58th Annual IEEE Symposium on Foundations of Computer Science

On the Local Structure of Stable Clustering Instances

Vincent Cohen-Addad
University of Copenhagen, Denmark

Chris Schwiegelshohn
Sapienza, University of Rome, Italy

***Abstract*—We study the classic k-median and k-means clustering objectives in the *beyond-worst-case* scenario. We consider three well-studied notions of structured data that aim at characterizing real-world inputs:**

- **Distribution Stability (introduced by Awasthi, Blum, and Sheffet, FOCS 2010)**
- **Spectral Separability (introduced by Kumar and Kannan, FOCS 2010)**
- **Perturbation Resilience (introduced by Bilu and Linial, ICS 2010)**

We prove structural results showing that inputs satisfying at least one of the conditions are inherently "local". Namely, for any such input, any local optimum is close both in term of structure and in term of objective value to the global optima.

As a corollary we obtain that the widely-used Local Search algorithm has strong performance guarantees for both the tasks of recovering the underlying optimal clustering and obtaining a clustering of small cost. This is a significant step toward understanding the success of local search heuristics in clustering applications.

I. Introduction

Clustering is a fundamental, routinely-used approach to extract information from datasets. Given a dataset and the most important features of the data, a clustering is a partition of the data such that data elements in the same part have common features. The problem of computing a clustering has received a considerable amount of attention in both practice and theory.

The variety of contexts in which clustering problems arise makes the problem of computing a "good" clustering hard to define formally. From a theoretician's perspective, clustering problems are often modeled by an objective function we wish to optimize (*e.g.,* the famous k-median or k-means objective functions). This *modeling step* is both needed and crucial since it provides a framework to quantitatively compare algorithms. Unfortunately, the most popular objectives for clustering, like the k-median and k-means objectives, are hard to approximate, even when restricted to Euclidean spaces.

This view is generally not shared by practitioners. Indeed, clustering is often used as a preprocessing step to simplify and speed up subsequent analysis, even if this analysis admits polynomial time algorithms. If the clustering itself is of independent interest, there are many heuristics with good running times and results on real-world inputs.

This induces a gap between theory and practice. On the one hand, the algorithms that are efficient in practice cannot be proven to achieve good approximation to the k-median and k-means objectives in the worst-case. Since approximation ratios are one of the main methods to evaluate algorithms, theory predicts that determining a good clustering is a difficult task. On the other hand, the best theoretical algorithms turn out to be noncompetitive in applications because they are designed to handle "unrealistically" hard instances with little importance for practitioners. To bridge the gap between theory and practice, it is necessary to go *beyond the worst-case analysis* by, for example, characterizing and focusing on inputs that arise in practice.

A. Real-world Inputs

Several approaches have been proposed to bridge the gap between theory and practice. For example, researchers have considered the average-case scenario (*e.g.,* [17]) where the running time of an algorithm is analyzed with respect to some probability distribution over the set of all inputs. Smooth analysis (*e.g.,* [38]) is another celebrated approach that analyzes the running time of an algorithm with respect to worst-case inputs subject to small random perturbations.

Another successful approach, the one we take in this paper, consists in focusing on *structured* inputs. In a seminal paper, Ostrovsky, Rabani, Schulman, and Swamy [37] introduced the idea that inputs that come from practice induce a *ground-truth* or a *meaningful* clustering. They argued that an input I contains a meaningful clustering into k clusters if the optimal k-median cost of a clustering using k centers, say $\mathrm{OPT}_k(I)$, is much smaller than the optimal cost of a clustering using $k-1$ centers $\mathrm{OPT}_{k-1}(I)$.

The project leading to this application has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 748094.

The second author acknowledges the support by Deutsche Forschungsgemeinschaft within the Collaborative Research Center SFB 876, project A2, and the Google Focused Award on Web Algorithmics for Large-scale Data Analysis.

The authors thank their dedicated advisor for this project: Claire Mathieu. Without her, this collaboration would not have been possible.

0272-5428/17 $31.00 © 2017 IEEE
DOI 10.1109/FOCS.2017.14

This is also motivated by the *elbow method*[1] used by practitioners to define the number of clusters. More formally, an instance I of k-median or k-means satisfies the *α-ORSS property* if $\text{OPT}_k(I)/\text{OPT}_{k-1}(I) \leq \alpha$.

The popular k-means++ algorithm [4] achieves an $O(1)$-approximation for these inputs [20], [30], [37]. ORSS-stability also implies several other conditions aiming to capture well-clusterable instances. Thus, the inputs satisfying the ORSS property arguably share some properties with the real-world inputs.

These results have opened new research directions and raised several questions. For example:

- Is it possible to obtain similar results for more general classes of inputs?
- How does the parameter α impact the approximation guarantee and running time?
- Is it possible to prove good performance guarantees for other popular heuristics?
- How close to the "ground-truth" clustering are the approximate clusterings?

Many lines of research have tried to address these questions. The arguably most important approaches and most relevant to this paper can be roughly grouped under one of the following stability notions.

Distribution Stability (Def. IV.1): Awasthi, Blum and Sheffet [6] have tackled the first two questions by introducing the notion of *distribution stable* instances. Distribution stable instances are a generalization of the ORSS instances (in other words, any instance satisfying the ORSS property is distribution stable). They also introduced a new algorithm tailored for distribution stable instances that achieves a $(1+\varepsilon)$-approximation for α-ORSS inputs (and more generally α-distribution stable instances) in time $n^{O(1/\varepsilon\alpha)}$. This was the first algorithm whose approximation guarantee was independent from the parameter α for α-ORSS inputs.

Spectral Separability (Def. VII.1): Kumar and Kannan [33] tackled the first and third questions by introducing the *proximity* condition[2]. This condition also generalizes the ORSS condition. It is motivated by the goal of learning a distribution mixture in a d-dimensional Euclidean space. Quoting [33], the message of their paper can loosely be stated as:

> If the projection of any data point onto the line joining its cluster center to any other cluster center is γk times standard deviations closer to its own center than the other center, then we can cluster correctly in polynomial time.

In addition, they have made a significant step toward understanding the success of the classic k-means by showing that it achieves a $1 + O(1/\gamma)$-approximation for instances that satisfy the proximity condition. This result has been further improved by Awasthi and Sheffet [9].

Perturbation Resilience (Def. VI.1): In a seminal work, Bilu and Linial [19] introduced a new condition to capture real-world instances. They argue that the optimal solution of a real-world instance is often much better than any other solution and so, a slight perturbation of the instance does not lead to a different optimal solution. Perturbation-resilient instances have been studied in various contexts (see *e.g.,* [7], [11], [12]). For clustering problems, an instance is said to be *α-perturbation resilient* if an adversary can change the distances between pairs of elements by a factor at most α and the optimal solution remains the same. Recently, Angelidakis, Makarychev, and Makarychev [3] have given a polynomial-time algorithm for solving 2-perturbation-resilient instances[3]. Balcan and Liang [12] have tackled the third question by showing that a classic algorithm for hierarchical clustering can solve $1 + \sqrt{2}$-perturbation-resilient instances. This very interesting result leaves open the question as whether classic algorithms for ("flat") clustering could also be proven to be efficient for perturbation-resilient instances.

Main Open Questions: Previous work has made important steps toward bridging the gap between theory and practice for clustering problems. However, we still do not have a complete understanding of the properties of "well-structured" inputs, nor do we know why the algorithms used in practice perform so well. Some of the most important open questions are the following:

- Do the different definitions of well-structured input have common properties?
- Do heuristics used in practice have strong approximation ratios for well-structured inputs?
- Do heuristics used in practice recover the "ground-truth" clustering on well-structured inputs?

B. Our Results: A unified approach via Local Search

We make a significant step toward answering the above open questions. We show that the classic Local Search heuristic (see Algorithm 1), that has found widespread application in practice (see Section II), achieves *good* approximation guarantees for distribution-stable, spectrally-separable, and perturbation-resilient instances (see Theorems IV.2, V.2, VII.2).

More concretely, we show that Local Search is a polynomial-time approximation scheme (PTAS) for both distribution-stable and spectrally-separable[4] instances. In the case of distribution stability, we also answer the above open question by showing that *most* of the structure of the optimal underlying clustering is recovered by the algorithm.

For γ-perturbation-resilient instances, we show that if $\gamma > 3$ then any solution is the optimal solution if it

[1] The elbow-method consists in running an (approximation) algorithm for an incrementally increasing number of clusters until the cost drops significantly.

[2] In this paper, we work with a slightly more general condition called *spectral separability* but the motivations behind the two conditions are similar.

[3] We note that it is NP-hard to recover the optimal clustering of any $2-\varepsilon$-perturbation-resilient instance [18].

[4] Assuming a standard preprocessing step consisting of a projection onto a subspace of lower dimension.

cannot be improved by adding or removing 2γ centers. We also show that the analysis is essentially tight.

These results show that well-structured inputs have the property that the local optima are close both qualitatively (in terms of structure) and quantitatively (in terms of objective value) to the global "ground-truth" optimum.

These results make a significant step toward explaining the success of Local Search approaches for solving clustering problems in practice.

Algorithm 1 Local Search(ε) for k-Median and k-Means

Input: point set A, candidate solutions F cost function cost, integer k
Parameter: ε
$S \leftarrow$ Arbitrary subset of F of cardinality at most k.
while $\exists\, S'$ s.t. $|S'| \leq k$ **and** $|S-S'|+|S'-S| \leq 2/\varepsilon$ **and** cost$(S') \leq (1-\varepsilon/n)$ cost(S)
do
    $S \leftarrow S'$
end while
Output: S

II. Related Work

Worst-Case Hardness: The problems we study are NP-hard: k-median and k-means are already NP-hard in the Euclidean plane [36], [35]. In terms of hardness of approximation, both problems are APX-hard, even in the Euclidean setting when both k and d are part of the input (see Guha and Khuller [25], Guruswami et al. [28], and Awasthi et al. [8]). On the positive side, constant factor approximations are known in metric space for both k-median and k-means (see [2], [21]). For Euclidean spaces we have a PTAS for both problems, either assuming d fixed and k arbitrary [23], [24], [29], [32], or assuming k fixed and d arbitrary [34].

Local Search: Local Search is an all-purpose heuristic that may be applied to any problem, see Aarts and Lenstra [1] for a general introduction. Arya et al. [5] showed that Local Search with a neighborhood size of $1/\varepsilon$ gives a $3+2\varepsilon$ approximation to k-median, see also [27]. Kanungo et al. [31] proved an approximation ratio of $9+\varepsilon$ for k-means clustering by Local Search, which was until very recently [2] the best known algorithm with a polynomial running time in metric and Euclidean spaces. Recently, Local Search with an appropriate neighborhood size was shown to be a PTAS for k-means and k-median in certain restricted metrics including constant dimensional Euclidean space [23], [24]. Due to its simplicity, Local Search is also a popular subroutine for clustering tasks in various more specialized computational models, *e.g.,* [26].

Stability Conditions: A further condition related to the aforementioned is *approximation stability*. Defined by Balcan et al. [10], [14] (see also robust-perturbation resilience [16]), it requires that any clustering with cost within a factor c of the optimum has a distance at most ε to the target clustering. Balcan et al. [10], [14] then showed that this condition is sufficient to both bypass worst-case lower bounds for the approximation factor, and to find a clustering with distance $O(\varepsilon)$ from the target clustering. The condition was extended to account for the presence of noisy data by Balcan et al. [13]. This approach was improved for other min-sum clustering objectives such as correlation clustering by Balcan and Braverman [15]. For constant c, (c,ε) approximation stability also implies the β-stability condition of Awasthi et al. [6] with constant β, if the target clusters are greater than εn.

III. Definitions and Notations

The problem we consider in this work is the following slightly more general version of the k-means and k-median problems.

Definition III.1 (k-Clustering)**.** *Let A be a set of clients, F a set of centers, both lying in a metric space $(\mathcal{X}, dist)$,* cost *a function $A \times F \to \mathbb{R}_+$, and k a non-negative integer. The* k-clustering problem *asks for a subset S of F, of cardinality at most k, that minimizes*

$$cost(S) = \sum_{x \in A} \min_{c \in S} cost(x, c).$$

The clustering of A induced by *S is the partition of A into subsets $C = \{C_1, \dots C_k\}$ such that $C_i = \{x \in A \mid c_i = \operatorname{argmin}_{c \in S} cost(x,c)\}$ (breaking ties arbitrarily).*

The well known k-median and k-means problems correspond to the special cases $\text{cost}(a,c) = \text{dist}(a,c)$ and $\text{cost}(a,c) = \text{dist}(a,c)^2$ respectively. Throughout the rest of this paper, let OPT denote the value of an optimal solution. To give slightly simpler proofs for β-distribution-stable and α-perturbation-resilient instances, we will assume that $\text{cost}(a,b) = \text{dist}(a,b)$. If $\text{cost}(a,b) = \text{dist}(a,b)^p$, then α depends exponentially on the p for perturbation resilience. For distribution stability, we still have a PTAS by introducing a dependency in $1/\varepsilon^{O(p)}$ in the neighborhood size of the algorithm. The analysis is unchanged save for various applications of the following lemma at different steps of the proof.

Lemma III.2. *Let $p \geq 0$ and $1/2 > \varepsilon > 0$. For any $a, b, c \in \mathcal{X}$, we have*

$$cost(a,b) \leq (1+\varepsilon)^p cost(a,c) + cost(c,b)(1+1/\varepsilon)^p.$$

IV. Distribution Stability

Definition IV.1. *Let $(A, F, cost, k)$ be an instance of k-clustering where $A \cup F$ lie in a metric space and let $S^* = \{c_1^*, \dots, c_k^*\} \subseteq F$ be a set of centers and $C^* = \{C_1^*, \dots, C_k^*\}$ be the clustering induced by S^*. Further, let $\beta > 0$. Then the pair $(A, F, cost, k), (C^*, S^*)$*

is a β-distribution stable instance *if, for any* i *and for any* $x \in C_i$*, for any* $j \neq i$*,*

$$cost(x, c_j^*) \geq \beta \frac{OPT}{|C_j^*|},$$

where $cost(x, c_j^*)$ *is the cost of assigning* x *to* c_j^**.*

For any instance (A, F, cost, k) that is β-distribution stable, we refer to (C^*, S^*) as a β-clustering of the instance. We show the following theorem for the k-median problem. For the k-clustering problem with parameter p, the constant η becomes a function of p.

Theorem IV.2. *Let* $p > 0$*,* $\beta > 0$*, and* $\varepsilon < 1/3$*. For a* β*-stable instance with* β *clustering* (C^*, S^*) *and an absolute constant* η*, the cost of the solution output by Local Search*$(4\varepsilon^{-3}\beta^{-1} + O(\varepsilon^{-2}\beta^{-1}))$ *(Algorithm 1) is at most* $(1 + \eta\varepsilon)cost(C^*)$*.*

Moreover, let $L = \{L_1, \ldots, L_k\}$ *denote the clusters of the solution output by Local Search*$(4\varepsilon^{-3}\beta^{-1} + O(\varepsilon^{-2}\beta^{-1}))$*. Then there exists a bijection* $\phi : L \mapsto C^*$ *such that for at least* $m = k - O(\varepsilon^{-3}\beta^{-1})$ *clusters* $L'_1, \ldots, L'_m \subseteq L$*, the following two statements hold.*

- *At least a* $(1 - \varepsilon)$ *fraction of* $IR_i^{\varepsilon^2} \cap C_i^*$ *are served by a unique center* $\mathcal{L}(i)$ *in solution* $\mathcal{L}$*.*
- *The total number of clients* $p \in \bigcup_{j \neq i} C_j^*$ *served by* $\mathcal{L}(i)$ *in* $\mathcal{L}$ *is at most* $\varepsilon|C_i^*|$*.*

We first give a high-level description of the analysis. Assume for simplicity that all the optimal clusters cost less than an ε^3 fraction of the total cost of the optimal solution. Combining this assumption with the β-distribution-stability property, one can show that the centers and points close to the center are far away from each other. Thus, guided by the objective function, the local search algorithm identifies most of these centers. In addition, we can show that for most of these good centers the corresponding cluster in the local solution is very similar to the optimal cluster. In total, only very few clusters (a function of ε and β) of the optimal solution are not present in the local solution. We conclude our proof by using local optimality. Our proof includes a few ingredients from [6] such as the notion of *inner-ring* (we work with a slightly more general definition) and distinguishes between *cheap* and *expensive* clusters. Nevertheless our analysis is slightly stronger as we consider a significantly weaker stability condition and can not only analyze the cost of the solution of the algorithm, but also the structure of its clusters.

Throughout this section, we consider a set of centers $S^* = \{c_1^*, \ldots, c_k^*\}$ whose induced clustering is $C^* = \{C_1^*, \ldots, C_k^*\}$ and such that the instance is β-stable with respect (C^*, S^*). We denote by *clusters* the parts of a partition $C^* = \{C_1^*, \ldots, C_k^*\}$. Let $\text{cost}(C^*) = \sum_{i=1}^k \sum_{x \in C_i^*} \text{cost}(x, c_i^*)$. Moreover, for any cluster C_i^*, for any client $x \in C_i^*$, denote by g_x the cost of client x in solution C^*: $g_x = \text{cost}(x, c_i^*) = \text{dist}(x, c_i^*)$ since we consider the k-median problem. Let $\mathcal{L}$ denote the output of LocalSearch$(\beta^{-1}\varepsilon^{-3})$ and l_x the cost induced by client x in solution $\mathcal{L}$, namely $l_x = \min_{\ell \in \mathcal{L}} \text{cost}(x, \ell)$, and $\text{cost}(\mathcal{L}) = \sum_{x \in A} l_x$. The following definition is a generalization of the inner-ring definition of [6].

Definition IV.3. *For any* ε_0*, we define the* inner ring *of cluster* i*,* $IR_i^{\varepsilon_0}$*, as the set of* $x \in A \cup F$ *such that* $dist(x, c_i^*) \leq \varepsilon_0 \beta OPT/|C_i^*|$*.*

We say that cluster i is *cheap* if $\sum_{x \in C_i^*} g_x \leq \varepsilon^3 \beta \text{OPT}$, and *expensive* otherwise. We aim at proving the following structural lemma.

Lemma IV.4. *There exists a set of clusters* $Z^* \subseteq C^*$ *of size at most* $2\varepsilon^{-3}\beta^{-1} + O(\varepsilon^{-2}\beta^{-1})$ *such that for any cluster* $C_i^* \in C^* - Z^*$*, we have the following properties*

1) C_i^* *is cheap.*
2) *At least a* $(1 - \varepsilon)$ *fraction of* $IR_i^{\varepsilon^2} \cap C_i^*$ *are served by a unique center* $\mathcal{L}(i)$ *in solution* $\mathcal{L}$*.*
3) *The total number of clients* $p \in \bigcup_{j \neq i} C_j$ *served by* $\mathcal{L}(i)$ *in* $\mathcal{L}$ *is at most* $\varepsilon|IR_i^{\varepsilon^2} \cap C_i^*|$*.*

We start with the following lemma which generalizes Fact 4.1 in [6].

Lemma IV.5. *Let* C_i^* *be a cheap cluster. For any* ε_0*, we have* $|IR_i^{\varepsilon_0} \cap C_i^*| > (1 - \varepsilon^3/\varepsilon_0)|C_i^*|$*.*

We then prove that the inner rings of cheap clusters are disjoint for $\frac{\varepsilon^3}{\varepsilon_0} < 1$ and $\varepsilon_0 < \frac{1}{3}$.

Lemma IV.6. *Let* $\frac{\varepsilon^3}{\varepsilon_0} < 1$ *and* $\varepsilon_0 < \frac{1}{3}$*. If* $C_i^* \neq C_j^*$ *are cheap clusters, then* $IR_i^{\varepsilon_0} \cap IR_j^{\varepsilon_0} = \emptyset$*.*

For each cheap cluster C_i^*, let $\mathcal{L}(i)$ denote a center of $\mathcal{L}$ that belongs to $\text{IR}_i^{\varepsilon}$ if there exists exactly such center and remain undefined otherwise. By Lemma IV.6, $\mathcal{L}(i) \neq \mathcal{L}(j)$ for $i \neq j$.

Lemma IV.7. *Let* $\varepsilon < \frac{1}{3}$*. Let* $C^* - Z_1$ *denote the set of clusters* C_i^* *that are cheap, such that* $\mathcal{L}(i)$ *is defined and such that at least* $(1 - \varepsilon)|IR_i^{\varepsilon^2} \cap C_i^*|$ *clients of* $IR_i^{\varepsilon^2} \cap C_i^*$ *are served in* $\mathcal{L}$ *by* $\mathcal{L}(i)$*. Then* $|Z_1| \leq (2\varepsilon^{-3} + 11.25 \cdot \varepsilon^{-2} + 22.5 \cdot \varepsilon^{-1})\beta^{-1}$*.*

Proof: There are five different types of clusters in C^*:

1) k_1 expensive clusters
2) k_2 cheap clusters with no center of $\mathcal{L}$ belonging to $\text{IR}_i^{\varepsilon}$
3) k_3 cheap clusters with at least two centers of $\mathcal{L}$ belonging to $\text{IR}_i^{\varepsilon}$
4) k_4 cheap clusters with $\mathcal{L}(i)$ being defined and less than $(1 - \varepsilon)|\text{IR}_i^{\varepsilon^2} \cap C_i^*|$ clients of $\text{IR}_i^{\varepsilon^2} \cap C_i^*$ are served in $\mathcal{L}$ by $\mathcal{L}(i)$
5) k_5 cheap clusters with $\mathcal{L}(i)$ being defined and at least $(1 - \varepsilon)|\text{IR}_i^{\varepsilon^2} \cap C_i^*|$ clients of $\text{IR}_i^{\varepsilon^2} \cap C_i^*$ are served in $\mathcal{L}$ by $\mathcal{L}(i)$

The definition of cheap clusters immediately yields $k_1 \leq \varepsilon^{-3}\beta^{-1}$.

Since $\mathcal{L}$ and C^* both have k clusters and the inner rings of cheap clusters are disjoint (Lemma IV.6), we have $c_1 k_1 + c_3 k_3 + k_4 + k_5 = k_1 + k_2 + k_3 + k_4 + k_5 =$

$|Z_1| + k_5 = k$ with $c_1 \geq 0$ and $c_3 \geq 2$ resulting in $k_3 \leq (c_3 - 1)k_3 = (1 - c_1)k_1 + k_2 \leq k_1 + k_2$.

Before bounding k_2 and k_4, we discuss the impact of a cheap cluster C_i^* with at least a p fraction of the clients of $\text{IR}_i^{\varepsilon^2} \cap C_i^*$ being served in $\mathcal{L}$ by some centers that are not in $\text{IR}_i^{\varepsilon}$. By the triangular inequality, the cost for any client x of this p fraction is at least $(\varepsilon - \varepsilon^2)\beta\text{cost}(C^*)/|C_i^*|$. Then the total cost of all clients of this p fraction in $\mathcal{L}$ is at least $p|\text{IR}_i^{\varepsilon^2} \cap C_i^*|(1-\varepsilon)\varepsilon\beta\text{cost}(C^*)/|C_i^*|$. By Lemma IV.5, substituting $|\text{IR}_i^{\varepsilon^2} \cap C_i^*|$ yields for this total cost

$$p|\text{IR}_i^{\varepsilon^2} \cap C_i^*|(1-\varepsilon)\varepsilon\beta\frac{\text{cost}(C^*)}{|C_i^*|} \geq$$

$$p(1-\varepsilon)^2|C_i^*|\varepsilon\beta\frac{\text{cost}(C^*)}{|C_i^*|} = p(1-\varepsilon)^2\varepsilon\beta\text{cost}(C^*).$$

To determine k_2, we must use $p = 1$ while we have $p > \varepsilon$ for k_4. Therefore, the total costs of all clients of the k_2 and the k_4 clusters in $\mathcal{L}$ are at least $k_2(1-\varepsilon)^2\varepsilon\beta\text{cost}(C^*)$ and $k_4(1-\varepsilon)^2\varepsilon^2\beta\text{cost}(C^*)$, respectively.

Now, since $\text{cost}(\mathcal{L}) \leq 5\text{OPT} \leq 5\text{cost}(C^*)$, we have $(k_2 + k_4\varepsilon)\varepsilon\beta \leq 5/(1-\varepsilon)^2 \leq 45/4$.

Therefore, we have $|Z_1| = k_1 + k_2 + k_3 + k_4 \leq 2k_1 + 2k_2 + k_4 \leq (2\varepsilon^{-3} + 11.25\cdot\varepsilon^{-2} + 22.5\cdot\varepsilon^{-1})\beta^{-1}$. ■

We continue with the following lemma, whose proof relies on similar arguments.

Lemma IV.8. *There exists a set $Z_2 \subseteq C^* - Z_1$ of size at most $11.25\varepsilon^{-1}\beta^{-1}$ such that for any cluster $C_j^* \in C^* - Z_2$, the total number of clients $x \in \bigcup_{i \neq j} C_i$, that are served by $\mathcal{L}(j)$ in $\mathcal{L}$ is at most $\varepsilon|IR_i^{\varepsilon^2} \cap C_i^*|$.*

Therefore, the proof of Lemma IV.4 follows from combining Lemmas IV.7 and IV.8.

We now turn to the analysis of the cost of $\mathcal{L}$. Let $C(Z^*) = \bigcup_{C_i^* \in Z^*} C_i^*$. For any cluster $C_i^* \in C^* - Z^*$, let $\mathcal{L}(i)$ be the unique center of $\mathcal{L}$ that serves at least $(1-\varepsilon)|\text{IR}_i^{\varepsilon^2} \cap C_i^*| > (1-\varepsilon)^2|C_i|$ clients of $\text{IR}_i^{\varepsilon^2} \cap C_i^*$, see Lemmas IV.4 and IV.5. Let $\widehat{\mathcal{L}} = \bigcup_{C_i^* \in C^* - Z^*} \mathcal{L}(i)$ and define $\widehat{A}$ to be the set of clients that are served in solution $\mathcal{L}$ by centers of $\widehat{\mathcal{L}}$. Finally, let $A(\mathcal{L}(i))$ be the set of clients that are served by $\mathcal{L}(i)$ in solution $\mathcal{L}$. Observe that the $A(\mathcal{L}(i))$ partition $\widehat{A}$.

Lemma IV.9. *We have*

$$-\varepsilon \cdot \textit{cost}(\mathcal{L})/n + \sum_{x(A-\widehat{A})\cup - C(Z^*)} l_x \leq$$

$$\sum_{x \in (A-\widehat{A})\cup - C(Z^*)} g_x + \frac{2\varepsilon}{(1-\varepsilon)^2} \cdot (\textit{cost}(C^*) + \textit{cost}(\mathcal{L})).$$

Proof: Consider the following mixed solution $\mathcal{M} = \widehat{\mathcal{L}} \cup \{c_i^* \mid C_i^* \in Z^*\}$. We start by bounding the cost of $\mathcal{M}$. For any client $x \in \widehat{A}$, the center that serves it in $\mathcal{L}$ belongs to $\mathcal{M}$. Thus its cost in $\mathcal{M}$ is at most l_x. Now, for any client $x \in C(Z^*)$, the center that serves it in C^* is in $\mathcal{M}$, so its cost in $\mathcal{M}$ is at most g_x.

Finally, we evaluate the cost of the clients in $A - (\widehat{A} \cup C(Z^*))$. Consider such a client x and let C_i^* be the cluster it belongs to in solution C^*. Since $C_i^* \in C^* - Z^*$, $\mathcal{L}(i)$ is defined and we have $\mathcal{L}(i) \in \widehat{\mathcal{L}} \subseteq \mathcal{M}$. Hence, the cost of x in $\mathcal{M}$ is at most $\text{cost}(x, \mathcal{L}(i))$. Observe that by the triangular inequality, $\text{cost}(x, \mathcal{L}(i)) \leq \text{cost}(x, c_i^*) + \text{cost}(c_i^*, \mathcal{L}(i)) = g_x + \text{cost}(c_i^*, \mathcal{L}(i))$.

Now consider a client $x' \in R_i := \text{IR}_i^{\varepsilon^2} \cap C_i^* \cap A(\mathcal{L}(i))$. By the triangular inequality, we have $\text{cost}(c_i^*, \mathcal{L}(i)) \leq \text{cost}(c_i^*, x') + \text{cost}(x', \mathcal{L}(i)) = g_{x'} + l_{x'}$. Hence,

$$\text{cost}(c_i^*, \mathcal{L}(i)) \leq \frac{1}{|R_i|} \sum_{x' \in R_i} (g_{x'} + l_{x'}).$$

It follows that assigning the clients of $C_i^* \cap (A - \widehat{A})$ to $\mathcal{L}(i)$ induces a cost of at most

$$\sum_{x \in C_i^* \cap (A - \widehat{A})} g_x + \frac{|C_i^* \cap (A - \widehat{A})|}{|R_i|} \sum_{x' \in R_i} (g_{x'} + l_{x'}).$$

Due to Lemma IV.4, we have $|R_i| = |\text{IR}_i^{\varepsilon^2} \cap C_i^* \cap A(\mathcal{L}(i))| \geq (1-\varepsilon) \cdot |\text{IR}_i^{\varepsilon^2} \cap C_i^*|$ and $|(\text{IR}_i^{\varepsilon^2} \cap C_i^*) \cap (A - \widehat{A})| \leq \varepsilon \cdot |\text{IR}_i^{\varepsilon^2} \cap C_i^*|$. Further, $|(C_i^* - \text{IR}_i^{\varepsilon^2}) \cap (A - \widehat{A})| \leq |(C_i^* - \text{IR}_i^{\varepsilon^2})| = |C_i^*| - |\text{IR}_i^{\varepsilon^2} \cap C_i^*|$. Combining these three bounds, we have

$$\begin{aligned}
\frac{|C_i^* \cap (A - \widehat{A})|}{|R_i|} &= \frac{|C_i^* \cap (A - \widehat{A})|}{|\text{IR}_i^{\varepsilon^2} \cap C_i^* \cap A(\mathcal{L}(i))|} \\
&\leq \frac{|C_i^*| - (1-\varepsilon)|\text{IR}_i^{\varepsilon^2} \cap C_i^*|}{(1-\varepsilon) \cdot |\text{IR}_i^{\varepsilon^2} \cap C_i^*|} \qquad (1) \\
&= \frac{|C_i^*|}{(1-\varepsilon) \cdot |\text{IR}_i^{\varepsilon^2} \cap C_i^*|} - 1 \\
&\leq \frac{|C_i^*|}{(1-\varepsilon)^2 \cdot |C_i^*|} - 1 \leq \frac{2\varepsilon - \varepsilon^2}{(1-\varepsilon)^2} < \frac{2\varepsilon}{(1-\varepsilon)^2} \qquad (2)
\end{aligned}$$

where the inequality in (2) follows from Lemma IV.5.

Summing over all clusters $C_i^* \in C^* - Z^*$, we obtain that the cost in $\mathcal{M}$ for the clients in $(A - \widehat{A}) \cap C_i^*$ is less than

$$\sum_{c \in A - (\widehat{A} \cup C(Z^*))} g_x + \frac{2\varepsilon}{(1-\varepsilon)^2} \cdot (\text{cost}(C^*) + \text{cost}(\mathcal{L})).$$

By Lemmas IV.7 and IV.8, we have $|\mathcal{M} - \mathcal{L}| + |\mathcal{L} - \mathcal{M}| = 2 \cdot |Z^*| \leq (4\varepsilon^{-3} + O(\varepsilon^{-2}))\beta^{-1}$. By selecting the neighborhood size of Local Search (Algorithm 1) to be greater than this value, we have $(1 - \varepsilon/n) \cdot \text{cost}(\mathcal{L}) \leq \text{cost}(\mathcal{M})$. Therefore, combining the above observations, we have

$$\begin{aligned}
&(1 - \tfrac{\varepsilon}{n}) \cdot \text{cost}(\mathcal{L}) \leq \\
&\sum_{x \in \widehat{A} - C(Z^*)} l_x + \sum_{x \in C(Z^*)} g_x + \\
&\sum_{x \in A - (\widehat{A} \cup C(Z^*))} g_x + \tfrac{2\varepsilon}{(1-\varepsilon)^2} \cdot (\text{cost}(C^*) + \text{cost}(\mathcal{L})).
\end{aligned}$$

By simple transformations, we then obtain

$$-\frac{\varepsilon}{n}\cdot \text{cost}(\mathcal{L}) + \sum_{x\in A-(\widehat{A})\cup C(Z^*)} l_x \leq \sum_{x\in A-(\widehat{A})\cup C(Z^*)} g_x + \frac{2\varepsilon}{(1-\varepsilon)^2}\cdot(\text{cost}(C^*)+\text{cost}(\mathcal{L})).$$

■

We now turn to evaluate the cost for the clients that are in $\widehat{A} - C(Z^*)$. For any cluster $C_i^* \in C^* - C(Z^*)$ and for any $x \in C_i^* - A(\mathcal{L}(i))$ define Reassign(x) to be the cost of x with respect to the center in $\mathcal{L}(i)$. Note that there exists only one center of $\mathcal{L}$ in IR_i^ε for any cluster $C_i^* \in C^* - C(Z^*)$. Before going deeper in the analysis, we need the following lemma.

Lemma IV.10. *For any $C_i^* \in C^* - C(Z^*)$, we have*

$$\sum_{x\in C_i^*-A(\mathcal{L}(i))} \textit{Reassign}(x) \leq \sum_{x\in C_i^*-A(\mathcal{L}(i))} g_x + \frac{2\varepsilon}{(1-\varepsilon)^2}\sum_{x\in C_i^*}(l_x+g_x).$$

We now partition the clients of cluster $C_i^* \in C^* - Z^*$. For any i, let B_i be the set of clients of C_i^* that are served in solution $\mathcal{L}$ by a center $\mathcal{L}(j)$ for some $j \neq i$ and $C_j^* \in C^* - Z^*$. Moreover, let $D_i = (A(\mathcal{L}(i)) \cap (\bigcup_{j\neq i} B_j))$. Finally, define $E_i = (C_i^* \cap \widehat{A}) - \bigcup_{j\neq i} D_j$.

Lemma IV.11. *Let C_i^* be a cluster in $C^* - Z^*$. Define the solution $\mathcal{M}^i = \mathcal{L} - \{\mathcal{L}(i)\} \cup \{c_i^*\}$ and denote by m_x^i the cost of client x in solution $\mathcal{M}^i$. Then*

$$\sum_{x\in A} m_x^i \leq \sum_{\substack{x\in A-\\(A(\mathcal{L}(i))\cup E_i)}} l_x + \sum_{x\in E_i} g_x + \sum_{x\in D_i} \textit{Reassign}(x) + \sum_{\substack{x\in A(\mathcal{L}(i))-\\(E_i\cup D_i)}} l_x + \frac{\varepsilon}{(1-\varepsilon)}\left(\sum_{x\in E_i} g_x + l_x\right).$$

We can thus prove the following lemma, which concludes the proof.

Lemma IV.12. *We have*

$$-\varepsilon\cdot \textit{cost}(\mathcal{L}) + \sum_{x\in\widehat{A}-C(Z^*)} l_x \leq \sum_{x\in\widehat{A}-C(Z^*)} g_x + \frac{3\varepsilon}{(1-\varepsilon)^2}\cdot(\textit{cost}(\mathcal{L})+\textit{cost}(C^*)).$$

The proof of Theorem IV.2 follows from (1) summing the equations from Lemmas IV.9 and IV.12 and noting that $((A - \hat{A}) \cup C(Z^*)) \cup (\hat{A} - C(Z^*)) = A$. The comparison of the structure of the local solution to the structure of C^* is an immediate corollary of Lemma IV.4.

V. Perturbation Resilience

We first give the definition of α-perturbation-resilient instances.

Definition V.1. *Let $I = (A, F, cost, k)$ be an instance for the k-clustering problem. For $\alpha \geq 1$, I is* α-perturbation-resilient *if there exists a unique optimal set of centers $C^* = \{c_1^*, \ldots, c_k^*\}$ and for any instance $I' = (A, F, cost', k, p)$, such that*

$$\forall\ a, b \in \mathcal{P},\ cost(a,b) \leq cost'(a,b) \leq \alpha cost(a,b),$$

the unique optimal set of centers is $C^* = \{c_1^*, \ldots, c_k^*\}$.

For ease of exposition, we assume that $\text{cost}(a,b) = \text{dist}(a,b)$ (*i.e.,* we work with the k-median problem). Given solution S_0, we say that S_0 is $1/\varepsilon$-*locally optimal* if any solution S_1 such that $|S_0 - S_1| + |S_1 - S_0| \leq 2/\varepsilon$ has at least cost(S_0).

Theorem V.2. *Let $\alpha > 3$. For any instance of the k-median problem that is α-perturbation-resilient, any $2(\alpha-3)^{-1}$-locally optimal solution is the optimal set of centers $\{c_1^*, \ldots, c_k^*\}$.*

Moreover, define l_c to be the cost for client c in solution $\mathcal{L}$ and g_c to be its cost in the optimal solution C^*. Finally, for any sets of centers S and $S_0 \subset S$, define $N_S(S_0)$ to be the set of clients served by a center of S_0 in solution S, i.e.: $N_S(S_0) = \{x \mid \exists s \in S_0, \text{dist}(x,s) = \min_{s'\in S}\text{dist}(x,s')\}$.

The proof of Theorem V.2 relies on the following theorem of particular interest.

Theorem V.3 (Local-Approximation Theorem.)**.** *Let $\mathcal{L}$ be a $1/\varepsilon$-locally optimal solution and C^* be any solution. Define $S = \mathcal{L}\cap C^*$ and $\tilde{\mathcal{L}} = \mathcal{L} - S$ and $\tilde{C^*} = C^* - S$. Then*

$$\sum_{c\in N_{C^*}(\tilde{C^*})-N_{\mathcal{L}}(\tilde{\mathcal{L}})} l_c + \sum_{c\in N_{\mathcal{L}}(\tilde{\mathcal{L}})} l_c \leq \sum_{c\in N_{C^*}(\tilde{C^*})-N_{\mathcal{L}}(\tilde{\mathcal{L}})} g_c + (3+2\varepsilon)\sum_{c\in N_{\mathcal{L}}(\tilde{\mathcal{L}})} g_c.$$

We first show how Theorem V.3 allows us to prove Theorem V.2.

Proof of Theorem V.2: Given an instance (A, F, dist, k), we define the following instance $I' = (A, F, \text{dist}', k)$, where dist$'(a,b)$ is a distance function defined over $A \cup F$ that we detail below. For each client $c \in N_{\mathcal{L}}(\tilde{\mathcal{L}}) \cup N_{C^*}(\tilde{C^*})$, let ℓ_c be the center of $\mathcal{L}$ that serves it in $\mathcal{L}$, for any point $p \neq \ell_c$, we define dist$'(c,p) = \alpha\text{dist}(c,p)$ and dist$'(c,\ell_c) = \text{dist}(c,\ell_c)$. For the other clients we set dist$'$ = dist. Observe that by local optimality, the clustering induced by $\mathcal{L}$ is $\{c_1^*, \ldots, c_k^*\}$ if and only if $\mathcal{L} = C^*$. Therefore, the cost of C^* in instance I' is equal to

$$\alpha\sum_{c\in N_{\mathcal{L}}(\tilde{\mathcal{L}})} g_c + \sum_{\substack{c\in N_{C^*}(\tilde{C^*})\\ -N_{\mathcal{L}}(\tilde{\mathcal{L}})}} \min(\alpha g_c, l_c) + \sum_{\substack{c\notin N_{C^*}(\tilde{C^*})\\ \cup N_{\mathcal{L}}(\tilde{\mathcal{L}})}} g_c.$$

On the other hand, the cost of $\mathcal{L}$ in I' is the same as in I. By Theorem V.3

$$\sum_{c\in N_{C^*}(\tilde{C^*})-N_{\mathcal{L}}(\tilde{\mathcal{L}})} l_c + \sum_{c\in N_{\mathcal{L}}(\tilde{\mathcal{L}})} l_c + \leq \sum_{c\in N_{C^*}(\tilde{C^*})-N_{\mathcal{L}}(\tilde{\mathcal{L}})} g_c + (3+\tfrac{2(\alpha-3)}{2}) \sum_{c\in N_{\mathcal{L}}(\tilde{\mathcal{L}})} g_c$$

and by definition of S we have, for each element $c \notin N_{C^*}(\tilde{C^*}) \cup N_{\mathcal{L}}(\tilde{\mathcal{L}})$, $l_c = g_c$.

Thus the cost of $\mathcal{L}$ in I' is at most

$$\left(3+\frac{2(\alpha-3)}{2}\right) \sum_{c\in N_{\mathcal{L}}(\tilde{\mathcal{L}})} g_c + \sum_{\substack{c\in N_{C^*}(\tilde{C^*}) \\ -N_{\mathcal{L}}(\tilde{\mathcal{L}})}} g_c + \sum_{\substack{c\notin N_{C^*}(\tilde{C^*}) \\ \cup N_{\mathcal{L}}(\tilde{\mathcal{L}})}} g_c$$

Now, observe that for the clients in $N_{C^*}(\tilde{C^*}) - N_{\mathcal{L}}(\tilde{\mathcal{L}}) = N_{C^*}(\tilde{C^*}) \cap N_{\mathcal{L}}(S)$, we have $l_c \geq g_c$.

Therefore, we have that the cost of $\mathcal{L}$ is at most the cost of C^* in I' and so by definition of α-perturbation-resilience, we have that the clustering $\{c_1^*, \ldots, c_k^*\}$ is the unique optimal solution in I'. Therefore $\mathcal{L} = C^*$ and the Theorem follows. ∎

We now turn to the proof of Theorem V.3.

Consider the following bipartite graph $\Gamma = (\tilde{\mathcal{L}} \cup \tilde{C^*}, \mathcal{E})$ where $\mathcal{E}$ is defined as follows. For any center $f \in \tilde{C^*}$, we have $(f, \ell) \in \mathcal{E}$ where ℓ is the center of $\tilde{\mathcal{L}}$ that is the closest to f. Denote $N_\Gamma(\ell)$ the neighbors of the point corresponding to center ℓ in Γ.

For each edge $(f, \ell) \in \mathcal{E}$, for any client $c \in N_{C^*}(f) - N_{\mathcal{L}}(\ell)$, we define Reassign_c as the cost of reassigning client c to ℓ. We derive the following lemma.

Lemma V.4. *For any client c, $\text{Reassign}_c \leq l_c + 2g_c$.*

Proof: By definition we have $\text{Reassign}_c = \text{dist}(c, \ell)$. By the triangle inequality $\text{dist}(c, \ell) \leq \text{dist}(c, f) + \text{dist}(f, \ell)$. Since f serves c in C^* we have $\text{dist}(c, f) = g_c$, hence $\text{dist}(c, \ell) \leq g_c + \text{dist}(f, \ell)$. We now bound $\text{dist}(f, \ell)$. Consider the center ℓ' that serves c in solution $\mathcal{L}$. By the triangle inequality we have $\text{dist}(f, \ell') \leq \text{dist}(f, c) + \text{dist}(c, \ell') = g_c + l_c$. Finally, since ℓ is the closest center of f in $\mathcal{L}$, we have $\text{dist}(f, \ell) \leq \text{dist}(f, \ell') \leq g_c + l_c$ and the lemma follows. ∎

We partition the centers of $\tilde{\mathcal{L}}$ as follows. Let $\tilde{\mathcal{L}}_0$ be the set of centers of $\tilde{\mathcal{L}}$ that have degree 0 in Γ. Let $\tilde{\mathcal{L}}_{\leq \varepsilon^{-1}}$ be the set of centers of $\tilde{\mathcal{L}}$ that have degree at least one and at most $1/\varepsilon$ in Γ. Let $\tilde{\mathcal{L}}_{>\varepsilon^{-1}}$ be the set of centers of $\tilde{\mathcal{L}}$ that have degree greater than $1/\varepsilon$ in Γ.

We now partition the centers of $\tilde{\mathcal{L}}$ and $\tilde{C^*}$ using the neighborhoods of the vertices of $\tilde{\mathcal{L}}$ in Γ. We start by iteratively constructing two set of pairs $S_{\leq \varepsilon^{-1}}$ and $S_{>\varepsilon^{-1}}$. For each center $\ell \in \tilde{\mathcal{L}}_{\leq \varepsilon^{-1}} \cup \tilde{\mathcal{L}}_{>\varepsilon^{-1}}$, we pick a set A_ℓ of $|N_\Gamma(\ell)| - 1$ centers of $\tilde{\mathcal{L}}_0$ and define a pair $(\{\ell\} \cup A_\ell, N_\Gamma(\ell))$. We then remove A_ℓ from $\tilde{\mathcal{L}}_0$ and repeat. Let $S_{\leq \varepsilon^{-1}}$ be the pairs that contain a center of $\tilde{\mathcal{L}}_{\leq \varepsilon^{-1}}$ and let $S_{>\varepsilon^{-1}}$ be the remaining pairs.

The following lemma follows from the definition of the pairs.

Lemma V.5. *Let $(R^{\tilde{\mathcal{L}}}, R^{\tilde{C^*}})$ be a pair in $S_{\leq \varepsilon^{-1}} \cup S_{>\varepsilon^{-1}}$. If $\ell \in R^{\tilde{\mathcal{L}}}$, then for any f such that $(f, \ell) \in \mathcal{E}$, $f \in R^{\tilde{C^*}}$.*

Lemma V.6. *For any pair $(R^{\tilde{\mathcal{L}}}, R^{\tilde{C^*}}) \in S_{\leq \varepsilon^{-1}}$ we have that*

$$\sum_{c\in N_{C^*}(R^{\tilde{C^*}})} l_c \leq \sum_{c\in N_{C^*}(R^{\tilde{C^*}})} g_c + 2 \sum_{N_{\mathcal{L}}(R^{\tilde{\mathcal{L}}})} g_c.$$

Proof: Consider the mixed solution $M = \mathcal{L} - R^{\tilde{\mathcal{L}}} \cup R^{\tilde{C^*}}$. For each point c, let m_c denote the cost of c in solution M. We have the following upper bounds

$$m_c \leq \begin{cases} g_c & \text{if } c \in N_{C^*}(R^{\tilde{C^*}}). \\ \text{Reassign}_c & \text{if } c \in N_{\mathcal{L}}(R^{\tilde{\mathcal{L}}}) - N_{C^*}(R^{\tilde{C^*}}). \\ l_c & \text{Otherwise.} \end{cases}$$

Now, observe that the solution M differs from $\mathcal{L}$ by at most $2/\varepsilon$ centers. Thus, by $1/\varepsilon$-local optimality we have $\text{cost}(\mathcal{L}) \leq \text{cost}(M)$. Summing over all clients and simplifying, we obtain

$$\sum_{c\in N_{C^*}(R^{\tilde{C^*}})} l_c + \sum_{c\in N_{\mathcal{L}}(R^{\tilde{\mathcal{L}}})-N_{C^*}(R^{\tilde{C^*}})} l_c \leq \sum_{c\in N_{C^*}(R^{\tilde{C^*}})} g_c + \sum_{c\in N_{\mathcal{L}}(R^{\tilde{\mathcal{L}}})-N_{C^*}(R^{\tilde{C^*}})} \text{Reassign}_c.$$

The lemma follows by combining with Lemma V.4. ∎

We now analyze the cost of the clients served by a center of $\mathcal{L}$ that has degree greater than ε^{-1} in Γ. The argument is very similar.

Lemma V.7. *For any pair $(R^{\tilde{\mathcal{L}}}, R^{\tilde{C^*}}) \in S_{>\varepsilon^{-1}}$ we have that*

$$\sum_{c\in N_{C^*}(R^{\tilde{C^*}})} l_c \leq \sum_{c\in N_{C^*}(R^{\tilde{C^*}})} g_c + 2(1+\varepsilon) \sum_{N_{\mathcal{L}}(R^{\tilde{\mathcal{L}}})} g_c.$$

Proof: Consider the center $\hat{\ell} \in R^{\tilde{\mathcal{L}}}$ that has in-degree greater than ε^{-1}. Let $\hat{L} = R^{\tilde{\mathcal{L}}} - \{\hat{\ell}\}$. For each $\ell \in \hat{L}$, we associate a center $f(\ell)$ in $R^{\tilde{C^*}}$ in such a way that each $f(\ell) \neq f(\ell')$, for $\ell \neq \ell'$. Note that this is possible since $|\hat{L}| = |R^{\tilde{C^*}}| - 1$. Let $\tilde{f}$ be the center of $R^{\tilde{C^*}}$ that is not associated with any center of $\hat{L}$.

Now, for each center ℓ of $\hat{L}$ we consider the mixed solution $M^\ell = \mathcal{L} - \{\ell\} \cup \{f(\ell)\}$. For each client c, we bound its cost m_c^ℓ in solution M^ℓ. We have

$$m_c^\ell = \begin{cases} g_c & \text{if } c \in N_{C^*}(f(\ell)). \\ \text{Reassign}_c & \text{if } c \in N_{\mathcal{L}}(\ell) - N_{C^*}(f(\ell)). \\ l_c & \text{Otherwise.} \end{cases}$$

Summing over all center $\ell \in \hat{L}$, we have by ε^{-1}-local optimality

$$\sum_{c\in N_{C^*}(R^{\tilde{C^*}})-N_{C^*}(\tilde{f})} l_c + \sum_{\ell\in R^{\tilde{\mathcal{L}}}} \sum_{c\in N_{\mathcal{L}}(\ell)} l_c \leq \sum_{c\in N_{C^*}(R^{\tilde{C^*}})-N_{C^*}(\tilde{f})} g_c + \sum_{\ell\in R^{\tilde{\mathcal{L}}}} \sum_{c\in N_{\mathcal{L}}(\ell)} \text{Reassign}_c. \quad (3)$$

We now complete the proof of the lemma by analyzing the cost of the clients in $N_{C^*}(\tilde{f})$. We consider the center $\ell^* \in \hat{L}$ that minimizes the reassignment cost of its clients. Namely, the center ℓ^* such that $\sum_{c \in N_{\mathcal{L}}(\ell^*)} \text{Reassign}_c$ is minimized. We then consider the solution $M^{(\ell^*,\tilde{f})} = \mathcal{L} - \{\ell^*\} \cup \{\tilde{f}\}$. For each client c, we bound its cost $m_c^{(\ell^*,\tilde{f})}$ in solution $M^{(\ell^*,\tilde{f})}$. We have

$$m_c^{(\ell^*,\tilde{f})} \leq \begin{cases} g_c & \text{if } c \in N_{C^*}(\tilde{f}). \\ \text{Reassign}_c & \text{if } c \in N_{\mathcal{L}}(\ell^*) - N_{C^*}(\tilde{f}). \\ l_c & \text{Otherwise.} \end{cases}$$

Thus, summing over all clients c, we have by local optimality

$$\sum_{c \in N_{C^*}(\tilde{f})} l_c + \sum_{c \in N_{\mathcal{L}}(\ell^*) - N_{C^*}(f(\ell^*))} l_c \leq \sum_{c \in N_{C^*}(\tilde{f})} g_c + \sum_{c \in N_{\mathcal{L}}(\ell^*) - N_{C^*}(f(\ell^*))} \text{Reassign}_c. \quad (4)$$

By Lemma V.4, combining Equations 3 and 4 and averaging over all centers of $\hat{L}$ we have

$$\sum_{c \in N_{C^*}(R^{\tilde{C}^*})} l_c \leq \sum_{c \in N_{C^*}(R^{\tilde{C}^*})} g_c + 2(1+\varepsilon) \sum_{N_{\mathcal{L}}(R^{\tilde{L}})} g_c.$$

■

We now turn to the proof of Theorem V.3.

Proof of Theorem V.3: Observe first that for any $c \in N_{\mathcal{L}}(\tilde{\mathcal{L}}) - N_{C^*}(\tilde{C^*})$, we have $l_c \leq g_c$. This follows from the fact that the center that serves c in C^* is in S and so in $\mathcal{L}$ and thus, we have $l_c \leq g_c$. Therefore

$$\sum_{c \in N_{\mathcal{L}}(\tilde{\mathcal{L}}) - N_{C^*}(\tilde{C^*})} l_c \leq \sum_{c \in N_{\mathcal{L}}(\tilde{\mathcal{L}}) - N_{C^*}(\tilde{C^*})} g_c. \quad (5)$$

We now sum the equations of Lemmas V.6 and V.7 over all pairs and obtain

$$\sum_{(R^{\tilde{L}}, R^{\tilde{C}^*})} \sum_{c \in N_{C^*}(R^{\tilde{C}^*}) \cup N_{\mathcal{L}}(R^{\tilde{L}})} l_c \leq \sum_{(R^{\tilde{L}}, R^{\tilde{C}^*})} \left(\sum_{c \in N_{C^*}(R^{\tilde{C}^*}) \cup N_{\mathcal{L}}(R^{\tilde{L}})} g_c + (2+2\varepsilon) \sum_{N_{\mathcal{L}}(R^{\tilde{L}})} g_c \right)$$

$$\Rightarrow \sum_{c \in N_{C^*}(\tilde{C^*}) \cup N_{\mathcal{L}}(\tilde{\mathcal{L}})} l_c \leq \sum_{c \in N_{C^*}(\tilde{C^*}) \cup N_{\mathcal{L}}(\tilde{\mathcal{L}})} g_c + (2+2\varepsilon) \sum_{c \in N_{\mathcal{L}}(\tilde{\mathcal{L}})} g_c.$$

Therefore,

$$\sum_{c \in N_{C^*}(\tilde{C^*}) - N_{\mathcal{L}}(\tilde{\mathcal{L}})} l_c + \sum_{N_{\mathcal{L}}(\tilde{\mathcal{L}})} l_c \leq \sum_{c \in N_{C^*}(\tilde{C^*}) - N_{\mathcal{L}}(\tilde{\mathcal{L}})} g_c + (3+2\varepsilon) \sum_{c \in N_{\mathcal{L}}(\tilde{\mathcal{L}})} g_c.$$

■

VI. Strong Perturbation-Resilience

Given a set of centers $C = \{c_1, ..., c_k\}$, we define the *induced* clustering of C to be the following partition of the clients $\{S_1, \ldots, S_k\}$ where $S_i = \{a \mid \text{dist}(a, c_i) \leq \text{dist}(a, C)\}$, where ties are broken arbitrarily.

Definition VI.1. *Let $I = (A, F, cost, k)$ be an instance for the k-clustering problem. For $\alpha \geq 1$, I is* α-strong-perturbation-resilient *if there exists a unique optimal set of centers $C^* = \{c_1^*, \ldots, c_k^*\}$ and a unique induced clustering $S^* = \{S_1, \ldots, S_k\}$ and for any instance $I' = (A, F, cost', k, p)$, such that*

$$\forall\ p, q \in \mathcal{P},\ cost(p,q) \leq cost'(p,q) \leq \alpha cost(p,q),$$

the unique optimal set of centers is $C^* = \{c_1^*, \ldots, c_k^*\}$ and the unique induced clustering of C^* is $S^* = \{S_1, \ldots, S_k\}$.

VII. Spectral Separability

In this section we will study the spectral separability condition for the Euclidean k-means problem.

Definition VII.1 (Spectral Separation [33][5])**.** *Let $(A, \mathbb{R}^d, ||\cdot||^2, k)$ be an input for k-means clustering in Euclidean space and let $\{C_1^*, \ldots C_k^*\}$ denote an optimal clustering of A with centers $S = \{c_1^*, \ldots c_k^*\}$. Denote by C an $n \times d$ matrix such that the row $C_i = \underset{c_j^* \in S}{\text{argmin}}\ ||A_i - c_j^*||^2$. Denote by $||\cdot||_2$ the spectral norm of a matrix. Then $\{C_1^*, \ldots C_k^*\}$ is* γ-spectrally separated*, if for any pair (i, j) the following condition holds:*

$$||c_i^* - c_j^*|| \geq \gamma \cdot \left(\frac{1}{\sqrt{|C_i^*|}} + \frac{1}{\sqrt{|C_j^*|}} \right) ||A - C||_2.$$

Nowadays, a standard preprocessing step in Euclidean k-means clustering is to project onto the subspace spanned by the rank k-approximation. Indeed, this is the first step of the algorithm by Kumar and Kannan [33] (see Algorithm 2).

Algorithm 2 k-means with spectral initialization [33]

1: Project points onto the best rank k subspace
2: Compute a clustering C with constant approximation factor on the projection
3: Initialize centroids of each cluster of C as centers in the original space
4: Run Lloyd's k-means until convergence

In general, projecting onto the best rank k subspace and computing a constant approximation on the projection results in a constant approximation in the original space. Kumar and Kannan [33] and later Awasthi and Sheffet [9] gave tighter bounds if the spectral separation is large

[5]The proximity condition of Kumar and Kannan [33] implies the spectral separation condition.

enough. Our algorithm omits steps 3 and 4. Instead, we project onto slightly more dimensions and subsequently use Local Search as the constant factor approximation in step 2. To utilize Local Search, we further require a candidate set of solutions. Due to space constraints and the fact that these types of techniques are fairly standard, we only give high-level details. Roughly speaking, we compute a γ-ball-cover for a sufficiently small γ depending only on ε and β. Such ball covers preserve cost and stability up to small multiplicative constant factors, and typically have size $O(\gamma^{-d})$. To reduce the dependency on the dimension, we then apply Johnson-Lindenstrauss type embeddings, leading to a set of candidate solutions of size $n^{\text{poly}(\varepsilon^{-1})}$. For pseudocode, we refer to Algorithm 3. Our main result is to show that, given spectral separability, this algorithm is PTAS for k-means (Theorem VII.2).

Algorithm 3 SpectralLS

1: Project points A onto the best rank k/ε subspace
2: Embed points into a random subspace of dimension $O(\varepsilon^{-2}\log n)$
3: Local Search($\Theta(\varepsilon^{-4})$)
4: Output clustering

Theorem VII.2. *Let $(A, \mathbb{R}^d, ||\cdot||^2, k)$ be an instance of Euclidean k-means clustering with optimal clustering $C = \{C_1^*, \dots C_k^*\}$ and centers $S = \{c_1^*, \dots c_k^*\}$. If C is more than $3\sqrt{k}$-spectrally separated, Algorithm 3 is a polynomial time approximation scheme.*

We first recall the basic notions and definitions for Euclidean k-means. Let $A \in \mathbb{R}^{n\times d}$ be a set of points in d-dimensional Euclidean space, where the row A_i contains the coordinates of the ith point. The singular value decomposition is defined as $A = U\Sigma V^T$, where $U \in \mathbb{R}^{n\times d}$ and $V \in \mathbb{R}^{d\times d}$ are orthogonal and $\Sigma \in \mathbb{R}^{d\times d}$ is a diagonal matrix containing the singular values where per convention the singular values are given in descending order, i.e. $\Sigma_{1,1} = \sigma_1 \geq \Sigma_{2,2} = \sigma_2 \geq \dots \Sigma_{d,d} = \sigma_d$. Denote the Euclidean norm of a d-dimensional vector x by $||x|| = \sqrt{\sum_{i=1}^d x_i^2}$. The spectral norm and Frobenius norm are defined as $||A||_2 = \sigma_1$ and $||A||_F = \sqrt{\sum_{i=1}^d \sigma_i^2}$, respectively.

The best rank k approximation $\min_{\text{rank}(X)=k} ||A - X||_F$ is given via $A_k = U_k\Sigma V^T = U\Sigma_k V^T = U\Sigma V_k^T$, where U_k, Σ_k and V_k^T consist of the first k columns of U, Σ and V^T, respectively, and are zero otherwise. The best rank k approximation also minimizes the spectral norm, that is $||A - A_k||_2 = \sigma_{k+1}$ is minimal among all matrices of rank k. The following fact is well known throughout k-means literature and will be used frequently throughout this section.

Fact VII.3. *Let A be a set of points in Euclidean space and denote by $c(A) = \frac{1}{|A|}\sum_{x\in A} x$ the centroid of A. Then the 1-means cost of any candidate center c can be decomposed via*

$$\sum_{x\in A} ||x - c||^2 = \sum_{x\in A} ||x - c(A)||^2 + |A| \cdot ||c(A) - c||^2$$

and

$$\sum_{x\in A} ||x - c(A)||^2 = \frac{1}{2\cdot|A|}\sum_{x\in A}\sum_{y\in A} ||x - y||^2.$$

Note that the centroid is the optimal 1-means center of A. For a clustering $C = \{C_1, \dots C_k\}$ of A with centers $S = \{c_1, \dots c_k\}$, the cost is then $\sum_{i=1}^k \sum_{p\in C_i} ||p - c_i||^2$. Further, if $c_i = \frac{1}{|C_i|}\sum_{p\in C_i} p$, we can rewrite the objective function in matrix form by associating the ith point with the ith row of some matrix A and using the cluster matrix $X \in \mathbb{R}^{n\times k}$ with $X_{i,j} = \begin{cases} \frac{1}{\sqrt{|C_j^*|}} & \text{if } A_i \in C_j^* \\ 0 & \text{else} \end{cases}$ to denote membership. Note that $X^TX = I$, i.e. X is an orthogonal projection and that $||A - XX^TA||_F^2$ is the cost of the optimal k-means clustering. k-means is therefore a constrained rank k-approximation problem.

We first restate the separation condition.

Definition VII.4 (Spectral Separation)**.** *Let A be a set of points and let $\{C_1, \dots C_k\}$ be a clustering of A with centers $\{c_1, \dots c_k\}$. Denote by C an $n \times d$ matrix such that $C_i = \underset{j\in\{1,\dots,k\}}{\text{argmin}} ||A_i - c_j||^2$. Then $\{C_1, \dots C_k\}$ is* γ spectrally separated*, if for any pair of centers c_i and c_j the following condition holds:*

$$||c_i - c_j|| \geq \gamma \cdot \left(\frac{1}{\sqrt{|C_i|}} + \frac{1}{\sqrt{|C_j|}}\right) ||A - C||_2.$$

The following crucial lemma relates spectral separation and distribution stability.

Lemma VII.5. *For a point set A, let $C = \{C_1, \dots, C_k\}$ be an optimal clustering with centers $S = \{c_1, \dots, c_k\}$ associated clustering matrix X that is at least $\gamma \cdot \sqrt{k}$ spectrally separated, where $\gamma > 3$. For $\varepsilon > 0$, let A_m be the best rank $m = k/\varepsilon$ approximation of A. Then there exists a clustering $K = \{C_1', \dots C_2'\}$ and a set of centers S_k, such that*

1) the cost of clustering A_m with centers S_k via the assignment of K is less than $||A_m - XX^TA_m||_F^2$ and
2) (K, S_k) is $\Omega((\gamma - 3)^2 \cdot \varepsilon)$-distribution stable.

We note that this lemma would also allow us to use the PTAS of Awasthi et al. [6]. Before giving the proof, we outline how Lemma VII.5 helps us prove Theorem VII.2. We first notice that if the rank of A is of order k, then elementary bounds on matrix norm show that spectral separability implies distribution stability. We aim to combine this observation with the following theorem due to Cohen et al. [22]. Informally, it states that for every rank k approximation, (an in particular for every constrained rank

k approximation such as k-means clustering), projecting to the best rank k/ε subspace is cost-preserving.

Theorem VII.6 (Theorem 7 of [22]). *For any $A \in \mathbb{R}^{n\times d}$, let A' be the rank $\lceil k/\varepsilon\rceil$-approximation of A. Then there exists some positive number c such that for any rank k orthogonal projection P,*

$$||A-PA||_F^2 \leq ||A'-PA'||_F^2+c \leq (1+\varepsilon)||A-PA||_F^2.$$

The combination of the low rank case and this theorem is not trivial as points may be closer to a wrong center after projecting. Lemma VII.5 determines the existence of a clustering whose cost for the projected points A_m is at most the cost of C^*. Moreover, this clustering has constant distribution stability as well which, allows us to use Local Search. Given that we can find a clustering with cost at most $(1+\varepsilon)\cdot||A_m-XX^TA_m||_F^2$, Theorem VII.6 implies that we will have a $(1+\varepsilon)^2$-approximation overall.

To prove the lemma, we will require the following steps:

- A lower bound on the distance of the projected centers $||c_iV_mV_m^T - c_jV_mV_m^T|| \approx ||c_i - c_j||$.
- Find a clustering K with centers $S_m^* = \{c_1V_mV_m^T, \ldots, c_k^*V_mV_m^T\}$ of A_m with cost less than $||A_m - XX^TA_m||_F^2$.
- Show that in a well-defined sense, K and C^* agree on a large fraction of points.
- For any point $x \in K_i$, show that the distance of x to any center not associated with K_i is large.

We first require a technical statement.

Lemma VII.7. *For a point set A, let $C = \{C_1, \ldots C_k\}$ be a clustering with associated clustering matrix X and let A' and A'' be optimal low rank approximations where without loss of generality $k \leq$ rank$(A') <$ rank(A''). Then for each cluster C_i*

$$\left|\left|\frac{1}{|C_i|}\sum_{j\in C_i}\left(A_j''-A_j'\right)\right|\right|_2 \leq \sqrt{\frac{k}{|C_i|}}\cdot||A-XX^TA||_2.$$

Proof: By Fact VII.3 $|C_i|\cdot||\frac{1}{|C_i|}\sum_{j\in C_i}(A_i''-A_i')||_2^2$ is, for a set of point indexes C_i, the cost of moving the centroid of the cluster computed on A'' to the centroid of the cluster computed on A'. For a clustering matrix X, $||XX^TA'-XX^TA'||_F^2$ is the sum of squared distances of moving the centroids computed on the point set A'' to the centroids computed on A'. We then have

$$\begin{aligned}&|C_i|\cdot\left|\left|\tfrac{1}{|C_i|}\textstyle\sum_{j\in C_i}(A_j''-A_j')\right|\right|_2^2\\ \leq\ &||XX^TA''-XX^TA'||_F^2 \leq ||X||_F^2\cdot||A''-A'||_2^2\\ \leq\ &k\cdot\sigma_{k+1}^2 \leq k\cdot||A-XX^TA||_2^2.\end{aligned}$$

■

Proof of Lemma VII.5: For any point p associated with some row of A, let $p^m = pV_mV_m^T$ be the corresponding row in A_m. Similarly, for some cluster C_i, denote the center in A by c_i and the center in A_m by c_i^m. Extend these notion analogously for projections p^k and c_i^k to the span of the best rank k approximation A_k.

We have for any $m \geq k$ $i \neq j$

$$\begin{aligned}&||c_i^m - c_j^m|| \geq\\ &||c_i - c_j|| - ||c_i - c_i^m|| - ||c_j - c_j^m|| \geq\\ &\gamma\cdot\left(\tfrac{1}{\sqrt{|C_i|}}+\tfrac{1}{\sqrt{|C_j|}}\right)\sqrt{k}||A-XX^TA||_2\\ &-\left(\tfrac{1}{\sqrt{|C_i|}}+\tfrac{1}{\sqrt{|C_j|}}\right)\sqrt{k}||A-XX^TA||_2 =\\ &(\gamma-1)\cdot\left(\tfrac{1}{\sqrt{|C_i|}}+\tfrac{1}{\sqrt{|C_j|}}\right)\sqrt{k}||A-XX^TA||_2, \quad (6)\end{aligned}$$

where the second inequality follows from Lemma VII.7.

In the following, let $\Delta_i = \frac{\sqrt{k}}{\sqrt{|C_i|}}||A-XX^TA||_2$. We will now construct our target clustering K. Note that we require this clustering (and its properties) only for the analysis. We distinguish between the following three cases.

Case 1: $p \in C_i$ **and** $c_i^m = \underset{j\in\{1,\ldots,k\}}{\textbf{argmin}}\ ||p^m - c_j||$

These points remain assigned to c_i^m. The distance between p_m and a different center c_j^m is at least $\frac{1}{2}||c_i^m - c_j^m|| \geq \frac{\gamma-1}{2}\varepsilon(\Delta_i+\Delta_j)$ due to Equation 6.

Case 2: $p \in C_i$, $c_i^m \neq \underset{j\in\{1,\ldots,k\}}{\textbf{argmin}}\ ||p^m - c_j||$,**and** $c_i^k \neq \underset{j\in\{1,\ldots,k\}}{\textbf{argmin}}\ ||p^k - c_j^k||$

These points will get reassigned to their closest center. The distance between p_m and a different center c_j^m is at least $\frac{1}{2}||c_i^m - c_j^m|| \geq \frac{\gamma-1}{2}(\Delta_i+\Delta_j)$ due to Equation 6.

Case 3: $p \in C_i$, $c_i^m \neq \underset{j\in\{1,\ldots,k\}}{\textbf{argmin}}\ ||p^m - c_j^m||$,**and** $c_i^k = \underset{j\in\{1,\ldots,k\}}{\textbf{argmin}}\ ||p^k - c_j^k||$

We assign p^m to c_i^m at the cost of a slightly weaker movement bound on the distance between p^m and c_j^m. Due to orthogonality of V, we have for $m > k$, $(V_m - V_k)^TV_k = V_k^T(V_m - V_k) = 0$. Hence $V_mV_m^TV_k = V_mV_k^TV_k + V_m(V_m - V_k)^TV_k = V_kV_k^TV_k + (V_m - V_k)V_k^TV_k = V_kV_k^TV_k = V_k$. Then $p^k = pV_kV_k^T = pV_mV_m^TV_kV_k^T = p_mV_kV_k^T$.

Further, $||p^k - c_j^k|| \geq \frac{1}{2}||c_j^k - c_i^k|| \geq \frac{\gamma-1}{2}(\Delta_i+\Delta_j)$ due to Equation 6. Then the distance between p_m and a different center c_j^m

$$\begin{aligned}&||p^m - c_j^m||\\ \geq\ &||p^m - c_j^k|| - ||c_j^m - c_j^k||\\ =\ &\sqrt{||p^m - p^k||^2 + ||p^k - c_j^k||^2} - ||c_j^m - c_j^k||\\ \geq\ &||p^k - c_j^k|| - \Delta_j \geq \frac{\gamma-3}{2}(\Delta_i+\Delta_j),\end{aligned}$$

where the equality follows from orthogonality and the second to last inequality follows from Lemma VII.7.

Now, given the centers $\{c_1^m, \ldots c_k^m\}$, we obtain a center matrix M_K where the ith row of M_K is the center according to the assignment of above.

Since both clusterings use the same centers but K improves locally on the assignments, we have $||A_m - M_K||_F^2 \leq ||A_m - XX^TA_m||_F^2$, which proves the first statement of the lemma. Additionally, due to the fact that $A_m - XX^TA_m$ has rank $m = k/\varepsilon$, we have

$$\begin{aligned} & ||A_m - M_K||_F^2 \leq ||A_m - XX^TA_m||_F^2 \\ \leq \; & m \cdot ||A_m - XX^TA_m||_2^2 \\ \leq \; & k/\varepsilon \cdot ||A - XX^TA||_F^2 \end{aligned} \tag{7}$$

To ensure stability, we will show that for each element of K there exists an element of C, such that both clusters agree on a large fraction of points. This can be proven by using techniques from Awasthi and Sheffet [9] (Theorem 3.1) and Kumar and Kannan [33] (Theorem 5.4), which we repeat for completeness.

Lemma VII.8. *Let $K = \{C'_1, \dots C'_k\}$ and $C = \{C_1, \dots C_k\}$ be defined as above. Then there exists a bijection $b : C \to K$ such that for any $i \in \{i, \dots, k\}$*

$$\left(1 - \frac{32}{(\gamma-1)^2}\right)|C_i| \leq b(|C_i|) \leq \left(1 + \frac{32}{(\gamma-1)^2}\right)|C_i|.$$

Proof: Denote by $T_{i\to j}$ the set of points from C_i such that $||c_i^k - p^k|| > ||c_j^k - p^k||$. We first note that $||A_k - XX^TA||_F^2 \leq 2k \cdot ||A_k - XX^TA||_2^2 \leq 2k \cdot \left(||A - A_k||_2 + ||A - XX^TA||_2\right)^2 \leq 8k \cdot ||A - XX^TA||_2^2 \leq 8 \cdot |C_i| \cdot \Delta_i^2$ for any $i \in \{1, \dots, k\}$. The distance $||p^k - c_i^k|| \geq \frac{1}{2}||c_i^k - c_j^k|| \geq \frac{\gamma-1}{2} \cdot \left(\frac{1}{\sqrt{C_i}} + \frac{1}{\sqrt{|C_j|}}\right)\sqrt{k}||A - XX^TA||_2^2$. Assigning these points to c_i^k, we can bound the total number of points added to and subtracted from cluster C_j by observing

$$\begin{aligned} \Delta_j^2 \textstyle\sum_{i\neq j} |T_{i\to j}| &\leq \textstyle\sum_{i\neq j} |T_{i\to j}| \cdot \left(\frac{\gamma-1}{2}\right)^2 \cdot (\Delta_i + \Delta_j)^2 \\ &\leq ||A_k - XX^TA||_F^2 \leq 8 \cdot |C_j| \cdot \Delta_j^2 \end{aligned}$$

and analogously

$$\Delta_j^2 \sum_{i\neq j} |T_{j\to i}| \leq 8 \cdot |C_j| \cdot \Delta_j^2.$$

Therefore, the cluster sizes are up to some multiplicative factor of $\left(1 \pm \frac{32}{(\gamma-1)^2}\right)$ identical. ■

We now have for each point $p^m \in C'_i$ a minimum distance of

$$\begin{aligned} & ||p^m - c_j^m|| \\ \geq \; & \frac{\gamma-3}{2} \cdot \left(\frac{1}{\sqrt{|C_i|}} + \frac{1}{\sqrt{|C_j|}}\right) \cdot \sqrt{k} \cdot ||A - XX^TA||_2 \\ \geq \; & \frac{\gamma-3}{2} \cdot \cdot \sqrt{k} \cdot ||A - XX^TA||_2 \cdot \\ & \left(\sqrt{\frac{1}{\left(1 + \frac{32}{(\gamma-1)^2}\right) \cdot |C'_i|}} + \sqrt{\frac{1}{\left(1 + \frac{32}{(\gamma-1)^2}\right) \cdot |C'_j|}}\right). \end{aligned}$$

where the first inequality holds due to Case 3, the second inequality holds due to Lemma VII.8. Finally, due to $\gamma > 3$ and Equation 7, this shows that the cost of assigning p^m to c_j^m is at least

$$\frac{4 \cdot (\gamma-3)^2}{81} \cdot \varepsilon \frac{||A_m - M_K||_F^2}{|C'_j|}.$$

This ensures that the distribution stability condition is satisfied. ■

Proof of Theorem VII.2: Given the optimal clustering C^* of A with clustering matrix X, Lemma VII.5 guarantees the existence of a clustering K with center matrix M_K such that $||A_m - M_K||_F^2 \leq ||A_m - XX^TA_m||$ and that C has constant distribution stability. If $||A_m - M_K||_F^2$ is not a constant factor approximation, we are already done, as Local Search is guaranteed to find a constant factor approximation. By Theorem IV.2, Local Search with appropriate (but constant) neighborhood size will find a clustering C' with cost at most $(1+\varepsilon)$ times the cost of K in $(A_m, F, ||\cdot||^2, k)$. Let Y be the clustering matrix of C'. We then have $||A_m - YY^TA_m||_F^2 + ||A - A_m||_F^2 \leq (1+\varepsilon)||A_m - M_K||_F^2 + ||A - A_m||_F^2 \leq (1+\varepsilon)^2||A_m - XX^TA_m||_F^2 + ||A - A_m||_F^2 \leq (1+\varepsilon)^3||A - XX^TA||_F^2$ due to Theorem VII.6. Rescaling ε completes the proof. ■

References

[1] E. Aarts and J. K. Lenstra, editors. *Local Search in Combinatorial Optimization*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1997.

[2] S. Ahmadian, A. Norouzi-Fard, O. Svensson, and J. Ward. Better guarantees for k-means and euclidean k-median by primal-dual algorithms. *CoRR*, abs/1612.07925, 2016.

[3] H. Angelidakis, K. Makarychev, and Y. Makarychev. Algorithms for stable and perturbation-resilient problems. In *Proc. of STOC*, pages 438–451, 2017.

[4] D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *Proc. of SODA*, pages 1027–1035, 2007.

[5] V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Munagala, and V. Pandit. Local search heuristics for k-median and facility location problems. *SIAM J. Comput.*, 33(3):544–562, 2004.

[6] P. Awasthi, A. Blum, and O. Sheffet. Stability yields a PTAS for k-median and k-means clustering. In *Proc. of FOCS*, pages 309–318, 2010.

[7] P. Awasthi, A. Blum, and O. Sheffet. Center-based clustering under perturbation stability. *Inf. Process. Lett.*, 112(1-2):49–54, 2012.

[8] P. Awasthi, M. Charikar, R. Krishnaswamy, and A. Kemal Sinop. The hardness of approximation of Euclidean k-means. In *Proc. of SoCG*, pages 754–767, 2015.

[9] P. Awasthi and O. Sheffet. Improved spectral-norm bounds for clustering. In *Proc. of APPROX*, pages 37–49, 2012.

[10] M.-F. Balcan, A. Blum, and A. Gupta. Approximate clustering without the approximation. In *Proc. of SODA*, pages 1068–1077, 2009.

[11] M.-F. Balcan, N. Haghtalab, and C. White. k-center clustering under perturbation resilience. In *Proc. of ICALP*, pages 68:1–68:14, 2016.

[12] M.-F. Balcan and Y. Liang. Clustering under perturbation resilience. *SIAM J. Comput.*, 45(1):102–155, 2016.

[13] M.-F. Balcan, H. Röglin, and S.-H. Teng. Agnostic clustering. In *Proc. of ALT*, pages 384–398, 2009.

[14] Maria-Florina Balcan, Avrim Blum, and Anupam Gupta. Clustering under approximation stability. *J. ACM*, 60(2):8, 2013.

[15] Maria-Florina Balcan and Mark Braverman. Finding low error clusterings. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

[16] Maria-Florina Balcan and Colin White. Clustering under local stability: Bridging the gap between worst-case and beyond worst-case analysis. *CoRR*, abs/1705.07157, 2017.

[17] S. Ben-David, B. Chor, O. Goldreich, and M. Luby. On the theory of average case complexity. *J. Comput. Syst. Sci.*, 44(2):193–219, 1992.

[18] S. Ben-David and L. Reyzin. Data stability in clustering: A closer look. *Theor. Comput. Sci.*, 558:51–61, 2014.

[19] Yonatan Bilu and Nathan Linial. Are stable instances easy? *Combinatorics, Probability & Computing*, 21(5):643–660, 2012.

[20] V. Braverman, A. Meyerson, R. Ostrovsky, A. Roytman, M. Shindler, and B. Tagiku. Streaming k-means on well-clusterable data. In *Proc. of SODA*, pages 26–40, 2011.

[21] J. Byrka, T. Pensyl, B. Rybicki, A. Srinivasan, and K. Trinh. An improved approximation for *k*-median, and positive correlation in budgeted optimization. In *Proc. of SODA*, pages 737–756, 2015.

[22] M. B. Cohen, S. Elder, C. Musco, C. Musco, and M. Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proc. of STOC*, pages 163–172, 2015.

[23] V. Cohen-Addad, P. N. Klein, and C. Mathieu. Local search yields approximation schemes for k-means and k-median in euclidean and minor-free metrics. In *Proc. of FOCS*, pages 353–364, 2016.

[24] Z. Friggstad, M. Rezapour, and M. R. Salavatipour. Local search yields a PTAS for k-means in doubling metrics. In *Proc. of FOCS*, pages 365–374, 2016.

[25] S. Guha and S. Khuller. Greedy strikes back: Improved facility location algorithms. *J. Algorithms*, 31(1):228–248, 1999.

[26] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams: Theory and practice. *IEEE Trans. Knowl. Data Eng.*, 15(3):515–528, 2003.

[27] A. Gupta and K. Tangwongsan. Simpler analyses of local search algorithms for facility location. *CoRR*, abs/0809.2554, 2008.

[28] V. Guruswami and P. Indyk. Embeddings and non-approximability of geometric problems. In *Proc. of SODA*, pages 537–538, 2003.

[29] S. Har-Peled and A. Kushal. Smaller coresets for k-median and k-means clustering. *Discrete & Computational Geometry*, 37(1):3–19, 2007.

[30] R. Jaiswal and N. Garg. Analysis of k-means++ for separable data. In *Proc. of APPROX*, pages 591–602, 2012.

[31] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. A local search approximation algorithm for k-means clustering. *Comput. Geom.*, 28(2-3):89–112, 2004.

[32] S. G. Kolliopoulos and S. Rao. A nearly linear-time approximation scheme for the euclidean k-median problem. *SIAM J. Comput.*, 37(3):757–782, June 2007.

[33] A. Kumar and R. Kannan. Clustering with spectral norm and the k-means algorithm. In *Proc. of FOCS*, pages 299–308, 2010.

[34] A. Kumar, Y. Sabharwal, and S. Sen. Linear-time approximation schemes for clustering problems in any dimensions. *J. ACM*, 57(2), 2010.

[35] M. Mahajan, P. Nimbhorkar, and K. R. Varadarajan. The planar k-means problem is NP-hard. *Theor. Comput. Sci.*, 442:13–21, 2012.

[36] N. Megiddo and K. J. Supowit. On the complexity of some common geometric location problems. *SIAM J. Comput.*, 13(1):182–196, 1984.

[37] R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy. The effectiveness of Lloyd-type methods for the k-means problem. *J. ACM*, 59(6):28, 2012.

[38] D. A. Spielman and S.-H. Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *J. ACM*, 51(3):385–463, 2004.