# A Bi-Criteria Approximation Algorithm for $k$-Means*

**Konstantin Makarychev[1], Yury Makarychev[2], Maxim Sviridenko[3], and Justin Ward[4]**

1 **Microsoft Research, Redmond, WA, USA**
`komakary@microsoft.com`
2 **Toyota Technological Institute at Chicago, Chicago, IL, USA**
`yury@ttic.edu`
3 **Yahoo Labs, New York, NY, USA**
`sviri@yahoo-inc.com`
4 **Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland**
`justin.ward@epfl.ch`

## Abstract

We consider the classical $k$-means clustering problem in the setting of bi-criteria approximation, in which an algorithm is allowed to output $\beta k > k$ clusters, and must produce a clustering with cost at most $\alpha$ times the to the cost of the optimal set of $k$ clusters. We argue that this approach is natural in many settings, for which the exact number of clusters is a priori unknown, or unimportant up to a constant factor. We give new bi-criteria approximation algorithms, based on linear programming and local search, respectively, which attain a guarantee $\alpha(\beta)$ depending on the number $\beta k$ of clusters that may be opened. Our guarantee $\alpha(\beta)$ is always at most $9+\epsilon$ and improves rapidly with $\beta$ (for example: $\alpha(2) < 2.59$, and $\alpha(3) < 1.4$). Moreover, our algorithms have only polynomial dependence on the dimension of the input data, and so are applicable in high-dimensional settings.

**1998 ACM Subject Classification** F.2 Analysis of Algorithms and Problem Complexity

**Keywords and phrases** k-means clustering, bicriteria approximation algorithms, linear programming, local search

**Digital Object Identifier** 10.4230/LIPIcs.APPROX-RANDOM.2016.14

## 1 Introduction

The $k$-means clustering problem is one of the most popular models for unsupervised machine learning. The problem is formally defined as follows.

▶ **Definition 1.** In the *$k$-means* problem, we are given a set $X$ of $n$ points $x_1, \dots, x_n$ in $\mathbb{R}^p$ and an integer parameter $k \geq 1$. Our goal is to partition $X$ into $k$ clusters $S_1, \dots, S_k$ and assign each cluster a center $a_i$ so as to minimize the cost $\sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - a_i\|^2$.

The most common heuristic for $k$-means is Lloyd's algorithm introduced in 1957 [22, 23]. Lloyd's algorithm starts with some initial solution and then iteratively improves it by alternating two steps: at the first step, the algorithm picks the optimal clustering for the current set of centers; at the second step, the algorithm picks the optimal set of centers for

---

[*] Work partially supported by EPSRC grant EP/J021814/1 and ERC Starting Grant 335288-OptApprox.

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2016).
Editors: Klaus Jansen, Claire Matthieu, José D. P. Rolim, and Chris Umans; Article No. 14; pp. 14:1–14:20
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

the current clustering. While we know that the algorithm performs well on well-clusterable data [26] it performs arbitrarily badly on general instances. There exist many variants of this algorithm and many heuristics for picking the initial solution. Unfortunately, none of them give a constant (not depending on $k$) factor approximation. One of the most popular ones is the $k$-means++ algorithm that has an $O(\log k)$-approximation factor [5].

There are several results showing that $k$-means is NP-hard even in restricted special cases [3, 24, 13]. The general $k$-means clustering problem has recently been shown to be APX-hard, ruling out a PTAS in the general case [7]. However, a variety of PTASes exist for special cases of the problem. Inaba, Katoh, and Imai [16] gave a $(1 + \varepsilon)$-approximation algorithm for the case in which the number of clusters, $k$, and the dimension of the space, $p$, are fixed. Since then many more PTASes were proposed for other special cases. Most recently, PTASes have been obtained via local search algorithms in the general setting in which distances come from a metric of constant doubling dimension [15] or from a graph with forbidden minors [12]. Both of these results can be specialized yield PTASes in the standard, Euclidean setting considered here whenever the dimension of the space is fixed.

In the general case, in which the dimension is not fixed, the best constant factor approximation algorithm was proposed by Kanungo et al. [18]. Their algorithm gives $9 + \varepsilon$ factor approximation. Previously, Jain and Vazirani [17] gave a (larger) constant-factor approximation for a discrete variant of $k$-means, via the primal-dual method. Using and connection with the $k$-median problem, Anagnostopoulos, Dasgupta, and Kumar [4] also designed a constant factor approximation algorithm for the general *co-clustering problem*, which includes $k$-means as a special case. Aggarwal, Deshpande, and Kanan [2] showed that running the $k$-means++ algorithm for more steps gives an $\alpha = 4 + \varepsilon$ factor approximation by opening $\lceil 16(k + \sqrt{k})\rceil$ centers, and also showed how to modify the resulting solution to obtain a set of $k$ centers attaining an $O(1)$ factor guarantee.

In most practical applications the target number $k$ of clusters is not fixed in advance. Rather, we would like to find a number $k$ that provides a well-clusterable solution. Here, we show how to substantially improve the approximation factor by slightly violating the constraint on the number of clusters. We present bi-criteria approximation algorithms for the general case of the problem. A $(\beta, \alpha)$ bi-criteria approximation algorithm finds a solution with $\beta k$ clusters, whose cost is at most $\alpha$ times the optimal cost of a solution using $k$ clusters. In contrast to the approach of Aggarwal, Deshpand, and Kanan [2], our algorithms find an approximate solution for every $\beta > 1$. Our approximation is always at most 9, and decreases rapidly with $\beta$. In particular, we obtain a 4-approximation by opening only $1.65k$ centers, improving over previous results [2] by a factor of nearly 10, and obtain improved approximation factors $\alpha(\beta)$ as $\beta$ continues to grow. For example, $\alpha(1.3) < 6.45$, $\alpha(1.5) < 4.8$; $\alpha(2) < 2.59$, and $\alpha(3) < 1.4$. In general, we argue that in many applications the number of clusters is not important as long as it approximately equals $k$. For these applications we can obtain an approximation factor very close to 1.

We give three bi-criteria algorithms – two based on linear programming and one based on local search. We show the algorithms' approximation factors as a function of $\beta$ in Figure 1. Note that our linear programming algorithm attains a better approximation $\alpha$ for large $\beta$, while the local search algorithm is better for $\beta$ near 1.

Both of our algorithms are based on a reduction from the general $k$-means problem, in which cluster centers may be placed at any point in $\mathbb{R}^p$, to the following problem, in which we are restricted to a given, discrete set of candidate cluster centers with specified distances from each point. As part of reduction, we utilize dimensionality reduction to ensure that the number of discrete candidate centers that must be considered is polynomial in both the number of points $n$ and in the dimension $p$.

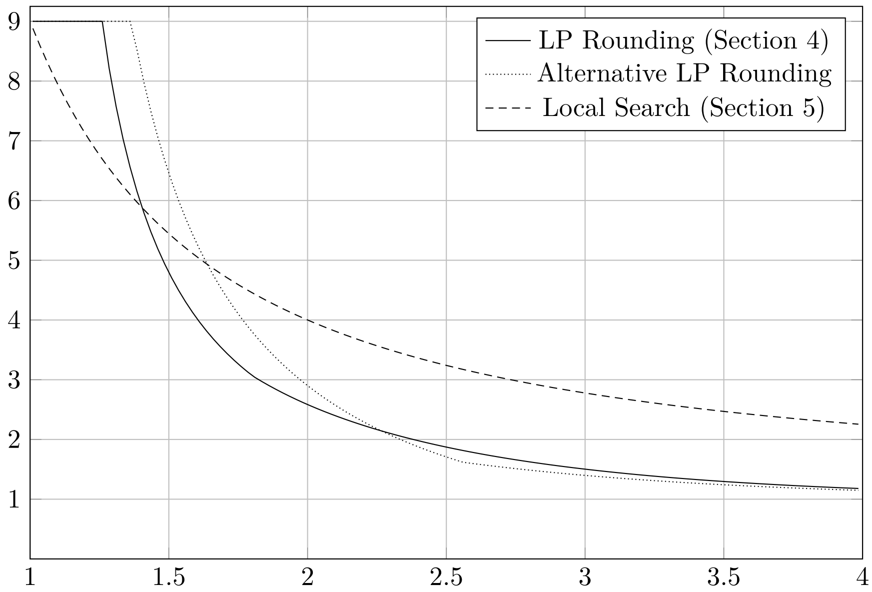

**Figure 1** Approximation ratios obtained from opening $\beta k$ centers.

▶ **Definition 2.** In the *$k$-median problem*, we are given a set of points $\mathcal{D}$, a set of potential center locations $\mathcal{C}$ and a distance function[1] $(d(i,j))_{i\in\mathcal{C},j\in\mathcal{D}}$. The cost of assigning point $j$ to center $i$ is $d(i,j)$. Our goal is to open at most $k$ centers and assign each point to a center so as to minimize the total cost.

The first approximation algorithms for the $k$-median problem were given by Lin and Vitter [21], who gave an LP-rounding algorithm that attains an approximation factor of $1+\varepsilon$ by opening $O(k\ln n)$ centers (i.e. a $(1+\epsilon, O(\ln n))$ bi-criteria approximation). In further work, Lin and Vitter [20] showed that if the distance function $d$ is a *metric*, it is possible to obtain a $2(1+\varepsilon)$ approximation algorithm by opening only $(1+1/\varepsilon)k$ centers. The first constant-factor approximation for the metric $k$-median problem using only $k$ centers was obtained by Arya et al. [6], who showed that a simple local search algorithm gives a $3+\varepsilon$ approximation. This remained the state of the art until recently, when Li and Svensson [19] gave a $2.732+\varepsilon$ approximation algorithm based on LP rounding. Subsequently, this has been improved to $2.675+\varepsilon$ by Byrka [8].

Unfortunately, our resulting $k$-median instance is non-metric, and so we must employ an alternative to the standard triangle inequality in our analysis. In the case of our LP-based algorithms, we use the fact that our reduction produces instances satisfying a 3-relaxed 3-hop triangle inequality, a concept that we define in Section 2. In the case of local search, we note that given any partition of points of $\mathbb{R}^p$ into clusters $S_1,\ldots,S_k$, the optimal location of each $k$-means cluster $S_i$'s center is the centroid of all points in $S_i$. This, combined with the fact that our reduction to $k$-median approximately preserves the $k$-means cluster costs allows us to employ a similar approach to that of Kanungo et al. [18].

---

[1] Here, and throughout, we do not require the distance function to be a metric, as in some alternative definitions of the $k$-median problem. In particular, in our $k$-median instance, the distances will be the *squared* distances from given $k$-means instance.

## 1.1 Our Results

We give three approximation algorithms. The first algorithm is based on linear programming. It gives an

$$\alpha_1(\beta) = 1 + e^{-\beta}\Big(\frac{6\beta}{1-\beta} + \frac{(\beta-1)^2}{\beta}\Big)$$

approximation (see also (14) for a slightly tighter bound). The second algorithm is based on local search. It gives an

$$\alpha_2(\beta) = (1 + O(\varepsilon))\Big(1 + \frac{2}{\beta}\Big)^2$$

approximation. The third algorithm is also based on linear programming. It gives an

$$\alpha_3(\beta) = \max\big(1 + 8e^{-\beta}, \frac{\beta(e^{-1} + 8e^{-\beta})}{\beta - 1}\big)$$

approximation. The algorithm is similar to the first algorithm, but it uses pipage rounding (see [1]) instead of randomized rounding. In the conference version of the paper, we omit the description of the third algorithm. The approximation factors are shown in Figure 1.

In Section 2, we introduce the notation that we shall use throughout the rest of the paper and review standard notions related to both the $k$-means and $k$-median problems. In Section 3, we give the details of our reduction to the $k$-median problem. Finally, in Sections 4 and 5, respectively, we present our main LP-based algorithm and local search algorithm for the resulting $k$-median instances. For the sake of presentation, we defer some technical details to the appendix.

# 2 Preliminaries

We now fix some notation, and recall some basic properties of $k$-means solutions and the standard linear program for the $k$-median problem. Additionally, we define the notion of an $\alpha$-relaxed 3-hop triangle inequality, which will be crucial to the analysis of our LP-rounding algorithms.

## 2.1 $k$-means

Consider a given instance of the $k$-means problem, specified by a set of points $X \in \mathbb{R}^p$. Given a partition $S = \langle S_1, \dots, S_k\rangle$ of $X$ and a set $C = \langle c_1, \dots, c_k\rangle$ of centers in $\mathbb{R}^p$, denote by $\mathrm{cost}_X(S, C)$ the total cost of the clustering that, for each $1 \le i \le k$ assigns each point of $S_i$ to the center $c_i$:

$$\mathrm{cost}_X(S, C) = \sum_{i=1}^{k} \sum_{x \in S_i} \|x - c_i\|^2.$$

Note that to describe an optimal solution to the $k$-means problem, it is sufficient to specify either all clusters or all centers in the solution. Indeed, given a list of clusters $S_1, \dots, S_k$, we can find the optimal assignment of centers $c_i$ for it: the optimal choice of center $c_i$ for $S_i$ is $\frac{1}{|S_i|}\sum_{x_j \in S_i} x_j$. For this choice of $c_i$, we have

$$\sum_{x \in S_i} \|x - c_i\|^2 = \frac{1}{2|S_i|} \sum_{x', x'' \in S_i} \|x' - x''\|^2. \tag{1}$$

Given a partition $S = \langle S_1, \ldots, S_k \rangle$ of $X$ into clusters, we then denote by $\mathrm{cost}_X(S)$ the cost of this optimal choice of centers. That is,

$$\mathrm{cost}_X(S) = \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{x', x'' \in S_i} \|x' - x''\|^2.$$

Similarly, given a list $C$ of centers $c_1, \ldots, c_k$, we can find the optimal partition $S = \langle S_1, \ldots, S_k \rangle$ of $X$ into clusters. For each $c \in C$, let $N_C(c)$ be the set of those points $x \in X$ that are closer to $c_i$ than to other centers $c_j \neq c_i$ (if a point $x$ is at the same distance from several centers, we break the ties arbitrarily). The optimal partition for $C$ then sets $S_i = N_C(c_i)$. Given a set $C$ of $k$ centers, we define $\mathrm{cost}_X(C) \equiv \mathrm{cost}_X(T)$, where $T = \langle N_C(c_1), \ldots, N_C(c_k) \rangle$ is the partition induced by $C$.

## 2.2 $k$-median

We will reduce a given instance $X$ of the $k$-means problem to an instance of the (non-metric) discrete $k$-median problem, specified by $\langle \mathcal{D}, \mathcal{C}, d \rangle$. By analogy with the $k$-means problem, we can consider a partition $S = S_1, \ldots, S_k$ of points from $\mathcal{D}$, and then consider the best choice of a single center for each partition. We denote the cost of this choice by $\mathrm{cost}_{\mathcal{D},d}(S)$:

$$\mathrm{cost}_{\mathcal{D},d}(S) = \sum_{i=1}^{k} \min_{x \in \mathcal{C}} \sum_{j \in S_i} d(x, j).$$

Similarly, given a list of $k$ centers $C = \langle c_1, \ldots, c_k \rangle$, let $N_C(c_i)$ be the set of those points $x \in \mathcal{D}$ that are closer (according to the distance function $d$) to $c_i$ than to any other center in $C$ (again, if a point $x$ is at the same distance from several facilities, we break ties arbitrarily). As in the case of $k$-means, we define $\mathrm{cost}_{\mathcal{D},d}(C) \equiv \mathrm{cost}_{\mathcal{D},d}(T)$ where $T = \langle N_C(c_1), \ldots, N_C(c_k) \rangle$ is the partition of $\mathcal{D}$ induced by $C$.

Although the distance function $d$ in our $k$-median instances will not satisfy the standard triangle inequality, we can show that it satisfies a relaxed variant of the following sort:

▶ **Definition 3.** We say that $d$ satisfies an *$\alpha$-relaxed 3-hop triangle inequality* on $\mathcal{D} \cup \mathcal{C}$ if, for any $j, j' \in \mathcal{D}$ and $i, i' \in \mathcal{C}$, we have

$$d(i, j) \leq \alpha \left( d(i, j') + d(i', j') + d(i', j) \right).$$

Specifically, we shall show that the distances produced by our reduction satisfy a 3-relaxed 3-hop triangle inequality.

# 3 Reduction from $k$-means to $k$-median

We now give the details of our reduction from the $k$-means to the $k$-median problem. In the $k$-median problem, a finite set $\mathcal{C}$ of candidate centers is specified, while in the $k$-means problem, the ideal center for each cluster $S_i$ of points is given by the centroid of $S_i$. Ideally, we want to ensure that for every possible centroid of the original $k$-means instance, there is some nearby candidate center in $\mathcal{C}$. The following notion of an $\varepsilon$-approximate centroid set, introduced by[2] Matoušek [25], captures this requirement.

---

[2] Matoušek's original definition requires that $\mathcal{C}$ contains some point in an *$\epsilon$-tolerance ball* centered at the centroid of each non-empty cluster of points from $X$. The condition presented here follows from this one (see, for example, the proof of Lemma 4.1, in [25]).

▶ **Definition 4.** A set of points $\mathcal{C} \subset \mathbb{R}^p$ is an $\varepsilon$-approximate centroid set for $X \subset \mathbb{R}^p$ if for every $S \subset X$,

$$\min_{c \in \mathcal{C}} \sum_{x \in S} \|x - c\|^2 \leq (1+\varepsilon) \min_{c \in \mathbb{R}^p} \sum_{x \in S} \|x - c\|^2.$$

Observe that if $\mathcal{C}$ is an $\varepsilon$-approximate centroid set for $X$, then for every set of $k$ centers $C$ (in particular, for the optimal set $C^*$), there exists a $k$-point subset $\widetilde{C} \subset \mathcal{C}$ such that

$$\mathrm{cost}_X(\widetilde{C}) = \sum_{x \in X} \min_{c \in \widetilde{C}} \|x - c\|^2 \leq (1+\varepsilon) \sum_{x \in X} \min_{c \in C} \|x - c\|^2 = (1+\varepsilon)\, \mathrm{cost}_X(C).$$

Thus, if we restrict our search for $k$ center points in the $k$-means problem to only those points of $\mathcal{C}$, we lose at most a factor of $(1+\varepsilon)$.

Matoušek showed that for every set $X$ in $\mathbb{R}^p$ and $\varepsilon > 0$, there exists an $\varepsilon$-approximate centroid set of size $O(|X|\varepsilon^{-p}\log(1/\varepsilon))$.

▶ **Theorem 5** (Theorem 4.4 in [25]). *Given an $n$-point set $X \subset \mathbb{R}^p$ and $\varepsilon > 0$, an $\varepsilon$-approximate centroid set for $X$ of size $O(n\varepsilon^{-p}\log(1/\varepsilon))$ can be computed in time $O(n \log n + n\varepsilon^{-p}\log(1/\varepsilon))$.*

Unfortunately, in our setting, the dimension $p$ of the space in which points $x_1, \dots, x_n$ lie may be as large as $n$. Thus, in order to apply Theorem 5, we first embed $X$ into a low-dimensional space using the Johnson–Lindenstrauss transform.

▶ **Theorem 6** (Johnson–Lindenstrauss Flattening Lemma). *For every set of points $X$ in $\mathbb{R}^p$ and $\varepsilon \in (0,1)$, there exists a map $\varphi$ of $X$ into $\tilde{p} = O(\log|X|/\varepsilon^2)$ dimensional space such that*

$$\|x - y\|_2^2 \leq \|\varphi(x) - \varphi(y)\|_2^2 \leq (1+\varepsilon)\|x - y\|_2^2. \tag{2}$$

*We say that the map $\varphi$ is a* dimension reduction transform *for $X$.*

Given an instance $X$ of $k$-means, we apply the dimension reduction transform to $X$, get a set $X' \subset \mathbb{R}^{\tilde{p}}$, and then find an $\varepsilon$-approximate centroid set $\mathcal{C}$ to $X'$. We obtain an instance $\langle X', \mathcal{C}, d\rangle$ of $k$-median with the squared Euclidean distance $d$. We show in Theorem 7 that the value of this instance is within a factor of $(1+\varepsilon)$ of the value of instance $X$ of $k$-means, and, moreover, that there is a one-to-one correspondence between solutions of instance $\langle X', \mathcal{C}, d\rangle$ and solutions of instance $X$. We defer the proof of Theorem 7 to Appendix A.

▶ **Theorem 7.** *The following hold:*

1. *For every $\varepsilon \in (0, 1/2)$, there exists a polynomial-time reduction from $k$-means to $k$-median with distance function that satisfies the 3-relaxed 3-hop triangle inequality. Specifically, given an instance $X$ of $k$-means, the reduction outputs an instance $\langle \mathcal{D}, \mathcal{C}, d\rangle$ of $k$-median with $|\mathcal{D}| = |X|$, $|\mathcal{C}| = n^{O(\log(1/\varepsilon)/\varepsilon^2)}$, and distance $d$ that satisfies the 3-relaxed 3-hop triangle inequality such that*

   $$\mathrm{OPT}_X \leq \mathrm{OPT}_{\langle \mathcal{D}, d\rangle} \leq (1+\varepsilon)\mathrm{OPT}_X,$$

   *where $\mathrm{OPT}_X$ is the value of the optimal solution to $X$ and $\mathrm{OPT}_{\langle \mathcal{D}, d\rangle}$ is the value of the optimal solution to $\langle \mathcal{D}, \mathcal{C}, d\rangle$. The reduction also gives a one-to-one correspondence $\psi : \mathcal{D} \to X$ such that*

   $$\mathrm{cost}_X(\psi(S)) \leq \mathrm{cost}_{\mathcal{D},d}(S) \leq (1+\varepsilon)\, \mathrm{cost}_X(\psi(S)),$$

   *where $S = \langle S_1, \dots, S_k\rangle$ is a partition of $\mathcal{D}$ and $\psi(S) = \langle \psi(S_1), \dots, \psi(S_k)\rangle$ is the corresponding partition of $X$. The reduction runs in time $n^{O(\log(1/\varepsilon)/\varepsilon^2)}$.*

2. *In instance* $\langle \mathcal{D}, \mathcal{C}, d\rangle$, $\mathcal{C} \subset \mathbb{R}^{\tilde{p}}$ *(for some* $\tilde{p}$*),* $\mathcal{C}$ *is an* $(\varepsilon/3)$*-approximate centroid set for* $\mathcal{D}$*, and* $d(c, x) = \|c - x\|^2$.

We remark, briefly, some elements of our reduction may be improved by using more sophisticated approaches for constructing approximate centroids and core sets (e.g. [14]), as well as recent specialized dimensionality reduction techniques [11]. Here we have chosen instead to present a more straightforward reduction.

## 4 Algorithm for $k$-Median with Relaxed Triangle Inequality

We now turn to the problem of approximating the $k$-median instance from Theorem 7. Our first algorithm is based on the following standard linear programming relaxation for the $k$-median problem:

$$\min \sum_{c\in\mathcal{C}} \sum_{x\in X} z_{xc} d(x, c), \tag{3}$$

$$\sum_{c\in\mathcal{C}} y_c = k, \tag{4}$$

$$\sum_{c\in\mathcal{C}} z_{xc} = 1, \qquad \forall x \in X, \tag{5}$$

$$z_{xc} \le y_c, \qquad \forall c \in \mathcal{C}, j \in X, \tag{6}$$

$$z_{xc}, y_c \ge 0. \tag{7}$$

In the integral solution, each variable $y_c$ indicates whether the center $c$ is open; and each variable $z_{xc}$ indicates whether the point $x$ is assigned to the center $c$. Constraint (4) asserts that we should open exactly $k$ centers; constraint (5) ensures that every point is assigned to exactly one center; finally, constraint (6) says that points can be assigned only to open centers. In a fractional LP solution, all $z_{xc}$ and $y_c$ lie in the interval $[0, 1]$. Note that in the integral solution, $z_{xc} = y_c$, if $z_{xc} > 0$ (as both $z_{xc}$ and $y_c$ must be equal to 1). We can slightly change any feasible LP solution so it also satisfies this property. Specifically, we split any center $c$ which does not satisfy $y_c = z_{xc}$ (for some $x \in X$) into two co-located centers $c_1$ and $c_2$: one with weight $z_{xc}$ and the other with weight $y_c - z_{xc}$. We distribute the weights $z_{x'c}$ among them as follows: we let $z_{x'c_1} = \min(z_{x'c}, y_{c_1})$; $z_{x'c_2} = y_{c_2} - \min(z_{x'c}, y_{c_1})$. Note that this is a standard assumption in the $k$-median literature. We refer the reader to [27] (see Lemma 1) and [10] for more details. The values $y_c$ define the measure $y$ on $\mathcal{C}$: $y(C) = \sum_{c\in C} y_c$. In the rounding algorithm and in the analysis, it will be convenient to think of this measure as a "continuous measure": That is, if needed we will split the centers into co-located centers to ensure that we can find a set of any given measure $\mu$.

For every point $x \in X$, let $C_x = \{c \in \mathcal{C} : z_{xc} > 0\}$. The set $C_x$ contains all centers that serve $x$ in the LP solution. Recall that we modify the solution so that $y_c = z_{xc}$ if $z_{xc} > 0$. Hence, $y_c = z_{xc}$ if $x \in C_x$. For every point $x \in X$, we define its LP radius $R_x$ as:

$$R_x = \sum_{c\in\mathcal{C}} z_{xc} d(x, c) = \sum_{c\in C_x} y_c d(x, c).$$

Observe, that the LP value, which we denote by $LP$, equals $\sum_{x\in X} R_x$.

**Algorithm.** We now describe our LP-rounding algorithm for the $k$-median problem with relaxed 3-hop triangle inequality.

▶ **Theorem 8.** *There exists a $(\beta, \alpha)$ bi-criteria approximation algorithm for $k$-means with*

$$\alpha(\beta) = 1 + e^{-\beta}\Big(\frac{6\beta}{1-\beta} + \frac{(\beta-1)^2}{\beta}\Big) \tag{8}$$

*for every $\beta > 1$.*

The algorithm first solves the LP problem and modifies the LP solution as described above if necessary. Then, it partitions all centers into $\beta k$ groups $Z \in \mathcal{Z}$, each with LP measure $1/\beta$. It picks one center $c$ at random from each group $Z$ with probability $\beta y_c$ (note that $\sum_{c \in Z} \beta y_c = 1$). The algorithm outputs the set of $\beta k$ chosen centers, and assigns every point to the closest center.

We now describe the construction of $\mathcal{Z}$ in more detail. We partition centers into $\beta k$ groups as follows. For every $x \in X$, we find the unique ball $B_x$ around $x$ whose LP weight exactly equals $1/\beta$ (To do so, we may split some centers, and pick some centers in $B_x$ at the boundary of the ball but not the others). We find a subset of points $\mathcal{W}$ such that balls $B_x$ with $x \in \mathcal{W}$ are disjoint, and for every point $x \in X$, we also define a "witness" $w(x) \in \mathcal{W}$. To this end, we sort all points $x \in X$ by the LP radius $R_x$ in the ascending order, and then consider them one by one. For each $x \in X$, if $B_x$ is disjoint from all previously chosen balls, then we add $x$ to the set $\mathcal{W}$ and set $w(x) = x$. Otherwise, if $B_x$ intersects some other ball $B_{x'}$ that is already chosen, we discard $B_x$ and set $w(x) = x'$. If there are several balls $B_{x'}$ intersecting $B_x$, we pick the first $x'$ according to our ordering as the witness. Note, that $R_{w(x)} \leq R_x$ for all $x$. Once we have found a disjoint collection of balls $\{B_x : x \in \mathcal{W}\}$, we add them to the set $\mathcal{Z}$. We partition centers not covered by $\cup_{x \in \mathcal{W}} B_x$ into groups of LP weight $1/\beta$ arbitrarily and add these groups to $\mathcal{Z}$. Thus, we obtain a partitioning $\mathcal{Z}$ of all centers into groups of LP weight $1/\beta$.

**Analysis.** We show that the algorithm returns a valid solution, and then prove an upper bound on its expected cost. For the sake of presentation, we defer some technical claims to Appendix B.

The algorithm picks exactly one center from each group, so it always picks $\beta k$ centers. Hence, it always outputs a valid solution. Let $S$ be the set of centers output by the algorithm. Denote the radius of the ball $B_x$ by $R_x^\beta$. For every center $x$, we estimate the expected distance from $x$ to the closest center $c$ in the solution $S$ i.e. $\mathbb{E}[d(x,S)]$. We show that $\mathbb{E}[d(x,S)] \leq \alpha(\beta)\, R_x$ for $\alpha(\beta)$ as in equation (8). Since $LP = \sum_x R_x$, we conclude that the algorithm has an approximation factor of $\alpha(\beta)$.

Fix $x \in X$. Recall, that $C_x = \{c : z_{xc} > 0\}$ is the set of all centers that serve $x$ in the LP solution. We upper bound $d(x,S)$ by $d(x, (C_x \cup B_{w(x)}) \cap S)$, which is the distance to the closest center in $C_x \cup B_{w(x)}$ chosen by the algorithm. Note that the solution $S$ always contains at least one center in $B_{w(x)}$, so $(C_x \cup B_{w(x)}) \cap S \neq \varnothing$. For the proof, we pick a particular (random) center $f(x) \in (C_x \cup B_{w(x)}) \cap S$.

We define $f(x)$ using the following randomized procedure. Consider the partitioning $\mathcal{Z}$ of all centers into groups of measure $1/\beta$ used by the algorithm. Let $\widetilde{\mathcal{Z}} = \{Z \cap C_x : Z \in \mathcal{Z};\ Z \cap C_x \neq \varnothing\}$ be the induced partitioning of the set $C_x$. For all $\widetilde{Z} \in \widetilde{\mathcal{Z}}$ we independently flip a coin and with probability $(1 - e^{-\beta y(\widetilde{Z})})/(\beta y(\widetilde{Z}))$ make the set $\widetilde{Z}$ *active*. We let $A \subset C_x$ to be the union of all active sets $\widetilde{Z}$; we say that centers in $A$ are active centers. Let $f(x)$ be the center in $A \cap S$ closest to $x$, if $A \cap S \neq \varnothing$ ; let $f(x)$ to be the unique center in $B_{w(x)} \cap S$, otherwise. We set $\mathcal{E} = 0$, if $A \cap S \neq \varnothing$; and $\mathcal{E} = 1$, otherwise. Roughly speaking, $\mathcal{E}$ indicates whether $f(x) \in C_x$ or $f(x) \in B_{w(x)}$: Specifically, if $\mathcal{E} = 0$, then $f(x) \in C_x$; if $\mathcal{E} = 1$, then $f(x) \in B_{w(x)}$. Note, however, that $C_x \cap B_{w(x)} \neq \varnothing$, and $f(x)$ may belong to $C_x \cap B_{w(x)}$.

The center $f(x)$ may not be the closest to $x$, but since $f(x) \in S$, we have

$$d(x, S) \leq d(x, (C_x \cup B_{w(x)}) \cap S) \leq d(x, f(x)).$$

Let us now derive bound on the expected distance of a single client $x$ to $f(x)$. We begin by considering the probability of the event $\mathcal{E}$.

▶ **Lemma 9.** $\Pr(\mathcal{E} = 0) = 1 - e^{-\beta}$.

**Proof.** Recall that the algorithm picks one center $c$ in every $Z \in \mathcal{Z}$ uniformly (with respect to the measure $y$) at random. Thus, the probability that the algorithm picks a center from $\tilde{Z}$ equals $\beta y(\tilde{Z})$. The probability that a given $\tilde{Z}$ contains a point from the solution $S$ and $\tilde{Z}$ is active equals $\beta y(\widetilde{Z}) \times (1 - e^{\beta y(\widetilde{Z})})/(\beta y(\widetilde{Z})) = (1 - e^{\beta y(\widetilde{Z})})$. The probability that no such $\widetilde{Z}$ exists equals

$$\prod_{\widetilde{Z} \in \widetilde{\mathcal{Z}}} e^{-\beta y(\widetilde{Z})} = e^{-\sum_{\widetilde{Z} \in \widetilde{\mathcal{Z}}} \beta y(\widetilde{Z})} = e^{-\beta y(C_x)} = e^{-\beta}. \qquad ◀$$

Using Lemma 9 we have:

$$\begin{aligned} E[d(x, f(x))] &= \Pr(\mathcal{E} = 0)\, \mathbb{E}\big[d(x, f(x)) \mid \mathcal{E} = 0\big] + \Pr(\mathcal{E} = 1)\, \mathbb{E}\big[d(x, f(x)) \mid \mathcal{E} = 1\big] \\ &= (1 - e^{-\beta})\, \mathbb{E}\big[d(x, f(x)) \mid \mathcal{E} = 0\big] + e^{-\beta}\, \mathbb{E}\big[d(x, f(x)) \mid \mathcal{E} = 1\big]. \end{aligned} \tag{9}$$

Let us now bound each remaining above, in turn.

▶ **Lemma 10.** $E[d(x, f(x)) \mid \mathcal{E} = 0] \leq R_x$.

**Proof.** We define two sets of random variables $P$ and $Q$, and then show that they are identically distributed. If the algorithm picks a center $c$ in $\widetilde{Z}$, and $\widetilde{Z}$ is active, let $P(\widetilde{Z}) = c$. Let $P(\widetilde{Z}) = \perp$, otherwise. The random variables $P(\widetilde{Z})$ are mutually independent for all $\widetilde{Z} \in \widetilde{\mathcal{Z}}$; and

$$\Pr(\widetilde{Z} = c) = \frac{(1 - e^{-\beta y(\widetilde{Z})})\, y_c}{y(\widetilde{Z})}$$

for $c \in \widetilde{Z}$.

To define $Q$, we introduce an auxiliary Poisson arrival process. At every point of time $t \in [0, \beta]$, we pick a center $c \in C_x$ with probability $y_c dt$ (i.e., with arrival rate $y_c$). For every $\widetilde{Z}$, let $Q(\widetilde{Z})$ be the first center chosen in $\widetilde{Z}$. If no centers in $\widetilde{Z}$ are chosen, we let $Q(\widetilde{Z}) = \perp$. Note that we pick two centers at exactly the same time with probability 0, hence $Q(\widetilde{Z})$ is well defined. Conditional on $Q(\widetilde{Z}) \neq \perp$, the random variable $Q(\widetilde{Z})$ is uniformly distributed in $\widetilde{Z}$ with respect to the LP measure $y$ (since at every given time $t$, the probability of arrival equals $y_c dt$). Then, $\Pr(Q(\widetilde{Z}) \neq \perp) = (1 - e^{-\beta y(\widetilde{Z})})$. Hence, $\Pr(Q(\widetilde{Z}) = c) = (1 - e^{-\beta y(\widetilde{Z})}) y_c / y(\widetilde{Z})$. Note that all random variables $Q$ are mutually independent. Thus, the random variables $Q$ have the same distribution as random variables $P$.

Note that if $\mathcal{E} = 0$, then $f(x)$ is the closest center in $\{P(\widetilde{Z}) : \widetilde{Z} \in \widetilde{\mathcal{Z}}; P(\widetilde{Z}) \neq \perp\}$ to $x$. If $\mathcal{E} = 1$, then all $P(\widetilde{Z})$ are equal to $\perp$. Let $U_Q = \{Q(\widetilde{Z}) : \widetilde{Z} \in \widetilde{\mathcal{Z}}; Q(\widetilde{Z}) \neq \perp\}$. Since $P$ and $Q$ have the same distribution, we have

$$\mathbb{E}[d(x, f(x)) \mid \mathcal{E} = 0] = \mathbb{E}[\min_{c \in U_Q} d(x, c) \mid U_Q \neq \varnothing].$$

Conditional on $U_Q \neq \varnothing$, the first center that arrives according to our stochastic process is uniformly distributed in $C_x$, according the measure $y$. The expected distance from it to $x$ equals $R_x$. Hence, $\mathbb{E}[\min_{c \in U_Q} d(x, c) \mid U_Q \neq \varnothing] \leq R_x$. ◀

Observe that for a random center $c$ distributed according to the LP measure $y$ in $C_x$ (i.e., $\Pr(c = c_0) = y(c_0)/y(C_x) = y(c_0)$), we have the exact equality $\mathbb{E}[d(x,c)] = R_x$. So Lemma 10 shows that the distribution of $f(x)$ given $\mathcal{E} = 0$ is "not worse" than the distribution according to $y$ in $C_x$.

We now bound the expected distance from $x$ to $f(x)$ given $\mathcal{E} = 1$.

▶ **Lemma 11.** *Let $\gamma = \beta y(D_x)$. Then,*

$$\mathbb{E}[d(x,f(x)) \mid \mathcal{E} = 1] \le \left(e^{\gamma}(1-\gamma) \times 3\Big(\frac{\beta}{\beta-1} + r_1 + r_2\Big) + (1 - e^{\gamma}(1-\gamma)) \times \frac{\beta}{\gamma}\right) R_x,$$

*for some non-negative numbers $r_1$ and $r_2$ such that $r_1 \le r_2$ and $\frac{1-\gamma}{\beta} r_1 + \frac{\beta-1}{\beta} r_2 \le R_x$.*

**Proof.** Recall, that $w(x)$ is the witness for $x$. Thus, the balls $B_x$ and $B_{w(x)}$ intersect and $R_{w(x)} \le R_x$. Let $c_\circ$ be an arbitrary center in $B_x \cap B_{w(x)}$. By the relaxed 3-hop triangle inequality,

$$\begin{aligned} d(x,f(x)) &\le 3\big(d(x,c_\circ) + d(w(x),c_\circ) + d(w(x),f(x)\big) \\ &\le 3\big(R_x^{\beta} + R_{w(x)}^{\beta} + d(w(x),f(x))\big). \end{aligned} \tag{10}$$

Here, we used that $R_x^{\beta}$ is the radius of $B_x$; $R_{w(x)}^{\beta}$ is the radius of $B_{w(x)}$. Now, let $D_x = B_{w(x)} \cap C_x$. In Lemmas 15 and 16 in Appendix B, we show that:

$$R_x^{\beta} \le \beta R_x/(\beta - 1) \tag{11}$$

and

$$R_{w(x)}^{\beta} + \mathbb{E}[d(w(x),f(x)) \mid f(x) \in B_{w(x)} \setminus D_x;\, \mathcal{E} = 1] \le r_1 + r_2, \tag{12}$$

for some pair of nonnegative numbers $r_1$ and $r_2$ ($r_1 \le r_2$) satisfying the conditions of the lemma. Taking expectations conditioned on $f(x) \in B_{w(x)} \setminus D_x$ and $\mathcal{E} = 1$ in (10), and then applying the bounds (11) and (12), we obtain:

$$\mathbb{E}[d(x,f(x)) \mid f(x) \in B_{w(x)} \setminus D_x;\, \mathcal{E} = 1] \le 3\Big(\frac{\beta R_x}{\beta - 1} + r_1 + r_2\Big). \tag{13}$$

In Lemmas 17 and 18 in the Appendix, we show, respectively, that

$$\Pr(f(x) \in B_{w(x)} \setminus D_x \mid \mathcal{E} = 1) = e^{\gamma}(1-\gamma),$$

and

$$\mathbb{E}[d(x,f(x)) \mid f(x) \in D_x;\, \mathcal{E} = 1] \le \frac{\beta R_x}{\gamma}.$$

Combining these bounds with (13), we obtain the desired inequality. ◀

Combining Lemmas 10 and 11 with (9), we have:

$$\mathbb{E}[d(x,f(x))] \le R_x\left((1-e^{-\beta}) + e^{-\beta}\left(e^{\gamma}(1-\gamma) \times 3\Big(\frac{\beta}{\beta-1} + \frac{r_1+r_2}{R_x}\Big) + (1-e^{\gamma}(1-\gamma)) \times \frac{\beta}{\gamma}\right)\right).$$

Now, we recall that $\gamma = \beta y(D_x)$, and note that $\gamma \in [0,1]$, since $y(B_{w(x)}) = 1/\beta$. In order to bound the right hand side, we take the maximum over all $\gamma \in [0,1]$ and $r_1, r_2 \ge 0$

satisfying the conditions given in Lemma 11. The right hand side is a linear function of $r_1$ and $r_2$. Hence, for a fixed $\gamma$ the maximum is attained at one of the two extreme points: $(r_1, r_2) = (0, \beta R_x/(\beta-1))$ or $(r_1, r_2) = (\beta R_x/(\beta-\gamma), \beta R_x/(\beta-\gamma))$. Substituting $r_1$ and $r_2$ in the previous inequality we get the following upper bound on the ratio $\mathbb{E}[d(x, f(x))]/R_x$:

$$\max_{\gamma\in[0,1]} \Big((1-e^{-\beta})+3e^{-(\beta-\gamma)}(1-\gamma)\Big(\frac{\beta}{\beta-1}+\max\big(\frac{\beta}{\beta-1},\frac{2\beta}{\beta-\gamma}\big)\Big)+\frac{\beta e^{-\beta}(1-e^{\gamma}(1-\gamma))}{\gamma}\Big). \tag{14}$$

This function can in turn be upper bounded by $\alpha(\beta)$, as defined in (8). We conclude that the approximation factor of the algorithm is upper bounded by $\alpha(\beta)$.

## 5 Local Search

For smaller values of $\beta$, we consider the standard local search algorithm (see, e.g., [6]) for the $k$-median problem using swaps of size $s$, but now allow the solution to contain $\beta k$ centers. The algorithm works as follows: we maintain a current solution $A$ comprising $\beta k$ centers in $\mathcal{C}$. We repeatedly attempt to reduce the cost of the current solution $A$ by closing a set of at most $s$ centers in $A$ and opening the same number of new centers from $\mathcal{C} \setminus A$. When no such local swap improves the cost of the solution $A$ we terminate and return $A$. In order to simplify our analysis, we do not worry about convergence time of the algorithm here. We note that by applying standard techniques (see [6, 9]), we can ensure that, for any $\delta > 0$, the algorithm converges in time polynomial in $n = |\mathcal{C} \cup \mathcal{D}|$ and $\frac{1}{\delta}$ by instead stopping when no local swap improves the cost of $A$ by a factor of $\left(1 - \frac{\delta}{\mathrm{poly}(n)}\right)$; the resulting algorithm's approximation ratio increases by only $\frac{1}{1-\delta}$.

Unfortunately standard analyses of local search algorithms for the $k$-median problem[6, 9] rely heavily on the triangle inequality, while the instances generated by Theorem 7 satisfy only a 3-relaxed 3-hop triangle inequality. Thus, we proceed as in Kanungo et al. [18]. Similarly to the previous section, we defer some technical details to Appendix C.

Let $O = \langle o_1, \dots, o_k\rangle$ be an optimal set of $k$ centers, and $A = \langle a_1, \dots, a_{\beta k}\rangle$ be the set of $\beta k$ centers produced by the local search algorithm. As in [18], we say that a center $a \in A$ *captures* a center $o \in O$ if $a$ is the center of $A$ that is closest to $o$. Note that each center in $A$ can potentially capture several centers in $O$, but each center in $O$ is captured by exactly one center of $A$. We now construct a set of local swaps to consider in our analysis. We say that a center in $A$ is "good" if it does not capture any center of $O$. Then, because each center of $O$ is captured by only one center of $A$, we must have at least $\beta k - k = (\beta - 1)k$ good centers in $A$. We fix some such set of $(\beta - 1)k$ good centers; we call them "auxiliary" centers and set them aside for now.

For the remaining $k$ centers $B \subseteq A$, we proceed exactly as in [18]: we assign each center in $O$ to the bad center of $B$ that captures it. This creates a partition $O_1, \dots, O_r$ of centers in $O$. We similarly partition the centers of $B$ into $r$ parts $B_1, \dots, B_r$ with $|B_i| = |O_i|$; for each $1 \le i \le r$, let $B_i$ contain the bad center of $B$ that captures all of $O_i$ together with $|B_i| - 1$ unique good centers of $B$. Note that the fact that each center of $O$ is captured only once ensures that there are indeed enough good centers in $B$ for our construction. Now, we use this partition of $B$ and $O$ to construct a set of swaps, each assigned some weight. If $|O_i| \le s$, we consider the $\langle B_i, O_i\rangle$ with weight 1. If $|O_i| = q > s$, we consider the group of all singleton swaps $\langle \{b\}, \{o\}\rangle$, where $o \in O_i$ and $b$ is a good center in $B_i$, each given weight $\frac{1}{q-1}$. At this point, note that every center in $O$ occurs in swaps of total weight 1, and every center in $B$

occurs in swaps of total weight at most $\frac{q}{q-1} \leq 1 + \frac{1}{s}$. Now, we add swaps involving auxiliary centers; for each of the $(\beta - 1)k$ auxiliary centers $a \in A \setminus B$ and each $o \in O$, we consider singleton swap $\langle \{a\}, \{o\} \rangle$, assigned weight $\frac{1}{k}$. Each center of $O$ now occurs in swaps of total weight $1 + (\beta - 1) = \beta$, while each center of $A \setminus B$ occurs in swaps of total weight 1.

Summarizing, our set of swaps satisfies the following properties: (1) each center of $O$ occurs in swaps of total weight $\beta$; (2) each center of $A$ occurs in swaps of total weight at most $1 + \frac{1}{s}$; (3) for any swap $\langle A', O' \rangle$ in our set, no center in $A'$ captures any center not in $O'$. We now give a brief sketch of how these properties lead to our desired approximation ratio (we give a full description of the analysis in the appendix). Our analysis closely follows that of [18].

Recall that for some set $C$ of centers and some $c \in C$, we denote by $N_C(c)$ the set of all points $x$ whose closest center in $C$ is $c$. As in [18], the total change $\text{cost}_{\mathcal{D},d}(A \setminus A' \cup O') - \text{cost}_{\mathcal{D},d}(A)$ due to performing a single swap $\langle A', O' \rangle$ is at most:

$$\sum_{o \in O'} \sum_{x \in N_O(o)} \big(d(x, o) - d(x, a_x)\big) + \sum_{a \in A'} \sum_{x \in N_A(A')} \big(d(x, a_{o_x}) - d(x, a_x)\big).$$

If $A$ is locally optimal, then we must have that $\text{cost}_{\mathcal{D},d}(A \setminus A' \cup O') - \text{cost}_{\mathcal{D},d}(A) \geq 0$ for all swaps $(A', O')$ considered by the algorithm. In particular, for each swap $\langle A', O' \rangle$ in our set, we have:

$$0 \leq \sum_{o \in O'} \sum_{x \in N_O(o)} \big(d(x, o) - d(x, a_x)\big) + \sum_{a \in A'} \sum_{x \in N_A(A')} \big(d(x, a_{o_x}) - d(x, a_x)\big). \tag{15}$$

Multiplying each inequality (15) by the weight of its swap and then adding the resulting inequalities we obtain:

$$0 \leq \beta \sum_{x \in \mathcal{D}} (d(x, o_x) - d(x, a_x)) + \left(1 + \frac{1}{s}\right) \sum_{x \in \mathcal{D}} (d(x, a_{o_x}) - d(x, a_x)),$$

due to properties (1) and (2) of our set of swaps. Theorem 7 part 2, which shows that our center set is an approximate $k$-means centroid set, then allows us to simplify the final term above as in [18], giving:

$$0 \leq \big(\beta + 2 + \tfrac{2}{s}\big) \text{cost}_{\mathcal{D},d}(O) - \left(\beta - \frac{2 + \frac{2}{s}}{\alpha}\right) \text{cost}_{\mathcal{D},d}(A) + O(\epsilon) \cdot \text{cost}_{\mathcal{D},d}(A),$$

where $\alpha^2 = \frac{\text{cost}_{\mathcal{D},d}(A)}{\text{cost}_{\mathcal{D},d}(O)}$ is the squared approximation ratio of our algorithm. Rearranging and simplifying (again, we give a detailed analysis in the appendix), we obtain

$$\alpha < \left(1 + \frac{2}{\beta} + \frac{2}{\beta s}\right) \frac{1}{1 - O(\epsilon)}.$$

Thus, we have the following theorem:

▶ **Theorem 12.** *There exists an algorithm that produces a solution for any instance of $\beta k$-median problem satisfying the properties of Theorem 7, where $\beta > 1$ is a fixed constant. For any $s \geq 1$ and any $\varepsilon \in (0, 1]$, the algorithm runs in time polynomial in $|\mathcal{C} \cup \mathcal{D}|$ and produces a solution $A$ satisfying:*

$$\text{cost}_{\mathcal{D},d}(A) \leq \left(1 + \frac{2}{\beta} + \frac{2}{\beta s}\right)^2 \frac{1}{1 - O(\epsilon)} \cdot \text{cost}_{\mathcal{D},d}(O)$$

*where $O$ is the optimal set of $k$ centers in $\mathcal{C}$.*

## References

1 Alexander A. Ageev and Maxim Sviridenko. Pipage rounding: A new method of constructing algorithms with proven performance guarantee. *J. of Combinatorial Optimization*, 8(3):307–328, 2004.

2 Ankit Aggarwal, Amit Deshpande, and Ravi Kannan. Adaptive sampling for k-means clustering. In *Proc. of the 12th International Workshop APPROX*, pages 15–28, 2009.

3 D. Aloise, A. Deshpande, P. Hansen, and P. Popat. NP-hardness of Euclidean sum-of-squares clustering. In *Machine Learning*, volume 75, pages 245–249, 2009.

4 Aris Anagnostopoulos, Anirban Dasgupta, and Ravi Kumar. A constant-factor approximation algorithm for co-clustering. In *Theory of Computing*, volume 8(1), pages 597–622, 2012.

5 D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *Proc. of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, 2007.

6 Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristics for k-median and facility location problems. *SIAM J. on Computing*, 33(3):544–562, 2004.

7 Pranjal Awasthi, Moses Charikar, Ravishankar Krishnaswamy, and Ali Kemal Sinop. The hardness of approximation of Euclidean k-means. In *Proc. of the 31st International Symposium on Computational Geometry*, pages 754–767, 2015.

8 Jarosław Byrka, Thomas Pensyl, Bartosz Rybicki, Aravind Srinivasan, and Khoa Trinh. An improved approximation for k-median, and positive correlation in budgeted optimization. In *Proc. of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 737–756, 2015. Updated version: `http://arxiv.org/abs/1406.2951`.

9 Moses Charikar and Sudipto Guha. Improved combinatorial algorithms for facility location problems. *SIAM J. on Computing*, 34(4):803–824, 2005.

10 Fabian A. Chudak and David B. Shmoys. Improved approximation algorithms for the uncapacitated facility location problem. *SIAM J. on Computing*, 33(1):1–25, 2003.

11 Michael B. Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proc. of the 47th Annual ACM on Symposium on Theory of Computing*, pages 163–172, 2015.

12 Vincent Cohen-Addad, Philip N. Klein, and Claire Mathieu. Local search yields approximation schemes for k-means and k-median in euclidean and minor-free metrics. *CoRR*, abs/1603.09535, 2016. URL: `http://arxiv.org/abs/1603.09535`.

13 S. Dasgupta. The hardness of k-means clustering. In *Technical Report CS2007-0890, University of California, San Diego.*, 2007.

14 Dan Feldman, Morteza Monemizadeh, and Christian Sohler. A PTAS for k-means clustering based on weak coresets. In *Proc. of the 23rd Annual Symposium on Computational Geometry*, pages 11–18, 2007.

15 Zachary Friggstad, Mohsen Rezapour, and Mohammad R. Salavatipour. Local search yields a PTAS for k-means in doubling metrics. *CoRR*, abs/1603.08976, 2016. URL: `http://arxiv.org/abs/1603.08976`.

16 Mary Inaba, Naoki Katoh, and Hiroshi Imai. Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering: (extended abstract). In *Proc. of the 10th Annual Symposium on Computational Geometry*, pages 332–339, 1994.

17 Kamal Jain and Vijay V. Vazirani. Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and Lagrangian relaxation. *J. ACM*, 48(2):274–296, March 2001.

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

**19** Shi Li and Ola Svensson. Approximating k-median via pseudo-approximation. In *Proc. of the 45th Annual ACM Symposium on Theory of Computing*, pages 901–910, 2013.

**20** Jyh-Han Lin and Jeffrey Scott Vitter. Approximation algorithms for geometric median problems. *Information Processing Lett.*, 44(5):245–249, December 1992.

**21** Jyh-Han Lin and Jeffrey Scott Vitter. $\epsilon$-approximations with minimum packing constraint violation. In *Proc. of the 24th Annual ACM Symposium on Theory of Computing*, pages 771–782, 1992.

**22** S. Lloyd. Least squares quantization in PCM. Technical report, Bell Laboratories, 1957.

**23** S. Lloyd. Least squares quantization in PCM. *IEEE Trans. on Information Theory*, 28(2):129–137, Mar 1982.

**24** Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The planar k-means problem is NP-hard. In *Proc. of the 3rd International Workshop on Algorithms and Computation*, pages 274–285, 2009.

**25** Jirı Matoušek. On approximate geometric k-clustering. *Discrete & Computational Geometry*, 24(1):61–84, 2000.

**26** R. Ostrovsky, Y. Rabani, L. Schulman, and C. Swamy. The effectiveness of Lloyd-type methods for the k-means problem. In *Proc. of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 165–174, 2006.

**27** Maxim Sviridenko. An improved approximation algorithm for the metric uncapacitated facility location problem. In *Integer Programming and Combinatorial Optimization: Proc. of the 9th IPCO Conference*, pages 240–257, 2002.

## A Proof of Theorem 7

In this section, we prove Theorem 7. Consider an instance of $k$-means with a set of points $X \subset \mathbb{R}^p$. Denote $n = |X|$. Let $\varepsilon' = \varepsilon/3$. Let $\varphi : \mathbb{R}^p \to \mathbb{R}^{\tilde{p}}$ be a dimension reduction transform for $X$ with distortion $(1+\varepsilon')$ as in Theorem 6. Note that $\tilde{p} = O(\log n/\varepsilon'^2) = O(\log n/\varepsilon^2)$.

Let $X' = \varphi(X) \subset \mathbb{R}^{\tilde{p}}$. Using the algorithm from Theorem 5, we compute an $\varepsilon'$-approximate centroid set $\mathcal{C} \subset \mathbb{R}^{\tilde{p}}$ for $X'$. The size of $\mathcal{C}$ is

$$O(n\varepsilon^{-\tilde{p}}\log(1/\varepsilon)) = n\varepsilon^{-O(\log n/\varepsilon^2)}\log(1/\varepsilon) = n \cdot n^{O(\log(1/\varepsilon)/\varepsilon^2)} = n^{O(\log(1/\varepsilon)/\varepsilon^2)};$$

we need time $O(n\log n + n\varepsilon^{-\tilde{p}}\log(1/\varepsilon)) = n^{O(\log(1/\varepsilon)/\varepsilon^2)}$ to compute it.

We first show that for every solution of the $k$-means problem on $X$ there is a corresponding solution of $k$-means problem on $X'$ in which all centers lie in $\mathcal{C}$, and vice versa.

▶ **Lemma 13.** *The following hold:*

1. *For every partition $S = \langle S_1, \dots, S_k\rangle$ of $X$, there is a corresponding clustering of $X'$ given by $S' = \langle \varphi(S_1), \dots, \varphi(S_k)\rangle$ and some centers $C' = \langle c'_1, \dots, c'_k\rangle \subseteq \mathcal{C}$ such that:*

$$\mathrm{cost}_{X'}(S', C') \leq (1+\varepsilon')^2\,\mathrm{cost}_X(S).$$

2. *For every partition $S' = \langle S'_1, \dots, S'_k\rangle$ of $X'$, there is a corresponding clustering $S = \big\langle \varphi^{-1}(S_1), \dots, \varphi^{-1}(S_k)\big\rangle$ of $X$ and some centers $C = \langle c_1, \dots, c_k\rangle \subseteq \mathbb{R}^p$ such that*

$$\mathrm{cost}_X(S, C) \leq \mathrm{cost}_{X'}(S').$$

**Proof.** Part 1: Consider a partition $S = \langle S_1, \dots, S_k\rangle$ of $X$ and the corresponding partition $S' = \langle S'_1, \dots, S'_k\rangle$ of $X'$, where $S'_i = \varphi(S_i)$. Let $c'_i = \arg\min_{c\in\mathcal{C}} \sum_{x\in S'_i} \|x' - c\|^2$ for

$i \in \{1, \dots, k\}$. Because $\mathcal{C}$ is an $\varepsilon'$-approximate centroid set for $X'$, we have, for each cluster $S'_i$,

$$\sum_{x \in S'_i} \|x - c'_i\|^2 \le (1+\varepsilon') \min_{c \in \mathbb{R}^{\tilde{p}}} \sum_{x \in S'_i} \|x - c\|^2 = (1+\varepsilon') \frac{1}{2|S'_i|} \sum_{x', x'' \in S'_i} \|x' - x''\|^2$$
$$= \frac{1+\varepsilon'}{2|S'_i|} \sum_{x', x'' \in S_i} \|\varphi(x') - \varphi(x'')\|^2 \le \frac{(1+\varepsilon')^2}{2|S_i|} \sum_{x', x'' \in S_i} \|x' - x''\|^2$$

Hence,

$$\mathrm{cost}_{X'}(S') = \sum_{i=1}^{k} \sum_{x \in S'_i} \|x - c'_i\|^2 \le (1+\varepsilon')^2 \, \mathrm{cost}_X(S).$$

Part 2: Consider a partition $S' = \langle S'_1, \dots, S'_k \rangle$ of $X'$ and the corresponding partition $S = \langle S_1, \dots, S_k \rangle$ of $X$, where $S_i = \varphi^{-1}(S'_i)$. Define the centers $c_i = \sum_{x \in S_i} x/|S_i|$. Then, for each cluster $S_i$, we have:

$$\sum_{x \in S_i} \|x - c_i\|^2 = \frac{1}{2|S_i|} \sum_{x', x'' \in S_i} \|x' - x''\|^2 \le \frac{1}{2|S_i|} \sum_{x', x'' \in S_i} \|\varphi(x') - \varphi(x'')\|^2$$
$$= \frac{1}{2|S'_i|} \sum_{x', x'' \in S'_i} \|x' - x''\|^2.$$

Hence,

$$\mathrm{cost}_X(S) = \sum_{i=1}^{k} \sum_{x \in S_i} \|x - c_i\|^2 \le \mathrm{cost}_{X'}(S'). \qquad \blacktriangleleft$$

Now we are ready to define instance $\langle \mathcal{D}, \mathcal{C}, d \rangle$. Let $\mathcal{D} = X'$, $\mathcal{C}$ be the $\varepsilon$-approximate centroid we defined above, and $d(c, x) = \|c - x\|^2$ for every $c \in \mathcal{C}$ and $x \in \mathcal{D}$. Define $\psi : \mathcal{D} \to X$ by $\psi(x) = \varphi^{-1}(x)$.

We prove that our reduction, which maps instance $X$ of $k$-means to instance $\langle \mathcal{D}, \mathcal{C}, d \rangle$ of $k$-median, satisfies the conditions of the theorem.

▶ **Lemma 14.** *Our reduction produces an instance that satisfies the following properties:*

1. *The distance function $d$ satisfies the 3-relaxed 3-hop triangle inequality on $\mathcal{D} \cup \mathcal{C}$.*
2. *For every partition $S = \langle S_1, \dots, S_k \rangle$ of $\mathcal{D}$ and the corresponding partition $\psi(S) = \langle \psi(S_1), \dots, \psi(S_k) \rangle$ of $X$, we have*

$$\mathrm{cost}_X(\psi(S)) \le \mathrm{cost}_{\mathcal{D},d}(S) \le (1+\varepsilon)\, \mathrm{cost}_X(\psi(S)).$$

3. *We have*

$$\mathrm{OPT}_X \le \mathrm{OPT}_{\langle \mathcal{D}, d \rangle} \le (1+\varepsilon)\mathrm{OPT}_X.$$

**Proof.** Claim 1 follows from the fact that:

$$\|x - w\|^2 \le (\|x - y\| + \|y - z\| + \|z - w\|)^2 \le 3(\|x - y\|^2 + \|y - z\|^2 + \|z - w\|^2).$$

for any $w, x, y, z \in \mathbb{R}^{\tilde{p}}$.

For claim 2, consider any partition $S$ of $\mathcal{D}$. Let $T = \psi(S)$ be the corresponding partition of $X$, given by $T_i = \psi(S_i)$. Then, from our definition of $d$, we have $\mathrm{cost}_{\mathcal{D},d}(S) = \mathrm{cost}_{X'}(S)$.

Moreover, by Lemma 13, we have $\text{cost}_{X'}(S)$ is between $\text{cost}_X(T)$ and $(1+\varepsilon')^2 \text{cost}_X(T)$. Thus,

$$\text{cost}_X(\psi(S)) \le \text{cost}_{\mathcal{D},d}(S) \le (1+\varepsilon')^2 \text{cost}_X(\psi(S)) \le (1+\varepsilon)\,\text{cost}_X(\psi(S)).$$

Since for every partition $S$ of $\mathcal{C}$ there is a corresponding partition $\psi(S)$ of $X$, and for every partition $T$ of $X$ there is a corresponding partition $\varphi(T)$ of $\mathcal{D}$, we immediately get from claim 2 that $\text{OPT}_X \le \text{OPT}_{\langle \mathcal{D},d\rangle} \le (1+\varepsilon)\text{OPT}_X$. ◀

## B Detailed Analysis of the LP Rounding Algorithm

Here we give a detailed proof of the necessary facts for the analysis of Section 4.

▶ **Lemma 15.** *The following inequality holds:* $R_x^\beta \le \beta R_x/(\beta-1)$.

**Proof.** We have

$$R_x = \sum_{c\in C_x} y_c d(x,c) \le \sum_{c\in C_x\setminus B_x} y_c d(x,c).$$

Every center $c \in C_x \setminus B_x$ is at distance at least $R_x^\beta$ from $x$. Hence,

$$R_x \le \sum_{c\in C_x\setminus B_x} y_c R_x^\beta = y(C_x\setminus B_x)\, R_x^\beta = \left(1-\frac{1}{\beta}\right) R_x^\beta.$$

The desired inequality follows. ◀

▶ **Lemma 16.** *There exist two nonnegative numbers* $r_1$ *and* $r_2$ *satisfying*

**1.** $\left(\frac{1-\gamma}{\beta}\right) r_1 + \left(\frac{\beta-1}{\beta}\right) r_2 \le R_x$,

**2.** $r_1 \le r_2$,

*such that* $R^\beta_{w(x)} + \mathbb{E}[d(w(x), f(x)) \mid f(x) \in B_{w(x)} \setminus D_x;\, \mathcal{E} = 1] \le r_1 + r_2$.

**Proof.** Denote the expected distance from a random center $c$ in $B_{w(x)} \setminus D_x$ to $w(x)$ by $r_1$ and distance from a random center $c$ in $C_{w(x)} \setminus B_{w(x)}$ to $w(x)$ by $r_2$:

$$r_1 = \frac{\sum_{c\in B_{w(x)}\setminus D_x} y_c d(w(x),c)}{y(B_{w(x)}\setminus D_x)} \qquad r_2 = \frac{\sum_{c\in C_{w(x)}\setminus B_{w(x)}} y_c d(w(x),c)}{y(C_{w(x)}\setminus B_{w(x)})}.$$

By the definition of $R_{w(x)}$, we have

$$\begin{aligned} R_{w(x)} &= \Big(\sum_{c\in D_x} y_c d(w(x),c)\Big) + y(B_{w(x)}\setminus D_x)\, r_1 + y(C_{w(x)}\setminus B_x)\, r_2 \\ &\ge \left(\frac{1-\gamma}{\beta}\right) r_1 + \left(\frac{\beta-1}{\beta}\right) r_2, \end{aligned}$$

since $y(B_x) = y(B_{w(x)}) = 1/\beta$, $y(C_{w(x)}) = 1$, and $y(D_x) = \gamma/\beta$. Note that $R_{w(x)} \le R_x$. Hence,

$$\left(\frac{1-\gamma}{\beta}\right) r_1 + \left(\frac{\beta-1}{\beta}\right) r_2 \le R_x.$$

Since all centers in $B_{w(x)} \setminus D_x$ lie inside of the ball of radius $R^\beta_{w(x)}$ around $w(x)$, and all centers in $C_{w(x)} \setminus B_{w(x)}$ lie outside of this ball, we have $r_1 \le R^\beta_{w(x)} \le r_2$. Hence,

$R^{\beta}_{w(x)} + d(w(x), f(x)) \le r_2 + d(w(x), f(x))$ and $r_1 \le r_2$. Conditional on $f(x) \in B_{w(x)} \setminus D_x$ and $\mathcal{E} = 1$, the random center $f(x)$ is distributed uniformly in $B_{w(x)} \setminus D_x$ with respect to the LP measure $y$. Hence, $\mathbb{E}[d(w(x), f(x)) \mid f(x) \in B_{w(x)} \setminus D_x; \mathcal{E} = 1] = r_1$. Consequently,

$$R^{\beta}_{w(x)} + \mathbb{E}[d(w(x), f(x)) \mid f(x) \in B_{w(x)} \setminus D_x; \mathcal{E} = 1] \le r_1 + r_2. \qquad ◀$$

▶ **Lemma 17.** *We have* $\Pr(f(x) \in D_x \mid \mathcal{E} = 1) = 1 - e^{\gamma}(1 - \gamma)$.

**Proof.** Observe that the set $D_x = B_{w(x)} \cap C_x$ is one of the sets in the partitioning $\widetilde{\mathcal{Z}}$ as $w(x) \in \mathcal{W}$ and $B_{w(x)} \in \mathcal{Z}$. Assume $f(x) \in D_x$ and $\mathcal{E} = 1$. Since $f(x) \in D_x$, we have $S \cap D_x \ne \varnothing$. Thus, $D_x$ must be inactive (otherwise, $\mathcal{E}$ would be 0). Moreover, for every $\widetilde{Z} \ne D_x$ ($\widetilde{Z} \in \widetilde{\mathcal{Z}}$), $\widetilde{Z}$ is inactive or $\widetilde{Z} \cap S = \varnothing$ (again, otherwise, $\mathcal{E}$ would be 0). Hence, the event

$$\{f(x) \in D_x \text{ and } \mathcal{E} = 1\}$$

can be represented as the intersection of the following three independent events: $\{S \cap D_x \ne \varnothing\}$, $\{D_x$ is not active$\}$, and $\{$there are no active centers in $(C_x \setminus D_x) \cap S\}$. The probability of the first event is $\beta y(D_x)$; the probability of the second event is $1 - (1 - e^{-\beta y(D_x)})/(\beta y(D_x))$; the probability of the third event is $e^{-\beta y(C_x \setminus D_x)}$ (this probability is computed as in Lemma 9). Thus,

$$\begin{aligned}\Pr(f(x) \in D_x \text{ and } \mathcal{E} = 1) &= \beta y(D_x) \times \Big(1 - \frac{1 - e^{-\beta y(D_x)}}{\beta y(D_x)}\Big) \times e^{-\beta y(C_x \setminus D_x)} \\ &= \big(\gamma - (1 - e^{-\gamma})\big) \times e^{-(\beta - \gamma)} = e^{-\beta}\big(1 - (1 - \gamma)e^{\gamma}\big).\end{aligned}$$

Combining this with Lemma 9, which shows that $\Pr(\mathcal{E} = 1) = e^{-\beta}$ completes the proof. ◀

▶ **Lemma 18.** *The following bound holds:* $\mathbb{E}[d(x, f(x)) \mid f(x) \in D_x; \mathcal{E} = 1] \le \frac{\beta R_x}{\gamma}$.

**Proof.** Given $f(x) \in D_x$ and $\mathcal{E} = 1$, the random center $f(x)$ is distributed uniformly in $D_x$ with respect to the LP measure $y$. Hence, $\Pr(f(x) = c) = y_c / y(D_x)$ for $c \in D_x$. We have

$$\mathbb{E}[d(x, f(x)) \mid f(x) \in D_x; \mathcal{E} = 1] = \frac{\sum_{c \in D_x} y_c d(x, c)}{y(D_x)} \le \frac{\sum_{c \in C_x} y_c d(x, c)}{y(D_x)} = \frac{R_x}{\gamma/\beta}. \qquad ◀$$

## C Detailed Analysis of the Local Search Algorithm

Here we give a detailed analysis of the local search algorithm from Section 5, closely following Kanungo et al. [18].

For a set of points $P \subseteq X$ and a point $c \in \mathbb{R}^p$, define the total distortion of $P$ with respect to $c$ as $\Delta(P, c) \equiv \sum_{c' \in P} \|c' - c\|^2$. We shall use the following Lemmas from [18]:

▶ **Lemma 19** (Lemma 2.1 in [18])**.** *Given a finite subset $P$ of points in $\mathbb{R}^p$, let $c$ be the centroid of $P$. Then, for any $c' \in \mathbb{R}^p$,* $\Delta(P, c') = \Delta(P, c) + |P| \cdot \|c - c'\|^2$

▶ **Lemma 20.** *Let $\langle \rho_i \rangle$ and $\langle \xi_i \rangle$ be two sequences of reals such that* $\alpha^2 = (\sum_i \rho_i^2)/(\sum_i \xi_i^2)$ *for some* $\alpha > 0$. *Then,*

$$\sum_{i=1}^{n} \rho_i \xi_i \le \frac{1}{\alpha} \sum_{i=1}^{n} \rho_i^2.$$

We now show how local optimality implies the desired inequality. For a point $x \in \mathcal{D}$, let $a_x$ and $o_x$ denote the closest facility to $x$ in $A$ and $O$, respectively. Recall that for for $a \in A$, $N_A(a)$ is precisely the set of all those points $x \in \mathcal{D}$ such that $a_x = a$, and, similarly, for $o \in O$, $N_O(o)$ is the set of all points $x \in \mathcal{D}$ such that $o_x = o$. Now, we upper bound the change in cost due to some swap $\langle A', O' \rangle$ in our set of swaps. We do this by constructing a feasible assignment of all points in $\mathcal{D}$ to centers in $A \setminus A' \cup O'$. For each $o \in O'$, we assign all the points in $N_O(o)$ to $o$. This changes the cost by

$$\sum_{o \in O'} \sum_{x \in N_O(o)} (d(x,o) - d(x,a_x)).$$

Now, fix a point $x \in N_A(A') \setminus N_O(O')$, and consider $x$'s closest optimal center $o_x$. We must have $o_x \notin O'$. Let $a_{o_x}$ be the closest center to $o_x$ in $A$. Then, by property (3) of our set of swaps, $a_{o_x} \notin A'$, since $a_{o_x}$ captures $o_x$ but $o_x \notin O'$. We reassign $x$ to $a_{o_x}$. The total cost of reassigning all such points $x$ is at most:

$$\sum_{a \in A'} \sum_{x \in N_A(A') \setminus N_O(O')} (d(x,a_{o_x}) - d(x,a_x)) \leq \sum_{a \in A'} \sum_{x \in N_A(A')} (d(x,a_{o_x}) - d(x,a_x)),$$

where the inequality follows from the fact that $a_x$ is the closest center to $x$ in $A$, and so $d(x,a_{o_x}) - d(x,a_x) \geq 0$ for all $x \in N_A(A') \cap N_O(O')$. Thus, the total change $\mathrm{cost}_{\mathcal{D},d}(A \setminus A' \cup O') - \mathrm{cost}_{\mathcal{D},d}(A)$ for each swap $\langle A', O' \rangle$ is at most:

$$\sum_{o \in O'} \sum_{x \in N_O(o)} (d(x,o) - d(x,a_x) + \sum_{a \in A'} \sum_{x \in N_A(A')} (d(x,a_{o_x}) - d(x,a_x)).$$

If $A$ is locally optimal, then we must have that $\mathrm{cost}_{\mathcal{D},d}(A \setminus A' \cup O') - \mathrm{cost}_{\mathcal{D},d}(A) \geq 0$ for all swaps $(A', O')$ considered by the algorithm. In particular, for each swap $\langle A', O' \rangle$ in our set, we have:

$$0 \leq \sum_{o \in O'} \sum_{x \in N_O(o)} \big(d(x,o) - d(x,a_x)\big) + \sum_{a \in A'} \sum_{x \in N_A(A')} \big(d(x,a_{o_x}) - d(x,a_x)\big).$$

Set $\gamma = 1 + \frac{1}{p}$. Then, multiplying each such inequality by the weight of its swap and then adding the resulting inequalities we obtain

$$\begin{aligned} 0 &\leq \beta \sum_{x \in \mathcal{D}} \big(d(x,o_x) - d(x,a_x)\big) + \gamma \sum_{x \in \mathcal{D}} \big(d(x,a_{o_x}) - d(x,a_x)\big) \\ &= \beta \,\mathrm{cost}_{\mathcal{D},d}(O) - (\beta + \gamma) \,\mathrm{cost}_{\mathcal{D},d}(A) + \gamma \sum_{x \in \mathcal{D}} d(x,a_{o_x}), \end{aligned} \tag{16}$$

where we have exploited properties (1) and (2) of our set of swaps to bound the number of times a given center in $O$ or $A$ is counted in our sum of inequalities.

It remains to bound the final term in (16). Consider some $o \in O$, and let $c$ be the centroid of $N_O(o)$. As above, we will let $a_o$ denote the closest center in $A$ to $O$. Then, note that:

$$\begin{aligned}
\Delta(N_O(o), a_o) &= \Delta(N_O(o), c) + |N_O(o)| \cdot \|c - a_o\|^2 && \text{(Lemma 19)} \\
&\le \Delta(N_O(o), o) + |N_O(o)| \cdot \|c - a_o\|^2 && \text{($c$ is the centroid of $N_O(o)$)} \\
&\le \sum_{x \in N_O(o)} \left[d(x,o) + (1+\varepsilon)\|o - a_o\|^2\right] \\
&\qquad \text{(Theorem 7 part 2, and the fact that $o$ is an optimal center for $N_O(o)$)} \\
&\le \sum_{x \in N_O(o)} \left[d(x,o) + (1+\varepsilon)\|o - a_x\|^2\right]. \\
&&& \text{($a_o$ is the closest center to $o$ in $A$)} \\
&\le (1+\varepsilon) \sum_{x \in N_O(o)} \left[d(x,o) + \|o - a_x\|^2\right].
\end{aligned}$$

Let $\alpha^2 = \frac{\mathrm{cost}_{\mathcal{D},d}(A)}{\mathrm{cost}_{\mathcal{D},d}(O)} = \frac{\sum_{x\in\mathcal{D}} d(x,a_x)}{\sum_{x\in\mathcal{D}} d(x,o_x)}$ be the approximation ratio attained by the algorithm. Summing over all $o \in O$, and recalling that for all $x \in N_O(o)$ we have $o_x = o$, we obtain:

$$\begin{aligned}
\sum_{x\in\mathcal{D}} d(x, a_{o_x}) &= \sum_{o \in O} \Delta(N_O(o), a_o) \\
&\le (1+\varepsilon) \sum_{o\in O} \sum_{x \in N_O(o)} \left[d(x,o) + \|o - a_x\|^2\right] \\
&= (1+\varepsilon) \sum_{x\in\mathcal{D}} \left[d(x,o_x) + \|o_x - a_x\|^2\right] \\
&\le (1+\varepsilon) \sum_{x\in\mathcal{D}} \left[d(x,o_x) + \|x - o_x\|^2 + \|x - a_x\|^2 + 2\|x - o_x\|\|x - a_x\|\right] \\
&= (1+\varepsilon) \sum_{x\in\mathcal{D}} \left[2d(x,o_x) + d(x,a_x) + 2\|x - o_x\|\|x - a_x\|\right] \\
&\le (1+\varepsilon) \sum_{x\in\mathcal{D}} \left[2d(x,o_x) + d(x,a_x) + \frac{2}{\alpha} d(x,a_x)\right] \\
&= (1+\varepsilon) \left[2\,\mathrm{cost}_{\mathcal{D},d}(O) + \left(1 + \frac{2}{\alpha}\right) \mathrm{cost}_{\mathcal{D},d}(A)\right]. \qquad (17)
\end{aligned}$$

Where in the last inequality, we have applied Lemma 20 to the sequences $\rho_i$ and $\xi_i$ defined by:

$$\alpha^2 = \frac{\sum_{x\in\mathcal{D}} d(x,a_x)}{\sum_{x\in\mathcal{D}} d(x,o_x)} = \frac{\sum_{i=1}^n \rho_i}{\sum_{i=1}^n \xi_i}.$$

Applying the upper bound (17) to the final term of (16), we obtain:

$$\begin{aligned}
0 &\le \beta\,\mathrm{cost}_{\mathcal{D},d}(O) - (\beta+\gamma)\,\mathrm{cost}_{\mathcal{D},d}(A) + \gamma(1+\varepsilon)\left[2\,\mathrm{cost}_{\mathcal{D},d}(O) + \left(1 + \tfrac{2}{\alpha}\right)\mathrm{cost}_{\mathcal{D},d}(A)\right] \\
&\le (\beta + 2\gamma)\,\mathrm{cost}_{\mathcal{D},d}(O) - \left(\beta - \tfrac{2\gamma}{\alpha}\right)\mathrm{cost}_{\mathcal{D},d}(A) + \left(3 + \tfrac{2}{\alpha}\right)\gamma\varepsilon\,\mathrm{cost}_{\mathcal{D},d}(A)
\end{aligned}$$

where we have used the fact that $\mathrm{cost}_{\mathcal{D},d}(O) \le \mathrm{cost}_{\mathcal{D},d}(A)$. Rearranging, we have

$$\begin{aligned}
(\beta + 2\gamma)\,\mathrm{cost}_{\mathcal{D},d}(O) &\ge \left(\beta - \frac{2\gamma}{\alpha} - \left(3 + \tfrac{2}{\alpha}\right)\gamma\epsilon\right)\mathrm{cost}_{\mathcal{D},d}(A) \\
&= \left(\beta - \frac{2\gamma}{\alpha} - \left(3 + \tfrac{2}{\alpha}\right)\gamma\epsilon\right)\alpha^2\,\mathrm{cost}_{\mathcal{D},d}(O),
\end{aligned}$$

which implies:

$$\begin{aligned}
\alpha^2\beta - 2\gamma\alpha - \beta - 2\gamma - \alpha^2\left(3 + \tfrac{2}{\alpha}\right)\gamma\epsilon &\le 0 \\
\alpha^2 - \frac{2\gamma\alpha}{\beta} - 1 - \frac{2\gamma}{\beta} - \frac{\alpha^2}{\beta}\left(3 + \tfrac{2}{\alpha}\right)\gamma\epsilon &\le 0 \\
(\alpha+1)\left(\alpha - 1 - \frac{2\gamma}{\beta}\right) - \frac{\alpha^2}{\beta}\left(3 + \tfrac{2}{\alpha}\right)\gamma\epsilon &\le 0 \\
\left(\alpha - 1 - \frac{2\gamma}{\beta}\right) - \frac{\alpha^2}{(\alpha+1)\beta}\left(3 + \tfrac{2}{\alpha}\right)\gamma\epsilon &\le 0.
\end{aligned}$$

Thus, we have:

$$(1 - O(\epsilon))\alpha \le 1 + \frac{2\gamma}{\beta} = 1 + \frac{2}{\beta} + \frac{2}{\beta p}.$$