







Parameterized Approximation For Robust Clustering in Discrete Geometric Spaces



Fateme Abbasi   
University of Wrocław, Poland

Jarosław Byrka   
University of Wrocław, Poland



Ameet Gadekar  
Bar-Ilan University, Ramat-Gan, Israel

Dániel Marx   
CISPA Helmholtz Center for Information Security,
Saarbrücken, Germany

Joachim Spoerhase   
University of Sheffield, UK

Sandip Banerjee  
IDSIA, USI-SUPSI, Lugano, Switzerland

Parinya Chalermsook   
Aalto University, Finland

Kamyar Khodamoradi  
University of Regina, Canada

Roohani Sharma 
University of Bergen, Norway

Abstract

We consider the well-studied ROBUST (k, z) -CLUSTERING problem, which generalizes the classic k -MEDIAN, k -MEANS, and k -CENTER problems and arises in the domains of robust optimization [Anthony, Goyal, Gupta, Nagarajan, Math. Oper. Res. 2010] and in algorithmic fairness [Abbasi, Bhaskara, Venkatasubramanian, 2021 & Ghadiri, Samadi, Vempala, 2022]. Given a constant $z \geq 1$, the input to ROBUST (k, z) -CLUSTERING is a set P of n points in a metric space (M, δ) , a weight function $w : P \rightarrow \mathbb{R}_{\geq 0}$ and a positive integer k . Further, each point belongs to one (or more) of the m many different groups $S_1, S_2, \dots, S_m \subseteq P$. Our goal is to find a set X of k centers such that $\max_{i \in [m]} \sum_{p \in S_i} w(p) \delta(p, X)^z$ is minimized.

Complementing recent work on this problem, we give a comprehensive understanding of the parameterized approximability of the problem in geometric spaces where the parameter is the number k of centers. We prove the following results:

- (i) For a universal constant $\eta_0 > 0.0006$, we devise a $3^z(1-\eta_0)$ -factor FPT approximation algorithm for ROBUST (k, z) -CLUSTERING in *discrete* high-dimensional Euclidean spaces where the set of potential centers is finite. This shows that the lower bound of 3^z for general metrics [Goyal, Jaiswal, Inf. Proc. Letters, 2023] no longer holds when the metric has geometric structure.
- (ii) We show that ROBUST (k, z) -CLUSTERING in discrete Euclidean spaces is $(\sqrt{3/2} - o(1))$ -hard to approximate for FPT algorithms, even if we consider the special case k -CENTER in logarithmic dimensions. This rules out a $(1 + \epsilon)$ -approximation algorithm running in time $f(k, \epsilon) \text{poly}(m, n)$ (also called efficient parameterized approximation scheme or EPAS), giving a striking contrast with the recent EPAS for the *continuous* setting where centers can be placed anywhere in the space [Abbasi et al., FOCS'23].
- (iii) However, we obtain an EPAS for ROBUST (k, z) -CLUSTERING in discrete Euclidean spaces when the dimension is sublogarithmic (for the discrete problem, earlier work [Abbasi et al., FOCS'23] provides an EPAS only in dimension $o(\log \log n)$). Our EPAS works also for metrics of sub-logarithmic doubling dimension.

2012 ACM Subject Classification Theory of computation \rightarrow Approximation algorithms analysis; Theory of computation \rightarrow Facility location and clustering

Keywords and phrases Clustering, approximation algorithms, parameterized complexity

Digital Object Identifier 10.4230/LIPIcs.ICALP.2024.6

Category Track A: Algorithms, Complexity and Games

Related Version *Full Version*: <https://arxiv.org/abs/2305.07316> [1]



© Fateme Abbasi, Sandip Banerjee, Jarosław Byrka, Parinya Chalermsook, Ameet Gadekar, Kamyar Khodamoradi, Dániel Marx, Roohani Sharma, and Joachim Spoerhase;
licensed under Creative Commons License CC-BY 4.0

51st International Colloquium on Automata, Languages, and Programming (ICALP 2024).
Editors: Karl Bringmann, Martin Grohe, Gabriele Puppis, and Ola Svensson;
Article No. 6; pp. 6:1–6:19



Leibniz International Proceedings in Informatics
LIPIcs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



Funding Fateme Abbasi and Jarosław Byrka were supported by the Polish National Science Centre (NCN) Grant 2020/39/B/ST6/01641. Sandip Banerjee acknowledges the support by SNSF Grant 200021 200731/1 and also the support of Polish National Science Centre (NCN) Grant 2020/39/B/ST6/01641 while at the University of Wrocław, Poland. Parinya Chalermsook, Kamyar Khodamoradi, and Joachim Spoerhase were supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 759557). Ameet Gadekar was supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 759557) while at Aalto University, and by the Israel Science Foundation (grant No. 1042/22).

1 Introduction

Clustering is a crucial method in the analysis of massive datasets and has widespread applications in operations research and machine learning. As a consequence, optimization problems related to clustering have received significant attention from the theoretical computer science community over the years. Within the framework of center-based clustering, k -CENTER, k -MEANS, and k -MEDIAN [25, 26, 10, 4, 27] are widely regarded as the most fundamental problems.

A general notion that captures various classic clustering problems is referred to as (k, z) -CLUSTERING in the literature, where $z \geq 1$ is a constant. In this type of problem, the input is a set P of data points (clients), a set F of centers (facilities), a metric δ on $P \cup F$, and a positive integer k . The goal is to find a set $C \subseteq F$ of k facilities that minimizes the following cost function:

$$\text{cost}(C) = \sum_{p \in P} \text{cost}(p, C)$$

where $\text{cost}(p, C) = \delta(p, C)^z$ and $\delta(p, C) = \min_{c \in C} \delta(p, c)$. Note that (k, z) -CLUSTERING encapsulates the classical k -MEDIAN, and k -MEANS for $z = 1$ and $z = 2$, respectively.

Center-based clustering has cemented its place as an unsupervised learning method that has proven effective in modeling a variety of real-world problem. In most of the practical machine learning applications however, it is observed that the input data is rarely of high quality.

To tackle this challenge, we study a robust version of (k, z) -CLUSTERING in this paper which can handle uncertainty in the input: Consider a situation where we do not have complete knowledge about the clients that will be served. In order to perform well despite this uncertainty, Anthony et al. [5] defined a concept of robustness for the k -MEDIAN problem, in which each possible scenario is represented by a group of clients and the goal is to find a solution that performs best possible even in the worst scenario. In this paper, we address the following robust version of the (k, z) -CLUSTERING problem (called ROBUST (k, z) -CLUSTERING):

ROBUST (k, z) -CLUSTERING

Input: Instance (P, F, δ) with δ being a metric on $P \cup F$, positive integer k , a weight function $w: P \rightarrow \mathbb{R}_+$, and m groups S_1, \dots, S_m such that $S_i \subseteq P$, $P = \cup_{i \in [m]} S_i$.

Output: A k -element subset $X \subseteq F$ that minimizes $\max_{i \in [m]} \sum_{p \in S_i} w(p) \delta(p, X)^z$.

Let $n = |P|$. We remark that, in addition to generalizing k -MEDIAN and k -MEANS, the ROBUST (k, z) -CLUSTERING problem encapsulates k -CENTER, when each group contains a distinct singleton. A similar objective has been studied in the context of *fairness*, in which

we aim to create a solution that will be appropriate for each of the specified groups of people. This problem is known in the literature as SOCIALLY FAIR k -MEDIAN, recently introduced independently by Abbasi et al. [3] and Ghadiri et al. [20]. Notice that Abbasi et al. [3] introduce fair clustering with client weights being inversely proportional to the group size as a normalization. On the other hand, Anthony et al. [5] introduce robust clustering with unweighted clients. Since our definition allows arbitrary client weights, we capture both of these settings.

While k -MEANS, k -MEDIAN, and k -CENTER admit constant-factor approximations, it is not very surprising that ROBUST (k, z) -CLUSTERING is harder due to its generality: Makarychev and Vakilian [29] design a polynomial-time $\mathcal{O}(\log m / \log \log m)$ -approximation algorithm, which is tight under a plausible complexity assumption [8]¹. As this precludes the existence of efficient constant-factor approximation algorithms, recent works have focused on designing constant factor *parameterized* (FPT) approximation algorithms². Along these lines, an FPT time $(3^z + \epsilon)$ -approximation algorithm has been proposed and shown to be tight under the Gap Exponential-Time Hypothesis (Gap-ETH) [22]. When allowing a parameterization on the number of groups m (instead of k), Ghadiri et al. designed a $(5 + 2\sqrt{6} + \epsilon)^z$ -approximation algorithm in $n^{\mathcal{O}(m^2)}$ time [21].

Motivated by the tight lower bounds for general discrete metrics, we focus on *geometric* spaces. Geometric spaces have a particular importance in real-world applications because data can often be represented via a (potentially large) collection of numerical attributes, that is, by vectors in a (possibly high-dimensional) geometric space. For example, in the bag-of-words model a document is represented by a vector where each coordinate specifies the frequency of a given word in that document. Such representations naturally lead to very high-dimensional data. A setting of particular interest is the high-dimensional *Euclidean space* where the metric is simply the Euclidean metric $\delta(x, y) = \|x - y\|_2$.

The study of clustering problems in high-dimensional Euclidean space is an important line of research that has received significant attention in the algorithms community. It may seem intuitive to believe that it should generally (for almost any problem) be possible to algorithmically leverage the geometric structure to separate high-dimensional Euclidean from general metrics. For clustering, however, this turns out to be either false or highly non-trivial in many cases. For example, it is a long-standing open question [19] whether k -CENTER admits a polynomial time $(2 - \epsilon)$ -approximation algorithm even in \mathbb{R}^2 , improving the tight bound of 2 in general metrics. Interestingly enough, for the more general Euclidean k -SUPPLIER problem, Nagarajan et al. [30] obtain an improvement over the tight bound of 3 in general metrics. The improved bounds for Euclidean k -MEDIAN and k -MEANS by Ahmadian et al. [4], Grandoni et al. [23], and recently by Cohen-Addad et al. [11] were breakthroughs. Concerning the more general ROBUST (k, z) -CLUSTERING, the tight inapproximability bound of $\Omega(\log m / \log \log m)$ in general metric continues to hold even in the line metric [8].

Similarly, the regime of FPT approximation algorithms for Euclidean clustering problems has received significant attention. Classic works design an Efficient Parameterized Approximation Scheme (EPAS), that is, a $(1 + \epsilon)$ -approximation in $f(k, \epsilon)\text{poly}(n)$ time, for k -CENTER [6] as well as for k -MEDIAN and k -MEANS [28]. Recent research focuses on the design of so-called coresets [31, 16] whose existence implies an EPAS if their size only depends on k and the error parameter ϵ .

¹ Note that they proved this factor for ROBUST k -MEDIAN, and the hardness result holds even in the line metric, unless $\text{NP} \subseteq \cap_{\delta > 0} \text{DTIME}(2^{n^\delta})$.

² Throughout the paper, parameterization refers to the natural parameter k .

In the real space \mathbb{R}^d , it is important to distinguish between the *discrete* and the *continuous* settings. In the discrete setting, both the point set P and the candidate center set F are finite subsets of \mathbb{R}^d while in the continuous setting, centers can be chosen anywhere in the metric space, that is, $F = \mathbb{R}^d$. A separate line of research has studied the contrast between continuous and discrete versions. For example, while discrete clustering variants are clearly polynomial-time solvable for constant k by trivial enumeration, the continuous versions of k -CENTER and k -MEDIAN are known to be NP-hard even for $k = 2$ [18] in high-dimensional Euclidean space. Also in terms of polynomial-time approximability, stronger lower bounds were shown by Cohen-Addad et al. [14] for the continuous versions. Indeed, there have been systematic research efforts in understanding these geometric clustering problems [15, 13, 14]. A recent result [2] implies an EPAS for ROBUST (k, z) -CLUSTERING in continuous Euclidean spaces (of any dimension), as well as in discrete Euclidean spaces in “relatively low” dimension, that is, dimension $o(\log \log n)$.

The main goal of this paper is to develop comprehensive understanding for ROBUST (k, z) -CLUSTERING in high-dimensional discrete Euclidean spaces, in particular, when the dimension is at least $\Omega(\log \log n)$.

1.1 Our contributions

First, motivated by a factor- $(3^z - o(1))$ hardness of FPT approximation for ROBUST (k, z) -CLUSTERING in general metrics [22], a natural question is whether the structures of Euclidean spaces can be leveraged to obtain better results in high dimensions. While it is intuitive to believe that such an improvement should generally (for almost any problem) be possible in geometric spaces, we note that this is sometimes not the case: The polynomial time inapproximability of ROBUST (k, z) -CLUSTERING remains $\Omega(\log m / \log \log m)$ even in the line metric [8].

Our first result gives an affirmative answer to this question.

► **Theorem 1.1 (High-Dimensional Euclidean Space).** *There exists a universal constant $\eta_0 > 0.0006$ such that for any constant positive integer z , there is a factor $3^z(1 - \eta_0)$ FPT approximation algorithm for ROBUST (k, z) -CLUSTERING in discrete Euclidean space \mathbb{R}^d that runs in time $2^{\mathcal{O}(k \log k)}$ $\text{poly}(m, n, d)$.*

We remark that, first, our running time has only a polynomial dependency on d . Secondly, the key take-home message for Theorem 1.1 is not about a concrete approximation factor, but rather a “proof of concept” that the factor of 3^z can be improved. Conceptually, this result shows that geometric spaces are indeed easier for ROBUST (k, z) -CLUSTERING than general metric spaces in the FPT world, in contrast to the polynomial-time world, where they seem to be equally hard [8]. The proof of this theorem relies on a new geometric insight that leverages the properties of Euclidean spaces (that do not hold in general metric spaces). The analysis of our algorithms “reduces” the global analysis of approximation factor to a “local” geometric instance, in which it suffices to merely analyze the behavior of three points in the Euclidean spaces.

Next, we focus on obtaining a complete characterization of the existence of EPAS in discrete Euclidean spaces. Recall that an EPAS exists in continuous Euclidean spaces of any dimensions and in discrete Euclidean spaces of dimension $o(\log \log n)$ [2], so to complete the landscape, we need to understand the discrete Euclidean spaces of dimension $\Omega(\log \log n)$.

In the next theorem, we prove that even the special case of k -CENTER does not admit an EPAS. This hardness holds for any ℓ_q metric and even in dimension $O(k \log n)$. More formally, we prove the following theorem.

► **Theorem 1.2** (Hardness in Discrete Euclidean Space). *For any constant positive integer q and any positive constant $\eta > 0$, there exists a function $d(k, n) = O(k \log n)$ such that there is no factor- $(3/2 - \eta)^{1/q}$ FPT approximation algorithm for the discrete k -CENTER problem in $\mathbb{R}^{d(k, n)}$ under the ℓ_q metric unless $W[1] = \text{FPT}$. Moreover, for the ℓ_2 metric this hardness holds even for some dimension $O(\log n)$, that is, independently of k .*

Our result therefore highlights the interesting contrast between the discrete and continuous settings in high-dimensional Euclidean spaces, which has been systematically studied in recent years [15, 13, 14]. As mentioned, the continuous setting admits an EPAS [2], so our hardness result implies that the discrete setting is harder than the continuous counterpart. This is contrast to the results Cohen-Addad et al. [14] mentioned earlier showing that continuous variants of k -MEDIAN and k -MEANS in geometric spaces are apparently harder to approximate (in polynomial time) than their discrete part as well as the different complexity status of continuous and discrete clustering in high-dimensional spaces even for $k = 2$ [18]. This shows a rather mysterious behavior of clustering problems in geometric spaces.

Our next theorem completes the FPT-approximability landscape by designing an EPAS for the problem in doubling metrics of dimension $d = o_k(\log n)^3$. We remark that the doubling dimension of the d -dimensional discrete Euclidean metric is $\Theta(d)$, that is, we obtain an EPAS for discrete Euclidean $o_k(\log n)$ -dimensional spaces in particular.

► **Theorem 1.3** (EPAS for Doubling Metric of Sub-Logarithmic Dimension). *There is an algorithm that computes $(1 + \epsilon)$ -approximate solution, for every $\epsilon > 0$, for ROBUST (k, z) -CLUSTERING in the metric of doubling dimension d in time $f(k, d, \epsilon, z) \text{poly}(m, n)$, where $f(k, d, \epsilon, z) = \left(\left(\frac{2^z}{\epsilon}\right)^d k \log k\right)^{\mathcal{O}(k)}$.*

Note that the above theorem yields an EPAS for ROBUST (k, z) -CLUSTERING when $d = o_k(\log n)$. Together with Theorem 1.2, this theorem gives (almost) a dichotomy result for the existence of EPAS: An EPAS exists for ROBUST (k, z) -CLUSTERING in $o_k(\log n)$ dimension, while obtaining an EPAS is $W[1]$ -hard in $\Omega_k(\log n)$ dimension. This leads to an almost complete understanding on the existence of EPAS in continuous and discrete Euclidean spaces.

2 Overview of Techniques

Improved FPT Approximation in High-Dimensional Discrete Euclidean Space

Our algorithm underlying Theorem 1.1 is a slight modification of the factor- $(3^z + \epsilon)$ FPT approximation algorithm for general metrics by Goyal and Jaiswal [22]. Our main technical contribution lies in the improved analysis. A key component of the analysis by Goyal and Jaiswal is a simple projection property of metric spaces (see Lemma 2.1 below). We argue that under minor additional assumptions, this property can be strengthened in Euclidean space. The resulting *assignment lemma* (see Lemma 3.1) is at the heart of our analysis and its proof relies on several new ideas and technically involved ingredients.

We briefly review the algorithm by Goyal and Jaiswal [22]. Their algorithm consists of two main steps. First, they compute a (κ, λ) -bicriteria solution $B \subseteq F$, that is, the cost of B is bounded by κOPT and the cardinality of B is bounded by λk . Specifically, they obtain guarantees $\kappa = 1 + \epsilon$ and $\lambda = \mathcal{O}(\log^2 n / \epsilon^2)$ for sufficiently small $\epsilon > 0$. In the second step, they extract a feasible solution from the (infeasible) bi-criteria solution B by enumerating all k -subsets of B and outputting the one of minimum cost.

³ We use notation $o_k(\cdot)$ to hide multiplicative factors depending only on k .

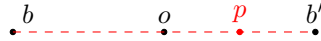
Their analysis is based on proving the existence of a k -subset of B whose cost is at most $(3^{z-1}(\kappa + 2))\text{OPT}$, which can be bounded by $(3^z + \epsilon)\text{OPT}$ assuming z being constant. Since the algorithm enumerates all k -subsets, this provides an upper bound on the cost of the algorithm. The key component of their existential argument is the following simple property of metric spaces, which we call *projection lemma*. It is convenient to think of O as an optimal solution and B as a bicriteria solution with $|B| > |O|$ but the lemma holds for any sets B, O .

► **Lemma 2.1** (Projection Lemma). *Let (Y, δ) be a metric space, and $B \subseteq Y$. Then for any set $O \subseteq Y$, there exists an assignment $\sigma: O \rightarrow B$ such that, for all $o \in O$ and $y \in Y$, we have*

$$\delta(y, \sigma(o)) \leq 2\delta(y, o) + \delta(y, B). \quad (1)$$

Intuitively, their lemma allows them to “project” the optimal solution O onto a k -subset $\sigma(O) \subseteq B$ of the bicriteria solution so that for any client $y \in Y$, the distance $\delta(y, \sigma(O))$ can be charged to $\delta(y, O)$ and $\delta(y, B)$. In fact, the number 3 in the approximation factor $3^z + \epsilon$ corresponds to the sum $(2 + 1)$ of the coefficients in front of $\delta(y, o)$ and $\delta(y, B)$.

In this paper, we study the setting where Y is a discrete Euclidean metric (P, F, δ) , that is, where P, F are finite subsets of \mathbb{R}^d and δ is the Euclidean distance. A natural attempt to improve the approximation factor in the Euclidean setting is to reduce the coefficients in front of the terms $\delta(y, o)$ and $\delta(y, B)$ in the projection lemma. Unfortunately, this straightforward approach fails: The projection lemma is tight even on the line metric; see Figure 1.



■ **Figure 1** This example shows that the projection lemma is tight even for the 1-dimensional Euclidean space. Let $o = 0$ be the optimum facility located at the origin and serving client $p = 1/2$. Let $b' = 1$ be the facility in B that serves p and let $b = \sigma(o) = -1$ be the facility in B nearest to o . We have $\text{OPT} = 1/2$, which also equals the cost of B . However $\delta(p, \sigma(o)) = 3/2 = 2 \times \delta(p, o) + 1 \times \delta(p, b')$. Combining multiple such examples in orthogonal directions and sharing facility b shows that the approximation ratio of the algorithm of Goyal and Jaiswal [22] approaches 3 in the discrete Euclidean space.

It turns out that slightly enlarging the projection space is already sufficient to bypass this obstacle. More specifically, we project onto the *midpoint closure*

$$\text{cl}(B) = B \cup \left\{ \pi_F \left(\frac{b + b'}{2} \right) : b, b' \in B \right\}, \quad (2)$$

of the bicriteria solution where $\pi_F(p)$ represents the closest facility in F to point p . This step exploits that the metric space is embedded into \mathbb{R}^d (so that the midpoints exist).

While on the algorithmic side a slight modification of the original algorithm is sufficient for the improvement, the analysis requires several new ideas and technically involved ingredients. To prove a strengthened version of the projection lemma (called assignment lemma) we set up a factor-revealing geometric optimization problem in the plane; see (3) in Definition 3.2 below. We call the optimum objective γ_β of this problem *displacement ratio*. Roughly speaking, this ratio corresponds to the maximum ratio between the left-hand and the right-hand side of (1) in Lemma 2.1. However, we project to $\text{cl}(B)$ rather than B and impose some additional minor restrictions. By a careful and technically involved analysis of this optimization problem we can upper bound the displacement ratio in the Euclidean setting by $1 - \epsilon_0$ for some universal constant $\epsilon_0 > 0$ as long as two obstructions are avoided. The first obstruction occurs in any configuration similar to the one in Figure 1 above where the bi-criteria solution contains two

facilities b, b' so that o is near to the mid-point of b and b' . However, in such a configuration facility o certifies that $b'' = \pi_F((b + b')/2)$ must be close to o allowing us to assign o to b'' contained in the mid-point closure. The second obstruction arises if p is β -near, that is, within a small distance β from o (but there is no facility in B such as b' as in the first obstruction). For β approaching 0, the displacement ratio of β -near points can approach 1 even if when projecting to the mid-point closure of B . To account for β -near points, we therefore cannot resort to the assignment lemma. However, the overall contribution of β -near points to the cost of the projected solution can be shown to be very small. More details of the algorithm and its analysis are provided in Section 3.1. The full proof of the assignment lemma is technically more involved and can be found in the full version [1].

Hardness of Discrete k -Center

Our proof constructs an instance of the discrete k -CENTER from an instance of MULTI-COLORED INDEPENDENT SET problem, which is known to be $W[1]$ -hard. In MULTI-COLORED INDEPENDENT SET, we are given a k -partite graph G with a k -partition of the vertices V_1, \dots, V_k , and the goal is to determine if there is an independent set that contains precisely one node from each set V_i , $i \in [k]$. The gadget in our construction is a set of nearly equidistant binary code words. Such code words with relative Hamming distance roughly $1/2$ and logarithmic length are known to exist (see Ta-Shma [32]). The high level idea is as follows. We associate each vertex of G with a unique code word of suitable length t . Then, we generate a data point in P for each vertex and edge of G by using code word(s) associated with the corresponding vertices. The construction guarantees the following crucial properties: (i) The Hamming distance between the data points of vertices is roughly t . (ii) The Hamming distance between a data point of vertex $v \in V_i$ and a data point of an edge e is roughly t if e is incident on $V_i \setminus \{v\}$ and is roughly $3t/2$ otherwise. (iii) The Hamming distance between the data points of edges is at least (close to) $3t/2$. Thus, the construction forces us to pick data points of vertices as centers in our solution and guarantees that the optimum cost of the k -CENTER instance is roughly t if and only if there is an independent set in G . As a result, approximating the cost of the k -CENTER instance better than a (roughly) $(3/2)^{1/q}$ factor would imply $W[1] = \text{FPT}$. That is because the cost of a k -CENTER instance is the maximum ℓ_q distance between a data point and its closest selected center, and hence, approximating this cost better than the mentioned factor allows us to distinguish between YES and NO cases of an arbitrary instance of MULTI-COLORED INDEPENDENT SET.

Approximation Scheme for Metrics of Sub-Logarithmic Doubling Dimension

Our algorithm comprises two main components, both based on standard techniques from the literature: instance compression and decomposition of the doubling metric into smaller balls. However, it becomes evident that a natural construction based on these standard techniques for ROBUST (k, z) -CLUSTERING faces serious information-theoretic limitations, as explained below. One natural idea for compressing a ROBUST (k, z) -CLUSTERING instance is to reduce the number of groups, as each group can be further compressed using a (k, z) -CLUSTERING coreset (such coresets exist [16]). This reduction yields a significantly smaller instance. If we could reduce the number of groups to $m' \ll m$ while approximately preserving the cost for every solution, we could obtain an EPAS as follows. First, apply a (k, z) -CLUSTERING coreset to every group of the compressed instance to obtain another ROBUST (k, z) -CLUSTERING instance with m' groups, each containing $g(k, \epsilon)$ points, where g is some function that represents the size of (k, z) -CLUSTERING coreset. It is essential to note that this compression

is acceptable for obtaining an EPAS since the coreset of a group approximately preserves the (k, z) -CLUSTERING cost of the group. Next, enumerate all k -partitions of the points within each group to find potential solutions. Finally, return the solution that has the minimum ROBUST (k, z) -CLUSTERING cost. Unfortunately, because ROBUST (k, z) -CLUSTERING captures k -CENTER (and consequently faces a coreset lower bound of $2^{\Omega(d)}$ in Euclidean space of dimension d [9]), the number of new groups must satisfy $m' \geq 2^{\Omega(d)}$. Consequently, the running time of this algorithm is $k^{2^{\Omega(d)}} \text{poly}(n, m)$, which is doubly exponential in d . It is worth noting that this algorithm matches the running time of [2] and does not yield an EPAS for sub-logarithmic dimension.

Furthermore, if we explore an alternative approach and utilize the coreset of k -CENTER, it is not immediately clear how to extend the coreset of k -CENTER to reduce the number of groups in an instance of ROBUST (k, z) -CLUSTERING. This is because, firstly, we would require a mapping between the old groups and the new groups, and secondly, this mapping should ideally approximately preserve the ROBUST (k, z) -CLUSTERING cost for every solution.

Another potential method for compressing the instance involves reducing the number of points in set P , rather than altering the groups, with the hope of designing an EPAS that can exploit the smaller P (without concern for the number of groups). However, for this approach to succeed, it is essential to establish a bijection between the old and new groups. Yet, it remains uncertain whether such a bijection exists. In typical coreset constructions, each point in the coreset P' of P has a weight that is the sum of the weights of the points in its local neighborhood in P which it is supposed to represent in P' . However, these points in P could potentially belong to different groups, making it challenging to establish the mapping between groups.

The core idea of our approach is to work with an alternative and more general definition of groups that permits a point to participate in different groups with varying weights. In this revised definition, instead of viewing groups as subsets of points, we treat each group as a weight function that assigns non-negative real values to points. This flexibility allows different weights to be assigned to the same point by different groups, which can, in fact, be of practical interest. Utilizing this new definition, we can devise an approach for compressing the points such that each point in the compressed instance can have a weight for group g that represents the sum of the weights of nearby points in g that were filtered out during compression. Essentially, this enables us to approximately preserve the group costs. With this approach and additional technical work that leverages the standard ball decomposition technique for doubling metrics, we derive a coreset for ROBUST (k, z) -CLUSTERING that can be employed to construct an EPAS for doubling metrics with sub-logarithmic dimension.

► **Remark 2.2.** Due to lack of space, we move some of the proofs to the full version [1]. The proofs of the Theorems and Lemmas with corresponding (\star) marked are provided in the full version [1].

3 High-Dimensional Discrete Euclidean Space

3.1 FPT Approximation Algorithm for Robust (k, z) -Clustering

In this section, we exploit non-trivial properties of the Euclidean metric to prove the following result that breaches the barrier of 3^z -approximation for ROBUST (k, z) -CLUSTERING in general metrics.

► **Theorem 1.1 (High-Dimensional Euclidean Space).** *There exists a universal constant $\eta_0 > 0.0006$ such that for any constant positive integer z , there is a factor $3^z(1 - \eta_0)$ FPT approximation algorithm for ROBUST (k, z) -CLUSTERING in discrete Euclidean space \mathbb{R}^d that runs in time $2^{\mathcal{O}(k \log k)} \text{poly}(m, n, d)$.*

Recall from Section 2 that our approach begins with computing a (κ, λ) -bicriteria solution B to the ROBUST (k, z) -CLUSTERING instance employing the algorithm proposed by Goyal-Jaiswal [22]. As we argued, it is sufficient to prove the existence of a k -subset of B whose cost is within a constant factor of optimal. The result by Goyal and Jaiswal [22] is based on the following simple projection lemma for general metrics whose proof we state here for the sake of later reference.

► **Lemma 2.1** (Projection Lemma). *Let (Y, δ) be a metric space, and $B \subseteq Y$. Then for any set $O \subseteq Y$, there exists an assignment $\sigma: O \rightarrow B$ such that, for all $o \in O$ and $y \in Y$, we have*

$$\delta(y, \sigma(o)) \leq 2\delta(y, o) + \delta(y, B). \quad (1)$$

Proof. For each $o \in O$, define $\sigma(o)$ as $\pi_B(o)$, the point in B closest in distance to o . Notice that for any $o \in O$, $y \in Y$, we have $\delta(y, \sigma(o)) \leq \delta(y, o) + \delta(o, \sigma(o))$ by triangle inequality. The lemma follows by combining this with $\delta(o, \sigma(o)) = \delta(o, B) \leq \delta(o, \pi_B(y)) \leq \delta(y, o) + \delta(y, B)$. ◀

This lemma itself is tight even in 1-dimensional Euclidean space (as we showed in Figure 1). In order to get around this issue, we make use of the property of our geometric space. Given the instance (P, F, δ) embedded into the Euclidean space and the bicriteria solution B , we project to the mid-point closure $\text{cl}(B)$ as defined in (2).

Notice that $|\text{cl}(B)| = \mathcal{O}(|B|^2)$. Let O be the optimal solution. For $\beta > 0$ we say that client $p \in P$ is β -far (from O w.r.t. B) if $\delta(p, O) \geq \beta \cdot \delta(p, B)$, and we say that client p is β -near otherwise. The key of our analysis is the following strengthening of the projection lemma for Euclidean space, which we call assignment lemma.

► **Lemma 3.1** (Assignment Lemma) (\star). Let $\beta_0 = 0.05$ and let $B \subseteq \mathbb{R}^d$. Then, for any $O \subseteq \mathbb{R}^d$, there exists an assignment $\sigma: O \rightarrow \text{cl}(B)$ such that, for all β_0 -far points $p \in \mathbb{R}^d$, we have $\delta(p, \sigma(O)) \leq (1 - \epsilon_0)(2\delta(p, O) + \delta(p, B))$ where $\epsilon_0 > 0.002$.

Proof Sketch. We start with defining the assignment function σ . Take any facility $o \in O$ and let $b = \pi_B(o)$. We assume w.l.o.g. that the instance is rotated so that p, b , and o lie in the plane spanned by the first two coordinates. For the sake of easier notation, we identify p, b, o by points in \mathbb{R}^2 . Further, by translation and scaling, we assume that o coincides with the origin and that $b = (-1, 0)$. Let $q = (0, 1)$ be the mirror image of b . Let α be a parameter to be fixed (we later set it to 0.6). We define $\sigma(o)$ based on the position of o relative to an α -ball. Specifically, $\sigma(o) = b$ if the α -ball centered at a point q contains no facility from B ; otherwise, $\sigma(o)$ is the projection $\pi_{\text{cl}(B)}(o)$ of o onto the mid-point closure of B .

Our goal is to analyze the displacement of a client p under the assignment rule σ . Recall from the proof of Lemma 2.1 that if $\sigma(o)$ is simply the projection onto B , then a client p , when served by facilities o and b' in sets O and B respectively, incurs a cost of at most $2\|p - o\| + \|p - b'\|$. We wish to show that the assignment cost in our algorithm is strictly smaller than this upper bound (under certain assumptions). We prove this by bounding the ratio of these two quantities.

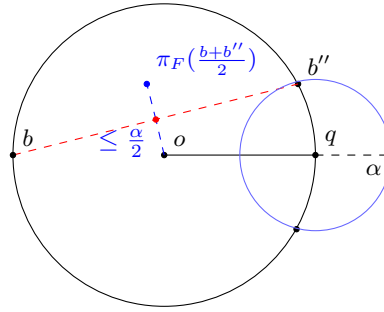
► **Definition 3.2** (Displacement Ratio). *For a given small constant $\beta > 0$, let the displacement ratio be defined as*

$$\gamma_\beta = \max_{\substack{p \in \mathbb{R}^d \setminus \text{ball}(o, \beta), \\ b' \in \mathbb{R}^d \setminus \text{ball}(o, 1)}} \left\{ \frac{\|p - \sigma(o)\|}{2\|p - o\| + \|p - b'\|} \right\}. \quad (3)$$

6:10 Parameterized Approximation for Robust Clustering in Discrete Geometric Spaces

Let S be the plane spanned by b , p , and o . After the appropriate rotations and translations we mentioned earlier, S would coincide with the x - y plane. In what follows, we also restrict b' to lie in \mathbb{R}^2 as well. We omit the argument why this assumption is without loss of generality from this sketch, and defer it to [1].

To show the lemma, we demonstrate that γ_β can be upper-bounded by $1 - f(\alpha, \beta)$ for some $f(\alpha, \beta) > 0$, where $f(\cdot)$ is a function dependent on α , β and the geometry of O and B . We distinguish two cases. First, suppose that B contains a facility b'' lying inside the α -ball around q . Recall that in this case $\sigma(o) = \pi_{\text{cl}(B)}(o)$. Hence $\sigma(o)$ is no farther from o than the facility $\pi_F((b + b'')/2)$ nearest to the midpoint of b and b'' . This allows us to bound the displacement ratio γ_β by $1 - \frac{1-\alpha}{2}$. See Figure 2 for an illustration. Notice that the optimal center o certifies the existence of a point in F nearby the mid-point of b and b'' .



■ **Figure 2** The midpoint of b and b'' is shown by red dot, $\|(b + b'')/2 - o\| \leq \frac{\alpha}{2}$ and thus $\|\sigma(o) - o\| \leq \alpha$.

In the second case, where the α -ball does not contain a facility from B , we argue that the points o , $\sigma(o) = b$, and b' are far enough from a co-linear position. This allows us to argue that the triangle inequality in the proof of Lemma 2.1 is not tight. Towards this, we divide the space into four regions R_1, R_2, R_3 and R_4 that could contain client p , we assume that p lies in the half plane above the x -axis (The case where p lies below the x -axis is symmetric.). Let q_1 be the intersection point of the surfaces of $\text{ball}(o, 1)$ and $\text{ball}(q, \alpha)$ above the x -axis. Let q_3 be the midpoint of q and q_1 , the region H is defined as the area above the lines passing through (q_3, o) and (o, b) , we define $R_1 = H \setminus \text{ball}(o, \beta)$. Next, consider $(1 - \omega)$ and $(1 + \omega)$ balls around o , H' is defined as the area below the line passing through (o, q_3) and above the line passing through (o, q) , we define $R_2 = (\text{ball}(o, 1 - \omega) \setminus \text{ball}(o, \beta)) \cap H'$, $R_3 = (\text{ball}(o, 1 + \omega) \setminus \text{ball}(o, 1 - \omega)) \cap H'$, and $R_4 = H' \setminus \text{ball}(o, 1 + \omega)$, the regions are indicated in Figure 3. Below, we provide full proof for one of these regions.

Assume that client p lies in region R_1 (see Figure 4). Let b'' be the closest point to p not in the interior of $\text{ball}(o, 1)$, and let p' be the point on the boundary of $\text{ball}(o, \beta)$ that is closet to p . Let p'' be the point where the segment (o, q_3) intersects the boundary of $\text{ball}(o, \beta)$, that is, $p'' = (\beta \cos \theta, \beta \sin \theta)$ where $\theta = \angle q_3 o q_1$. Notice that $\cos \theta = 1 - \frac{\alpha^2}{4}$. First, we assume p is inside $\text{ball}(o, 1 + 2\beta)$ in the region of R_1 .

► **Observation 3.3.** For any $\epsilon_1, \epsilon_2, X, Y \geq 0$:

$$\frac{X - \epsilon_1 + Y}{X + Y} \leq \frac{X - \epsilon_1 + Y + \epsilon_2}{X + Y + \epsilon_2}$$

Consider assigning p via p' to b . We bound the displacement cost as follows:

6:12 Parameterized Approximation for Robust Clustering in Discrete Geometric Spaces

We assume $\|p - p'\| \leq 1 + \beta$, and by observation 3.3, we obtain:

$$\begin{aligned} \gamma_\beta &\leq \frac{(1 + \beta)(1 + \sqrt{1 - \frac{\beta\alpha^2}{2}})}{2(1 + \beta)} \leq \frac{1}{2} + \frac{\sqrt{1^2 - \frac{2\beta\alpha^2}{4} + \frac{\beta^2\alpha^4}{16}}}{2} \leq \frac{1}{2} + \frac{1 - \frac{\beta\alpha^2}{4}}{2} \\ &= 1 - \frac{4 + \beta\alpha^2}{8} \end{aligned}$$

Second, let's assume that the client p is distant from o and positioned within region R_1 outside $\text{ball}(o, 1 + 2\beta)$, we can bound γ_β as follows:

$$\gamma_\beta \leq \frac{1 + \|o - p\|}{2\|o - p\|} \leq \frac{1 + 1 + 2\beta}{2(1 + 2\beta)} = \frac{1 + \beta}{1 + 2\beta} = 1 - \frac{\beta}{1 + 2\beta}$$

Therefore, by examining the position of p in the regions, we establish that γ_β is upper-bounded by $1 - f(\alpha, \beta)$. Consequently, Lemma 3.1 is substantiated by showing the existence of an $\alpha_0 \leq 0.6$ and a sufficiently small $\beta_0 \leq 0.05$ such that $\gamma_{\beta_0} \leq 1 - f(\alpha_0, \beta_0) = 1 - \epsilon_0 \leq 0.9978$. The proofs for the other regions as well as the full details of the rest of the argument are provided in the full version [1]. \blacktriangleleft

In the proof of Theorem 1.1, we show that this new assignment property is enough to derive an improved FPT approximation for ROBUST (k, z) -CLUSTERING in Euclidean space. Since the assignment σ maps every facility in O uniquely to a facility in $\text{cl}(B)$, this implies that $\sigma(O)$ is a feasible solution of cost at most $(3^z \cdot (1 - \eta_0))\text{OPT}$. This certifies the existence of a feasible solution being a subset of $\text{cl}(B)$ with the desired approximation factor. Hence, we can find such a solution in FPT time by enumeration. The complete proof of Theorem 1.1 is provided in the full version [1].

3.2 Hardness of Discrete k -Center

For this section, we use the following explicit construction of the so-called η -balanced error-correcting codes from a recent result of Ta-Shma [32] which we rephrase for our purposes as follows:

► Theorem 3.1. *Let $\eta \in (0, 1/2)$ be a positive constant. Then there is an algorithm that computes, for any given number $s \in \mathbb{N}$, an s -element set $B \subseteq \{0, 1\}^t$ of binary vectors of dimension $t = \mathcal{O}(\log s / \eta^{2+o(1)})$ such that for any $b \in B$, its Hamming weight $\|b\|_1$ and for any $b' \in B \setminus \{b\}$, the Hamming distance $\|b - b'\|_1$ both lie in the interval $[(1/2 - \eta)t, (1/2 + \eta)t]$. The running time of the algorithm is $\mathcal{O}(st)$.*

Proof. Ta-Shma [32] gives an explicit construction of a $t \times \lceil \log_2 s \rceil$ binary matrix generating a linear, binary, error-correcting code of message length $\lceil \log_2 s \rceil$, block length $t = \mathcal{O}(\log s / \eta^{2+o(1)})$, and pairwise Hamming distance between $(1/2 - \eta)t$ and $(1/2 + \eta)t$. Since the code is linear, it contains the zero code word. Hence each code word has Hamming weight in $[(1/2 - \eta)t, (1/2 + \eta)t]$. The time for constructing the matrix is polynomial in $\log s$ and t . Using the generating matrix, at least s many non-zero code words can be enumerated in time $\mathcal{O}(st)$, which dominates the time for computing the matrix. \blacktriangleleft

We leverage balanced error correcting codes as gadget in our hardness proof for discrete k -CENTER. For any binary vector $b \in \{0, 1\}^t$, we denote by \bar{b} the binary vector obtained by flipping each coordinate in b .

► **Theorem 1.2** (Hardness in Discrete Euclidean Space). *For any constant positive integer q and any positive constant $\eta > 0$, there exists a function $d(k, n) = O(k \log n)$ such that there is no factor- $(3/2 - \eta)^{1/q}$ FPT approximation algorithm for the discrete k -CENTER problem in $\mathbb{R}^{d(k, n)}$ under the ℓ_q metric unless $W[1] = \text{FPT}$. Moreover, for the ℓ_2 metric this hardness holds even for some dimension $O(\log n)$, that is, independently of k .*

Proof. We show a reduction from MULTI-COLORED INDEPENDENT SET, which is known to be $W[1]$ -hard [17]. The input is a k -partite graph $G = (V, E)$ with k -partition V_1, \dots, V_k . The question is if there is an independent set that is *multi-colored*, that is, it has precisely one node from each set V_i , $i \in [k]$. W.l.o.g. we assume that each V_i contains at least one node that is adjacent to all nodes $V \setminus V_i$. Adding such nodes, we can additionally assume that $|V_i| = n/k$ for each $i \in [k]$ where $n = |V|$.

Fix some constant $\eta \in (0, 1/2)$. Using Theorem 3.1, we construct a set $B \subseteq \{0, 1\}^t$ of n nearly equidistant code words of dimension $t = \mathcal{O}(\log n / \eta^{2+o(1)})$. We map each node $u \in V$ uniquely to some non-zero code word $b(u) \in B$. We construct a k -CENTER instance in $\mathbb{R}^{k \cdot t}$ as follows. We subdivide the coordinates of each point in $\mathbb{R}^{k \cdot t}$ into k blocks each containing t consecutive coordinates. In our set P of data points, we introduce for each node $v_i \in V_i$, $i \in [k]$, the point $p(v_i) \in P$ in which the i th block equals $b(v_i)$ and all other coordinates are zero. For each edge $(v_i, v_j) \in E$, $v_i \in V_i$, $v_j \in V_j$ for distinct $i, j \in [k]$ we create a point $p(v_i, v_j) \in P$ in which the i th block equals $\overline{b(v_i)}$, the j th block equals $\overline{b(v_j)}$, and all other coordinates are zero. No further points are added to P . We set the number of centers to be k completing the construction of the k -CENTER instance.

Let $i \in [k]$ and $v_i, v'_i \in V_i$ be distinct vertices. We have that $\|p(v_i) - p(v'_i)\|_q^q \leq \|b(v_i) - b(v'_i)\|_1 \leq (1/2 + \eta)t$ by Theorem 3.1. Let $v_j \in V_j$, $j \in [k]$ such that $(v_i, v_j) \in E$. By Theorem 3.1, we have that

$$\begin{aligned} \|p(v'_i) - p(v_i, v_j)\|_q^q &\leq \|b(v'_i) - \overline{b(v_i)}\|_1 + \|\overline{b(v_j)}\|_1 \\ &\leq (t - \|b(v'_i) - b(v_i)\|_1) + (t - (1/2 - \eta)t) \\ &\leq (t - (1/2 - \eta)t) + (1/2 + \eta)t \\ &\leq (1 + 2\eta)t. \end{aligned}$$

Hence if there is a multi-colored independent set I for G then $X = \{p(u) \mid u \in I\}$ is a k -element set such that $\delta(p, X)^q \leq (1 + 2\eta)t$ for any $p \in P$ under the ℓ_q metric, which gives an upper bound of $(1 + 2\eta)t$ on the k -CENTER objective in the completeness case.

For analyzing the soundness case, assume that there is no multi-colored independent set for G . Consider an arbitrary k -element set $X \subseteq V$. We say that $x \in X$ covers $p \in P$ if $\delta(p, x)^q < (3/2 - 3\eta)t$. We claim that there is some $p \in P$ not covered by any center in X . The correctness of this claim implies that any parameterized approximation algorithm with approximation ratio strictly better than $((3/2 - 3\eta)/(1 + 2\eta))^{1/q}$ implies that $W[1] = \text{FPT}$ and thus the theorem.

In order to prove this claim, we assume for the sake of contradiction, that all $p \in P$ are covered by some center in X . First, we argue that w.l.o.g. X contains no point of the form $p(v_i, v_j)$ where $(v_i, v_j) \in E$. In fact, for any $g \notin \{i, j\}$, we have that

$$\begin{aligned} \|p(v'_g) - p(v_i, v_j)\|_q^q &\geq \|b(v'_g)\|_1 + \|\overline{b(v_i)}\|_1 + \|\overline{b(v_j)}\|_1 \\ &\geq (1/2 - \eta)t + 2(t - (1/2 + \eta)t) \\ &= (3/2 - 3\eta)t. \end{aligned} \tag{4}$$

Hence $p(v_i, v_j)$ can cover $p(v'_g)$ only if $g = i$ or $g = j$. Similarly, $p(v_i, v_j)$ can cover $p(v'_g, v'_h)$ only if $i = g$ and $j = h$. But then these points would be covered by $p(v_i)$ as well and hence we could replace $p(v_i, v_j)$ with $p(v_i)$. We therefore assume that X contains only points of the form $p(v_i)$.

We claim that X is multi-colored. Otherwise, there would be some V_i that contains no point from X . By our initial assumption, V_i contains some point v_i that is adjacent to all points $V \setminus V_i$. Assuming $k \geq 3$ there exists at least one V_j , $j \neq i$ that contains at most one node from X . If V_j intersects X then let $v_j \in V_j \cap X$, and otherwise let v_j be an arbitrary node in V_j . By our assumption $(v_i, v_j) \in E$. If $v_j \in X$ then

$$\begin{aligned} \|p(v_j, v_i) - p(v_j)\|_q^q &\geq \|\overline{b(v_j)} - b(v_j)\|_1 + \|\overline{b(v_i)}\|_1 \\ &\geq t + (t - (1/2 + \eta)t) \\ &= (3/2 - \eta)t \end{aligned} \tag{5}$$

as the j th block of $p(v_j)$ equals $b(v_j)$ and the i th block of $p(v_j, v_i)$ equals $\overline{b(v_j)}$. If $v_j \notin X$ then for any $iv_h \in X$ we have $h \notin \{i, j\}$. Thus $\|p(v_i, v_j) - p(v_h)\|_q^q \geq (3/2 - 3\eta)t$, which follows as in (4). Hence $p(v_i, v_j)$ would not be covered showing that X is multi-colored. Since X is multi-colored it can not be an independent set. Hence there exists some edge (v_i, v_j) such that $v_i, v_j \in X$ but then $\|p(v_i) - p(v_i, v_j)\|_q^q \geq (3/2 - \eta)t$, $\|p(v_j) - p(v_i, v_j)\|_q^q \geq (3/2 - \eta)t$, and $\|p(v_h) - p(v_i, v_j)\|_q^q \geq (3/2 - 3\eta)t$ for any $v_h \in X$, $h \notin \{i, j\}$, which follows as in (5) and (4), respectively. Hence $\delta(p(v_i, v_j), X) \geq (3/2 - 3\eta)t$, implies that $p(v_i, v_j)$ is not covered.

We complete the proof by noting that the dimension of the instance can be reduced to $O(\log n)$ for Euclidean metrics by using the Johnson-Lindenstrauss transform with sufficiently small (constant) error parameter. \blacktriangleleft

4 EPAS for Metrics of Sub-Logarithmic Doubling Dimension

In this section, we show an EPAS for ROBUST (k, z) -CLUSTERING in metrics of sub-logarithmic doubling dimension. This result complements the hardness result of Section 3 (Theorem 1.2). Towards our goal, we prove the following result.

► **Theorem 1.3** (EPAS for Doubling Metric of Sub-Logarithmic Dimension). *There is an algorithm that computes $(1 + \epsilon)$ -approximate solution, for every $\epsilon > 0$, for ROBUST (k, z) -CLUSTERING in the metric of doubling dimension d in time $f(k, d, \epsilon, z)\text{poly}(m, n)$, where $f(k, d, \epsilon, z) = \left(\left(\frac{2z}{\epsilon}\right)^d k \log k\right)^{O(k)}$.*

Note that the above algorithm runs in FPT time for $d = o(\log n)$. We also remark that the above result can be extended to the continuous \mathbb{R}^d . Throughout this section, we assume that the weight aspect ratio $\frac{\max_{p \in P} w(p)}{\min_{p' \in P} w(p')}$ and the distance aspect ratio $\frac{\max_{p, p' \in P} \delta(p, p')}{\min_{p \neq p' \in P} \delta(p, p')}$ are bounded by $\text{poly}(n)$, some polynomial in n . For $p \in P$ and any number $r \geq 0$, denote by $\text{ball}(p, r)$ to be the closed ball centered at p of radius r . We prove the theorem in two steps: first, in Section 4.1 we show an algorithm to obtain a coreset for the problem, and then, in Section 4.2 we show how to use this coreset to get the algorithm of Theorem 1.3.

4.1 Coreset for Robust (k, z) -Clustering

The key idea for constructing coresets for ROBUST (k, z) -CLUSTERING crucially relies on the following alternate but equivalent definition of the problem. In this definition, we are given $\mathcal{I} = (F, P \subset \mathcal{M}, \mathcal{W})$, where either $F = \mathcal{M}$ or $F \subseteq \mathcal{M}$, where \mathcal{M} is doubling metric of dimension d , defined by the metric function δ . A *group* is a weight vector $\mathbf{w} \in \mathcal{W}$ such that $\mathbf{w} : P \rightarrow \mathbb{R}_{\geq 0}$. Given $X \subseteq F$, the distance vector $\boldsymbol{\delta}_P(X)$ is defined as $\boldsymbol{\delta}_P(X)[p] = \delta(p, X)^z$, for each $p \in P$. The cost of X for a group $\mathbf{w} \in \mathcal{W}$ is defined as $c(\mathbf{w}, X) = \mathbf{w} \cdot \boldsymbol{\delta}_P(X)$. For a ROBUST (k, z) -CLUSTERING instance $\mathcal{I} = (F, P, \mathcal{W})$, the cost of X is defined as $\text{cost}(\mathcal{I}, X) = \max_{\mathbf{w} \in \mathcal{W}} c(\mathbf{w}, X)$. The cost of the instance $\mathcal{I} = (F, P, \mathcal{W})$ is

$$\text{OPT}(\mathcal{I}) = \min_{X \subseteq F, |X|=k} \max_{\mathbf{w} \in \mathcal{W}} \text{cost}(\mathbf{w}, X)$$

Whenever the instance \mathcal{I} is clear from context, we will just write OPT . Notice that, in the original $\text{ROBUST}(k, z)\text{-CLUSTERING}$, a group is given by $S \subseteq P$, and this can be captured by weight vector $\mathbf{w}[p] = 0$ for $p \notin S$ and $w(p)$ otherwise. We prove the following coreset exists for $\text{ROBUST}(k, z)\text{-CLUSTERING}$.

► **Theorem 4.1** (Coreset for $\text{ROBUST}(k, z)\text{-CLUSTERING}$) (\star). Given an instance $\mathcal{I} = (F, P, \mathcal{W})$ of $\text{ROBUST}(k, z)\text{-CLUSTERING}$ in doubling metric of dimension d and $0 < \epsilon \leq 1$, there is an algorithm that, in time $(\frac{2z}{\epsilon})^{O(d)} \text{poly}(n, m)$, computes another instance $\mathcal{I}' = (F, P', \mathcal{W}')$ of $\text{ROBUST}(k, z)\text{-CLUSTERING}$ with $P' \subseteq P : |P'| = (\frac{2z}{\epsilon})^{O(d)} kz \log n$ such that for any $X \subseteq F$ with $|X| = k$,

$$(1 - \epsilon)\text{cost}(\mathcal{I}, X) \leq \text{cost}(\mathcal{I}', X) \leq (1 + \epsilon)\text{cost}(\mathcal{I}, X).$$

We remark that the above theorem yields a coreset of clients, and not of groups, and hence, the total size of coreset is comparable to the original instance. However, we will show later that such coreset is sufficient to get a parameterized approximation scheme with parameters k and d . We would also like to point out that the exponential dependency on d on the point set size of the coreset is inevitable since $\text{ROBUST}(k, z)\text{-CLUSTERING}$ captures $k\text{-CENTER}$, for which such a lower bound is known [9, 7]. To see that our notion of coreset for $\text{ROBUST}(k, z)\text{-CLUSTERING}$ coincides with the regular notion of coreset for $k\text{-CENTER}$, note that in this setting each group contains a single distinct point.

In the next section, we describe the algorithm of Theorem 4.1. Due to space constraints, we defer the analysis of our algorithm to the full version [1].

The Algorithm

Our algorithm is inspired by the grid construction approach of [24] that yields coresets for $k\text{-MEDIAN}$ and $k\text{-MEANS}$. Given an instance $\mathcal{I} = (F, P, \mathcal{W})$ of $\text{ROBUST}(k, z)\text{-CLUSTERING}$, the first step is to start with an (α, β) -bicriteria solution $B = \{b_i\}_{i \in [\beta k]}$ that opens at most βk facilities with the guarantee that $\text{cost}(\mathcal{I}, B) \leq \alpha \cdot \text{OPT}$, for some constants $\alpha, \beta \geq 1$. Let $R = \sqrt[z]{\frac{\text{cost}(\mathcal{I}, B)}{\alpha \tau}}$, where $\tau := \max_{\mathbf{w} \in \mathcal{W}} \|\mathbf{w}\|_1$. Let $\Delta = \frac{\max_{p \in P, \mathbf{w} \in \mathcal{W}} \mathbf{w}[p]}{\min_{p \in P, \mathbf{w} \in \mathcal{W}} \mathbf{w}[p]}$ be the weight aspect ratio of \mathcal{I} . Then, for each $b_i \in B$, consider the balls $\mathcal{B}_i^j := \text{ball}(b_i, 2^j R)$, for $j \in \{0, \dots, \lceil 2 \log(\alpha n \Delta) \rceil\}$. Note that, for $\mathbf{w} \in \mathcal{W}$ and $p \in P$ with $\mathbf{w}[p] > 0$, it holds that $\delta(p, B) \leq R \sqrt[z]{\alpha n \tau}$, since $\delta(p, B) \leq \sqrt[z]{\frac{\text{cost}(\mathcal{I}, B)}{\mathbf{w}[p]}} \leq \sqrt[z]{\frac{\alpha \tau}{\mathbf{w}[p]}} R \leq R \sqrt[z]{\alpha n \Delta}$. Hence, we have that every point $p \in P$ is contained in some ball \mathcal{B}_i^j . For $b_i \in B$, let $\mathcal{Q}_i^j = \mathcal{B}_i^j - \mathcal{B}_i^{j-1}$, for $j = \{1, \dots, \lceil 2 \log(\alpha \Delta) \rceil\}$, be the ring between \mathcal{B}_i^j and \mathcal{B}_i^{j-1} , with $\mathcal{Q}_i^0 = \mathcal{B}_i^0$. Decompose every ball \mathcal{B}_i^j into smaller balls each of radius $\frac{\epsilon}{40\alpha} R 2^j$ using the fact that the metric is a doubling metric. These balls can intersect, so we assign every point $p \in P$ to exactly one ball (for example, by associating p to the smallest ball containing p , breaking ties arbitrarily). For every ball \mathcal{B}_i^j and every smaller ball t of \mathcal{B}_i^j with $|t \cap \mathcal{Q}_i^j| \neq \emptyset$, pick an arbitrary point $p' \in t \cap \mathcal{Q}_i^j$ as the *representative* of (the points in) $t \cap \mathcal{Q}_i^j$, and add p' to the coreset P' with group weight vectors as follows. Corresponding to every group vector $\mathbf{w} \in \mathcal{W}$, create a new group vector $\mathbf{w}' \in \mathcal{W}'$. Then, $\mathbf{w}'[p'] := \sum_{p \in t \cap \mathcal{Q}_i^j} \mathbf{w}(p)$. Intuitively, $\mathbf{w}'[p']$ captures the total weight of points of \mathbf{w} in $t \cap \mathcal{Q}_i^j$. This concludes the coreset construction. For detailed pseudocode of the algorithm, please refer to the full version of the paper [1].

The high-level idea above is to decompose each ball \mathcal{B}_i^j into smaller balls and pick a distinct point as the representative of points in the non-empty decomposed ball. Additionally, such representative p' participates in the group \mathbf{w}' with weight which is sum of the weights of points in \mathbf{w} that are represented by p' . However, we want to decompose the ball \mathcal{B}_i^j into smaller balls in a way that the total number of balls remains the same, irrespective of the radius of the ball. This is necessary as for higher values of j , this number would depend on n , if we are not careful. While this does not seem to help much, as the radius of the decomposed balls is much larger for higher values j , it actually does the trick: since the points in these balls are far from b_i , and hence their connection cost to b_i is also large. This allows us to represent the radii of larger balls in terms of the connection cost of its points to B , thus bounding the error in terms of the cost of B , which in turn is bounded by αOPT , which gives us the desired guarantee.

4.2 EPAS for Robust (k, z) -Clustering

In this section, we show how to use the coreset obtained from Theorem 4.1 to get a $(1 + \epsilon)$ -approximate solution to the ROBUST (k, z) -CLUSTERING problem and provide an EPAS with respect to k and d , when $|P|$ is small. By scaling the distances in the instance of ROBUST (k, z) -CLUSTERING, we assume that the distances are between 1 and Δ' , for some number Δ' . Our algorithm (see Algorithm 1) uses the leader guessing idea of [12]. In the leader guessing approach, we guess the leader of every partition of a fixed optimal solution, where the leader of a partition is a closest point (client) in P to the corresponding optimal center. However, each point can participate in multiple groups, resulting in the total number of points being dependent on the number of groups, $|\mathcal{W}|$. In the full version [1], we show that guessing the leaders from P without considering the groups in \mathcal{W} is, in fact, sufficient. Further, to get a $(1 + \epsilon)$ -approximate solution, we use a standard ball decomposition lemma (for e.g., see the full version [1]).

► **Theorem 4.2.** *For any $0 < \epsilon \leq 1$, Algorithm 1, on input $\mathcal{I} = (F, P, \mathcal{W})$, computes $X \subseteq F : |X| \leq k$ such that $\text{cost}(\mathcal{I}, X) \leq (1 + \epsilon)\text{OPT}(\mathcal{I})$ in time $\left(\frac{z}{\epsilon}\right)^d \log n \binom{\mathcal{O}(k)}{|P|^k} \text{poly}(n, m)$.*

We conclude this section by proving the main claim of this section (Theorem 1.3) by using the results of Theorem 4.1 and Theorem 4.2 as follows.

Proof of Theorem 1.3. Given an instance $\mathcal{I} = (F, P, \mathcal{W})$ of ROBUST (k, z) -CLUSTERING, and the accuracy parameter $\epsilon > 0$, we invoke Theorem 4.1 on \mathcal{I} with parameter $\epsilon/10$ to obtain a coreset (P', \mathcal{W}') such that $P' \subseteq P : |P'| = \left(\frac{2z}{\epsilon}\right)^{\mathcal{O}(d)} kz \log n$. Let $\mathcal{I}' = (F, P', \mathcal{W}')$ be the resulting instance. Then, we invoke Theorem 4.2 on \mathcal{I}' with parameter $\epsilon/10$ to obtain $X \subseteq F : |X| \leq k$ such that $\text{cost}(\mathcal{I}', X) \leq (1 + \epsilon/10)\text{OPT}(\mathcal{I}')$.

First, we analyze the overall running time. With $|P'| = \left(\frac{2z}{\epsilon}\right)^{\mathcal{O}(d)} kz \log n$, Theorem 4.2 runs in time $\left(\left(\frac{2z}{\epsilon}\right)^d kz \log n\right)^{\mathcal{O}(k)} \text{poly}(n, m)$, leading to $\left(\left(\frac{2z}{\epsilon}\right)^d zk \log k\right)^{\mathcal{O}(k)} \text{poly}(n, m)$ as the overall running time as desired. For correctness, consider

$$\begin{aligned} \text{cost}(\mathcal{I}, X) &\leq (1 + \epsilon/10)\text{cost}(\mathcal{I}', X) && \text{by the coreset property} \\ &\leq (1 + \epsilon/10)^2\text{OPT}(\mathcal{I}') && \text{by Algorithm 1} \\ &\leq (1 + \epsilon/10)^3\text{OPT}(\mathcal{I}) && \text{by the coreset property} \\ &\leq (1 + \epsilon)\text{OPT}(\mathcal{I}). \end{aligned}$$

◀

■ **Algorithm 1** $(1 + \epsilon)$ -approximation algorithm for ROBUST (k, z) -CLUSTERING.

<p>Data: Instance $\mathcal{I} = (F, P, \mathcal{W})$ of ROBUST (k, z)-CLUSTERING Result: $(1 + \epsilon)$-approximate solution $X \subseteq F$</p> <pre> 1 Let $X \leftarrow \emptyset$; 2 forall k-tuples (ℓ_1, \dots, ℓ_k) of P do 3 forall k-tuples $(\lambda_1, \dots, \lambda_k)$ radii of (ℓ_1, \dots, ℓ_k) that are power of $(1 + \epsilon/10z)$ do 4 for $i \in [k]$ do 5 $\mathcal{B}_i \leftarrow \{\frac{\epsilon}{20z}$-ball decomposition of $\text{ball}(\ell_i, \lambda_i)\}$; 6 end 7 $T_i \leftarrow \{f \in F \mid f \text{ is an arbitrary facility in ball } b \in \mathcal{B}_i\}^a$; 8 forall k-tuples (t_1, \dots, t_k) of $T_1 \times \dots \times T_k$ do 9 if $\text{cost}(\mathcal{I}, \{t_1, \dots, t_k\}) < \text{cost}(\mathcal{I}, X)$ then 10 $X \leftarrow \{t_1, \dots, t_k\}$ 11 end 12 end 13 end 14 end 15 return X </pre>

^a If $F = \mathbb{R}^d$ then $T_i \leftarrow \{x_b \in F \mid x_b \text{ is the center of ball } b \in \mathcal{B}_i\}$

References

- 1 Fateme Abbasi, Sandip Banerjee, Jarosław Byrka, Parinya Chalermsook, Ameet Gadekar, Kamyar Khodamoradi, Dániel Marx, Roohani Sharma, and Joachim Spoerhase. Parameterized approximation for robust clustering in discrete geometric spaces, 2023. [arXiv:2305.07316](#).
- 2 Fateme Abbasi, Sandip Banerjee, Jarosław Byrka, Parinya Chalermsook, Ameet Gadekar, Kamyar Khodamoradi, Dániel Marx, Roohani Sharma, and Joachim Spoerhase. Parameterized approximation schemes for clustering with general norm objectives. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1377–1399, 2023. doi:10.1109/FOCS57990.2023.00085.
- 3 Mohsen Abbasi, Aditya Bhaskara, and Suresh Venkatasubramanian. Fair clustering via equitable group representations. In *Proc. ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 504–514, 2021.
- 4 Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. Better guarantees for k-means and euclidean k-median by primal-dual algorithms. In *Proc. 58th IEEE Annual Symposium on Foundations of Computer Science (FOCS'17)*, pages 61–72, 2017.
- 5 Barbara M. Anthony, Vineet Goyal, Anupam Gupta, and Viswanath Nagarajan. A plant location guide for the unsure: Approximation algorithms for min-max location problems. *Math. Oper. Res.*, 35(1):79–101, 2010. doi:10.1287/moor.1090.0428.
- 6 Mihai Badoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *Proc. 34th Annual ACM Symposium on Theory of Computing (STOC'04)*, pages 250–257, 2002.
- 7 Daniel Baker, Vladimir Braverman, Lingxiao Huang, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering in graphs of bounded treewidth. In *Proc. 37th International Conference on Machine Learning (ICML'20)*, volume 119, pages 569–579, 2020.
- 8 Sayan Bhattacharya, Parinya Chalermsook, Kurt Mehlhorn, and Adrian Neumann. New approximability results for the robust k -median problem. In *Proc. Scandinavian Workshop on Algorithm Theory (SWAT'14)*, pages 50–61, 2014.
- 9 Vladimir Braverman, Shaofeng H-C Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for ordered weighted clustering. In *Proc. International Conference on Machine Learning (ICML'19)*, pages 744–753, 2019.
- 10 T.W. Byrka, J.and Pensyl, B. Rybicki, A. Srinivasan, and K. Trinh. An improved approximation algorithm for k-median and positive correlation in budgeted optimization. *ACM Trans. Algorithms*, 13(2)(23):1–31, 2013.

- 11 Vincent Cohen-Addad, Hossein Esfandiari, Vahab S. Mirrokni, and Shyam Narayanan. Improved approximations for Euclidean k -means and k -median, via nested quasi-independent sets. In *Proc. 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC'22)*, pages 1621–1628, 2022.
- 12 Vincent Cohen-Addad, Anupam Gupta, Amit Kumar, Euiwoong Lee, and Jason Li. Tight FPT Approximations for k -Median and k -Means. In *Proc. 46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, volume 132, pages 42:1–42:14, 2019.
- 13 Vincent Cohen-Addad and CS Karthik. Inapproximability of clustering in l_p metrics. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 519–539. IEEE, 2019.
- 14 Vincent Cohen-Addad, CS Karthik, and Euiwoong Lee. On approximability of clustering problems without candidate centers. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2635–2648. SIAM, 2021.
- 15 Vincent Cohen-Addad and Euiwoong Lee. Johnson coverage hypothesis: Inapproximability of k -means and k -median in l_p -metrics. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1493–1530. SIAM, 2022.
- 16 Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. A new coresets framework for clustering. In Samir Khuller and Virginia Vassilevska Williams, editors, *Proc. 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC'21)*, pages 169–182, 2021.
- 17 Marek Cygan, Fedor V. Fomin, Lukasz Kowalik, Daniel Lokshtanov, Dániel Marx, Marcin Pilipczuk, Michal Pilipczuk, and Saket Saurabh. *Parameterized Algorithms*. Springer, 2015. doi:10.1007/978-3-319-21275-3.
- 18 Petros Drineas, Alan M. Frieze, Ravi Kannan, Santosh S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Mach. Learn.*, 56(1-3):9–33, 2004.
- 19 Tomás Feder and Daniel H. Greene. Optimal algorithms for approximate clustering. In *Proc. 20th Annual ACM Symposium on Theory of Computing (STOC'88)*, pages 434–444, 1988.
- 20 Mehrdad Ghadiri, Samira Samadi, and Santosh Vempala. Socially fair k -means clustering. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 438–448, 2021.
- 21 Mehrdad Ghadiri, Mohit Singh, and Santosh S Vempala. Constant-factor approximation algorithms for socially fair k -clustering. *arXiv preprint arXiv:2206.11210*, 2022.
- 22 Dishant Goyal and Ragesh Jaiswal. Tight fpt approximation for socially fair clustering. *Information Processing Letters*, 182:106383, 2023. doi:10.1016/j.ipl.2023.106383.
- 23 Fabrizio Grandoni, Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Rakesh Venkat. A refined approximation for euclidean k -means. *Inf. Process. Lett.*, 176:106251, 2022. doi:10.1016/j.ipl.2022.106251.
- 24 Sarel Har-Peled and Soham Mazumdar. On coresets for k -means and k -median clustering. In *Proc. 36th Annual ACM Symposium on Theory of Computing (STOC'04)*, page 291–300, 2004.
- 25 D.S. Hochbaum and D. Shmoys. A best possible heuristic for the k -center problem. *Mathematics of Operation Research*, 10(2):180–184, 1985.
- 26 K. Jain and V. 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, 2001.
- 27 Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. A local search approximation algorithm for k -means clustering. *Computational Geometry*, 28(2-3):89–112, 2004.
- 28 Amit Kumar, Yogish Sabharwal, and Sandeep Sen. Linear-time approximation schemes for clustering problems in any dimensions. *Journal of the ACM (JACM)*, 57(2):1–32, 2010.
- 29 Yury Makarychev and Ali Vakilian. Approximation algorithms for socially fair clustering. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3246–3264.

- PMLR, 15–19 August 2021. URL: <https://proceedings.mlr.press/v134/makarychev21a.html>.
- 30 Viswanath Nagarajan, Baruch Schieber, and Hadas Shachnai. The Euclidean k -supplier problem. *Math. Oper. Res.*, 45(1):1–14, 2020.
 - 31 Christian Sohler and David P. Woodruff. Strong coresets for k -median and subspace approximation: Goodbye dimension. In *Proc. 59th IEEE Annual Symposium on Foundations of Computer Science (FOCS'18)*, pages 802–813, 2018.
 - 32 Amnon Ta-Shma. Explicit, almost optimal, ϵ -balanced codes. In Hamed Hatami, Pierre McKenzie, and Valerie King, editors, *Proc. 49th Annual ACM SIGACT Symposium on Theory of Computing (STOC'17)*, pages 238–251. ACM, 2017. doi:10.1145/3055399.3055408.