Preclustering Algorithms for Imprecise Points

Mohammad Ali Abam
Department of Computer Engineering, Sharif University of Technology, Iran
abam@sharif.edu

Mark de Berg
Department of Mathematics and Computer Science, TU Eindhoven, the Netherlands
m.t.d.berg@tue.nl

Sina Farahzad
Department of Computer Engineering, Sharif University of Technology, Iran
farahzad@ce.sharif.edu

Mir Omid Haji Mirsadeghi
Department of Mathematical Sciences, Sharif University of Technology, Iran
mirsadeghi@sharif.edu

Morteza Saghafian
Department of Mathematical Sciences, Sharif University of Technology, Iran
morteza.saghafian65@student.sharif.edu

Abstract

We study the problem of *preclustering* a set B of imprecise points in $\mathbb{R}^d$: we wish to cluster the regions specifying the potential locations of the points such that, no matter where the points are located within their regions, the resulting clustering approximates the optimal clustering for those locations. We consider k-center, k-median, and k-means clustering, and obtain the following results.

Let $B := \{b_1, \ldots, b_n\}$ be a collection of disjoint balls in $\mathbb{R}^d$, where each ball b_i specifies the possible locations of an input point p_i. A partition $\mathcal{C}$ of B into subsets is called an $(f(k), \alpha)$-preclustering (with respect to the specific k-clustering variant under consideration) if (i) $\mathcal{C}$ consists of $f(k)$ preclusters, and (ii) for any realization P of the points p_i inside their respective balls, the cost of the clustering on P induced by $\mathcal{C}$ is at most α times the cost of an optimal k-clustering on P. We call $f(k)$ the *size* of the preclustering and we call α its *approximation ratio*. We prove that, even in $\mathbb{R}^1$, one may need at least $3k-3$ preclusters to obtain a bounded approximation ratio – this holds for the k-center, the k-median, and the k-means problem – and we present a $(3k, 1)$ preclustering for the k-center problem in $\mathbb{R}^1$. We also present various preclusterings for balls in $\mathbb{R}^d$ with $d \geqslant 2$, including a $(3k, \alpha)$-preclustering with $\alpha \approx 13.9$ for the k-center and the k-median problem, and $\alpha \approx 254.7$ for the k-means problem.

2012 ACM Subject Classification Theory of computation → Design and analysis of algorithms

Keywords and phrases Geometric clustering, k-center, k-means, k-median, imprecise points, approximation algorithms

Digital Object Identifier 10.4230/LIPIcs.SWAT.2020.3

Funding The work in this paper is supported by the Netherlands Organisation for Scientific Research (NWO) through Gravitation-grant NETWORKS-024.002.003.

1 Introduction

Clustering is one of the most important and widely studied problems in unsupervised learning. It comes in many different flavors, depending on the type of data to be clustered, the measure used to assess the quality of a clustering, and so on. In this paper we are interested in geometric clustering, where the data are points in $\mathbb{R}^d$, and we consider three well-known centroid-based clustering methods, namely k-center, k-median, and k-means, on so-called imprecise points.

17th Scandinavian Symposium and Workshops on Algorithm Theory (SWAT 2020).
Editor: Susanne Albers; Article No. 3; pp. 3:1–3:12
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

In (the geometric version of) centroid-based clustering one is given a set P of n points in $\mathbb{R}^d$, where d is a fixed constant, and an integer k. The goal is to partition P into k subsets $P_1, \ldots, P_k$ and assign a centroid q_i to each cluster P_i such that the cost of the resulting clustering is minimized. In the k-center problem the cost of the clustering is defined as $\max_{1 \leqslant i \leqslant k} \max_{p \in P_i} |pq_i|$, where $|pq|$ denotes the Euclidean distance between two points p and q. In the k-median problem the cost of a clustering is defined as $\sum_{1 \leqslant i \leqslant k} \sum_{p \in P_i} |pq_i|$, and in the k-means problem it is defined as $\sum_{1 \leqslant i \leqslant k} \sum_{p \in P_i} |pq_i|^2$. Given a collection of centroids it is always optimal to define the clusters by assigning each point in P to its nearest centroid. Thus an equivalent definition of the k-center problem, for instance, is to find a collection of $\{q_1, \ldots, q_k\}$ as centroids that minimizes $\max_{p \in P} \min_{1 \leqslant i \leqslant k} |pq_i|$. In other words, we want to find k congruent balls of minimum radius that together cover all points in P.

The k-center problem in $\mathbb{R}^d$ is NP-hard for $d \geqslant 2$ when k is part of the input. For the Euclidean k-center problem a PTAS exists, as shown by Agarwal and Procopiuc [1]. (For the k-center problem in general metric spaces, a PTAS does not exists; for this case an r-approximation algorithm with $r < 2$ is not possible unless P=NP, and several 2-approximation algorithms are known [5, 14].) The k-median and k-means problems are also NP-hard for $d \geqslant 2$, and they admit a PTAS as well [2, 4, 6, 8].

In the traditional setting the locations of the input points are known exactly. In practice this may not always be the case: typically locations are measured using GPS or other devices that are not completely accurate, or the points may move around inside a given region. This leads to the study of geometric algorithms on so-called *imprecise points*. Here, instead of specifying the exact coordinates of each input point, we specify a region for each point where it may be located. For points in the plane the regions are typically disks or squares. Over the past decade, many problems have been studied for imprecise points, including convex hulls (compute the smallest (or largest) possible convex hull of a set of imprecise points [7, 11]), Delaunay triangulations (preprocess a set of imprecise points such that for any given instantiation of the points in the given regions we can compute the Delaunay triangulation quickly [3]), separability problems [13], and more [9, 10, 12].

Problem statement and notation. In this paper we study the k-center, k-median, and k-means problem for imprecise points. The input is a set $B := \{b_1, \ldots, b_n\}$ of (closed) balls in $\mathbb{R}^d$, each representing the possible locations of an input point. Our goal is to compute a *preclustering* of the imprecise points, that is, a partition of B into a collection $\mathcal{C}$ of subsets called *preclusters* that gives a good clustering for any possible realization of the points inside the input balls. Next we define this more formally.

For a (precise) point set P, let $\textsc{Opt}_\infty(P, k)$ denote the cost of an optimal k-center clustering on P, that is,

$$\textsc{Opt}_\infty(P, k) := \min_{q_1, \ldots, q_k \in \mathbb{R}^d} \max_{p \in P} \min_{1 \leqslant i \leqslant k} |pq_i|.$$

The cost of an optimal solution for the k-median and k-means problem on a set P are denoted by $\textsc{Opt}_1(P, k)$ and $\textsc{Opt}_2(P, k)$, respectively.[1] Now consider an imprecise point set specified by a set $B = \{b_1, \ldots, b_n\}$ of balls. A point set $P := \{p_1, \ldots, p_n\}$ such that $p_i \in b_i$ for all $1 \leqslant i \leqslant n$ is called a B-*instance*. A preclustering $\mathcal{C}$ of the set B into preclusters B_i

[1] The subscript ∞ in $\textsc{Opt}_\infty$ refers to the fact that if d_i denotes the distance of point $p_i \in P$ to its nearest center, then we are minimizing the norm of the vector $\langle d_1, \ldots, d_n \rangle$ in the ℓ_∞-metric. For k-median and k-means we are minimizing the norm in the ℓ_1-metric and in the squared ℓ_2-metric, respectively.

induces a clustering on any B-instance P in a natural manner, namely by creating a cluster $P_i := \{p \in P : p \in B_i\}$ for every precluster $B_i \in \mathcal{C}$. The cost of the preclustering $\mathcal{C}$ on P, denoted by $\mathcal{C}\text{-}\textsc{Cost}_\infty(P)$ for the k-center problem, is defined as the cost of the induced clustering on P if we choose the centroid of each cluster P_i optimally, namely by solving the 1-clustering problem on P_i. So for the k-center problem we have

$$\mathcal{C}\text{-}\textsc{Cost}_\infty(P) := \max_{B_i \in \mathcal{C}} \min_{q \in \mathbb{R}^d} \max_{p \in P_i} |pq|.$$

The preclustering costs for the k-median and k-means problem are denoted by $\mathcal{C}\text{-}\textsc{Cost}_1(P)$ and $\mathcal{C}\text{-}\textsc{Cost}_2(P)$, respectively, and they are defined similarly. To quantify the quality of a preclustering $\mathcal{C}$ on B (with respect to the k-clustering problem under consideration) we define $\mathcal{C}$ to be a $(f(k), \alpha)$-*preclustering* if

- $\mathcal{C}$ consists of $f(k)$ preclusters,
- $\mathcal{C}\text{-}\textsc{Cost}(P) \leqslant \alpha \cdot \textsc{Opt}(P, k)$ for any B-instance P.

We call $f(k)$ the *size* of the preclustering and we call α its *approximation ratio*. Ideally, we would like to have a $(k, 1)$-preclustering, but this is not always possible. If the balls in B have a non-empty common intersection, then any preclustering with fewer than n preclusters may have an arbitrarily bad approximation ratio, even for the 2-center problem. Hence, we assume (as is often done in papers on imprecise points) that the balls in B are disjoint.

Our results. As mentioned, obtaining a $(k, 1)$-preclustering is not always possible. This leads to the question: what is the smallest value for $f(k)$ such that we can always obtain an $(f(k), 1)$-preclustering? More generally, which trade-offs are possible between the size $f(k)$ of the preclustering and its approximation ratio α?

In Section 2 we study this problem in $\mathbb{R}^1$. We show that there are input sets B that require at least $3k - 3$ preclusters to get a bounded approximation ratio; this holds for the k-center problem, the k-median problem, as well as the k-means problem. We complement this result by proving that any set B of intervals in $\mathbb{R}^1$ admits a $(3k, 1)$-preclustering for the k-center problem. This preclustering can be computed in polynomial time.

In Section 3 we consider the d-dimensional version of the problem for $d \geqslant 2$. We give an example showing that here a $(3k, 1)$-preclustering does not always exist, and we present a $(3k, \alpha)$-preclustering with $\alpha \approx 13.9$ for the k-center and k-median problem, and $\alpha \approx 254.7$ for the k-means problem. A different parameterization of the strategy gives a $(6k, 3)$-preclustering for k-center and k-median, and a $(6k, 10)$-preclustering for k-means in $\mathbb{R}^2$.

Finally, in Section 4 we obtain tight asymptotic bounds on the size of the preclustering needed to obtain any given approximation ratio $\varepsilon > 0$ for the k-center problem. In particular, we prove that $\Theta(\lceil 1/\varepsilon^d \rceil \cdot k)$ preclusters are always sufficient and sometimes necessary to obtain approximation ratio ε.

2 The 1-dimensional problem

We begin by proving that even in $\mathbb{R}^1$ – here the input balls are disjoint intervals on the line – preclusterings with only k preclusters cannot always guarantee a good approximation ratio. In fact, we sometimes need as much as $3k - 3$ preclusters in any preclustering with bounded approximation ratio.

▶ **Theorem 1.** *For any integer $k \geqslant 2$ and any given α, there is a set B of disjoint intervals in $\mathbb{R}^1$ that does not admit a (k', α)-preclustering with $k' < 3k - 3$. This holds for k-center, k-median, as well as k-means clustering.*

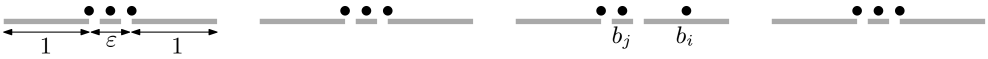


Figure 1 Illustration of the lower-bound construction for $k = 5$: a collection of $k - 1$ groups of three intervals (in grey), each group consisting of a left and right interval of length 1 separated by a gap of length ε, and a middle interval in inside this gap. The points in the B-instance used in the proof are shown slightly above intervals for clarity.

Proof. Let B be a collection of $3k - 3$ disjoint intervals in $\mathbb{R}^1$ consisting of $k - 1$ groups of three intervals each. The left and right interval in each group have length 1 and are at distance ε from each other, where ε is a sufficiently small number that will be specified later. The middle interval from the group lies in the gap between the left and right interval with its center at the center of the gap; see Fig. 1. Now consider a preclustering $\mathcal{C} = \{B_1, \dots, B_{k'}\}$. If $k' < 3k - 3$, then there is at least one precluster containing two intervals, b_i and b_j. Assume without loss of generality that $\text{length}(b_i) \geqslant \text{length}(b_j)$, and consider the B-instance in which each point p_t is placed in its interval $b_t \in B$ as follows.

- If $t = i$ or b_t is a middle interval, then p_t lies at the center of b_t.
- If $t \neq i$ and b_t is a left interval, then p_t lies at the right endpoint of b_t.
- If $t \neq i$ and b_t is a right interval, then p_t lies at the left endpoint of b_t.

Note that with this placement we have $|p_i p_j| \geqslant 1/2$. We will argue that by choosing ε appropriately we get the desired result.

First consider the k-center problem. Note that $\textsc{Opt}_\infty(P, k) \leqslant \varepsilon/2$. Indeed, by putting a centroid at the center of each of the $k - 1$ gaps and one centroid at p_i, all points in P are at distance at most $\varepsilon/2$ from a centroid. On the other hand, $\mathcal{C}\text{-}\textsc{Cost}_\infty(P) \geqslant 1/4$ since the centroid for the cluster containing p_i and p_j is at distance at least $1/4$ from p_i or p_j. Hence,

$$\frac{\mathcal{C}\text{-}\textsc{Cost}_\infty(P)}{\textsc{Opt}_\infty(P,k)} \geqslant \frac{1/4}{\varepsilon/2} = \frac{1}{2\varepsilon}.$$

For $\varepsilon < 1/(2\alpha)$ we thus enforce an approximation ratio greater than α.

The argument for k-median and k-means is similar. For k-median we have $\textsc{Opt}_1(P, k) \leqslant 2(k-1)(\varepsilon/2)$ and $\mathcal{C}\text{-}\textsc{Cost}_1(P) \geqslant 1/2$, so $\varepsilon < 1/(2(k-1)\alpha)$ enforces an approximation ratio greater than α, while for k-means we have $\textsc{Opt}_2(P, k) \leqslant 2(k-1)(\varepsilon/2)^2$ and $\mathcal{C}\text{-}\textsc{Cost}_2(P) \geqslant 2(1/4)^2$, so $\varepsilon < \sqrt{1/(4(k-1)\alpha)}$ suffices. ◀

▶ Remark 2. The construction in the proof of Theorem 1 uses an input set B of size $3k - 3$. We can easily generate an input set with the same behavior for any $n \geqslant 3k - 3$, by adding another $n - 3k + 3$ tiny intervals inside one of the gaps between a left and a right interval from the same group.

Theorem 1 states that for some problem instances any preclustering with fewer than $3k - 3$ preclusters has arbitrarily large approximation ratio. We now show how to obtain a 1-approximation with only $3k$ preclusters for the k-center problem. We assume from now on that $n > 3k$, otherwise we can trivially create a zero-cost solution with at most $3k$ preclusters.

Before we describe our preclustering strategy, we first generalize the k-center problem in $\mathbb{R}^1$ from points to intervals. In this generalization the input is a collection B of n intervals, and the goal is to find a collection $\mathcal{I} := \{I_1, \dots, I_k\}$ of intervals that together cover all intervals in B and such that the maximum radius of the intervals in $\mathcal{I}$ is minimized. (The radius of an interval is half its length.) We denote the value of an optimal solution $\mathcal{I}$ to the k-center problem on B by $\textsc{Opt}_\infty(B, k)$, so $\textsc{Opt}_\infty(B, k) := \max_{I_i \in \mathcal{I}} \text{radius}(I_i)$.

Our preclustering algorithm is now as follows.

```
PreClustering-1D$(B, k)$
1. Sort the intervals in $B$ by radius, such that $\text{radius}(b_1) \geqslant \cdots \geqslant \text{radius}(b_n)$.
2. For each $k' \in \{0, \ldots, 2k\}$ do the following.
   a. Let $\{B_1, \ldots, B_{(3k-k')}\}$ be an optimal $(3k-k')$-center clustering on $\{b_{k'+1}, \ldots, b_n\}$,
      and let $\text{Opt}_\infty(\{b_{k'+1}, \ldots, b_n\}, 3k-k')$ be its cost.
   b. Let $\mathcal{C}(k')$ be the preclustering $\{\{b_1\}, \ldots, \{b_{k'}\}, B_1, \ldots, B_{(3k-k')}\}$.
3. Of all preclusterings $\mathcal{C}(0), \ldots, \mathcal{C}(2k)$ found in Step 2, let $\mathcal{C}(k')$ be the one that minimizes
   $\text{Opt}_\infty(\{b_{k'+1}, \ldots, b_n\}, 3k-k')$. Let $\mathcal{C} := \mathcal{C}(k')$ and return $\mathcal{C}$.
```

▶ **Theorem 3.** *Any set B of disjoint intervals in $\mathbb{R}^1$ admits a $(3k, 1)$-preclustering for the k-center problem and this algorithm can be executed in polynomial time.*

Proof. Obviously PreClustering-1D(B, k) gives a preclustering $\mathcal{C}$ with $3k$ preclusters. It remains to prove that $\mathcal{C}$ has approximation ratio 1. Let P be a B-instance, and let $Q \in \{q_1, \ldots, q_k\}$ be an optimal set of centroids for the k-center problem on P. Thus by placing an interval of radius $\text{Opt}_\infty(P, k)$ centered at each centroid $q_i \in Q$, we cover all points in P. By assigning each point in P to its nearest centroid in Q, with ties broken arbitrarily, we obtain a partition of P into k clusters. This partition induces a preclustering $\mathcal{C}^*$ of size k on B. We use $\mathcal{C}^*$ to define two types of intervals: *outer intervals*, which are the leftmost or rightmost interval in any of the preclusters $B_i \in \mathcal{C}^*$, and *inner intervals*, which are the remaining intervals. Note that the number of outer intervals is at most $2k$. Define k^* as the largest k' such that $b_1, \ldots, b_{k'}$ are all outer intervals, where $b_1, \ldots, b_n$ is the sorted set of intervals obtained in Step 1 of the algorithm. Since b_{k^*+1} is an inner interval, we have

$$\text{Opt}_\infty(P, k) \geqslant \text{radius}(b_{k^*+1}). \tag{1}$$

The preclustering $\mathcal{C} := \mathcal{C}(k')$ returned by our algorithm minimizes $\text{Opt}_\infty(\{b_{k'+1}, \ldots, b_n\}, 3k-k')$. Note that $\mathcal{C}\text{-Cost}_\infty(P) \leqslant \text{Opt}_\infty(\{b_{k'+1}, \ldots, b_n\}, 3k-k')$, since the intervals $b_1, \ldots, b_{k'}$ are all in singleton preclusters and an interval covering all intervals in a precluster B_i obviously covers all points from P in those interval. Hence,

$$\mathcal{C}\text{-Cost}_\infty(P) \leqslant \text{Opt}_\infty(\{b_{k^*+1}, \ldots, b_n\}, 3k-k^*).$$

It remains to argue that $\text{Opt}_\infty(P, k) \geqslant \text{Opt}_\infty(\{b_{k^*+1}, \ldots, b_n\}, 3k-k^*)$. To this end, we create a collection $\mathcal{I}$ of intervals as follows.

- For each outer interval b_j with $j > k^*$ we create an interval equal to b_j.
- For each precluster $B_i \in \mathcal{C}^*$ that has at least one inner interval, we create a minimum-length interval covering all inner intervals of B_i.

Note that $\mathcal{I}$ contains at most $3k-k^*$ intervals, and that these intervals together cover all intervals in $\{b_{k^*+1}, \ldots, b_n\}$. Hence,

$$\max_{I \in \mathcal{I}} \text{radius}(I) \geqslant \text{Opt}_\infty(\{b_{k^*+1}, \ldots, b_n\}, 3k-k^*).$$

Moreover, $\text{Opt}_\infty(P, k) \geqslant \text{radius}(I)$ for any $I \in \mathcal{I}$. Indeed, if I is equal to an outer interval b_j with $j > k^*$ then $\text{Opt}_\infty(P, k) \geqslant \text{radius}(b_j)$ by Inequality (1), and otherwise I is the minimum-length interval covering all inner intervals of some precluster B_i. (In the latter case we also have $\text{Opt}_\infty(P, k) \geqslant \text{radius}(I)$ because in any B-instance the cluster of B_I includes a point in both outer intervals) We conclude that

$$\text{Opt}_\infty(P, k) \geqslant \max_{I \in \mathcal{I}} \text{radius}(I) \geqslant \text{Opt}_\infty(\{b_{k^*+1}, \ldots, b_n\}, 3k-k^*).$$

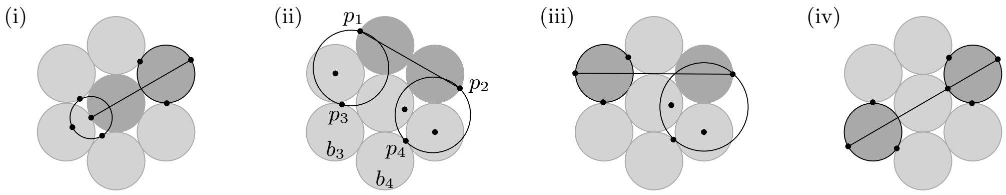


Figure 2 The seven balls shown in the figure do not admit a $(3k, 1)$-preclustering for $k = 2$.

It remains to argue that PreClustering-1D(B, k) can be implemented to run in polynomial time. The most time-consuming step is Step 2a, which can be implemented to run in $O(n^2k)$ time using dynamic programming in a straightforward manner. ◀

Theorem 3 only holds for the k-center problem. In the next section we present a more general algorithm, which not only works in higher dimensions but also for k-median and k-means. The approximation ratio will not be as good as the one provided by Theorem 2, however.

3 The d-dimensional problem

In the previous section we saw that for some problem instances any preclustering with fewer than $3k - 3$ preclusters has an arbitrarily large approximation ratio. The result is stated for $\mathbb{R}^1$ but it also holds in $\mathbb{R}^d$ for $d > 1$: we can use exactly the same construction, replacing the intervals by d-dimensional balls whose centers lie on the x_1-axis. We also presented an algorithm giving a $(3k, 1)$-preclustering for intervals in $\mathbb{R}^1$, for the k-center problem.

Fig. 2 shows that a $(3k, 1)$-preclustering is not always possible for the k-center problem in $\mathbb{R}^2$. The figure shows a set B of seven unit balls, with one central ball touching the other six balls. For $k = 2$ a preclustering of size $3k$ would use five singleton preclusters and one precluster with two balls. There are four combinatorially distinct ways of choosing the precluster of two balls, indicated by the dark grey balls in parts (i)–(iv) of the figure. For each case, a B-instance is shown (the black dots), and the optimal solution to the 2-center problem for the instance is shown (the two black circles). The best preclustering is the one in part (ii). Here the two points p_1, p_2 in the dark grey balls are placed at distance 4 from each other, so $\mathcal{C}$-Cost$_\infty(P) = 2$. The point p_3 inside the ball b_3 is placed as close to p_1 as possible, while p_4 is placed as close to p_2 as possible. The other points are placed such that they are either contained in the ball with diameter p_1p_3 or in the ball with diameter p_2p_4. Hence, Opt$_\infty(P) = (\sqrt{13} - 1)/2$. The balls in this construction are not disjoint, but we can scale them by a factor $(1 - \varepsilon)$ to obtain an instance where any $(3k, \alpha)$-preclustering has $\alpha \geqslant 2/((\sqrt{13} - 1)/2) = 4/(\sqrt{13} - 1) \approx 1.54$.

We now present a preclustering strategy that works for k-center, k-means and k-median in any dimension. It is similar to, and actually somewhat simpler than, the preclustering algorithm we presented for the 1-dimensional k-center problem.

PreClustering-dD(B, k)

1. Sort the balls in B by radius, such that radius$(b_1) \geqslant \cdots \geqslant$ radius(b_n).
2. Define $B_{\text{small}} := \{b_{2k+1}, \ldots, b_n\}$; we call the balls in B_{small} *small.* Let $\{P_1, \ldots, P_k\}$ be an optimal k-center (or k-median, or k-means) clustering on the point set centers$(B_{\text{small}}) := \{c_j : 2k + 1 \leqslant j \leqslant n\}$, where c_j is the center of the ball b_j. Let $\{B_1, \ldots, B_k\}$ be the preclustering on B_{small} induced by it.
3. Return the preclustering $\mathcal{C} := \{\{b_1\}, \ldots, \{b_{2k}\}, B_1, \ldots, B_k\}$.

Before we analyze the algorithm's approximation ratio, we note that, depending on the dimension d and the value of k, we may not be able to implement Step 2 efficiently. However, instead of computing an optimal k-clustering on the centers of the small balls, we can also compute a $(1+\varepsilon')$-approximation of the optimal clustering. For an appropriate $\varepsilon' = O(\varepsilon)$ this increases the approximation ratio by only a factor $1+\varepsilon$, as explained later.

Obviously $\textsc{PreClustering-dD}(B,k)$ gives a preclustering of size $3k$. To analyze the approximation ratio, we use the following lemma.

▶ **Lemma 4.** *For any B-instance P the preclustering $\mathcal{C} := \{\{b_1\},\ldots,\{b_{2k}\},B_1,\ldots,B_k\}$ computed by the algorithm satisfies:*

(i) $\mathcal{C}\text{-}\textsc{Cost}_\infty(P) \leqslant \textsc{Opt}_\infty(P,k) + 2\cdot \mathrm{radius}(b_{2k+1})$

(ii) $\mathcal{C}\text{-}\textsc{Cost}_1(P) \leqslant \textsc{Opt}_1(P,k) + 2\sum_{j=2k+1}^{n} \mathrm{radius}(b_j)$

(iii) $\mathcal{C}\text{-}\textsc{Cost}_2(P) \leqslant 4\cdot \textsc{Opt}_2(P,k) + 6\sum_{j=2k+1}^{n} \mathrm{radius}(b_j)^2$.

Proof. We first prove part (i) of the lemma. Let P be any B-instance, let $p_j \in P$ denote the point inside b_j, and let c_j be the center of b_j. Recall that $P_i \subset P$ is the subset of points in the instance corresponding to the precluster B_i. Define $P_{\text{small}} := \{p_{2k+1},\ldots,p_n\}$ to be the set of points from P in the small balls, and define $C_{\text{small}} := \{c_{2k+1},\ldots,c_n\}$. Note that $P_{\text{small}} = P_1 \cup \cdots \cup P_k$ and that

$$|p_j c_j| \leqslant \mathrm{radius}(b_j) \leqslant \mathrm{radius}(b_{2k+1}) \tag{2}$$

for all $p_j \in P_{\text{small}}$. We define the following sets of centroids:

- Let $Q := \{q_1,\ldots,q_k\}$ be the set of centroids in an optimal k-center solution for the entire point set P. We have

$$\max_{p_j\in P_{\text{small}}} \min_{q_i\in Q} |p_j q_i| \leqslant \max_{p_j\in P} \min_{q_i\in Q} |p_j q_i| = \textsc{Opt}_\infty(P,k). \tag{3}$$

- Let $Q' := \{q'_1,\ldots,q'_k\}$ be the set of centroids in the optimal k-center clustering on C_{small} used in Step 2 of the algorithm. Thus

$$\max_{c_i\in C_{\text{small}}} \min_{q'_j\in Q'} |c_i q'_j| = \textsc{Opt}_\infty(C_{\text{small}},k) \leqslant \max_{c_i\in C_{\text{small}}} \min_{q_j\in Q} |c_i q'_j|. \tag{4}$$

- Let $Q'' := \{q''_1,\ldots,q''_k\}$, where q''_i is the optimal centroid for P_i. Note that for all P_i we have

$$\max_{p_j\in P_i} |p_j q''_j| \leqslant \max_{p_j\in P_i} |p_j q'_j|. \tag{5}$$

Since the total cost of the singleton preclusters is trivially zero, we have

$$\begin{aligned}
&\mathcal{C}\text{-}\textsc{Cost}_\infty(P) \\
&= \max_{1\leqslant i\leqslant k} \max_{p_j\in P_i} |p_j q''_i| \\
&\leqslant \max_{1\leqslant i\leqslant k} \max_{p_j\in P_i} |p_j q'_i| && \text{(Inequality (5))}\\
&\leqslant \max_{1\leqslant i\leqslant k} \max_{p_j\in P_i} \big(|p_j c_j| + |c_j q'_i|\big) && \text{(triangle inequality)}\\
&\leqslant \mathrm{radius}(b_{2k+1}) + \max_{1\leqslant i\leqslant k} \max_{p_j\in P_i} |c_j q'_i| && \text{(Inequality (2))}\\
&\leqslant \mathrm{radius}(b_{2k+1}) + \max_{c_j\in C_{\text{small}}} \min_{q'_i\in Q'} |c_j q'_i| && \text{(definition of } C_{\text{small}})\\
&\leqslant \mathrm{radius}(b_{2k+1}) + \max_{c_j\in C_{\text{small}}} \min_{q_i\in Q} |c_j q_i| && \text{(Inequality (4))}\\
&\leqslant \mathrm{radius}(b_{2k+1}) + \max_{p_j\in P_{\text{small}}} \min_{q_i\in Q} \big(|c_j p_j| + |p_j q_i|\big) && \text{(triangle inequality)}\\
&\leqslant 2\cdot \mathrm{radius}(b_{2k+1}) + \max_{p_j\in P_{\text{small}}} \min_{q_i\in Q} |p_j q_i| && \text{(Inequality (2))}\\
&\leqslant 2\cdot \mathrm{radius}(b_{2k+1}) + \textsc{Opt}_\infty(P,k) && \text{(Inequality (3))}
\end{aligned}$$

To prove part (ii) of the lemma, which deals with the k-median problem, note that Inequality (2) still holds while Inequalities (3)–(5) hold if we replace the max-operator by a summation. Part (ii) can thus be derived using a similar derivation as for part (i).

To prove part (iii), which deals with the k-means problem, we need to work with squared distances. Note that Inequality (2) still holds, while Inequalities (3)–(5) hold if we replace the max-operator with a summation and all distance values with their squared values. For squared distances the triangle inequality does not hold. Instead we use the Cauchy-Schwarz inequality, which implies that if a, b, c are positive reals with $a \leqslant b + c$, then $a^2 \leqslant 2b^2 + 2c^2$. A similar computation as above can now be used to prove part (iii), we have

$$\begin{aligned}
&\mathcal{C}\text{-}\textsc{Cost}_2(P) && \\
&= \textstyle\sum_{i=1}^{k} \sum_{p_j \in P_i} |p_j q_i''|^2 && \\
&\leqslant \textstyle\sum_{i=1}^{k} \sum_{p_j \in P_i} |p_j q_i'|^2 && \text{(Inequality (5))} \\
&\leqslant \textstyle\sum_{i=1}^{k} \sum_{p_j \in P_i} \left(2|p_j c_j|^2 + 2|c_j q_i'|^2\right) && \text{(Cauchy-Schwarz)} \\
&\leqslant 2\textstyle\sum_{j=2k+1}^{n} \text{radius}(b_j)^2 + 2\sum_{i=1}^{k} \sum_{p_j \in P_i} |c_j q_i'|^2 && \text{(Inequality (2))} \\
&\leqslant 2\textstyle\sum_{j=2k+1}^{n} \text{radius}(b_j)^2 + 2\sum_{c_j \in C_{\text{small}}} \min_{q_i' \in Q'} |c_j q_i'|^2 && \text{(definition of } C_{\text{small}}\text{)} \\
&\leqslant 2\textstyle\sum_{j=2k+1}^{n} \text{radius}(b_j)^2 + 2\sum_{c_j \in C_{\text{small}}} \min_{q_i \in Q} |c_j q_i|^2 && \text{(Inequality (4))} \\
&\leqslant 2\textstyle\sum_{j=2k+1}^{n} \text{radius}(b_j)^2 + 2\sum_{p_j \in P_{\text{small}}} \min_{q_i \in Q} \left(2|c_j p_j|^2 + 2|p_j q_i|^2\right) && \text{(Cauchy-Schwarz)} \\
&\leqslant 6\textstyle\sum_{j=2k+1}^{n} \text{radius}(b_j)^2 + 4\sum_{p_j \in P_{\text{small}}} \min_{q_i \in Q} |p_j q_i|^2 && \text{(Inequality (2))} \\
&\leqslant 6\textstyle\sum_{j=2k+1}^{n} \text{radius}(b_j)^2 + 4 \cdot \textsc{Opt}_2(P, k) && \text{(Inequality (3))}
\end{aligned}$$

◀

The lemma above shows that our preclustering gives an additive error that depends on the radii of the small balls. The following two lemmas will be used to turn this into a multiplicative error. Let r_d^* be the smallest possible radius of any ball that intersects three disjoint unit balls in $\mathbb{R}^d$.

▶ **Lemma 5.** *We have*

(i) $\textit{Opt}_\infty(P, k) \geqslant r_d^* \cdot \text{radius}(b_{2k+1})$

(ii) $\textit{Opt}_1(P, k) \geqslant r_d^* \cdot \sum_{j=2k+1}^{n} \text{radius}(b_j)$

(iii) $\textit{Opt}_2(P, k) \geqslant (r_d^*)^2 \cdot \sum_{j=2k+1}^{n} \text{radius}(b_j)^2$

Proof. For part (i) notice that by the Pigeonhole Principle an optimal clustering must have a cluster containing at least three points from $\{p_1, \ldots, p_{2k+1}\}$. The cost of this cluster is lower bounded by the radius of the smallest ball intersecting three balls of radius at least b_{2k+1}, which is in turn lower bounded by $r_d^* \cdot \text{radius}(b_{2k+1})$.

For part (ii) let $P_1, P_2, \ldots, P_k$ be the clusters in an optimal k-median clustering on P, and let q_i be the centroid of P_i in this clustering. Let B_i be the set of balls corresponding the points in P_i. We claim that

$$\sum_{p_j \in P_i} |p_j q_i| \geqslant r_d^* \cdot \Big(\Big(\sum_{b_j \in B_i} \text{radius}(b_j) \Big) - \text{ sum of the radii of the two largest balls in } B_i \Big). \tag{6}$$

To show this, let $b(q_i, r)$ be the ball of radius r centered at q_i, let $P_i(r) := \{p_j \in P_i : b_j \cap b(q_i, r) \neq \emptyset\}$ be the set of points in P_i whose associated ball intersects $b(q_i, r)$, and let $B_i(r)$ be the corresponding set of balls. Since for sufficiently large r we have $P_i = P_i(r)$, it suffices to show that for all $r > 0$ we have

$$\sum_{p_j \in P_i(r)} |p_j q_i| \geqslant r_d^* \cdot \Big(\Big(\sum_{p_j \in B_i(r)} \text{radius}(b_j) \Big) - \text{ sum of the radii of the two largest balls in } B_i(r) \Big).$$

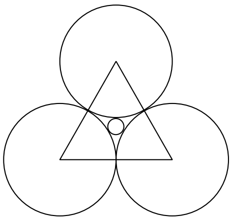


Figure 3 The figure shows the smallest possible ball intersecting three disjoint unit balls in 2D. The larger balls are the unit balls and the radius of the small ball is $r_2^* = \frac{2}{\sqrt{3}} - 1$.

To prove this, consider this inequality as r increases from $r = 0$ to $r = \infty$. As long as $|P_i(0)| \leqslant 2$ the right-hand side is zero and so the inequality is obviously true. As we increase r further, $b(q_i, r)$ starts intersecting more and more balls from B_i. Consider what happens to the inequality when $b(q_i, r)$ starts intersecting another ball $b_\ell \in B_i$. Then p_ℓ is added to $P_i(r)$, so the left-hand side of the inequality increases by $|p_\ell q_i|$, which is at least r. The right-hand side increases by at most r_d^* times the radius of the third-largest ball in B_i. By definition of r_d^*, if three balls intersect a ball of radius r then the smallest has radius at most r/r_d^*. Hence, the right-hand side increases by at most r and the inequality remains true.

Recall that $b_1, \ldots, b_{2k}$ are the $2k$ largest balls in B. Hence, summing Inequality (6) over all clusters $P_1, \ldots, P_k$ gives

$$\textsc{Opt}_1(P,k) = \sum_{i=1}^{k} \sum_{p_j \in P_i} |p_j q_i| \geqslant r_d^* \cdot \left(\sum_{i=1}^{k} \sum_{b_j \in B_i} \text{radius}(b_j) - \sum_{j=1}^{2k} \text{radius}(b_j) \right) = r_d^* \cdot \sum_{j=2k+1}^{n} \text{radius}(b_j).$$

For part (iii) the same proof as (ii) works if we replace all distances with squared distances. ◀

▶ **Lemma 6.** *For all $d \geqslant 2$ we have $r_d^* = 2/\sqrt{3} - 1$.*

Proof. It is easy to see that $r_d^* \leqslant r_2^*$, since any configuration of three disjoint unit disks in the plane, with a fourth disk intersecting all three, can be extended to $\mathbb{R}^d$ by embedding the centers of the balls on a 2-dimensional plane in $\mathbb{R}^d$. Next we show that $r_d^* \geqslant r_2^*$ for all $d \geqslant 2$, which implies that $r_d^* = r_2^*$.

Let $d \geqslant 2$ and let b, b', b'' be three disjoint unit balls in $\mathbb{R}^d$. Let c, c', c'' denote the centers of b, b', and b'', respectively, and let h be a 2-dimensional plane containing c, c', c''. Let D be a smallest ball that intersects b, b', b'' and whose center is restricted to lie on h. Then radius$(D) \geqslant r_2^*$. We claim that D is in fact a smallest ball intersecting b, b', b'' even if we do not restrict the center of this ball to be on h. Indeed, if a ball D' with center $q \notin h$ intersects b, b', b'', then the ball of the same radius as D' and whose center is the orthogonal projection of q onto h also intersects b, b', b''.

It remains to show that $r_2^* = 2/\sqrt{3} - 1$. The configuration minimizing the radius of the smallest ball intersecting b, b', b'' is where b, b', b'' are pairwise touching, resulting in the claimed bound – see Fig. 3. ◀

We are now ready to prove the following theorem.

▶ **Theorem 7.** *Let B be a set of disjoint balls in $\mathbb{R}^d$ with $d \geqslant 2$. Then*

(i) *there exists a $(3k, 7 + 4\sqrt{3})$-preclustering for the k-center and the k-median problem,*

(ii) *there exists a $(3k, 130 + 72\sqrt{3})$-preclustering for the k-means problem.*

Moreover, a $(3k, 7 + 4\sqrt{3} + \varepsilon)$-preclustering for the k-center and the k-median problem, and a $(3k, 130 + 72\sqrt{3} + \varepsilon)$-preclustering for the k-means problem can be computed in polynomial time.

Proof. Parts (i) and (ii) follow immediately by putting together Lemmas 4–6. It remains to argue that we can compute a preclustering whose approximation ratio is as claimed in polynomial time. Recall that each of the three clustering problems admits a PTAS [1, 2, 4, 6, 8], that is, for any given $\varepsilon' > 0$ we can compute a $(1+\varepsilon')$-approximation to an optimal clustering in polynomial time. To obtain the result, we set $\varepsilon' := \varepsilon/(1 + \frac{1}{r_d^*})$ for the k-center and k-median problem and $\varepsilon' := \varepsilon/(2 + \frac{2}{(r_d^*)^2})$ for the k-means problem. Then in Step 2 of PreClustering-dD(B, k) we compute a $(1 + \varepsilon')$-approximation of the optimal clustering. The resulting algorithm runs in polynomial time. The only change in the analysis will appear in Inequality (4) of Lemma 4, where we get an extra multiplicative factor $1 + \varepsilon'$. With the above choice of ε' the approximation ratio for the whole algorithm will increase by ε. ◀

Generalizing the solution. We generalize the above theorem in order to control the number of preclusters for various approximations. Let r_d^p be the minimum possible value for the radius of a ball being tangent to p disjoint unit balls in $\mathbb{R}^d$ for $d \geqslant 2$. Notice that $r_d^3 = r_d^*$. We can generalize the above result for appropriate p as follows.

The algorithm here is similar to PreClustering-dD, but in Step 2 we replace b_{2k+1} by $b_{(p-1)k+1}$ and in Step 3 we return the preclustering $\mathcal{C} := \{\{b_1\}, \ldots, \{b_{(p-1)k}\}, B_1, \ldots, B_k\}$. Note that Lemmas 4, 5 still hold if we replace $2k + 1$ with $(p - 1)k + 1$ and r_d^* with r_d^p.

▶ **Theorem 8.** *Let B be a set of disjoint balls in $\mathbb{R}^d$ with $d \geqslant 2$. Then*

(i) *there exists a $(pk, 1 + \frac{2}{r_d^p})$-preclusterings for the k-center and the k-median problem.*

(ii) *there exists a $(pk, 4 + \frac{6}{(r_d^p)^2})$-preclustering for the k-means problem.*

Moreover, a $(pk, 1 + \frac{2}{r_d^p} + \varepsilon)$-preclustering for the k-center and the k-median problem, and a $(pk, 4 + \frac{6}{(r_d^p)^2} + \varepsilon)$-preclustering for the k-means problem can be computed in polynomial time.

For instance, for $d = 2$ and $p = 6$ we have $r_2^6 = 1$ – indeed, any ball intersecting six disjoint unit balls in $\mathbb{R}^2$ has at least unit radius itself – leading to the following corollary. (For other bounds on r_d^p, see at [15].)

▶ **Corollary 9.** *Any set of disjoint balls in $\mathbb{R}^2$ admits a $(6k, 3)$-preclustering for the k-center and the k-median problem, and a $(6k, 10)$-preclustering for k-means problem.*

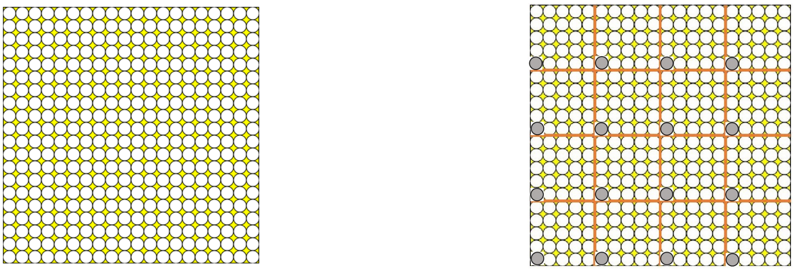
4 Asymptotically tight trade-offs for the k-center problem

Next, we explain how to obtain a $(\Theta(\lceil 1/\varepsilon^d \rceil \cdot k), \varepsilon)$-preclustering for the k-center problem, by adding more steps to the algorithm PreClustering-dD(B, k).

▶ **Lemma 10.** *For any point set P in $\mathbb{R}^d$, any integer $k \geqslant 1$, and any $\varepsilon > 0$ we have*

$$\textsc{Opt}_\infty(P, c_d(\varepsilon) \cdot k) \leqslant \varepsilon \cdot \textsc{Opt}_\infty(P, k)$$

for $c_d(\varepsilon) = \lceil \sqrt{d}/\varepsilon \rceil^d$.



(a) n unit balls forming a square in 2D. **(b)** clustering the circles into square-shaped clusters.

Figure 4 Illustration for the proof Theorem 12.

Proof. First consider the case $k = 1$. Let Q be the optimal centroid for P and let S be the smallest hypercube centered at Q and containing P. Note that the edge length of S is at most $2\text{OPT}_\infty(P,k)$. Partition S into $\lceil\sqrt{d}/\varepsilon\rceil^d$ smaller hypercubes of edge length at most $2\varepsilon \cdot \text{OPT}_\infty(P,k)/\sqrt{d}$, and for each such hypercube make a cluster containing all points in it. Note that each such cluster can be covered by a ball of radius $\varepsilon \cdot \text{OPT}_\infty(P,k)$. Hence,

$$\text{OPT}_\infty(P, \lceil\sqrt{d}/\varepsilon\rceil^d \cdot k) \leqslant \varepsilon \cdot \text{OPT}_\infty(P,k).$$

For $k > 1$ we can simply apply the result for $k = 1$ to each of the k clusters in an optimal k-center clustering on P. ◀

With this lemma in hand we can now run algorithm PRECLUSTERING-DD(B, k') with the appropriate value of k, namely $k' = c_d(\varepsilon/(7+4\sqrt{3}))\cdot k$, and then by Theorem 7 we get a $(3k', \varepsilon)$-preclustering with $k' = \Theta(\lceil 1/\varepsilon\rceil^d \cdot k)$.

▶ **Theorem 11.** *Let B be a set of disjoint balls in $\mathbb{R}^d$ with $d \geqslant 2$. Then there exists a $(\Theta(\lceil 1/\varepsilon^d\rceil \cdot k), \varepsilon)$-preclustering for B for any positive constant ε.*

Finally, we show that this number of preclusters is asymptotically the best number we can achieve.

▶ **Theorem 12.** *There exists a set B of n disjoint balls in $\mathbb{R}^d$ such that in any $(f(k), \varepsilon)$-preclustering of B for the k-center problem, we have $f(k) = \Omega(\lceil 1/\varepsilon^d\rceil \cdot k)$.*

Proof. Observe that it suffices to prove the lower bound for $k = 1$; for larger k we can simply copy the construction k times and put the copies sufficiently far from each other. Now, for $k = 1$ consider a set B of $n^{1/d} \times \cdots \times n^{1/d}$ unit balls arranged in a grid-like pattern, as in Fig. 4a. Note that $\text{OPT}_\infty(P,1) \leqslant \sqrt{d}(n^{1/d}+1)$ for any B-instance P. Now partition the "grid" into $(\sqrt{d}/\varepsilon)^d$ "subgrids" as in Fig. 4b. For each subgrid, select the ball with the lexicographically smallest center (shaded in Fig. 4b), and let $B^* \subset B$ be the set of selected balls. If a preclustering uses fewer than $(\sqrt{d}/\varepsilon)^d$ preclusters, two of the balls from B^* will end up in the same precluster. But then there is a B-instance P where $\mathcal{C}\text{-COST}_\infty(P) > \varepsilon \cdot \sqrt{d} \cdot n^{1/d} + 1$. Hence, any $(f(1), \varepsilon)$-precluster must have $\Omega(\lceil 1/\varepsilon^d\rceil)$ preclusters. ◀

5 Concluding remarks

In this paper, we introduced the concept of preclustering for imprecise points and studied it for k-center,k-median and k-means problems. It would be interesting if one can fill the gap between lower and upper bounds for the number of preclusters needed in order to approximate the optimum solution. Also one can try to generalize the ideas used in section 4 for the k-median and k-means versions. It would also be interesting to study non-disjoint balls, and try to obtain preclusterings whose size and approximation ratio depend on the amount of overlap between the balls.

References

1 Pankaj K. Agarwal and Cecilia Magdalena Procopiuc. Exact and approximation algorithms for clustering. *Algorithmica*, 33(2):201–226, 2002. doi:10.1007/s00453-001-0110-y.

2 Sanjeev Arora, Prabhakar Raghavan, and Satish Rao. Approximation schemes for euclidean k-medians and related problems. In *Proceedings of the ACM Symposium on the Theory of Computing*, pages 106–113, 1998. doi:10.1145/276698.276718.

3 Kevin Buchin, Maarten Löffler, Pat Morin, and Wolfgang Mulzer. Preprocessing imprecise points for delaunay triangulation: Simplified and extended. *Algorithmica*, 61(3):674–693, 2011. doi:10.1007/s00453-010-9430-0.

4 Dan Feldman, Morteza Monemizadeh, and Christian Sohler. A PTAS for k-means clustering based on weak coresets. In *Proceedings of ACM Symposium on Computational Geometry*, pages 11–18, 2007. doi:10.1145/1247069.1247072.

5 Dorit S. Hochbaum and David B. Shmoys. A best possible heuristic for the k-center problem. *Math. Oper. Res.*, 10(2):180–184, 1985. doi:10.1287/moor.10.2.180.

6 Ragesh Jaiswal, Amit Kumar, and Sandeep Sen. A simple D 2-sampling based PTAS for k-means and other clustering problems. In *Computing and Combinatorics COCOON 2012*, volume 7434 of *Lecture Notes in Computer Science*, pages 13–24. Springer, 2012. doi:10.1007/978-3-642-32241-9_2.

7 Wenqi Ju, Jun Luo, Binhai Zhu, and Ovidiu Daescu. Largest area convex hull of imprecise data based on axis-aligned squares. *J. Comb. Optim.*, 26(4):832–859, 2013. doi:10.1007/s10878-012-9488-5.

8 Stavros G. Kolliopoulos and Satish Rao. A nearly linear-time approximation scheme for the euclidean kappa-median problem. In *Algorithms - ESA Proceedings*, volume 1643 of *Lecture Notes in Computer Science*, pages 378–389. Springer, 1999. doi:10.1007/3-540-48481-7_33.

9 Chih-Hung Liu and Sandro Montanari. Minimizing the diameter of a spanning tree for imprecise points. *Algorithmica*, 80(2):801–826, 2018. doi:10.1007/s00453-017-0292-6.

10 Maarten Löffler. *Data Imprecision in Computational Geometry*. PhD thesis, Utrecht University, Netherlands, 2009.

11 Maarten Löffler and Marc J. van Kreveld. Largest and smallest convex hulls for imprecise points. *Algorithmica*, 56(2):235–269, 2010. doi:10.1007/s00453-008-9174-2.

12 Takayuki Nagai and Nobuki Tokura. Tight error bounds of geometric problems on convex objects with imprecise coordinates. In *JCDCG*, volume 2098 of *Lecture Notes in Computer Science*, pages 252–263. Springer, 2000. doi:10.1007/3-540-47738-1_24.

13 Farnaz Sheikhi, Ali Mohades, Mark de Berg, and Ali D. Mehrabi. Separability of imprecise points. *Comput. Geom.*, 61:24–37, 2017. doi:10.1016/j.comgeo.2016.10.001.

14 David B. Shmoys and Éva Tardos. An approximation algorithm for the generalized assignment problem. *Math. Program.*, 62:461–474, 1993. doi:10.1007/BF01585178.

15 István Talata. Exponential lower bound for the translative kissing numbers of d -dimensional convex bodies. *Discrete and Computational Geometry*, 19(3):447–455, 1998. doi:10.1007/PL00009362.