ and $v = (x - y)$. Then we obtain

$$\begin{aligned}\int_0^b \int_0^b |x - y|^\alpha f_i(x) f_i(y) dx dy &= \frac{1}{2(1 - e^{-b})^2} \cdot \int_{-b}^b \int_{|v|}^{2b - |v|} |v|^\alpha e^{-u} du dv \\ &= \frac{1}{(1 - e^{-b})^2} \cdot \int_0^b \int_v^{2b - v} v^\alpha e^{-u} du dv \\ &= \frac{1}{(1 - e^{-b})^2} \cdot \int_0^b v^\alpha (e^{-v} - e^{v - 2b}) du dv \\ &\leq 2 \int_0^\infty v^\alpha e^{-v} dv \\ &= 2\Gamma(\alpha + 1) .\end{aligned}$$

Using the fact that we consider $b \geq 10$, we obtain the desired upperbound. □

In light of Claim D.2, it is is clear that for any fixed $\alpha > 2$, the D^α seeding algorithm will guarantee a constant factor approximation in expectation.

However, we use the instances we build to prove Theorem 2.4. To this end, we will need to look at the cost of a group of 4 clusters after the greedy k-means++ algorithm selected exactly one center inside a group j. We distinguish two key cases: (1) If the first center x in the cluster $C_i \in G_j$ such that the distance between x and the centroid μ_i of the cluster is less than $\log(b)$, and (2) if the first center is at distance at most $b/100$ from the centroid M_j of the corresponding square. Let us denote by K_1 and K_2 the total cost of the 4 clusters in G_j after case (1) or (2) happened respectively.

Claim D.3. For k a big enough constant, we have that

$$K_1 \geq 11b^2/2 ,$$

and

$$K_2 \leq 5b^2 .$$

Proof. Assume w.l.o.g. that the first center belongs to center C_1. In the first case then the total cost is lower bounded by the cost of the other three clusters C_2, C_3, C_4. Assume C_3 is the center whose centroid μ_3 is the furthest away from μ_1. Then the cost of this center is lower bounded by

$$\int_0^b (2b - \log(b) - x)^2 e^{-x} dx \geq 4b^2 + O(b \log b) .$$

For the other two clusters C_2 and C_4, we use Pythagore theorem. Since they lie on a perpendicular line to C_1 and we are in case (1), the distance of any point in C_2 or C_4 is at distance at least $b - \log b$ from the sampled center. Hence we can lower bound the cost of each of the other two clusters with the following quantity

$$\int_0^b (b - \log(b))^2 f(x)dx \geq b^2 + O(b \log b) \ .$$

Hence the total cost is at least

$$6b^2 + O(b \log b)$$

which is more than $11b^2/2$ for b big enough.

In the second case, the cost of each cluster is upperbounded by

$$\int_0^b (1.01b - x)^2 f(x)dx \leq 1.03b^2 + O(b) \ .$$

Hence the total cost is upperbounded by

$$4.12b^2 + O(b) \ ,$$

which is less than $5b^2$ for b big enough. □

Next, we define as $\mathcal{L}$ the event that the greedy k-means++ algorithm samples one of its $f(k)$ samples at distance at most $b/100$ from the centroid M_j of some still undiscovered group G_j.

Claim D.4. Assume we are at iteration $t < k/4$. Then

$$\mathbb{P}[\mathcal{L}] \geq 1/2.$$

Proof. At iteration $t < k/4$, it must be that one group G_j is still undiscovered. Then we have that (by triangle inequality)

$$\text{cost}^{(2)}(G_j) \geq 4(\Delta - 2a)^2 > \Delta^2 \ ,$$

for k a big enough constant. The total cost of the discovered clusters is at most

$$k \cdot (2a)^2 \leq k(2 \log f(k))^2 \leq \text{cost}^{(2)}(G_j)/(10 f(k)) \ .$$

By union-bound we obtain that with probability at least $9/10$, none of the sampled candidate centers belong to an already discovered group. Conditionned on that event, we can simply compute by triangle inequality that the probability that a specific sample is at distance at most $b/100$ from the centroid is at least (assuming b and $f(k)$ are some big enough constant)

$$\frac{9}{10} \cdot \left(\int_{99b/100}^b \frac{e^{-x}}{1 - e^{-b}} dx \right) \geq \frac{9}{10} \cdot (e^{-99b/100} - e^{-b}) \geq \frac{9}{10} e^{-99b/100} \geq \frac{9}{10(f(k))^{1/\sqrt{2}}} \geq \frac{1}{f(k)} \ .$$

Therefore, with probability at most $(1 - 1/f(k))^{f(k)} \leq 1/e$, we have that none of the sampled candidate centers are close the centroid M_j of a square. Therefore with probability $1/2$ at least one of them is close to a centroid M_j of an undiscovered group. □

When the event $\mathcal{L}$ happens, it must be (by Claim D.3) that the selected center is at distance at least $\log b$ from the center it belongs to. By Claim D.4, this happens in expectation at least $k/8$ times. This means that the expected cost of the output clustering is at least

$$(k/8) \cdot \int_0^{\log b} (x - \log b)^2 e^{-x} dx \geq (k/8) \cdot ((\log b)^2 + O(\log b)) = \Omega((k/8) \cdot (\log \log(f(k)))^2 \ .$$

Since by Claim D.1 we have that the optimum cost is equivalent to k (for k growing to infinity), this concludes the proof of Theorem 2.4.