Clustering under Perturbation Stability in Near-Linear Time*

Pankaj K. Agarwal† Hsien-Chih Chang* Kamesh Munagala‡
Erin Taylor* Emo Welzl§

July 9, 2019

Abstract

We consider the problem of center-based clustering in low-dimensional Euclidean spaces under the perturbation stability assumption. An instance is α-stable if the underlying optimal clustering continues to remain optimal even when all pairwise distances are arbitrarily perturbed by a factor of at most α. Our main contribution is in presenting efficient exact algorithms for α-stable instances whose running times depend near-linearly on the size of the data set for all clustering objectives when $\alpha \geq 2+\sqrt{3}$. For k-center and k-means problems, our algorithms also achieve polynomial dependence on the number of clusters, k, when $\alpha \geq 2+\sqrt{3}+\Omega(1)$ in any fixed number of dimensions. For k-medians, our algorithms have polynomial dependence on k when $\alpha > 5$ in any fixed dimensions; and when $\alpha \geq 2+\sqrt{3}$ in two dimensions. Our algorithms are simple, and only require applying techniques such as local search or dynamic programming to a suitably modified metric space, combined with careful choice of data structures.

*Work on this paper was supported by NSF under grants CCF-14-08784, CCF-15-13816, CCF-15-46392, CCF-16-37397, IIS-14-08846, and IIS-1447554, by an ARO grant W911NF-15-1-0408, by an BSF Grant 2012/229 from the U.S.-Israel Binational Science Foundation, and research awards from Adobe and Facebook.

†Department of Computer Science, Duke University, USA.

‡Department of Computer Science, Duke University, USA. Email: kamesh@cs.duke.edu

§Department of Computer Science, ETH Zürich, Switzerland.

1 Introduction

Clustering is a fundamental problem in unsupervised learning and data summarization, with wide ranging applications that span myriad areas. Typically, the data points are assumed to lie in a Euclidean space, and the goal in center-based clustering is to open a set of k centers to minimize the objective cost, usually a function over the distance from each data point to its closest center. The k-medians objective minimizes the sum of distances; the k-means objective minimizes the sum of squares; and the k-center objective minimizes the longest distance. In the worst case, all these objectives are NP-hard to optimize even in the plane [43, 45].

A substantial body of work has focused on developing polynomial-time approximation algorithms and analyzing natural heuristics for these problems. Given the sheer size of data sets we are confronted with these days, such as those generated in genomics or mapping applications, even a polynomial-time algorithm may be too slow to be useful in practice—even computing all pairs of distances can be computationally burdensome. What we need are algorithms whose running time is near-linear in the input size and polynomial in the number of clusters. For instance, the reason Lloyd's heuristic [42] for k-means is widely used is precisely because it circumvents the all-pairs distance bottleneck. Consequently, near-linear time algorithms for clustering have been widely studied.

An alternative approach to addressing the worst-case intractability of clustering is the "beyond worst-case" analysis approach that makes reasonable assumptions on the data itself. Here we consider the notion of *stability* that has emerged as a popular way of circumventing the worst-case intractability of the problem. In these notions, the assumption is that any "small perturbation" of input points does not change the optimal solution. This is natural in real datasets, where often, the optimal clustering is clearly demarcated, and the distances are obtained heuristically. Different notions of stability differ in how "small perturbation" is defined. In this paper, we are particularly interested in the notions of stability along the line of Bilu and Linial [20] and Awasthi, Blum, and Sheffet [12]. A clustering instance is *α-perturbation resilient* or *α-stable* if the optimal clustering does not change when all distances are perturbed by a factor of at most α. Similarly, a clustering instance is *α-center proximal* if any point is at least a factor of α closer to its own optimal center than any other optimal center. The focus of this line of work has been to design algorithms that recover the *exact* optimal clustering—the ground truth—in polynomial time for α-stable instances and making α as small as possible.

We focus on recovering the optimal clustering for stable clustering instances as well. But instead of optimizing the value of α, we ask the question: *Can algorithms be designed that compute exact solutions to stable instances of Euclidean center-based clustering that run in time near-linear in the input size?* We note that an $(1+\varepsilon)$-approximation solution, for an arbitrarily small constant $\varepsilon > 0$, may differ significantly from an optimal solution (the ground truth) even for stable instances, so one cannot hope to use an approximation algorithm to recover the optimal clustering.

1.1 Our Results

In this paper, we answer the above question in the affirmative, and present near-linear time algorithms for finding optimal solutions of stable clustering instances. In particular, we show the following meta-theorem:

Theorem 1.1. *Let X be a set of n points in $\mathbb{R}^d$ for some constant d, let $k \geq 1$ be an integer, and let $\alpha \geq 2+\sqrt{3}$ be a parameter. If the k-medians, k-means, or k-center clustering instance for X is α-stable, then the optimal solution can be computed in $\tilde{O}(n \operatorname{poly} k + f(k))$ time.*

In the above theorem, the $\tilde{O}$ notation suppresses logarithmic terms in n and the spread of the point set. The function $f(k)$ depends on the choice of algorithm, and we present the exact dependence below.

We also omit terms depending solely the dimension, d. Furthermore, the above theorem is robust in the sense that the algorithm is not restricted to choosing the input points as centers (*discrete setting*), and can potentially choose arbitrary points in the Euclidean plane as centers (*continuous setting*)—indeed, we show that these notions are identical under reasonable assumption on stability.

At a more fine-grained level, we present several algorithms that require mild assumptions on the stability condition. In the results below, as well as throughout the paper, we present our results both for the Euclidean plane, as well as generalizations to higher (but fixed number of) dimensions.

Dynamic Programming. In Section 3, we present a dynamic programming algorithm that computes the optimal clustering in $O(nk^2 + n \operatorname{polylog} n)$ time for α-stable k-means and k-center in any fixed dimension, provided that $\alpha \geq 2 + \sqrt{3} + \Omega(1)$.

Local Search. In Sections 4 and 5, we show that the standard 1-swap local search algorithm, which iteratively swaps out a center in the current solution for a new center as long as the resulting total cost improves, computes an optimal clustering for α-stable instances of k-medians assuming $\alpha > 5$. We also show that it can be implemented in $O(nk^2 \log^3 n \log \Delta)$ for $d = 2$ and in $O(nk^{2d-1} \operatorname{polylog} n \log \Delta)$ for $d > 2$, where Δ is the spread of the point set.

Coresets. In Section 6, we use multiplicative core-sets to compute the optimal clustering for k-medians in any fixed dimension, when $\alpha \geq 2 + \sqrt{3}$. The running time is $O(nk^2 + f(k))$ where $f(k)$ is an exponential function of k. The same result also holds for k-means and k-center objectives.

Techniques. The key difficulty with developing fast algorithms for computing the optimal clustering is that some clusters could have a very small size compared to others. This issue persists even when the instances are stable. Imagine a scenario where there are multiple small clusters, and an algorithm must decide whether to merge these into one cluster while splitting some large cluster, or keep them intact. Now imagine this situation happening recursively, so that that the algorithm has multiple choices about which clusters to recursively split. The difference in cost between these many options, and the size of the small clusters, can be small enough that any $(1+\varepsilon)$-approximation can be agnostic, while an exact solution cannot. This issue has caused existing work on finding exact optima either to use dynamic programming [8] or local search with large number of swaps [23, 32]; or to make assumptions lower-bounding the size of the optimal clusters or the spread of their centers [30].

Our main technical insight for the first two results is simple in hindsight, yet powerful: For a stable instance, if the Euclidean metric is replaced by another metric that is a good approximation, then the optimal clustering does not change under the new metric and in fact the instance remains stable albeit with a smaller perturbation parameter.

In d-dimensions, the L_1-metric is within a $\sqrt{d}$ factor of the Euclidean metric, the advantage of using the L_1-metric is that it is amenable to efficient data structures. To avoid the $\sqrt{2}$ loss in 2-dimensions and the $\sqrt{d}$ loss in α in higher dimensions, we replace the Euclidean metric with an appropriate *polyhedral metric*—that is, a convex distance function where each unit ball is a regular polyhedron—yielding efficient procedures for the following primitives:

- **Cost of 1-swap.** Given a candidate set of centers S, maintain a data structure that efficiently updates the total cost if center $x \in S$ is replaced by center $y \notin S$.
- **Cost of 1-clustering.** Given a partition of the data points, maintain a data structure where the cost of 1-clustering (under any objectives) can be efficiently updated as partitions are merged.

The first primitive is required for implementing the 1-swap local search, while the second primitive arises in the dynamic programming based algorithm.

We next combine the insight of changing the metrics with additional techniques. For local search, we build on the approach in [23, 27, 32] that shows local search with t-swaps for large enough constant t finds an optimal solution for stable instances in polynomial time for any fixed-dimension Euclidean space. Their analysis does not extend as is to 1-swap, which is critical in achieving near-linear running time—note that even when $t = 2$ there are a quadratic number of candidate swaps per step. As mentioned in the introduction of Friggstad *et al.* [32], such extension is possible if one is satisfied with merely polynomial time; we provide a complete proof of the 1-swap case based on the candidate swaps defined in Arya *et al.* [10], supplied with an efficient data structure to answer the cost of any candidate swap in time linear in k and logarithmic in n.

For the dynamic programming algorithm, we use the following insight: In Euclidean spaces, for $\alpha \geq 2 + \sqrt{3}$, the longest edge of the minimum spanning tree over the input points partitions the data set in two, such that any optimal cluster lies completely in one of the two sides of the partition. Combined with the change of metrics one can achieve near-linear running time.

Our analysis of the 1-swap local search for k-medians requires $\alpha > 5$. To handle smaller settings of α, we develop a core-set based algorithm that computes the optimal clustering in near-linear time provided that $\alpha \geq 2 + \sqrt{3}$. We note that core-sets have been used extensively to develop near-linear time approximation algorithms for clustering, but it is surprising that they can also be used to compute the optimal clustering for stable instances. The advantage of this approach is that it provides a unified approach to handle all of k-medians, k-center, and k-means, albeit with running time that depends exponentially on k.

We conclude this discussion by commenting that our analysis of all three approaches might not be tight, and it is possibly that variants of all three approaches work for smaller values of α.

1.2 Related Work

All of k-medians, k-means, and k-center are widely studied from the perspective of approximation algorithms and are known to be hard to approximate. Indeed, for general metric spaces, k-center is hard to approximate to within a factor of $2-\varepsilon$ [37]; k-medians is hard to $(1+2/e)$-approximate [38]; and k-means is hard to approximate within a factor of 1.0013 [41]. Even when the metric space is Euclidean, k-means is still NP-hard when $k = 2$ [6, 28], and there is an $n^{\Omega(k)}$ lower bound on running time for k-medians and k-means in 4-dimensional Euclidean space under the exponential-time hypothesis [24].

There is a long line of work in developing $(1+\varepsilon)$-approximations for these problems in Euclidean spaces. The holy grail of this work has been the development of algorithms that are near-linear time in n, and several techniques are now known to achieve this. This includes randomly shifted quad-trees [9], coresets [2, 35, 36] and local search [23, 25, 26], among others.

There are many notions of clustering stability that have been considered in literature [1, 5, 11, 14, 15, 19, 31, 40, 47]. The exact definition of stability we study here was first introduced in Awasthi *et al.* [12]; their definition in particular resembles the one of Bilu and Linial [20] for max-cut problem, which later has been adapted to other optimization problems [7, 8, 16, 44, 46]. Building on a long line of work [12, 13, 17, 18], which gradually reduced the stability parameter, Angelidakis *et al.* [8] present a dynamic programming based exact polynomial time algorithm for discrete 2-stable instances for all center-based objectives.

Chekuri and Gupta [22] show that natural LP-relaxation is integral for 2-stable k-center problem. Recent work by Cohen-Addad [27] provides a framework for analyzing local search algorithms for stable instances. This work shows that for an α-stable instance when $\alpha > 3$, any solution is the optimal

solution if it cannot be improved by swapping $\lceil 2/(\alpha-3)\rceil$ centers. Focusing on Euclidean spaces of fixed dimensions, Friggstad *et al.* [32] show that local search algorithm with $O(1)$-swaps runs in polynomial time under even an extremely mild $(1+\delta)$-stable assumption for any $\delta > 0$. However, none of the algorithms for stable instances of clustering so far have running time near-linear in n, even when the stability parameter α is large, points lie in $\mathbb{R}^2$, and the underlying metric is Euclidean.

On the hardness side, solving $(3-\delta)$-center proximal k-medians instances in general metric spaces is NP-hard for any $\delta > 0$ [12]. When restricted to Euclidean spaces in arbitrary dimensions, Ben-David and Reyzin [19] showed that for every $\delta > 0$, the problem for discrete $(2-\delta)$-center proximal k-medians instances is NP-hard. Similarly, the clustering problem for discrete k-center remains hard for α-stable instances when $\alpha < 2$, assuming standard complexity assumption that NP $\neq$ RP [17]. Under the same complexity assumption, discrete α-stable k-means is also hard when $\alpha < 1+\delta_0$ for some positive constant δ_0 [32]. Deshpande *et al.* [30] showed it is NP-hard to $(1+\varepsilon)$-approximate $(2-\delta)$-center proximal k-means instances.

2 Definitions and Preliminaries

Due to constraint on the length of the paper, most proofs in this section have been moved to Appendix A.

Clustering. Let $X = \{p_1, \ldots, p_n\}$ be a set of n points in $\mathbb{R}^d$, and let $\delta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ be a distance function (not necessarily a metric satisfying triangle inequality). For a set $Y \subseteq \mathbb{R}^d$, we define $\boldsymbol{\delta(p, Y)} := \min_{y \in Y} \delta(p, y)$. A ***k*-clustering** of X is a partition of X into k non-empty **clusters** $X_1, \ldots, X_k$. We focus on center-based clusterings that are induced by a set $S := \{c_1, \ldots, c_k\}$ of k **centers**; each X_i is the subset of points of X that are closest to c_i in S under δ, that is, $X_i := \{p \in X \mid \delta(p, c_i) \leq \delta(p, c_j)\}$ (ties are broken arbitrarily). Assuming the nearest neighbor of each point of X in S is unique (under distance function δ), S defines a k-clustering of X. Sometimes it is more convenient to denote a k-clustering by its set of centers S.

The quality of a clustering S of X is defined using a cost function $\boldsymbol{\$(X, S)}$; cost function $\$$ depends on the distance function δ, so sometimes we may use the notation $\$_\delta$ to emphasize the underlying distance function. The goal is to compute $S^* := \arg\min_S \$(X, S)$ where the minimum is taken over all subsets $S \subset \mathbb{R}^d$ of k points. Several different cost functions have been proposed, leading to various optimization problems. We consider the following three popular variants:

- ***k*-medians clustering**: the cost function is $\$(X, S) = \sum_{p \in X} \delta(p, S)$.
- ***k*-means clustering**: the cost function is $\$(X, S) = \sum_{p \in X} (\delta(p, S))^2$.
- ***k*-center clustering**: the cost function is $\$(X, S) = \max_{p \in X} \delta(p, S)$.

In some cases we wish S to be a subset of X, in which case we refer to the problem as the **discrete *k*-clustering** problem. For example, the discrete k-medians problem is to compute

$$\arg\min_{\substack{S \subseteq X \\ |S| = k}} \sum_{p \in X} \delta(p, S).$$

The discrete k-means and discrete k-center problems are defined analogously.

Given point set X, distance function δ, and cost function $\$$, we refer to $(X, \delta, \$)$ as a **clustering instance**. If $\$$ is defined directly by the distance function δ, we use (X, δ) to denote a clustering instance. Note that a center of a set of points may not be unique (e.g. when δ is defined by the L_1-metric and $\$$ is the sum of distances) or it may not be easy to compute (e.g. when δ is defined by the L_2-metric and $\$$ is the sum of distances).

Stability. Let X be a point set in Euclidean space $\mathbb{R}^d$. For $\alpha \geq 1$, a clustering instance $(X, \delta, \$_\delta)$ is ***α*-stable** if for any *perturbed distance function* $\tilde{\delta}$ (not necessary a metric) satisfying $\delta(p,q) \leq \tilde{\delta}(p,q) \leq \alpha \cdot \delta(p,q)$ for all $p, q \in \mathbb{R}^d$, any optimal clustering of $(X, \delta, \$_\delta)$ is also an optimal clustering of $(X, \tilde{\delta}, \$_{\tilde{\delta}})$. Note that the cluster centers as well as the cost of the optimal clustering may be different for the two instances.

We will exploit the following property of stability, which follows directly from its definition. First we ensure that it is safe to assume that the optimal clustering of an α-stable instance (X, δ) is *unique*.

Lemma 2.1. *Let (X, δ) be an α-stable clustering instance with $\alpha > 1$. Then the optimal clustering O of (X, δ) is unique.*

Metric approximations. The next lemma, which we rely on heavily throughout the paper, is the observation that a change of metric preserves the optimal clustering as long as the new metric is a β-approximation of the original metric satisfying $\beta < \alpha$.

Lemma 2.2. *Given point set X, let δ and δ' be two metrics satisfying $\delta(p,q) \leq \delta'(p,q) \leq \beta \cdot \delta(p,q)$ for all p and q in X for some β. Let (X, δ) be an α-stable clustering instance with $\alpha > \beta$. Then the optimal clustering of (X, δ) is also the optimal clustering of (X, δ'), and vice versa. Furthermore, (X, δ') is an (α/β)-stable clustering instance.*

Polyhedral metric. In light of the metric approximation lemma, we would like to approximate the Euclidean metric without losing too much stability. This brings up the following definition of a collection of convex distance functions generalizing the L_1-metric in Euclidean space. Let $N \subseteq S^{d-1}$ be a centrally-symmetric set of γ unit vectors (that is, if $u \in N$ then $-u \in N$) such that for any unit vector $v \in S^{d-1}$, there is a vector $u \in N$ within angle $\arccos(1-\varepsilon) = O(\sqrt{\varepsilon})$. The number of vectors needed in N is known to be $O(\varepsilon^{-(d-1)/2})$.

Definition 2.3. *We define the **polyhedral metric** $\delta_N : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ as follows:*

$$\boldsymbol{\delta_N(p,q) := \max_{u \in N} \langle p - q, u \rangle.}$$

Since N is centrally symmetric, δ_N is symmetric and thus a metric. The unit ball under δ_N is a convex polyhedron, thus the name polyhedral metric. By construction, an easy calculation shows that for any p and q in $\mathbb{R}^d$,

$$\|p - q\| \geq \delta_N(p,q) \geq (1-\varepsilon) \cdot \|p - q\|.^{1}$$

By taking ε to be small enough, the optimal clustering for α-stable clustering instance $(X, \|\cdot\|, \$)$ is the same as that for $(X, \delta_N, \$)$ by Lemma 2.2, and the new instance $(X, \delta_N, \$)$ is $(1-\varepsilon)\alpha$-stable if the original instance $(X, \|\cdot\|, \$)$ is α-stable.

Center proximity. A clustering instance (X, δ) satisfies *α-center proximity property* [12] if for any distinct optimal clusters X_i and X_j with centers c_i and c_j respectively and any point $p \in X_i$, one has $\alpha \cdot \delta(p, c_i) < \delta(p, c_j)$. Awasthi, Blum, and Sheffet showed that any α-stable instance satisfies α-center proximity [12, Fact 2.2] ([8, Theorem 3.1] under metric perturbation).

Optimal solutions of α-stable instances satisfy the following separation properties.[2]

[1] By scaling each vector in N by an α factor, we can ensure that $(1+\varepsilon) \cdot \|p - q\| \geq \delta_N(p,q) \geq \|p - q\|$.

[2] We give an additional list of known separation properties in Appendix A.

- α-center proximity implies that

$$(\alpha-1)\cdot\delta(p,c_i)<\delta(p,q) \text{ for any } p\in X_i \text{ and any } q\notin X_i.$$

  For $\alpha\geq 2$, this implies a point is closer to its own center than to any point of another cluster.[3]

- For $\alpha\geq 2+\sqrt{3}$, α-center proximity implies that

$$\delta(p,p')<\delta(p,q) \text{ for any } p,p'\in X_i \text{ and any } q\notin X_i.$$

  In other words, from any point p in X, any *intra*-cluster distance to a point p' is shorter than any *inter*-cluster distance to a point q.[4]

- We make use of the following stronger intra-inter distance property on α-stable instances, which allows us to compare *any* intra-distance between two points in X_i and *any* inter-distance between a point in X_i and a point in X_j.

Lemma 2.4. *Let (X,δ) be an α-stable instance, $\alpha>1$, and let X_1 be a cluster in an optimal clustering with $q\in X\setminus X_1$ and $p,p',p''\in X_1$.*

- *If δ is a metric, then $\delta(p,p')\leq\delta(p'',q)$ for $\alpha\geq 2+\sqrt{5}$.*
- *If δ is the Euclidean metric in $\mathbb{R}^d$, then $\delta(p,p')\leq\delta(p'',q)$ for $\alpha\geq 2+\sqrt{3}$.*

Discrete vs continuous clustering. We now show that for stable instances, it is enough to consider the discrete version of the clustering problem.

Lemma 2.5. *For any α-stable instance $(X,\delta,\$_\delta)$ with $\alpha\geq 2+\sqrt{3}$, any partition of a continuous optimal k-clustering is also the partition of the discrete optimal k-clustering.*

3 k-means and k-center: Efficient Dynamic Programming

We now describe a simple, efficient algorithm for computing the optimal clustering for the k-means and k-center problem assuming the given instance is α-stable for $\alpha\geq 2+\sqrt{3}$. Roughly speaking, we make the following observation: if there are at least two clusters, then the two endpoints of the longest edge of the minimum spanning tree of X belong to different clusters, and no cluster has points in both subtrees of the minimum spanning tree delimited by the longest edge. This enables us to perform simple dynamic programming, and we need a procedure to compute the costs of various clusters. We describe the dynamic programming algorithm in Section 3.1 and then describe the procedure for computing cluster costs in Section 3.2. We summarize the results in this section by the following theorem.

Theorem 3.1. *Let X be a set with n points lying in $\mathbb{R}^d$ and $k\geq 1$ an integer. If the k-means or k-center instance for X under the Euclidean metric is α-stable for $\alpha\geq 2+\sqrt{3}+\Omega(1)$ then the optimal k-means or k-center clustering can be computed in $O(nk^2+n\operatorname{polylog} n)$ time. For $d=2$ the assumption can be relaxed to $\alpha\geq 2+\sqrt{3}$, and the algorithm also works for k-median.*

[3] Known as the *weak center proximity* [17].
[4] Known as the *strict separation property* [15, 19].

3.1 Fast Dynamic Programing

The following lemma is the key observation for our algorithm.

Lemma 3.2. *Let $(X, \delta, \$)$ be an α-stable k-clustering instance with $\alpha \geq 2+\sqrt{3}$ and $k \geq 2$, and let T be the minimum spanning tree of X under Euclidean metric δ. Then*

1. *The two endpoints u and v of the longest edge e in T do not belong to the same cluster; and*
2. *each cluster lies in the same connected component of $T \setminus \{e\}$.*

Proof: Assume for contradiction that the longest spanning tree edge uv belongs to the same cluster X_i in the optimal k-clustering O. Since $k > 1$, there is at least one other cluster X_j of O with a spanning tree edge xy connecting X_i to X_j. Given $\alpha \geq 2+\sqrt{3}$, $d(u,v) < d(x,y)$ by Lemma 2.4, a contradiction. The second statement follows from Angelidakis *et al.* [8, Lemma 4.1]. □

Algorithm. We fix the metric δ and the cost function \$. In two dimensions, one can safely assume that the metric here is Euclidean, and therefore the minimum requirement on α is indeed $2+\sqrt{3}$ in order to apply Lemma 3.2. However when $d > 2$ we have to switch to a polygonal metric that closely approximates the Euclidean metric. Therefore, in light of Lemma 2.2, α has to be bounded strictly away from $2+\sqrt{3}$ to ensure enough stability after changing to the polygonal metric. We postpone the details to Section 3.2.

For a subset $Y \subseteq X$ and for an integer j between 1 and $k-1$, let $\boldsymbol{\mu(Y;\ell)}$ denote the optimal cost of an ℓ-clustering on Y (under δ and \$). Recall that our definition of j-clustering required all clusters to be non-empty, so it is not defined for $|Y| < j$. For simplicity, we assume that $\mu(Y;j) = \infty$ for $|Y| < j$. Let $\boldsymbol{T}$ be the minimum spanning tree on X under δ, let uv be the longest edge in T; let X_u and X_v be the set of vertices of the two components of $T \setminus \{uv\}$. Then $\mu(X;k)$ satisfies the following recurrence relation:

$$\mu(X;k) = \begin{cases} \mu(X;1) & \text{if } k = 1, \\ \infty & \text{if } k > |X|, \\ \min_{1 \leq i < k} \{\mu(X_u;i) + \mu(X_v;k-i)\} & \text{if } |X| > 1 \text{ and } k > 1. \end{cases} \tag{1}$$

Using (1), we compute $\mu(X;k)$ as follows. Let $\mathscr{R}$ be a *recursion tree*, a binary tree where each node v in $\mathscr{R}$ is associated with a subtree T_v of T. If v is the root of $\mathscr{R}$, then $T_v = T$. Recursion tree R is defined recursively as follows. Let $X_v \subseteq X$ be the set of vertices of T in T_v. If $|X_v| = 1$, then v is a leaf. Each interior node v of T is also associated with the longest edge e_v of T_v. Removal of e_v decomposes T_v into two connected components, each of which is associated with one of the children of v. After having computed T, $\mathscr{R}$ can be computed in $O(n \log n)$ time by sorting the edges in decreasing order of their costs.[5]

For each node $v \in \mathscr{R}$ and for every i between 1 and $k-1$, we compute $\mu(X_v;i)$ as follows. If v is a leaf, we set $\mu(X_v;1) = 0$ and $\mu(X_v;i) = \infty$ otherwise. For all interior nodes v, we compute $\mu(X_v;i)$ using the algorithms described in Section 3.2. Finally, if v is an interior node and $i > 1$, we compute $\mu(X_v;i)$ using the recurrence relation (1). Recall that if w and z are the children of v, then $\mu(X_w;\ell)$ and $\mu(X_z;r)$ for all ℓ and r have been computed before we compute $\mu(X_v;i)$.

Let $\tau(n)$ be the time spent in computing T plus the total time spent in computing $\mu(X_v,1)$ for all nodes $v \in \mathscr{R}$. Then the overall time taken by the algorithm is $O(nk^2 + \tau(n))$. What is left is to compute the minimum spanning tree T and all $\mu(X_v,1)$ efficiently.

[5] Tree $\mathscr{R}$ is nothing but the minimum spanning tree constructed by the Kruskal's algorithm.

3.2 Efficient Implementation

In this section, we show how to obtain the minimum spanning tree and compute $\mu(X_v;1)$ efficiently for 1-mean and 1-center when $X \subseteq \mathbb{R}^d$ and 1-median when $X \subseteq \mathbb{R}^2$. We can compute the Euclidean minimum spanning tree T in $O(n \log n)$ time in $\mathbb{R}^2$ [49]. We can then compute $\mu(X_v;1)$ efficiently either under Euclidean metric (for 1-mean), or switch to the L_1-metric and compute $\mu(X_v;1)$ efficiently using Lemma 2.2 (for 1-center and 1-median).

There are two difficulties in extending the 2D data structures to higher dimensions. No near-linear time algorithm is known for computing the Euclidean minimum spanning tree for $d \geq 3$, and we can work with the L_1-metric only if $\alpha \geq \sqrt{d}$ (Lemma 2.2). We address both of these difficulties by working with a polyhedral metric δ_N. Let $\alpha \geq 2+\sqrt{3}+\Omega(1)$ be the stability parameter. First we compute the minimum spanning tree under polyhedral metric; by taking the number of vectors in N (defined by the polyhedral metric) to be large enough, δ_N $(1+\varepsilon)$-approximates the Euclidean metric. Thus Lemma 3.2 still holds under polyhedral metric using Lemma 2.2. Afterwards, we compute $\mu(X_v;1)$ for 1-mean by switching to Euclidean metric, and for 1-center in higher dimensions by using the same polyhedral metric.

Computing the MST in higher dimensions. By Lemma 2.2, we only need to compute the minimum spanning tree of X under the polyhedral metric, δ_N. Point set X can be preprocessed in $O(n \operatorname{polylog} n)$ time so that for a query point $q \in \mathbb{R}^d$, its nearest-neighbor under δ_N can be computed in $O(\operatorname{polylog} n)$ time. Therefore using the results of Callahan and Kosaraju [21], the minimum spanning tree of X under δ_N can be computed in $O(n \operatorname{polylog}(n))$ time.

Data structure. We compute $\mu(X_v;1)$ in a bottom-up manner. When processing a node v of $\mathscr{R}$, we maintain a dynamic data structure $\boldsymbol{\Psi_v}$ on X_v from which $\mu(X_v;1)$ can be computed quickly. The exact form of Ψ_v depends on the cost function to be described below. Before that, we analyze the running time $\tau(n)$ spent on computing every $\mu(X_v;1)$. Let w and z be the two children of v. Suppose we have Ψ_w and Ψ_z at our disposal and suppose $|X_w| \leq |X_z|$. We insert the points of X_w into Ψ_z one by one and obtain Ψ_v from which we compute $\mu(X_v;1)$. Suppose $Q(n)$ is the update time of Ψ_v as well as the time taken to compute $\mu(X_v;1)$ from Ψ_v. The total number of insert operations performed over all nodes of $\mathscr{R}$ is $O(n \log n)$ because we insert the points of the smaller set into the larger set at each node of $\mathscr{R}$. Hence $\tau(n) = O(Q(n) \cdot n \log n)$. We now describe the data structure for each specific clustering problem.

1-mean. We work with the L_2-metric. In this case the center of a single cluster consisting of X_v is the centroid $\sigma_v = \left(\sum_{p \in X_v} p\right)/|X_v|$, and

$$\mu(X_v;1) = \sum_{p \in X_v} \|p\|^2 - |X_v| \cdot \|\sigma_v\|^2.$$

At each node v, we maintain $\sum_{p \in X_v} p$ and $\sum_{p \in X_v} \|p\|^2$. Point insertion takes $O(1)$ time, so $Q(n) = 1$.

1-center in 2D. We work with the L_1-metric. First we look at the case when $d = 2$. We wish to find the smallest L_1-disc (a diamond) that contains X_v. Let $e^+ = (-1,1)$ and $e^- = (-1,1)$. Then the radius ρ_v of the smaller L_1-disc containing X_v is

$$\rho_v = \frac{1}{2} \max\left\{ \max_{p \in X_v} \langle p, e^+ \rangle - \min_{p \in X_v} \langle p, e^+ \rangle, \max_{p \in X_v} \langle p, e^- \rangle - \min_{p \in X_v} \langle p, e^- \rangle \right\}. \tag{2}$$

We maintain the four terms $\max_{p\in X_v}\langle p, e^+\rangle$, $\min_{p\in X_v}\langle p, e^+\rangle$, $\max_{p\in X_v}\langle p, e^-\rangle$, and $\min_{p\in X_v}\langle p, e^-\rangle$ at v. A point can be inserted in $O(1)$ time and ρ_v can be computed from these four terms in $O(1)$ time. Therefore, $Q(n) = O(1)$.

1-center in higher dimensions. For a node v, we need to compute the smallest ball $B(X_v)$ under δ_N that contains X_v. We need a few geometric observations to compute the smallest enclosing ball efficiently.

For each $u \in N$, let H_u be the halfspace $\langle x, u\rangle \le 1$, that is, the halfspace bounded by the hyperplane tangent to S^{d-1} at u and containing the origin. Define $Q := \bigcap_{u\in N} H_u$. A ball of radius λ centered at p under δ_N is $P + \lambda Q$. For a vector $u \in N$, let $\overline{p}_u := \arg\max_{p\in x_v}\langle p, u\rangle$ be the maximal point in direction u. Set $\overline{X}_v := \{\overline{p}_u \mid u \in N\}$. The following simple lemma is the key to computing $B(X_v)$.

Lemma 3.3. *Any δ_N-ball that contains $\overline{X}_v$ also contains X_v.*

By Lemma 3.3, it suffices to compute $B(\overline{X}_v)$. The next observation is that $B(\overline{X}_v)$ has a basis of size $d+1$, i.e. there is a subset Y of $d+1$ points of X_v such that $B(Y) = B(\overline{X}_v) = B(X_v)$. One can try all possible subsets of $\overline{X}_v$ in $O(\gamma^{d+1}) = 2^{O(d^2)}$ time.[6] We note that $\overline{X}_v$ can be maintained under insertion in $O(\gamma) = 2^{O(d)}$ time, and we then re-compute $B(\overline{X}_v)$ in $2^{O(d^2)}$ time. Hence, $Q(n) = O(1)$.

1-median in 2D. We consider the k-medians problem in two dimensions. We work with the L_1 metric. Our goal is to find a center c that minimizes the sum of distances from points in X_v to c. The 1-median under 1-norm can be computed coordinate-wise: Let (x, y) be the coordinates of center c, then

$$\sum_i \big(|x_i - x| + |y_i - y|\big) = \sum_i |x_i - x| + \sum_i |y_i - y|.$$

Dynamic median for one-dimensional values can be easily maintained using two heaps, so that preprocessing takes $O(|X_v|)$ time. Each query can be done in $O(1)$ time and each point insertion can be done in $O(\log n)$ time.

Remark. The dynamic programming algorithm works for k-medians in 2D because 1-median of a set of points can be maintained under the L_1 metric in $O(\log n)$ time per insertion. However, we do not know how to maintain the 1-median efficiently for the Euclidean or polyhedral metric, so this approach does not extend to higher dimensions. In Section 4 we give an efficient algorithm for k-medians in higher dimensions.

4 k-Medians: Single-Swap Local Search

Let (X, δ) be an instance of α-stable k-medians in $\mathbb{R}^2$ for $\alpha > 5\sqrt{2}$. By Lemma 2.5, it suffices to consider the *discrete* k-medians problem, where cluster centers are subset of input points. In Section 4.1, we describe a simple local-search algorithm for finding the optimal clustering (X, δ). In Section 4.2 we show that the algorithm terminates within $O(k \log(n\Delta))$ iterations where Δ is the spread of X.[7] We obtain the following.

Theorem 4.1. *Let (X, δ) be an α-stable instance of the k-medians problem for some $\alpha > 5\sqrt{2}$ where X is a set of n points in $\mathbb{R}^2$ equipped with L_p-metric δ. The 1-swap local search algorithm terminates with the optimal in $O(k \log(n\Delta))$ iterations for any metric space.*

[6] A more complex algorithm can compute $B(\overline{X}_v)$ in $\gamma \cdot 2^{O(d)} = 2^{O(d)}$ time, but we ignore this improvement.

[7] The *spread* of a point set is the ratio between the longest and shortest pairwise distances.

Next, in Section 5 we show that 1-swap local search can be implemented efficiently. A naïve implementation of the algorithm requires checking $O(nk)$ 1-swaps, and computing the cost of clustering after each 1-swap in $O(nk)$ time, thus taking $O(n^2k^2)$ time at each step of the algorithm. If δ is the L_1-metric and $X \subset \mathbb{R}^2$, each step of the local search can be performed in $\tilde{O}(nk^2)$ time. By Lemma 2.2, the algorithm also finds the optimal clustering even if δ is the L_2-metric (or for that matter, any L_p-metric) when $\alpha > 5\sqrt{2}$. We conclude this section by sketching how to implement the algorithm efficiently in higher dimensions while lowering the assumption on stability to $\alpha > 5$.

4.1 Local-search Algorithm

We customize the standard local-search framework for the k-clustering problem [26,27,33]. At each step we have a k-clustering induced by a set S of k cluster centers. We find a pair of points $x \in X$ and $y \in S$ such that $\$(X, S+x-y)$ is minimized. If $\$(X, S+x-y) \geq \$(X, S)$, we stop and return the k-clustering induced by S. Otherwise we replace S with $S+x-y$ and repeat the above step. The pair (x, y) will be referred to as a **1-swap**.

Remark. If we were computing an approximate solution of the discrete k-medians problem, we could terminate the algorithm earlier, once the improvement after the swap is sufficiently small. However, we continue until no improvement is possible; in Section 4.2 we show that in fact a global minimum can be reached efficiently.

4.2 Local-search Analysis

The high-level structure of our analysis follows Friggstad *et al.* [33], however new ideas are needed to make it work for 1-swap.

In this subsection, we denote a k-clustering by the set of its cluster centers. Let S be a fixed k-clustering, and let O be the optimal clustering. For a subset $Y \subseteq X$, we use $\mathbf{\$(Y)}$ and $\mathbf{\$^*(Y)}$ to denote $\$(Y, S)$ and $\$(Y, O)$, respectively. Similarly, for a point $p \in X$, we use $\mathbf{nn(p)}$ and $\mathbf{nn^*(p)}$ to denote the nearest neighbor of p in S and in O, respectively; define $\boldsymbol{\delta(p)}$ to be $\delta(p, S)$ and $\boldsymbol{\delta^*(p)}$ to be $\delta(p, O)$.

We partition X into four subsets as follows:

- $\boldsymbol{X_{00}} := \{p \in X \mid \mathrm{nn}(p) \in S \setminus O, \mathrm{nn}^*(p) \in O \setminus S\}$;
- $\boldsymbol{X_{01}} := \{p \in X \mid \mathrm{nn}(p) \in S \setminus O, \mathrm{nn}^*(p) \in S \cap O\}$;
- $\boldsymbol{X_{10}} := \{p \in X \mid \mathrm{nn}(p) \in S \cap O, \mathrm{nn}^*(p) \in O \setminus S\}$;
- $\boldsymbol{X_{11}} := \{p \in X \mid \mathrm{nn}(p) \in S \cap O, \mathrm{nn}^*(p) \in S \cap O\}$.

Observe that for any point p in X_{11}, $\$(p) = \$^*(p)$; for any point p in X_{01}, one has $\$(p) \leq \$^*(p)$; and for any point p in X_{10}, one has $\$(p) \geq \$^*(p)$. Costs $\delta(p)$ and $\delta^*(p)$ are not directly comparable for point p in X_{00}. A k-clustering S is **C-good** for some parameter $C \geq 0$ if

$$\$(X) \leq \$^*(X) + C \cdot \$^*(X_{00}). \tag{3}$$

Let $\mathbf{\Psi(X_{00})}$ denote the terms $C \cdot \$^*(X_{00})$ in inequality (3).

Lemma 4.2. *Any C-good clustering S for an α-stable clustering instance $(X, \delta, \$)$ must be optimal for $\alpha \geq C+1$.*

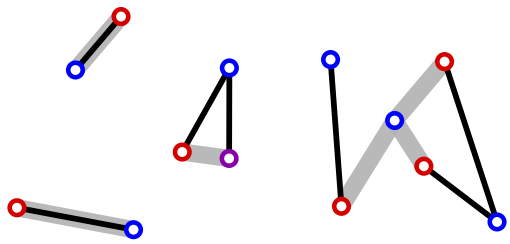

Figure 4.1. Illustration of candidate swaps $\mathscr{S}$ in $\mathbb{R}^2$. The blue dots belongs to set S, the red dots belongs to set O; the only purple dot is in $S \cap O$. The thick gray segments indicate pairs inside the stars; each star has exact one blue dot as its center. The black pairs are the candidate swaps. Notice that the partitions of S and O form connected components.

Proof: Define a perturbed distance function $\tilde{\delta}: X \times X \to \mathbb{R}_{\geq 0}$ with respect to the given clustering S as follows:

$$\tilde{\delta}(p', p) := \begin{cases} \alpha \cdot \delta(p', p) & \text{if } p \neq \mathrm{nn}(p'), \\ \delta(p, p') & \text{otherwise.} \end{cases}$$

Note that $\tilde{\delta}$ is not symmetric. Let $\tilde{\$}(\cdot,\cdot)$ denote the cost function under the perturbed distance function $\tilde{\delta}$. The optimal clustering under perturbed cost function is the same as the original optimal clustering O by the stability assumption. Consider the cost of O under the perturbed cost:

$$\tilde{\$}(X, O) = \alpha \cdot \$(X_{00}, O) + \alpha \cdot \$(X_{01}, O) + \alpha \cdot \$(X_{10}, O) + \$(X_{11}, O).$$

By definition of perturbed distance $\tilde{\delta}$, $\tilde{\$}(X, S) = \$(X, S)$. Now, by the assumption that clustering S is C-good,

$$\begin{aligned} \tilde{\$}(X, S) = \$(X, S) &\leq \$(X, O) + C \cdot \$(X_{00}, O) \\ &\leq (C+1) \cdot \$(X_{00}, O) + \$(X_{01}, O) + \$(X_{10}, O) + \$(X_{11}, O) \\ &\leq \tilde{\$}(X, O); \end{aligned}$$

the last inequality follows by taking $\alpha \geq C + 1$. This implies that S is an optimal clustering for $(X, \tilde{\delta})$, and thus is equal to O. □

Next, we prove a lower bound on the improvement in the cost of clustering after performing a 1-swap. Following Arya *et al.* [10], define the set of **candidate swaps** $\mathscr{S}$ as follows: For each center i in S, consider the **star Σ_i** centered at i defined as the collection of pairs

$$\Sigma_i := \{(i, j) \in S \times O \mid \mathrm{nn}(j) = i\}.$$

Denote **center(j)** to be the center of the star where j belongs; in other words, $\mathrm{center}(j) = i$ if j belongs to Σ_i. Partition O based on the stars, and partition S into parts of equal size based on the partition of O; denote that parts of S and O as S_Is and O_Is accordingly. For each pair S_I and O_I, there is only one star Σ_i whose center i lies in S_I. The rest of $|S_I| - 1$ centers in S_I does not belong to any (nonempty) stars. Add an arbitrary maximal matching between $S_I - i$ and O_I as candidate swaps into $\mathscr{S}$, as well as a pair between the only unmatched center in O_I and an arbitrary center in $S_I - i$. Notice that all centers but one in $S_I - i$ and O_I belongs to exactly one candidate swap, and exactly one center in $S_I - i$ belongs to two candidate swaps. See Figure 4.1 for an illustration. The number of candidate swaps in $\mathscr{S}$ is exactly k.

Lemma 4.3. *For each point p in X_{01}, X_{10}, or X_{11}, the set of candidate swaps $\mathscr{S}$ satisfies*

$$\sum_{S' \in \mathscr{S}} (\delta(p) - \delta'(p)) \geq \delta(p) - \delta^*(p); \tag{4}$$

and for each point p in X_{00}, the set of candidate swaps $\mathscr{S}$ satisfies

$$\sum_{S' \in \mathscr{S}} (\delta(p) - \delta'(p)) \geq (\delta(p) - \delta^*(p)) - \Psi(p), \tag{5}$$

where $\Psi(p) := 4 \cdot \delta^*(p)$, and $\$'$ is the cost function on X defined with respect to S'.

Proof: For point p in X_{11}, one has $\delta(p) = \delta^*(p)$ and inequality (4) immediately holds. For point p in X_{01}, one has $\delta(p) \leq \delta^*(p)$; when $\mathrm{nn}(p)$ is being swapped out by some in 1-swap S', $\mathrm{nn}^*(p)$ must be in S', and therefore inequality (4) holds. For point p in X_{10}, one has $\delta(p) \geq \delta^*(p)$; center $\mathrm{nn}(p)$ will never be swapped out by any 1-swap in $\mathscr{S}$. By construction of $\mathscr{S}$, there is exactly one S' in $\mathscr{S}$ that swaps $\mathrm{nn}^*(p)$ in; for that particular swap we have $\delta'(p) = \delta^*(p)$. This implies inequality (4).

Our final goal is to prove inequality (5); that is, for each point p in X_{00},

$$\sum_{S' \in \mathscr{S}} (\delta(p) - \delta'(p)) \geq (\delta(p) - \delta^*(p)) - 4 \cdot \delta^*(p) = \delta(p) - 5 \cdot \delta^*(p). \tag{6}$$

Consider the reassignment cost of p under the candidate swap (i, j) in $\mathscr{S}$:

$$\delta'(p) = \begin{cases} \min\{\delta(p), \delta^*(p)\} & \text{if } j = \mathrm{nn}^*(p), \\ \delta(p) & \text{if } j \neq \mathrm{nn}^*(p) \text{ and } i \neq \mathrm{nn}(p), \\ \delta(p, \mathrm{center}(\mathrm{nn}^*(p))) & \text{if } j \neq \mathrm{nn}^*(p) \text{ and } i = \mathrm{nn}(p). \end{cases}$$

First we argue that $\mathrm{center}(\mathrm{nn}^*(p))$ is a valid center in $S - i + j$ that can be swapped to. Under this scenario $\mathrm{nn}(p)$ is swapped out but $\mathrm{nn}^*(p)$ is not swapped in. From the construction of candidate swaps $\mathscr{S}$, either the star Σ_i has size 1 and $\mathrm{nn}^*(p)$ belongs to a different star, in which case $\mathrm{center}(\mathrm{nn}^*(p))$ is a different star center that must be in $S - i + j$; or the star Σ_i has size greater than 1, in which case $\mathrm{center}(\mathrm{nn}^*(p))$ can be in a different star or be in the same star Σ_i that does not belong to any candidate swaps. Either case, $\mathrm{center}(\mathrm{nn}^*(p))$ must exist in $S - i + j$.

Next, we upper bound $\delta(p, \mathrm{center}(\mathrm{nn}^*(p)))$ by a linear combination of $\delta(p)$ and $\delta^*(p)$:

$$\begin{aligned} \delta(p, \mathrm{center}(\mathrm{nn}^*(p))) &\leq \delta(p, \mathrm{nn}^*(p)) + \delta(\mathrm{nn}^*(p), \mathrm{center}(\mathrm{nn}^*(p))) \\ &\leq \delta^*(p) + \delta(\mathrm{nn}^*(p), \mathrm{nn}(p)) \\ &\leq \delta^*(p) + (\delta^*(p) + \delta(p)) = \delta(p) + 2\delta^*(p). \end{aligned}$$

Consider the sum in inequality (6) and δ'. In the first case, we might swap in $\mathrm{nn}^*(p)$, in which case $\delta'(p) \leq \delta^*(p)$. Once the optimal center is swapped in, it will never be swapped out so this case occurs exactly once. In the second case, we do not swap out $\mathrm{nn}(p)$ and do not swap in $\mathrm{nn}^*(p)$, so $\delta'(p) = \delta(p)$. In the final case, $\mathrm{nn}(p)$ is swapped out, but $\mathrm{nn}^*(p)$ is not swapped in. In this case, we have shown that $\delta'(p) \leq \delta(p) + 2\delta^*(p)$. This case occurs at most two times as every center in S, in particular $\mathrm{nn}(p)$, belongs to at most two candidate swaps. Thus we have,

$$\sum_{S' \in \mathscr{S}} (\delta(p) - \delta'(p)) \geq (\delta(p) - \delta^*(p)) + 0 + 2(\delta(p) - \delta(p) - 2\delta^*(p)) = \delta(p) - 5\delta^*(p));$$

this concludes the proof. □

Lemma 4.4. *Let S be a k-clustering of (X,δ) that is not C-good for some fixed constant $C > 4+\varepsilon$ with arbitrarily small $\varepsilon > 0$. There is always a 1-swap $S' := S - x + y$ such that $\$'(X) - \$^*(X) \le (1-\varepsilon/(1+\varepsilon)k)\cdot(\$(X)-\$^*(X))$, where $\$'$ is the cost function on X defined with respect to S'.*

Proof: By Lemma 4.3 one has $\$(X)-\$'(X) \ge (\$(X)-\$^*(X)-\Psi(X_{00}))/k$ for some 1-swap S' and its corresponding cost function $\$'(\cdot)$. Since S is not C-good, $\$(X)-\$^*(X) > C\cdot\$^*(X_{00})$. Rearranging and plugging the definition of $\Psi(\cdot)$, we have

$$\begin{aligned}
\$'(X)-\$^*(X) &\le \$(X)-\$^*(X)-(\$(X)-\$^*(X)-\Psi(X_{00}))/k \\
&\le \$(X)-\$^*(X)-(\$(X)-\$^*(X)-4\cdot\$^*(X_{00}))/k \\
&\le \$(X)-\$^*(X)-(\$(X)-\$^*(X)+(M-1)\cdot(\$(X)-\$^*(X))-4M\cdot\$^*(X_{00}))/Mk \\
&\le \left(1-\frac{\varepsilon}{(1+\varepsilon)k}\right)\cdot(\$(X)-\$^*(X)),
\end{aligned}$$

where the last inequality holds by taking M to be arbitrarily large (say $M > 1+1/\varepsilon$). □

5 Efficient Implementation of Local Search

In this section we show how the best 1-swap at each step of the local search algorithm can be computed in $O(nk^2\log^3 n)$ time if δ is the L_1-metric. We describe the implementation for $d=2$ in Section 5.1 and then discuss the extension to higher dimensions in Section 5.2. For $\alpha > 5\cdot\sqrt{d}$, by Lemma 2.2, this procedure also computes the optimal k-median clustering under the Euclidean distance. For larger value of d, the optimal clustering for an $O(1)$-stable instance cannot be computed using the L_1-metric. Instead we use a polyhedral distance metric and show that each step of 1-swap can be implemented in $O(nk^{2d-1}\operatorname{polylog} n)$ time under the assumption that $\alpha > 5$. We obtain the following:

Theorem 5.1. *Let (X,δ) be an α-stable instance of the k-median problem where $X\subset\mathbb{R}^d$ and δ is the Euclidean metric. For $\alpha > 5$, the 1-swap local search algorithm computes the optimal k-clustering of (X,δ) in $O(nk^{2d-1}\operatorname{polylog} n)$ time.*

Even in 2d, using the polyhedral metric improves the requirement of α, from $\alpha > 5\sqrt{2}$ to $\alpha > 5$. We first present the slightly weaker result in Section 5.1 under the L_1-metric, as it is straightforward to implement and more intuitive. Then we describe the extension to higher dimensional Euclidean space using the polyhedral metric in Section 5.2.

5.1 Swap Cost Computation

Given a clustering instance $(X,\delta,\$)$ and a current k-clustering S, we want to find the 1-swap giving the best cost improvement:

$$\arg\min_{\substack{x\in X\setminus S\\ y\in S}} \$(X, S+x-y).$$

Again by Lemma 2.2 we assume the distances and costs are defined under 1-norm.

Voronoi diagram under L_1 norm. First, we fix a point $x\in X\setminus S$ to insert and a center $y\in S$ to drop. Define $S' := S+x-y$. We build the L_1 Voronoi diagram V of S'. The cells of V may not be convex, but they are *star-shaped*: for any $c\in S'$ and for any point $x\in \mathrm{Vor}(c)$, the segment cx lies completely in $\mathrm{Vor}(c)$. Furthermore, all line segments on the cell boundaries of V must have slopes belonging to one of the four possible values: vertical, horizontal, diagonal, or antidiagonal.

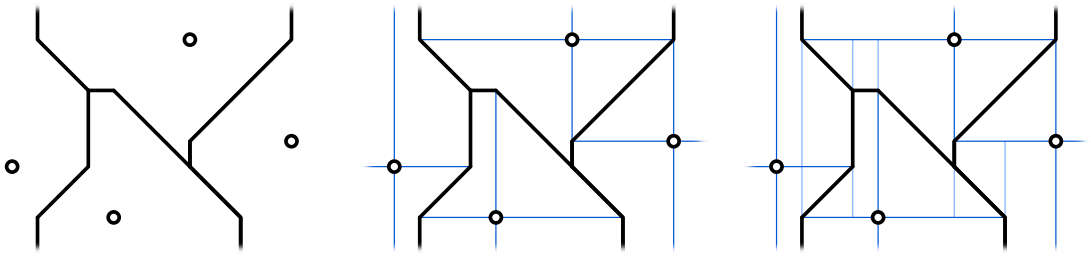

Figure 5.1. L_1 Voronoi diagram V, quadrant decomposition $\tilde{V}$, and trapezoid decomposition $V^{\parallel}$.

Next, decompose each Voronoi cell $\mathrm{Vor}(c)$ into four *quadrants* centered at c. Denote the resulting subdivision of V as $\tilde{V}$. We compute a *trapezoidal decomposition* of the diagram $\tilde{V}$ by drawing a vertical segment from each vertex of $\tilde{V}$ in both directions until it meets an edge of V; see Figure 5.1. Trapezoidal decomposition $V^{\parallel}$ has $O(k)$ trapezoids. For each trapezoid $\tau \in V^{\parallel}$, let $X_\tau := X \cap \tau$. The cost of the new clustering S' can be computed as $\$(X,S') = \sum_{\tau\in V^{\parallel}} \$(X_\tau,S')$.

Range-sum queries. Now we discuss how to compute $\$(X_\tau,S')$. Each trapezoid τ in cells $\mathrm{Vor}(c)$ is associated with a vector $u(\tau) \in \{\pm 1\}^2$, depending on which of the four quadrants τ belongs to with respect to the axis-parallel segments drawn passing through the center c of the cell. If τ lies in the top-right quadrant then $u(\tau) = (1,1)$. Similarly if τ lies in the top-left (resp. bottom-left, bottom-right) then $u(\tau) = (-1,1)$ (resp. $(-1,-1)$, $(1,-1)$). With this notation,

$$\$(X_\tau,S') = \sum_{x\in X_\tau} \|x-c\|_1 = \sum_{x\in X_\tau} \langle x-c, u(\tau)\rangle = \sum_{x\in X_\tau} \langle x, u(\tau)\rangle - |X_\tau|\cdot\langle c, u(\tau)\rangle. \tag{7}$$

We preprocess X into a data structure independent of S', that answers the following query:

- TrapezoidSum(τ,u): Given a trapezoid τ and a vector $u \in \{\pm 1\}^2$, return $|X\cap\tau|$ as well as $\sum_{x\in X\cap\tau}\langle x,u\rangle$.

Each trapezoid τ in $V^{\parallel}$ (for any $S' \subseteq S$) has at most four edges with at most three directions, that is, $\tau = \bigcap_{i=1}^{4} H_i$ where H_i is a halfplane normal to one of the four directions, namely horizontal, vertical, and diagonal or antidiagonal. Let us assume that τ has diagonal direction. We can build a similar data structure if τ has any of the three other directions. The above query can be viewed as a 3-oriented polygonal range query [29]. We construct a 3-level range tree Ψ on X. Roughly speaking, each level filters the points of X lying in the halfplanes in one of the directions. Omitting the details (which can be found in [29]), Ψ can be constructed in $O(n\log^2 n)$ time and uses $O(n\log^2 n)$ space. Each node ξ at the third level of Ψ is associated with a subset $X_\xi \subseteq X$. We store $w(\xi,u) := \sum_{x\in X_\xi}\langle x,u\rangle$ for each $u\in\{\pm1\}^2$ and $|X_\xi|$ at ξ.

For a trapezoid τ, the query procedure identifies a set Ξ_τ of $O(\log^3 n)$ third-level nodes such that $X\cap\tau = \cup_{\xi\in\Xi_\tau} X_\xi$ and each point of $X\cap\tau$ appears as exactly one node of Ξ_τ. Furthermore, Ξ_τ can be computed in $O(\log^3 n)$ time. Then

$$\sum_{x\in X_\tau}\langle x,u\rangle = \sum_{\xi\in\Xi_\tau} w(\xi,u) \qquad \text{and} \qquad |X_\tau| = \sum_{\xi\in\Xi_\tau}|X_\xi|.$$

With the information stored at the nodes in Ξ_τ, TrapezoidSum(τ,u) query can be answered in $O(\log^3 n)$ time. By performing TrapezoidSum$(\tau,u(\tau))$ query for all $\tau\in V^{\parallel}$, $\$(X_\tau,S')$ can be computed in $O(k\log^3 n)$ time since $V^{\parallel}$ has a total of $O(k)$ trapezoids.

1-SWAP(X, S):
  input: Point set X and centers S
  for each point $x \in X \setminus S$ and center $y \in S$:
    $S' \leftarrow S + x - y$
    $V \leftarrow L_1$ Voronoi diagram of S'
    $\tilde{V} \leftarrow$ decompose each cell Vor(c) into four quadrants centered at c
    $V^{\parallel} \leftarrow$ trapezoidal decomposition of $\tilde{V}$
    for each trapezoid $\tau \in V^{\parallel}$:
      $\$(X_\tau, S') \leftarrow$ TRAPEZOIDSUM$(\tau, u(\tau))$
    $\$(X, S') \leftarrow \sum_{\tau \in V^{\parallel}} \$(X_\tau, S')$
  return (x, y) with the lowest $\$(X, S + x - y)$

Figure 5.2. Efficient implementation of 1-swap under 1-norm.

We summarize the implementation of 1-swap algorithm in Figure 5.2. The 1-swap procedure considers at most nk different k-clusterings. Therefore we obtain the following.

Lemma 5.2. *Let $(X, \delta, \$)$ be a given clustering instance where δ is the L_1 metric, and let S be a given k-clustering. After $O(n \log n)$ time preprocessing, we find a k-clustering $S' := S + x - y$ minimizing $\$(X, S')$ among all choices of (x, y) in $O(nk^2 \log^3 n)$ time.*

5.2 Extending to Higher Dimensions

The 1-SWAP algorithm can be extended to higher dimensions using the theory of geometric arrangements [3, 4, 48]. The extension is technical, so we only sketch its proof here.

As in Section 3.2, instead of working with the L_1 metric, we work with a polyhedral metric as defined by Equation (2.3). Let the centrally-symmetric set $N \subseteq S^{d-1}$ and the convex polyhedron Q be defined as in Section 3.2. The set N partitions $\mathbb{R}^d$ into a set of $O(1)$ polyhedral cones denoted by $\mathscr{C} := \{C_1, \ldots, C_\gamma\}$, each with 0 as its apex so that all points (when viewed as vectors) in a cone have the same vector u of N as the nearest neighbor under the cosine distance, i.e. the polyhedral distance $\delta_N(0, u)$ is realized by u. The total complexity of $\mathscr{C}$ is $O(\gamma) = O(1)$.

We show that X can be preprocessed into a data structure so that $\$(X, S)$, the cost of the k-clustering induced by any k-point subset S of X under δ_N can be computed in $O(k^{2d} \operatorname{polylog}(n))$ time.

Let $S \subset X$ be a set of k points. We compute the Voronoi diagram VD(S) of S under the distance function δ_N. More precisely, for a point $c \in S$, let $f_c : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ be the function $f_c(x) := \delta_N(c, x)$. The graph of f_c is a polyhedral cone in $\mathbb{R}^{d+1}$ whose level set for the value λ is the *homothet* copy of Q, $c + \lambda Q$. Voronoi diagram VD(S) is the minimization diagram of function f_c over every points c in S; that is, the projection of the lower envelope $f(x) := \min_c f_c(x)$ onto the hyperplane $X_{d+1} = 0$ (identified with $\mathbb{R}^d$). We further decompose each Voronoi cell Vor(c) of VD(S) by drawing the family of cones in $\mathscr{C}$ from c; put it differently, by drawing the cone $c + C_j$ for $1 \leq j \leq k$, within the cell Vor(c). Each cell τ in the refined subdivision of Vor(c) has the property that for all points $x \in \tau$, $\delta_N(x, c)$ is realized by the vector of N—by u_j if $x \in c + C_j$. Let $\tilde{V}$ denote the resulting refinement of VD(S).

Finally, we compute the vertical decomposition of each cell in $\tilde{V}$, which is the extension of the trapezoidal decomposition to higher dimensions; see [39, 48] for details. Let $V^{\parallel}$ denote the resulting convex subdivision of $\mathbb{R}^d$. It is known that $V^{\parallel}$ has $O(k^{2d-2})$ cells, that it can be computed in $O(k^{2d-2})$ time, and that each cell of $V^{\parallel}$ is convex and bounded by at most $2d$ facets, namely it is the intersection of at most $2d$ halfspaces. Using the same structure of the distance function δ_N, we can show that there is a set U of $O(\gamma^d) = O(1)$ unit vectors such that each facet of a cell in $V^{\parallel}$ is normal to a vector in U, and that U depends only on N and not on S.

With these observations at hand, we preprocess X into a data structure as follows: we fix a $2d$-tuple $\bar{u} := (u_1, \ldots, u_{2d}) \in U^{2d}$. Let $R_{\bar{u}}$ be the set of all convex polyhedra formed by the intersection of at most $2d$ halfspaces each of which is normal to a vector in $\bar{u}$. Using a multi-level range tree (consisting of $2d$ levels), we preprocess X in $O(n \log^{2d} n)$ time into a data structure $\Psi_{\bar{u}}$ of size $O(n \log^{2d-1} n)$ for each $\bar{u}$, so that for a query cell $\tau \in R_{\bar{u}}$ and for a vector $u \in N$, we can quickly compute the total weight $w(\tau, u) = \sum_{p \in X \cap \tau} \langle p, u \rangle$ in $O(\log^{2d} n)$ time.

For a given S, we compute the cost $\$(X, S)$ as follows. We first compute $\mathrm{VD}(S)$ and $V^{\parallel}(S)$. For each cell $\tau \in V^{\parallel}(S)$ lying in $\mathrm{Vor}(c)$, let $u(\tau) \in N$ be the vector u_j such that $\tau \subseteq c + C_j$. As in the 2d case,

$$\$(X,S) = \sum_c \sum_{p \in \mathrm{Vor}(c)} \delta_N(p,c) = \sum_c \sum_{\tau \in V^{\parallel}(S) \cap \mathrm{Vor}(c)} \sum_{p \in X \cap \tau} \langle p - c, u(\tau) \rangle$$

$$= \sum_c \sum_{\tau \in V^{\parallel}(S) \cap \mathrm{Vor}(c)} \left(\sum_{p \in X \cap \tau} \langle p, u(\tau) \rangle - |X \cap \tau| \cdot \langle c, u(\tau) \rangle \right).$$

Fix a cell $\tau \in V^{\parallel}(S) \cap \mathrm{Vor}(c)$. Suppose $\tau \in R_{\bar{u}}$. Then by querying the data structure $\Psi_{\bar{u}}$ with τ and $u(\tau)$, we can compute $w(\tau, u) = \sum_{p \in X \cap \tau} \langle p, u(\tau) \rangle$ in $O(\log^d n)$ time. Repeating this procedure over all cells of $V^{\parallel}(S)$, $\$(X, S)$ can be computed in $O(k^{2d-1} \log^{2d} n)$ time, after an initial preprocessing of $O(n \log^{2d} n)$ time.

6 Coresets and an Alternative Linear Time Algorithm

In this section we provide an alternative way to compute the optimal k-clustering, where the objective can be any of k-center, k-means, or k-medians. Here we are aiming for a running time linear in n, but potentially with exponential dependence on k. With such a goal we can further relax the stability requirement using the idea of *coresets*. When there is strict separation between clusters (when $\alpha \geq 2+\sqrt{3}$), we can recover the optimal clustering. We note that this provides a significant improvement to the stability parameter needed for k-medians over the local search approach, albeit with worse running time dependence on k.

Coresets. Let (X, δ) be a clustering instance. The **radius** of a cluster X_i is the maximum distance between its center and any point in X_i. Let S be a given k-clustering of (X, δ), with clusters $X_1, \ldots, X_k$, centers $c_1, \ldots, c_k$, and radius $r_1, \ldots, r_k$, respectively. Let O be the optimal k-clustering of (X, δ), with clusters $X_1^*, \ldots, X_k^*$, centers $c_1^*, \ldots, c_k^*$ and radius $r_1^*, \ldots, r_k^*$, respectively. Let $B(c, r)$ denote the ball centered at c with radius r under δ.

A point set $Q \subseteq X$ is a **multiplicative ε-coreset** of X if every k-clustering S of Q satisfies

$$X \subseteq \bigcup_i B\left(c_i, (1+\varepsilon) \cdot r_i\right).$$

Lemma 6.1. *Let (X, δ) be a $(1+\varepsilon)$-stable clustering instance with optimal k-clustering O. A multiplicative ε-coreset of X contains at least one point from each cluster of O.*

Proof: Let Q be a multiplicative ε-coreset of X. We start by defining a k-clustering S_Q of Q. For each point q in Q, assign q to its cluster in the optimal clustering O. This results in some clustering with at most k clusters. Insert additional empty clusters to create a valid k-clustering S_Q of Q.

Assume that Q does not contain any points from some optimal cluster X_i^* of O. Consider the center point c_i^* of X_i^*. By the fact that Q is a multiplicative ε-coreset, c_i^* must be contained in a ball resulting

from the expansion of each cluster of S_Q by an ε-fraction of its radius. In notation, let the cluster of S_Q whose expansion covers c_i^* be X_j, with center c_j and radius r_j. Then one has $\delta(c_j, c_i^*) \le (1+\varepsilon)\cdot r_j$.

Because S_Q is constructed by restricting the optimal clustering O on Q, cluster X_j is a subset of some optimal cluster in O: $X_j \subseteq X_j^*$. This implies $r_j \le r_j^*$. Additionally, c_j and c_i^* lie in different optimal clusters, as c_j is in Q and therefore does not lie in X_i^*. So by $(1+\varepsilon)$-center proximity:

$$\delta(c_j, c_i^*) > (1+\varepsilon)\cdot\delta(c_j, c_j^*) = (1+\varepsilon)\cdot r_j^* \ge (1+\varepsilon)\cdot r_j,$$

contradicting to $\delta(c_j, c_i^*) \le (1+\varepsilon)\cdot r_j$. Therefore, Q must contain at least one point from each optimal cluster. □

Algorithm. We first compute a constant approximation to the clustering problem instance (X, δ). We then recursively construct a multiplicative coreset of size $O(k!/\varepsilon^{dk})$ [36]. By taking $\varepsilon = 1$, the coreset has size $O(k!)$ and for 2-stable instances, the multiplicative 2-coreset Q contains at least one point from every optimal cluster by Lemma 6.1. After obtaining this coreset, we then need to reconstruct the optimal clustering. By [19, Corollary 9], when our instance satisfies strict separation ($\alpha \ge 2+\sqrt{3}$), taking any point from each optimal cluster induces the optimal partitioning of X. To find k such points, each from a different optimal cluster, we try all possible k subsets of Q as the candidate k centers. For each set of centers we compute its cost (under polygonal metric) using the cost computation scheme of Section 5, then take the clustering with minimum cost. Finally, recompute the optimal centers using any 1-clustering algorithm on each cluster of O.

It is known constant approximation to any of the k-means, k-median, or k-center instance can be computed in $O(nk)$ time [34] and even in $\tilde{O}(n)$ time [25, 36] in constant-dimensional Euclidean spaces. Thus computing the multiplicative 2-coreset takes $O(nk^2 + k!)$ time [35]. Using the cost computation scheme from Section 5, after $O(n\log^{2d} n)$ preprocessing time, the cost of each clustering can be computed in $\tilde{O}(k^{2d-1})$ time. There are at most $O((k!)^k)$ possible choices of center set of size k. Therefore we conclude the section with the following theorem.

Theorem 6.2. *Let X be a set with n points lying in $\mathbb{R}^d$ and $k \ge 1$ an integer. If the k-means, k-median, or k-center instance for X under the Euclidean distance is α-stable for $\alpha \ge 2+\sqrt{3}$ then the optimal clustering can be computed in $\tilde{O}(nk^2 + k^{2d-1}\cdot(k!)^k)$ time.*

References

[1] M. Ackerman and S. Ben-David. Clusterability: A theoretical study. In *Proceedings of of the 12th International Conference on Artificial Intelligence and Statistics*, volume 5 of *JMLR Proceedings*, pages 1–8, 2009.

[2] P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan. Geometric approximation via coresets. *Combinatorial and computational geometry*, 52:1–30, 2005.

[3] P. K. Agarwal, J. Pach, and M. Sharir. State of the union (of geometric objects). In J. E. Goodman, J. Pach, and R. Pollack, editors, *Surveys on Discrete and Computational Geometry: Twenty Years Later*, volume 453 of *Contemporary Mathematics*, pages 9–48. American Mathematical Society, 2008.

[4] P. K. Agarwal and M. Sharir. Arrangements and their applications. In J.-R. Sack and J. Urrutia, editors, *Handbook of computational geometry*, pages 49–119. Elsevier, 2000.

[5] N. Ailon, A. Bhattacharya, R. Jaiswal, and A. Kumar. Approximate clustering with same-cluster queries. In A. R. Karlin, editor, *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, volume 94 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 40:1–40:21, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[6] D. Aloise, A. Deshpande, P. Hansen, and P. Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine learning*, 75(2):245–248, 2009.

[7] H. Angelidakis, P. Awasthi, A. Blum, V. Chatziafratis, and C. Dan. Bilu-Linial stability, certified algorithms and the independent set problem. Preprint, October 2018.

[8] H. Angelidakis, K. Makarychev, and Y. Makarychev. Algorithms for stable and perturbation-resilient problems. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 438–451. ACM, 2017.

[9] S. Arora, P. Raghavan, and S. Rao. Approximation schemes for Euclidean k-medians and related problems. In *STOC*, volume 98, pages 106–113, 1998.

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

[11] H. Ashtiani, S. Kushagra, and S. Ben-David. Clustering with same-cluster queries. In *Advances in neural information processing systems*, pages 3216–3224, 2016.

[12] P. Awasthi, A. Blum, and O. Sheffet. Center-based clustering under perturbation stability. *Information Processing Letters*, 112(1–2):49–54, 2012.

[13] A. Bakshi and N. Chepurko. Polynomial time algorithm for 2-stable clustering instances. Preprint, July 2016.

[14] M.-F. Balcan, A. Blum, and A. Gupta. Approximate clustering without the approximation. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pages 1068–1077. Society for Industrial and Applied Mathematics, 2009.

[15] M.-F. Balcan, A. Blum, and S. Vempala. A discriminative framework for clustering via similarity functions. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 671–680. ACM, 2008.

[16] M.-F. Balcan and M. Braverman. Nash equilibria in perturbation-stable games. *Theory of Computing*, 13(1):1–31, 2017.

[17] M.-F. Balcan, N. Haghtalab, and C. White. k-center clustering under perturbation resilience. In I. Chatzigiannakis, M. Mitzenmacher, Y. Rabani, and D. Sangiorgi, editors, *43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016)*, volume 55 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 68:1–68:14, Dagstuhl, Germany, 2016. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[18] M. F. Balcan and Y. Liang. Clustering under perturbation resilience. *SIAM Journal on Computing*, 45(1):102–155, 2016.

[19] S. Ben-David and L. Reyzin. Data stability in clustering: A closer look. *Theoretical Computer Science*, 558(1):51–61, 2014.

[20] Y. Bilu and N. Linial. Are stable instances easy? *Combinatorics, Probability and Computing*, 21(5):643–660, 2012.

[21] P. B. Callahan and S. R. Kosaraju. Faster algorithms for some geometric graph problems in higher dimensions. In *Proceedings of the fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 291–300. Society for Industrial and Applied Mathematics, 1993.

[22] C. Chekuri and S. Gupta. Perturbation resilient clustering for k-center and related problems via LP relaxations. In E. Blais, K. Jansen, J. D. P. Rolim, and D. Steurer, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018)*, volume 116 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 9:1–9:16, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[23] V. Cohen-Addad. A fast approximation scheme for low-dimensional k-means. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, pages 430–440, Philadelphia, PA, USA, 2018. Society for Industrial and Applied Mathematics.

[24] V. Cohen-Addad, A. de Mesmay, E. Rotenberg, and A. Roytman. The bane of low-dimensionality clustering. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, pages 441–456, Philadelphia, PA, USA, 2018. Society for Industrial and Applied Mathematics.

[25] V. Cohen-Addad, A. E. Feldmann, and D. Saulpic. Near-linear time approximation schemes for clustering in doubling metrics. Preprint, June 2019.

[26] 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. *SIAM Journal on Computing*, 48(2):644–667, 2019.

[27] V. Cohen-Addad and C. Schwiegelshohn. On the local structure of stable clustering instances. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 49–60. IEEE, 2017.

[28] S. Dasgupta. The hardness of k-means clustering. Technical report, Department of Computer Science and Engineering, University of California, 09 2008.

[29] M. De Berg, M. Van Kreveld, M. Overmars, and O. Schwarzkopf. Computational geometry. In *Computational geometry*, pages 1–17. Springer, 1997.

[30] A. Deshpande, A. Louis, and A. Vikram Singh. On Euclidean k-means clustering with α-center proximity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2087–2095, 2019.

[31] A. Dutta, A. Vijayaraghavan, and A. Wang. Clustering stable instances of Euclidean k-means. In *Advances in Neural Information Processing Systems*, pages 6500–6509, 2017.

[32] Z. Friggstad, K. Khodamoradi, and M. R. Salavatipour. Exact algorithms and lower bounds for stable instances of Euclidean k-means. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '19, pages 2958–2972, Philadelphia, PA, USA, 2019. Society for Industrial and Applied Mathematics.

[33] Z. Friggstad, M. Rezapour, and M. R. Salavatipour. Local search yields a PTAS for k-means in doubling metrics. *SIAM Journal on Computing*, 48(2):452–480, 2019.

[34] T. F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.

[35] S. Har-Peled. No, coreset, no cry. In *International Conference on Foundations of Software Technology and Theoretical Computer Science*, pages 324–335. Springer, 2004.

[36] S. Har-Peled and S. Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 291–300. ACM, 2004.

[37] D. S. Hochbaum and D. B. Shmoys. A best possible heuristic for the k-center problem. *Math. Oper. Res.*, 10(2):180–184, May 1985.

[38] K. Jain, M. Mahdian, and A. Saberi. A new greedy approach for facility location problems. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 731–740. ACM, 2002.

[39] V. Koltun. Almost tight upper bounds for vertical decompositions in four dimensions. *Journal of the ACM (JACM)*, 51(5):699–730, 2004.

[40] A. Kumar and R. Kannan. Clustering with spectral norm and the k-means algorithm. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 299–308. IEEE, 2010.

[41] E. Lee, M. Schmidt, and J. Wright. Improved and simplified inapproximability for k-means. *Information Processing Letters*, 120:40–43, 2017.

[42] S. P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[43] M. Mahajan, P. Nimbhorkar, and K. Varadarajan. The planar k-means problem is NP-hard. *Theoretical Computer Science*, 442:13–21, 2012.

[44] K. Makarychev, Y. Makarychev, and A. Vijayaraghavan. Bilu-Linial stable instances of max cut and minimum multiway cut. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 890–906. SIAM, 2014.

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

[46] M. Mihalák, M. Schöngens, R. Šrámek, and P. Widmayer. On the complexity of the metric TSP under stability considerations. In *International Conference on Current Trends in Theory and Practice of Computer Science*, pages 382–393. Springer, 2011.

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

[48] M. Sharir and P. K. Agarwal. *Davenport-Schinzel Sequences and Their Geometric Applications*. Cambridge University Press, New York, NY, USA, 1995.

[49] C. D. Toth, J. O'Rourke, and J. E. Goodman, editors. *Handbook of discrete and computational geometry*. Chapman and Hall/CRC, 2017.

A Appendix: Missing Proofs

Properties of α-center proximity. Let (X,δ) be a clustering instance satisfying α-center proximity, where δ is a metric and $\alpha > 1$. Let X_1 be a cluster with center c_1 in an optimal clustering. Let $p, p', p'' \in X_1$ and $q \in X \setminus X_1$ with c_2 the center of q's cluster. Then,

(1) $(\alpha-1)\cdot\delta(p,c_1) < \delta(p,q)$;

(2) $(\alpha-1)\cdot\delta(p,c_1) < \delta(c_1,c_1)$;

(3) $(\alpha-1)\cdot\delta(c_1,c_1) < (\alpha+1)\cdot\delta(p,q)$;

(4) $(\alpha-1)\cdot\delta(p,p') < \frac{2\alpha}{\alpha-1}\cdot\delta(p,q)$; $\quad\delta(p,p') < \delta(p,q)$ for $\alpha \geq 2+\sqrt{3}$

(5) $(\alpha-1)\cdot\delta(p',p'') < \frac{2(\alpha+1)}{\alpha-1}\cdot\delta(p,q)$. $\quad\delta(p',p'') < \delta(p,q)$ for $\alpha \geq 2+\sqrt{5}$

Proof:

(1) $\delta(p,q) \leq (\alpha-1)\cdot\delta(p,c_1)$ yields the following contradiction.

$$\alpha\cdot\delta(q,c_2) < \delta(q,c_1) \leq \delta(p,c_1)+\delta(p,q) \leq \alpha\cdot\delta(p,c_1) \quad\Rightarrow\quad \delta(q,c_2) < \delta(p,c_1)$$
$$\alpha\cdot\delta(p,c_1) < \delta(p,c_2) \leq \delta(q,c_2)+\delta(p,q) \leq \delta(q,c_2)+(\alpha-1)\cdot\delta(p,c_1) \quad\Rightarrow\quad \delta(p,c_1) < \delta(q,c_2)$$

(2) Follows from $\alpha\cdot\delta(p,c_1) < \delta(p,c_2) \leq \delta(p,c_1)+\delta(c_1,c_2)$.

(3) Follows by $\delta(c_1,c_2) \leq \delta(c_1,p)+\delta(p,q)+\delta(q,c_2) \overset{(1)}{<} \left(\frac{2}{\alpha-1}+1\right)\cdot\delta(p,q) = \frac{\alpha+1}{\alpha-1}\cdot\delta(p,q)$.

(4) Follows by

$$\begin{aligned}(\alpha-1)\cdot\delta(p,p') &\leq (\alpha-1)\cdot\delta(p,c_1)+(\alpha-1)\cdot\delta(p',c_1) \overset{(1),(2)}{<} \delta(p,q)+\delta(c_1,c_2)\\ &\overset{(3)}{<} \delta(p,q)+\frac{\alpha+1}{\alpha-1}\cdot\delta(p,q) = \frac{2\alpha}{\alpha-1}\cdot\delta(p,q).\end{aligned}$$

(5) Follows by

$$(\alpha-1)\cdot\delta(p',p'') \leq (\alpha-1)\cdot\delta(p',c_1)+(\alpha-1)\cdot\delta(p'',c_1) \overset{(2)}{<} 2\cdot\delta(c_1,c_2) \overset{(3)}{<} \frac{2(\alpha+1)}{\alpha-1}\cdot\delta(p,q).$$

□

Proof (of Lemma 2.1): Assume for contradiction that there are two optimal clusterings O and O'. There must be a point p in X that belongs to a cluster centered at c in O but is assigned to a different center c' in O'. Consider the perturbed distance $\tilde{\delta}$ by scaling inter-cluster distances in O by an α factor while preserving all intra-cluster distances unchanged. Then

$$\alpha\cdot\delta(p,c) \leq \alpha\cdot\delta(p,c') = \tilde{\delta}(p,c') \leq \tilde{\delta}(p,c) = \delta(p,c),$$

where the first inequality is by definition of clustering O, the second inequality is by definition of clustering O' still being optimal under $\tilde{\delta}$ by α-stability, and the two equalities are follows from how the perturbed distance is defined. This give a contradiction as long as $\alpha > 1$. □

Proof (of Lemma 2.2): Because (X,δ) is α-stable for $\alpha > \beta$, the optimal clustering of (X,δ) is also an optimal clustering of (X,δ') by taking δ' to be the perturbed distance. Now, for an arbitrary perturbed distance $\tilde{\delta}'$ satisfying $\delta'(p,q) \leq \tilde{\delta}'(p,q) \leq (\alpha/\beta)\cdot\delta'(p,q)$ for all $p,q \in X$, one has

$$\delta(p,q) \leq \delta'(p,q) \leq \tilde{\delta}'(p,q) \leq \frac{\alpha}{\beta}\cdot\delta'(p,q) \leq \alpha\cdot\delta(p,q),$$

and therefore the optimal clustering O of (X,δ) and (X,δ') is must be an optimal clustering of $(X,\tilde{\delta}')$, proving that (X,δ') is (α/β)-stable. Providing $\alpha>\beta$, the optimal clustering of (X,δ') is again unique by Lemma 2.1; in other words, the optimal clustering of (X,δ') is by definition equal to the optimal clustering of (X,δ). □

Proof (of Lemma 2.4): Let c_1 and c_2 be the centers of cluster X_1 and q's cluster, respectively.

(i) First we show that $\delta(c_1,c_2)<\frac{\alpha+1}{\alpha-1}\cdot\delta(p'',q)$:

$$\begin{aligned}\delta(p'',q)&\geq\delta(q,c_1)-\delta(p'',c_1)\\&\geq\delta(c_1,c_2)-\delta(q,c_2)-\delta(p'',c_1)\\&>\delta(c_1,c_2)-\big(\delta(q,p'')+\delta(p'',q)\big)/(\alpha-1).\end{aligned}$$

Rearranging the inequality proves the claim.

Now $\alpha\cdot\delta(p,c_1)<\delta(p,c_2)\leq\delta(p,c_1)+\delta(c_1,c_2)$, which implies that $(\alpha-1)\cdot\delta(p,c_1)<\delta(c_1,c_2)$. It follows that

$$\begin{aligned}\delta(p,p')&\leq\delta(p,c_1)+\delta(p',c_1)\\&<2\cdot\delta(c_1,c_2)/(\alpha-1)\\&<\frac{2(\alpha+1)}{(\alpha-1)^2}\cdot\delta(p'',q)\\&\leq\delta(p'',q),\end{aligned}$$

where the last inequality holds when $\alpha\geq 2+\sqrt{5}$.

(ii) Let us assume $\delta(c_1,c_2)=1$. We know that $\delta(p,c_2)>\alpha\cdot\delta(p,c_1)$ for all $p\in X_1$. The set of points p with $\delta(p,c_2)=\alpha\cdot\delta(p,c_1)$ is known as Apollonian Circle A_1 (with c_1 inside, but not centered at c_1!), see Fig. A.1. X_1 must be contained inside this circle A_1, or sphere in higher dimensions. Similarly, there is a sphere A_2 enclosing q's cluster (relative to X_1).

We take the classical fact that these are circles as given, but we want to understand the involved parameters. Of course, the circle A_1 has to be centered on the line ℓ through c_1 and c_2. Let a and b be the intersections of A_1 with ℓ, with b on the segment c_1c_2. $\delta(c_1,b)=\alpha\delta(c_2,b)=\alpha(1-\delta(c_1,b))$, hence $\delta(c_1,b)=\frac{1}{\alpha+1}$. Similarly, $\delta(c_1,a)=\alpha\delta(c_2,a)=\alpha(1+\delta(c_1,a))$, hence $\delta(c_1,a)=\frac{1}{\alpha-1}$. This sets the diameter of A_1 to $\frac{1}{\alpha+1}+\frac{1}{\alpha-1}=\frac{2\alpha}{\alpha^2-1}$, and the distance between A_1 and A_2 to $1-2\cdot\frac{1}{\alpha+1}=\frac{\alpha-1}{\alpha+1}$. It follows that $\delta(p',p'')/\delta(p''',q)<\frac{2\alpha}{\alpha^2-1}/\frac{\alpha-1}{\alpha+1}=\frac{2\alpha}{(\alpha-1)^2}$.

□

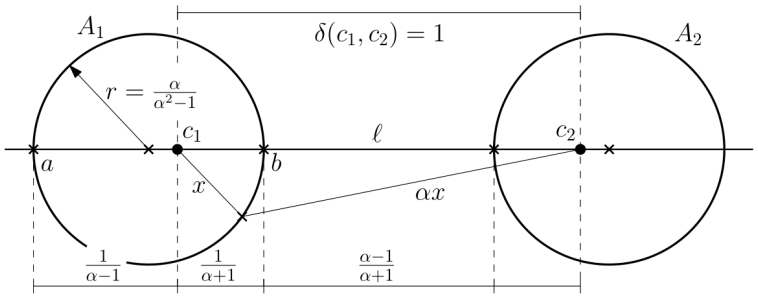

Figure A.1. The Apollonian Circles (with parameter α) for clusters centered at c_1 and c_2.

Proof (of Lemma 2.5): Consider O to be an optimal solution of an arbitrary α-stable instance (X, δ) in the continuous setting; denote the centers in O as $o_1, \ldots, o_k$. Define solution O' to be the set of centers $\text{nn}_1, \ldots, \text{nn}_k$, where nn_i is defined to be the nearest point of o_i in X. By definition O' is a discrete solution as all centers nn_i lie in X. We now argue that O' is in fact an optimal solution of the k-clustering instance (X, δ).

First we show that nn_i must be a point that was assigned to o_i in clustering O. Assume for contradiction that nn_i was in a different cluster with center o_j. Let p be an arbitrary point in cluster O_i. By center proximity one has $\delta(p, \text{nn}_i) > (\alpha - 1) \cdot \delta(p, o_i)$. But then this implies $(\alpha - 1) \cdot \delta(p, o_i) < \delta(p, \text{nn}_i) \leq \delta(p, o_i) + \delta(o_i, \text{nn}_i)$, that is, $\delta(p, o_i) \leq (\alpha - 2) \cdot \delta(p, o_i) < \delta(o_i, \text{nn}_i)$ given $\alpha \geq 3$, a contradiction.

Now again take any point p in some arbitrary cluster O_i. Compare the distances $\delta(p, \text{nn}_i)$ and $\delta(p, \text{nn}_j)$ for any other center nn_j in O'. By [19, Theorem 8] for $\alpha > 2 + \sqrt{3}$ any intra-cluster distance is smaller than any inter-cluster distance. Thus, $\delta(p, \text{nn}_i) < \delta(p, \text{nn}_j)$ since nn_i lies in O_i and nn_j lies in O_j. Therefore the clustering formed by the centers in O' is identical to the clustering of O, thus proving that O' is an optimal solution of (X, δ). □