



Fast Approximate ℓ -Center Clustering in High-Dimensional Spaces

Mirośław Kowaluk ¹, Andrzej Lingas ^{2,*} and Mia Persson ³¹ Institute of Informatics, University of Warsaw, 02-097 Warsaw, Poland; kowaluk@mimuw.edu.pl² Department of Computer Science, Lund University, Box 118, 221 00 Lund, Sweden³ Department of Computer Science and Media Technology, Malmö University, 205 06 Malmö, Sweden; mia.persson@mau.se

* Correspondence: andrzej.lingas@cs.lth.se

Abstract

We study the design of efficient approximation algorithms for the ℓ -center clustering and minimum-diameter ℓ -clustering problems in high-dimensional Euclidean and Hamming spaces. Our main tool is randomized dimension reduction. First, we present a general method of reducing the dependency of the running time of a hypothetical algorithm for the ℓ -center problem in a high-dimensional Euclidean space on the dimension. Utilizing this method in part, we provide $(2 + \epsilon)$ -approximation algorithms for the ℓ -center clustering and minimum-diameter ℓ -clustering problems in Euclidean and Hamming spaces that are substantially faster than the known 2-approximation algorithms when both ℓ and the dimension are super-logarithmic. Next, we apply the general method to the recent fast approximation algorithms with higher approximation guarantees for the ℓ -center clustering problem in a high-dimensional Euclidean space. Finally, we provide a speed-up of the known $O(1)$ -approximation method for the generalization of the ℓ -center clustering problem that allows z outliers (i.e., z input points may be ignored when computing the maximum distance from an input point to a center) in high-dimensional Euclidean and Hamming spaces.

Keywords: ℓ_2 distance; Euclidean space; Hamming distance; Hamming space; clustering; approximation algorithm; time complexity

1. Introduction

Clustering is nowadays a standard tool in the data analysis of computational biology, medical sciences, computer vision, and machine learning. One of the most popular variants of clustering is the ℓ -center clustering problem and the related minimum-diameter ℓ -clustering problem in metric spaces (including, among others, Euclidean and Hamming spaces). Given a finite set P of points in a metric space, the first problem asks for a set of ℓ points in the metric space, called *centers*, such that the maximum distance from a point in P to its nearest center is minimized. The second problem asks for a partition of the input point set P into ℓ clusters such that the maximum of cluster diameters is minimized. See Figure 1. Both problems are known to be NP-hard and even NP-hard to approximate within $2 - \epsilon$ for any constant $\epsilon > 0$ [1].

González provided a simple 2-approximation method for ℓ -center clustering that also yields a 2-approximation for minimum-diameter ℓ -clustering [1]. Hochbaum and Shmoys obtained analogous approximation results in a graph-based setting of the ℓ -center problem, where the input is not merely a set of points in a metric space but an edge-weighted complete graph with weights satisfying the triangle inequality, and centers can be selected only



Academic Editor: Yu-Chen Hu

Received: 21 January 2026

Revised: 24 February 2026

Accepted: 19 March 2026

Published: 23 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

among the graph’s vertices. In case of a d -dimensional space, González’s method takes $O(nd\ell)$ time, where n is the number of input points and d is the dimension. For Euclidean spaces of bounded dimensions, and more generally, for metric spaces of bounded doubling dimensions, there exist faster 2-approximation algorithms for the ℓ -center problem, although their running times hide an exponential dependence on the dimension; see [2] and [3], respectively. There are also several more recent works on speeding up ℓ -center approximation algorithms in Euclidean spaces by allowing worse approximation guarantees (e.g., [4–6], and for a bicriteria variant [7]). In particular, trade-offs between approximation guarantees of the form $O(\alpha)$ and the running times $\text{poly}(d \log n)(n + \ell^{1+1/\alpha^2} n^{O(1/\alpha^{2/3})})$ and $\text{poly}(d \log n)n\ell^{1/\alpha^2}$ have been obtained in [5] (in [5], the notation $\tilde{O}(\cdot)$ suppresses $\text{poly}(d \log n)$ factors) by refining an earlier result from [4].

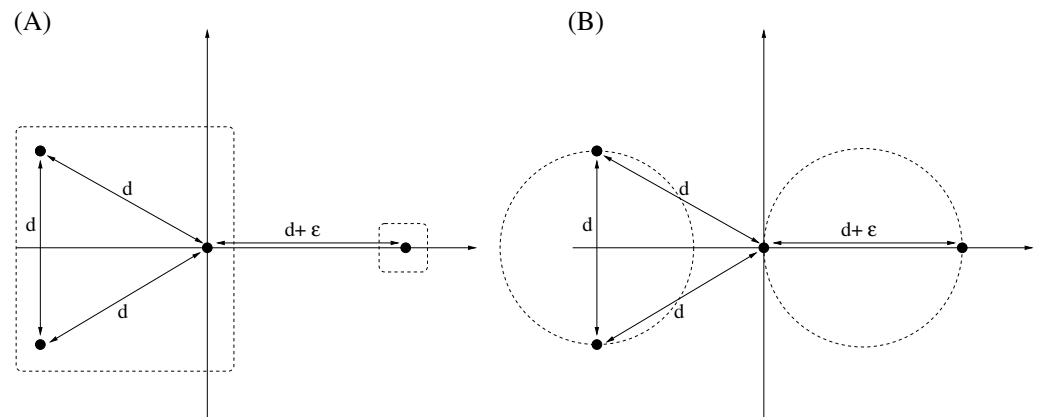


Figure 1. Figure (A) presents a minimum diameter 2-clustering of four points in the plane. Figure (B) presents the 2-clustering of the same points that is induced by an optimal solution to the 2-center problem for these points. The optimal solution consists of the midpoint of the vertical segment between the two points on the left-hand side and the midpoint of the horizontal segment connecting the two points on the right-hand side.

In some of the aforementioned applications of both ℓ -clustering problems, massive datasets in a high-dimensional metric space combined with a large value of the parameter ℓ may arise. For instance, in recent combinatorial, clustering-based algorithms for Boolean $n \times n$ matrix multiplication [8] and 0-1 $n \times n$ matrix multiplication [9], the rows of the first matrix and/or the columns of the second matrix are interpreted as points in an n -dimensional Hamming space. They are subject to an approximate ℓ -center clustering, where ℓ is merely expected to be sublinear in n . In such situations, neither the $O(nd\ell)$ -time method nor the algorithms with substantially worse approximation guarantees or time complexity heavily dependent on d are sufficiently useful.

1.1. Our Contributions

We focus on the design of efficient approximation algorithms for the ℓ -center clustering and minimum-diameter ℓ -clustering problems in high-dimensional Euclidean and Hamming spaces. First, we present a general method of reducing the dependency of the running time of a hypothetical algorithm for the ℓ -center problem in a high-dimensional Euclidean space on the dimension. The algorithm is required to be *conservative*, i.e., it must always return centers that belong to the input point set. The method relies on randomized dimension reduction and almost preserves the approximation guarantee of the original algorithm. Utilizing this method in part, we provide $(2 + \epsilon)$ -approximation algorithms for the ℓ -center clustering and minimum-diameter ℓ -clustering problems in Euclidean and Hamming spaces that are substantially faster than the known 2-approximation algorithms when both ℓ and the dimension are super-logarithmic. Next, we apply the general method

to the recent fast approximation algorithms with higher approximation guarantees for the ℓ -center clustering problem in a high-dimensional Euclidean space. Finally, we provide a speed-up of the known $O(1)$ -approximation method of Charikar et al. [10] for the generalization of the ℓ -center clustering problem that allows z outliers (i.e., z input points may be ignored while computing the maximum distance from an input point to a center) in high-dimensional Euclidean and Hamming spaces [10]. The speed-up is also based on randomized dimension reduction and the resulting approximation guarantee is only slightly larger than the original one. See also Table 1 for a summary of our contributions.

Table 1. A summary of speed-ups of approximation algorithms for ℓ -center problems obtained by randomized dimension reduction presented in this paper. Importantly, the approximation guarantees for our algorithms hold with high probability forming. The approximation guarantee of 3 for the ℓ -center problem with outliers in an arbitrary metric space in [10] as well as our guarantee of $3 + \epsilon$ in a Euclidean or Hamming space are derived with respect to an optimal conservative solution, where the centers belong to input points. The authors of [10] claim that they can remove this assumption in case of a Euclidean space.

Problem	Metric	Approx.	Time Complexity	Reference
ℓ -center/min-diam.	arbitrary	2	$O(nd\ell)$	[1]
ℓ -center/min-diam.	Euclid., Ham.	$2 + \epsilon$	$O(n \log n(d + \ell)/\epsilon^2)$	this paper
ℓ -center	Euclidean	$O(\alpha)$	$\frac{\text{poly}(d \log n)}{(n + \ell^{1+1/\alpha^2} n^{O(1/\alpha^{2/3})})}$	[4,5]
ℓ -center	Euclidean	$O(\alpha)$	$\tilde{O}(nd/\epsilon^2 + \ell^{1/\alpha^2} n^{O(1/\alpha^{2/3})})$	this paper
ℓ -center	Euclidean	$O(\alpha)$	$\text{poly}(d \log n) n \ell^{1/\alpha^2}$	[4,5]
ℓ -center	Euclidean	$O(\alpha)$	$\tilde{O}(nd/\epsilon^2 + n \ell^{1/\alpha^2})$	this paper
ℓ -center + outliers	arbitrary	3	$\text{poly}(n, d, \ell)$	[10]
ℓ -center + outliers	Euclid., Ham.	$3 + \epsilon$	$\tilde{O}(n^2(\epsilon^{-2} + \ell))$	this paper

1.2. Techniques

Our fast randomized algorithms for approximate ℓ -center clustering and minimum-diameter ℓ -clustering problems in high-dimensional Euclidean and Hamming spaces are based on a variant of randomized dimension reduction in Euclidean spaces given by Achlioptas in [11], together with the observation that the Hamming distance between two 0-1 vectors is equal to their squared ℓ_2 distance. The main idea of a randomized dimension reduction is to provide a uniform random map from a d -dimensional metric space to its k -dimensional subspace that preserves distances up to $1 \pm \epsilon$ factors, where $k = O(\log n)$, with high probability. Johnson and Lindenstrauss (JL) were first to provide such maps from \mathbb{R}^d to \mathbb{R}^k for the ℓ_2 norm [12]. The advantage of Achlioptas' variant of JL dimension reduction is that such a random map can be generated using only binary random variables [11].

Our main algorithmic tool is the elegant furthest-point traversal method due to González [1]. Following Charikar et al. [10], in our speed-up of their approximation algorithm for the ℓ -center problem with outliers, we reduce the approximation task to finding a specific set cover of the input point set, using at most ℓ sets, each consisting of input points within a given threshold distance from candidate centers. As [10], we use a greedy method to find such a set cover for a fixed distance threshold.

1.3. Paper Organization

The next section contains basic definitions and facts on randomized dimension reduction. Section 3 presents our general method for reducing the dependence on the dimension. Section 4 provides our fast $(2 + \epsilon)$ -approximation algorithms for the ℓ -center clustering and minimum-diameter ℓ -clustering problems in high-dimensional Euclidean and Hamming spaces. Section 5 presents applications of our method from Section 3 to recent fast approximation algorithms for ℓ -center clustering in high-dimensional Euclidean spaces. Section 6 provides a speed-up of the known greedy $O(1)$ -approximation method for the ℓ -center clustering problem with outliers in high-dimensional Euclidean space. We conclude with final remarks.

2. Preliminaries

For a positive integer r , $[r]$ stands for the set of positive integers not exceeding r . The cardinality of a finite set S is denoted by $|S|$.

The transpose of a matrix D is denoted by D^\top .

The *Hamming distance* between two points a, b (vectors) in $\{0, 1\}^d$ is the number of the coordinates in which the two points differ. Alternatively, it can be defined as the distance between a and b in the ℓ_1 metric over $\{0, 1\}^d$. It is denoted by $\text{ham}(a, b)$. The distance between two real vectors a, b in the ℓ_2 metric is denoted by $\|a - b\|_2$.

An event is said to hold *with high probability* (w.h.p. for short) in terms of a parameter N related to the input size if it holds with probability of at least $1 - \frac{1}{N^\alpha}$, where α is any constant not less than 1.

The following fact and corollaries enable an efficient randomized dimension reduction in high-dimensional Euclidean and Hamming spaces.

Fact 1 (Achlioptas [11]). *Let P be an arbitrary set of n points in \mathbb{R}^d , represented as an $n \times d$ matrix A . Given $\epsilon, \beta > 0$, let $k_0 = \frac{4+2\beta}{\epsilon^2/2-\epsilon^3/3} \log n$. For an integer $k \geq k_0$, let R be a $d \times k$ random matrix (R_{ij}) , where $R_{ij} = 1$ with probability $\frac{1}{2}$ and $R_{ij} = -1$ otherwise. Let $E = \frac{1}{\sqrt{k}}AR$ and let $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ map the i -th row of A on the i -th row of E . With probability of at least $1 - n^{-\beta}$, for all $u, v \in P$,*

$$(1 - \epsilon)(\|u - v\|_2)^2 \leq (\|f(u) - f(v)\|_2)^2 \leq (1 + \epsilon)(\|u - v\|_2)^2.$$

Since $1 - \epsilon \leq \sqrt{1 - \epsilon}$ and $\sqrt{1 + \epsilon} \leq 1 + \epsilon$ for $\epsilon \in (0, 1)$, we immediately obtain the following corollary from Fact 1.

Corollary 1. *Assume the notation from Fact 1. Suppose that $\epsilon \in (0, 1)$ and $P \subset \mathbb{R}^d$. Then, w.h.p. for all $u, v \in P$,*

$$(1 - \epsilon)\|u - v\|_2 \leq \|f(u) - f(v)\|_2 \leq (1 + \epsilon)\|u - v\|_2.$$

Observe that if $u, v \in \{0, 1\}^d$, then $\text{ham}(u, v) = (\|u - v\|_2)^2$. Hence, we also obtain the following corollary from Fact 1.

Corollary 2. *Assume the notation from Fact 1. Suppose that $\epsilon > 0$ and $P \subset \{0, 1\}^d \subset \mathbb{R}^d$. Then, w.h.p. for all $u, v \in P$,*

$$(1 - \epsilon)\text{ham}(u, v) \leq (\|f(u) - f(v)\|_2)^2 \leq (1 + \epsilon)\text{ham}(u, v).$$

3. Approximate ℓ -Center Clustering in High-Dimensional Spaces

The ℓ -center clustering problem in the Euclidean \mathbb{R}^d is defined as follows: given a set P of n points in \mathbb{R}^d , find a set T of ℓ points in \mathbb{R}^d that minimizes $\max_{v \in P} \min_{u \in T} \|v - u\|_2$. The minimum-diameter ℓ -clustering problem in the Euclidean \mathbb{R}^d is defined as follows: given a set P of n points in \mathbb{R}^d , find a partition of P into ℓ subsets P_1, P_2, \dots, P_ℓ that minimizes $\max_{i \in [\ell]} \max_{v, u \in P_i} \|v - u\|_2$. The ℓ -center clustering problem could also be termed as the minimum-radius ℓ -clustering problem. See Figure 1. Both problems are known to be NP-hard to approximate within $2 - \epsilon$, where $\epsilon > 0$, in metric spaces [1,13].

González’s simple 2-approximation method for ℓ -center clustering also yields a 2-approximation for minimum-diameter ℓ -clustering [1]. It picks an arbitrary input point as the first center and repeatedly extends the current center set by selecting an input point that maximizes its distance from the current center set, until ℓ centers are found. In case of d -dimensional Euclidean or Hamming space, his method takes $O(nd\ell)$ time, where n is the number of input points. By forming for each of the ℓ centers, the cluster consisting of all input points for which this center is the closest one (with ties solved arbitrarily), one obtains an ℓ -clustering whose maximum cluster diameter is within a factor of two of the minimum one [1].

Recall that an approximation algorithm for the ℓ -center clustering problem is conservative if it always returns centers belonging to the input point set. In this section, we present a general randomized method for decreasing the dependence on d in the running time of a hypothetical conservative approximation algorithm for the ℓ -center clustering problem in the Euclidean \mathbb{R}^d , using the randomized dimension reduction given in Fact 1.

procedure DIMREDCENTER(ℓ, P, ϵ, SR)

Input: A positive integer ℓ , a set P of points $p_1, \dots, p_n \in \mathbb{R}^d, n > \ell$, a real $\epsilon \in (0, \frac{1}{2})$, and a conservative approximation subroutine SR for the m -center clustering problem in \mathbb{R}^q , where $m \leq \ell$ and $q \leq d$.

Output: A set T of ℓ centers of P .

1. Set n to the number of input points and k to $O(\log n / \epsilon^2)$.
2. Generate a random $d \times k$ matrix R with entries in $\{-1, 1\}$, defining the function $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ by $f(x) = \frac{1}{\sqrt{k}}xR$ (see Fact 1).
3. Compute the values of the function f for each point $p_i \in P$, i.e., for $i = 1, \dots, n$, compute $\frac{1}{\sqrt{k}}p_iR$. Also, for $i = 1, \dots, n$, if the value of f^{-1} is not yet defined on $f(p_i)$ then set it to p_i .
4. Set $\ell' = \min\{\ell, |f(P)|\}$ and compute a set T' of ℓ' centers of $f(P) = \{f(p_1), \dots, f(p_n)\}$ by running the subroutine SR for the ℓ' -center clustering problem on $f(P)$.
5. Set T to $\{f^{-1}(u') | u' \in T'\}$, if $\ell' < \ell$ then extend T by $\ell - \ell'$ arbitrary points in $P \setminus T$, and return it.

By Corollary 1, we immediately obtain the following lemma.

Lemma 1. Assume the notation from Fact 1. Let T be a set of ℓ centers of the input point set P . For any $\epsilon > 0$, the following inequalities hold w.h.p.:

$$(1 - \epsilon) \max_{v \in P} \min_{u \in T} \|v - u\|_2 \leq \max_{v \in P} \min_{u \in T} \|f(v) - f(u)\|_2,$$

$$\max_{v \in P} \min_{u \in T} \|f(v) - f(u)\|_2 \leq (1 + \epsilon) \max_{v \in P} \min_{u \in T} \|v - u\|_2.$$

Lemma 2. Assume the notation from Fact 1. For any $\epsilon \in (0, 1/2)$, w.h.p. for all $v, u \in P$, the following inequalities hold:

$$(1 - \epsilon)\|f(v) - f(u)\|_2 \leq \|v - u\|_2,$$

$$\|v - u\|_2 \leq (1 + 2\epsilon)\|f(v) - f(u)\|_2.$$

Proof. By the right-hand inequality in Corollary 1, we obtain for any $v, u \in P$, and $\epsilon \in (0, 1/2)$, $\frac{\|f(v)-f(u)\|_2}{1+\epsilon} \leq \|v - u\|_2$. It follows that $\|f(v) - f(u)\|_2 - \frac{\epsilon}{1+\epsilon}\|f(v) - f(u)\|_2 \leq \|v - u\|_2$. Consequently, we obtain the first inequality in this lemma.

Similarly, by the left-hand inequality in Corollary 1, we infer that for any $v, u \in P$, and $\epsilon \in (0, 1/2)$, $\|v - u\|_2 \leq \frac{\|f(v)-f(u)\|_2}{1-\epsilon}$. This yields $\|v - u\|_2 \leq \|f(v) - f(u)\|_2 + \frac{\epsilon}{1-\epsilon}\|f(v) - f(u)\|_2$. Since $\epsilon \in (0, 1/2)$, the second inequality in this lemma follows. \square

Lemma 2 immediately yields the following lemma.

Lemma 3. Assume the notation from Fact 1. Let $f(P) = \{f(p_1), \dots, f(p_n)\}$, $\ell' = \min\{\ell, |f(P)|\}$, and let T' be a set of ℓ' centers for the point set $f(P)$ in \mathbb{R}^k . For $v' \in f(P)$, let $f^{-1}(v') = p_q$, where $q = \min\{i | f(p_i) = v'\}$. For any $\epsilon \in (0, 1/2)$, the following inequalities hold w.h.p.:

$$(1 - \epsilon) \max_{v' \in f(P)} \min_{u' \in T'} \|v' - u'\|_2 \leq \max_{v \in P} \min_{u' \in T'} \|v - f^{-1}(u')\|_2,$$

$$\max_{v \in P} \min_{u' \in T'} \|v - f^{-1}(u')\|_2 \leq (1 + 2\epsilon) \max_{v' \in f(P)} \min_{u' \in T'} \|v' - u'\|_2.$$

Lemma 4. Assume the notation from Fact 1. If $\epsilon \in (0, 1/2)$ and the conservative subroutine SR provides an α -approximation, then w.h.p. DIMREDCENTER(ℓ, P, ϵ, SR) returns a $(1 + \epsilon)(1 + 2\epsilon)\alpha$ -approximate conservative solution to the ℓ -center clustering problem for P in the Euclidean \mathbb{R}^d .

Proof. Let T_1 be an optimal ℓ -center solution to the ℓ -center clustering problem for $P \subset \mathbb{R}^d$. Similarly, let T_2 be an optimal ℓ' -center solution to the ℓ' -center clustering problem for $f(P) \subset \mathbb{R}^k$. Next, let $r_1 = \max_{v \in P} \min_{u \in T_1} \|v - u\|_2$, $r_2 = \max_{v' \in f(P)} \min_{u' \in T_2} \|v' - u'\|_2$. By Lemma 1, we have $(1 - \epsilon)r_1 \leq r_2 \leq (1 + \epsilon)r_1$. The procedure first computes an α -approximate ℓ' -center solution T' to the ℓ' -center clustering problem for $f(P)$ in \mathbb{R}^k . It follows that $\max_{v' \in f(P)} \min_{u' \in T'} \|v' - u'\|_2 \leq (1 + \epsilon)\alpha r_1$. The procedure returns $T = \{f^{-1}(u') | u' \in T'\}$ extended by $\ell - \ell'$ arbitrary points in $P \setminus T$ as an approximate ℓ -center solution to the ℓ -center clustering problem for the input point set $P \subset \mathbb{R}^d$. The second inequality in Lemma 3 implies $\max_{v \in P} \min_{u' \in T'} \|v - f^{-1}(u')\|_2 \leq (1 + 2\epsilon) \max_{v' \in f(P)} \min_{u' \in T'} \|v' - u'\|_2$. We conclude that T yields a $(1 + 2\epsilon)(1 + \epsilon)\alpha$ approximation to the ℓ -center clustering problem for P . \square

Lemma 5. All steps of DIMREDCENTER(ℓ, P, ϵ, SR) except for Step 4 can be implemented in $O((nd \log n)/\epsilon^2)$ time.

Proof. Step 1 can be done in $O(nd)$ time. Step 2 takes $O(dk) = O((d \log n)/\epsilon^2)$ time. The preprocessing in Step 3 requires $O(ndk)$, i.e., $O((nd \log n)/\epsilon^2)$ time. Finally, Step 5 takes $O(nd)$ time. \square

4. Fast $(2 + \epsilon)$ -Approximation for ℓ -Center Clustering

In this section, we provide faster randomized $(2 + \epsilon)$ -approximation methods for the ℓ -center clustering and minimum-diameter ℓ -clustering problems in the Euclidean \mathbb{R}^d and

the Hamming space $\{0, 1\}^d$. These methods are faster than the previously known methods with approximation guarantee close to 2 when $d = \omega(\log n)$ and $\ell = \omega(\log n)$. Our method for the Euclidean \mathbb{R}^d is obtained by plugging González’s method as the subroutine in the procedure *DIMREDCENTER*.

Theorem 1. *Let P be a set of n points $p_1, \dots, p_n \in \mathbb{R}^d$, ℓ an integer smaller than n , and let $\epsilon \in (0, 1/2)$. The ℓ -center clustering problem for P admits a conservative randomized approximation algorithm that, with high probability, provides a $(2 + \epsilon)$ approximation of an optimal ℓ -center clustering of P and runs in time $O(n \log n(d + \ell)/\epsilon^2)$. A slight modification of the algorithm yields a $(2 + \epsilon)$ approximation to an ℓ -clustering of P with a minimum cluster diameter, also with high probability, and with the same asymptotic running time.*

Proof. To prove the first part, we run *DIMREDCENTER*($\ell, P, \epsilon/8, GO$), where *GO* denotes the conservative González’s 2-approximation algorithm for the ℓ -center clustering problem. By Lemma 4, this yields a $(1 + \epsilon/8)(1 + 2\epsilon/8)2 \leq (2 + \epsilon)$ approximation of an optimal solution. Since Step 4 takes $O(n\ell k) = O((n\ell \log n)/\epsilon^2)$ time, the entire procedure can be implemented in $O(n \log n(d + \ell)/\epsilon^2)$ time by Lemma 5.

To prove the second part, we extend *DIMREDCENTER*($\ell, P, \epsilon/8, GO$) slightly. Let T' be the set of ℓ' centers of $f(P)$ constructed by González’s algorithm in Step 4 of *DIMREDCENTER*($\ell, P, \epsilon/8, GO$). In Step 5, we additionally form the ℓ' clusters $P_1, \dots, P_{\ell'}$ by assigning each point $v \in P$ to the center in T' that is closest to $f(v)$. This extension also takes $O(n \log n(d + \ell)/\epsilon^2)$ time. Let s be a point in $f(P) \setminus T'$ that maximizes the distance from T' ; denote this distance by r' . Note that each point in $f(P)$ is within distance $\leq r'$ from its nearest center in T' , and crucially, any two points in $T' \cup \{s\}$ are at least r' apart due to the furthest-point traversal used by González’s algorithm. By Lemma 1, any two points in the set $f^{-1}(T') \cup \{f^{-1}(s)\}$ are at least $\frac{r'}{1+\epsilon/8}$ apart. Assume that $\ell' = \ell$. Then the latter set contains $\ell + 1$ points, so at least two of them have to lie in the same cluster in any ℓ -clustering of P . Consequently, the maximum diameter of a cluster in any ℓ -clustering of P is at least $\frac{r'}{1+\epsilon/8}$. On the other hand, the diameter of any cluster P_i is at most $2\frac{r'}{1-\epsilon/8}$ by Lemma 1. Consequently, the ratio between the maximum diameter in our ℓ -clustering $P_i, i \in [\ell]$, and that in a minimum diameter ℓ -clustering is at most $\frac{2(1+\epsilon/8)}{1-\epsilon/8} \leq 2 + \epsilon$.

To complete the proof note that we may assume, w.l.o.g., that $\ell = \ell'$ w.h.p. Indeed, we may assume w.l.o.g. that there are at least $\ell + 1$ non-overlapping points in P , since otherwise, the problem admits a trivial solution. Let t be the minimum distance between a pair of these $\ell + 1$ non-overlapping points. W.h.p., each pair of f -images of these $\ell + 1$ points is apart at least by $(1 - \epsilon)t$ by Corollary 1. We therefore conclude that $|f(P)| > \ell$ w.h.p. \square

Let us recall the definitions of the ℓ -center clustering and minimum-diameter ℓ -clustering problems in a Hamming space. The ℓ -center clustering problem in the Hamming space $\{0, 1\}^d$ is defined as follows: given a set P of n points in $\{0, 1\}^d$, find a set T of ℓ points in \mathbb{R}^d that minimizes $\max_{v \in P} \min_{u \in T} \text{ham}(v, u)$. The minimum-diameter ℓ -clustering problem in the Hamming space $\{0, 1\}^d$ is defined as follows: given a set P of n points in $\{0, 1\}^d$, find a partition of P into ℓ subsets P_1, P_2, \dots, P_ℓ that minimizes $\max_{i \in [\ell]} \max_{v, u \in P_i} \text{ham}(v, u)$.

To derive a result analogous to Theorem 1 for Hamming spaces, we cannot directly apply the general method for Euclidean spaces from Section 3. Instead, we use the following procedure.

procedure DIMREDHAMCENTER(ℓ, P, ϵ)

Input: A positive integer ℓ , a set P of points $p_1, \dots, p_n \in \{0, 1\}^d$, $n > \ell$, and a real $\epsilon \in (0, \frac{1}{2})$.

Output: A set $T \subset P$ of ℓ centers of P .

1. Set n to the number of input points and k to $O(\log n / \epsilon^2)$.
2. Generate a random $d \times k$ matrix R with entries in $\{-1, 1\}$, defining the function $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ by $f(x) = \frac{1}{\sqrt{k}}xR$ (see Fact 1).
3. Compute the values of the function f on each point $p_i \in P$, i.e., for $i \in [n]$, compute $\frac{1}{\sqrt{k}}p_iR$.
4. Set T to $\{p_1\}$, and for $j \in [n] \setminus \{1\}$, set W_{1j} to $(\|f(p_1) - f(p_j)\|_2)^2$ (see Corollary 2).
5. $\ell - 2$ times iterate the following three steps:
 - (a) Find $p_m \in P \setminus T$ that maximizes $\min_{p_q \in T} W_{qm}$ and extend T to $T \cup \{p_m\}$.
 - (b) For each $p_j \in P \setminus T$, set W_{mj} to $(\|f(p_m) - f(p_j)\|_2)^2$.
 - (c) For each $p_j \in P \setminus T$, update $\min_{p_i \in T} W_{ij}$.
6. Find $p_m \in P \setminus T$ that maximizes $\min_{p_q \in T} W_{qm}$ and extend T to $T \cup \{p_m\}$.
7. Return T .

By the specification of W_{ij} in DIMREDHAMCENTER(ℓ, P, ϵ) and Corollary 2, we immediately obtain the following lemma.

Lemma 6. Assume the notation from DIMREDHAMCENTER(ℓ, P, ϵ). For $i \in [n]$ and $j \in [n]$, the following inequalities hold w.h.p.:

$$W_{ij} \leq (1 + \epsilon)\text{ham}(p_i, o_j),$$

$$(1 - \epsilon)\text{ham}(p_i, p_j) \leq W_{ij}.$$

Lemma 7. DIMREDHAMCENTER(ℓ, P, ϵ) runs in time $O(n \log n(d + \ell) / \epsilon^2)$.

Proof. Step 1 can be done in $O(nd)$ time. Step 2 takes $O(dk) = O((d \log n) / \epsilon^2)$ time. The preprocessing in Step 3 requires $O(ndk)$, i.e., $O((nd \log n) / \epsilon^2)$ time. Step 4 can be done in $O((n \log n) / \epsilon^2)$ time. Steps 5(a) and 5(c) take $O(n)$ time while Step 5(b) as Step 4 can be done in $O((n \log n) / \epsilon^2)$ time. Consequently, the whole Step 5 requires $O((n \ell \log n) / \epsilon^2)$ time. Finally, Step 6 takes $O((n \log n) / \epsilon^2)$ time similarly as Steps 4 and 5(b). It remains to observe that the overall running time of DIMREDHAMCENTER(ℓ, P, ϵ) is dominated by those of Step 3 and Step 5. \square

Theorem 2. Let P be a set of n points $p_1, \dots, p_n \in \{0, 1\}^d$, ℓ an integer smaller than n , and let $\epsilon \in (0, 1/2)$. DIMREDHAMCENTER($\ell, P, \epsilon/5$), with high probability, provides a $(2 + \epsilon)$ -approximation of an optimal ℓ -center clustering of P in the Hamming space $\{0, 1\}^d$ in time $O(n \log n(d + \ell) / \epsilon^2)$. Its slight modification yields a $(2 + \epsilon)$ approximation to an ℓ -clustering of P with a minimum cluster diameter w.h.p.

Proof. Let T be the set of ℓ centers output by DIMREDHAMCENTER($\ell, P, \epsilon/5$). Next, let $r = \max_{v \in P} \min_{u \in T} \text{ham}(v, u)$, $r_w = \max_{p_i \in P} \min_{p_j \in T} W_{ij}$, and let p_q be a point for which the latter maximum is achieved. It follows from Lemma 6 that $r_w \geq r(1 - \epsilon)$ w.h.p. By the specification of the procedure DIMREDHAMCENTER and the definition of p_q , the set $T \cup \{p_q\}$ consists of $\ell + 1$ points such that for any pair p_v, p_u of points in this set, we

have $W_{vu} \geq r_w$. Consequently, w.h.p. these $\ell + 1$ points are at the Hamming distance at least $r_w / (1 + \epsilon/5)$ from each other by Lemma 6. Let T^* be an optimal set of ℓ centers of P . Two of these $\ell + 1$ points in $T \cup \{p_q\}$ must share the same nearest center in T^* . It follows that the Hamming distance from at least one of these two points to its nearest center in T^* is at least $\frac{r_w}{2(1+\epsilon/5)}$ w.h.p. Since $r_w \geq r(1 - \epsilon/5)$ w.h.p. by Lemma 6, we infer that $\max_{p_i \in P} \min_{p_j \in T^*} \text{ham}(p_i, p_j)$ is at least $r \frac{1-\epsilon/5}{2(1+\epsilon/5)}$ w.h.p. Hence, the ratio between the maximum distance from an input point to a center in T and that to a center in T^* is at most $2 \frac{1+\epsilon/5}{1-\epsilon/5} \leq 2 + \epsilon$ by $\epsilon \leq 1/2$. Together with Lemma 7, this completes the proof of the first part.

To prove the second part, we can slightly modify $\text{DIMREDHAMCENTER}(\ell, P, \epsilon/5)$ so that it returns a partition of P into clusters $P_i, i \in [\ell]$, where P_i consists of all points in P whose closest center (in terms of the approximate distances W_{ij}) is the i -th center. To implement the modification, in Step 5(c), we update not only $\min_{p_i \in T} W_{ij}$ but also the identity of the current center $p_i \in T$ that minimizes W_{ij} . This slight modification does not affect the asymptotic running time of $\text{DIMREDHAMCENTER}(\ell, P, \epsilon/5)$.

The proof of the $2 + \epsilon$ approximation guarantee in the second part proceeds by arguments analogous to those seen in the first part. Consider the $\ell + 1$ points identified in the proof of the first part. Recall that, w.h.p., they pairwise are at least $r_w / (1 + \epsilon/5)$ apart. Two of the $\ell + 1$ points must share the same cluster in any ℓ -clustering, in particular, an ℓ -clustering that minimizes the diameter. Hence, the minimum possible diameter is at least $r_w / (1 + \epsilon/5)$ w.h.p. On the other hand, the diameter of the clusters $P_i, i \in [\ell]$, induced by the ℓ centers returned by $\text{DIMREDHAMCENTER}(\ell, P, \epsilon/5)$ is at most $2r_w / (1 - \epsilon/5)$ w.h.p. by Lemma 6. Consequently, the ratio between the maximum diameter in our ℓ -clustering $P_i, i \in [\ell]$, and that in a minimum-diameter ℓ -clustering is at most $\frac{2r_w / (1-\epsilon/5)}{r_w / (1+\epsilon/5)} \leq 2 + \epsilon$ w.h.p., where the final inequality holds for $\epsilon < \frac{1}{2}$. This completes the proof of the second part. \square

5. Faster $O(\alpha)$ -Approximations for ℓ -Center Clustering

In this section, we shall plug recent fast approximation methods for the ℓ -center clustering problem in Euclidean spaces into the procedure DIMREDCENTER in order to decrease the heavy dependence of their running times on the dimension, while worsening their approximation guarantees only minimally. The following fact is a recent refinement of an earlier result from [4].

Fact 2 ([4,5]). For $\alpha \geq 1$, the ℓ -center clustering problem for a set P of n points in the Euclidean \mathbb{R}^d admits an $O(\alpha)$ -approximation in time $\text{poly}(d \log n)(n + \ell^{1+1/\alpha^2} n^{O(1/\alpha^{2/3})})$, or alternatively, in time $\text{poly}(d \log n)n^{\ell^{1/\alpha^2}}$.

Importantly, we may assume w.l.o.g. that the $O(\alpha)$ -approximation method in Fact 2 is conservative. Otherwise, we can simply replace the current centers with closest input points, which increases the distance from any input point to its nearest center by at most a factor of two. By setting the subroutine SR in $\text{DIMREDCENTER}(\ell, P, \epsilon, SR)$ to the method of Fact 2 and using Lemmata 4 and 5 with $\epsilon = \Omega(1)$, we can substantially reduce the dependence of the running time on d .

Theorem 3. For $\alpha \geq 1$, the ℓ -center clustering problem for a set P of n points in the Euclidean \mathbb{R}^d admits, with high probability, an $O(\alpha)$ -approximation in time $\tilde{O}(nd/\epsilon^2 + \ell^{1+1/\alpha^2} n^{O(1/\alpha^{2/3})})$, or alternatively, in time $\tilde{O}(nd/\epsilon^2 + n\ell^{1/\alpha^2})$.

6. Fast $O(1)$ -Approximation for ℓ -Center Clustering with Outliers

In the variants of the ℓ -center clustering and minimum-diameter ℓ -clustering problems with outliers, a given number z of input points may be discarded as outliers when attempting to minimize the maximum distance to the nearest center or the maximum cluster diameter [10,14]. Charikar et al. were the first to provide a polynomial-time $O(1)$ -approximation to the ℓ -center clustering problem with outliers [10]. They used a greedy method.

Fact 3 ([10]). *Given a set P of n points from an arbitrary metric, an integer $\ell \leq n$, and an integer z , there is a polynomial-time 3-approximation algorithm for the ℓ -center clustering problem with z outliers in that metric.*

In the proof of Fact 3, the authors assume that the optimal solution considered is conservative (i.e., the centers are input points), but they claim that this assumption can be removed in case of a Euclidean space while preserving the approximation guarantee.

The general method of dimension reduction given in Section 3 cannot be applied directly to the ℓ -center clustering problem with outliers. For this reason, we just modify the original greedy method from [10], using the approximate interdistances based on the reduction instead of the true distances. In this way, we can significantly speed-up the greedy method at the cost of slightly increasing the approximation guarantee. In particular, we can replace an $O(n^2d)$ component of the time complexity of the method with $\tilde{O}(n^2)$.

procedure DIMREDCENOUT(ℓ, P, ϵ, z)

Input: A positive integers ℓ, z , a set P of points $p_1, \dots, p_n \in \mathbb{R}^d$, where $n > \ell, z$, a real $\epsilon \in (0, \frac{1}{2})$.

Output: A set T of ℓ centers for a subset of P of cardinality $n - z$.

1. Set n to the number of input points and k to $O(\log n/\epsilon^2)$.
2. Generate a random $d \times k$ matrix R with entries in $\{-1, 1\}$, defining the function $f: \mathbb{R}^d \rightarrow \mathbb{R}^k$ by $f(x) = \frac{1}{\sqrt{k}}xR$ (see Fact 1).
3. Compute the values of the function f for each point $p_i \in P$, i.e., for $i = 1, \dots, n$, compute $\frac{1}{\sqrt{k}}p_iR$.
4. For $i, j \in [n]$, compute $W'_{ij} = \|f(p_i) - f(p_j)\|_2$ and set W' to the matrix (W'_{ij}) .
5. (a) Compute and sort the set $B = \{W'_{ij} | i, j \in [n]\} \cup \{W'_{ij}(1 + 2\epsilon) | i, j \in [n]\}$.
(b) By binary search, find the smallest r in the sorted B such that GREEDY(ℓ, W', ϵ, z, r) returns YES.
6. Set T to the set of centers returned by the successful call of GREEDY with the smallest r .

procedure *GREEDY*(ℓ, W', ϵ, z, r)
Input: The input parameters $\ell, P, \epsilon, z, W' = (W'_{ij})$ are specified as in *DIMREDCENOUT*(ℓ, P, ϵ, z); r is a positive real number.
Output: YES if there is an ℓ -center clustering of a $(n - z)$ -point subset of P such that the maximum ℓ_2 distance from a point in the subset to its nearest center does not exceed $3(1 + \epsilon)r$ otherwise NO.

1. For $i \in [n]$, compute the set G_i of points $p_j \in P$ s.t. $W'_{ij} \leq r(1 + \epsilon)$ and the set E_i of points $p_j \in P$ s.t. $W'_{ij} \leq 3r(1 + \epsilon)$. Set S to $[n]$.
2. ℓ times iterate the following block:
 - (a) Select a set G_j of the largest cardinality among (the not yet selected) sets $G_i, i \in S$. Select also E_j . Set S to $S \setminus \{j\}$.
 - (b) Mark the points in E_j and remove the newly marked points from all (the not yet selected) sets G_i, E_i , for $i \in S$.
3. If the total number of marked points is at least $n - z$, i.e., $|\bigcup_{i=1}^{\ell} E_i| \geq n - z$, then return YES along with $E_i, i \in [\ell]$ else output NO.

Lemma 8. *If there is a set T of ℓ centers in $P \subset \mathbb{R}^d$ such that for at least $n - z$ points $p \in P$, $\min_{t \in T} \|t - p\|_2 \leq r$ then *GREEDY*($\ell, P, W', \epsilon, z, r$) returns YES w.h.p.*

Proof. We may assume w.l.o.g. that p_1, \dots, p_ℓ are the centers in the sets G_i , and their extensions $E_i, i \in [\ell]$, produced by *GREEDY*($\ell, P, W', \epsilon, z, r$). Let $T = \{t_1, \dots, t_\ell\}$ and for $i \in [\ell]$, let $T_i = \{p \in P \mid \|t_i - p\|_2 \leq r\}$. Following [10], to prove that the lemma is sufficient to show induction on $i \in [\ell]$, one can order the sets T_i so that $E_1 \cup E_2 \dots E_i$ cover at least as many points in P as $T_1 \cup T_2 \dots T_i$ w.h.p. The proof is obtained by assigning to each point in the latter set union a distinct point in the former set union.

By the inductive hypothesis, we may assume that the sets T_1, T_2, \dots, T_{i-1} and the assignment of a distinct point in $E_1 \cup E_2 \dots E_{i-1}$ to each point in $T_1 \cup T_2 \dots T_{i-1}$ have been determined. Suppose that the set G_i intersects some remaining set T_j . Then, E_i covers all points in T_j not covered by $E_1 \cup E_2 \dots E_{i-1}$ w.h.p. by Corollary 1. Therefore, we can assign to each point in T_j not covered by $E_1 \cup E_2 \dots E_{i-1}$ the point itself. Otherwise, by the greedy choice of G_i and Corollary 1, G_i and consequently also E_i contain at least as many points outside $E_1 \cup E_2 \dots E_{i-1}$ as T_j . Therefore, we can assign to each point in T_j not covered by $E_1 \cup E_2 \dots E_{i-1}$ a distinct point in E_i . Consequently, we can further rearrange the order of the remaining sets $T_q, i \leq q \leq \ell$, so T_j becomes T_i . Importantly, no point can be doubly assigned since in the i -th iteration of the *GREEDY* procedure, the updated sets E_i are disjoint from the sets $E_q, q < i$. □

Lemma 9. *Assume the notation from Lemma 8. *GREEDY*($\ell, P, W', \epsilon, z, r$) can be implemented in $\tilde{O}(n^2\ell)$ time.*

Proof. In Step 1, we compute representations of the sets G_i, E_i in the form of dictionaries keeping the indices of the points belonging to these sets. The dictionaries can be formed in $\tilde{O}(n^2)$ time. A single iteration of the block in Step 2 takes $\tilde{O}(n^2)$ time. Hence, the whole Step 2 takes $\tilde{O}(n^2\ell)$ time. Finally, Step 3 can be done in $O(n)$ time. □

Lemma 10. **DIMREDCENOUT*(ℓ, P, ϵ, z) can be implemented in time $\tilde{O}(n^2(\epsilon^{-2} + \ell))$.*

Proof. Steps 1–3 are analogous to those in *DIMREDCENTER*(ℓ, P, ϵ, SR), and hence, they can be implemented in time $O((nd \log n)/\epsilon^2)$ by Lemma 5. Step 4 takes $O(n^2k) =$

$O((n^2 \log n) / \epsilon^2)$ time. Step 5(a) requires $\tilde{O}(n^2)$ time. By Lemma 9, Step 5(b) can be done in $\tilde{O}(n^2 \ell)$ time. Finally, Step 6 can be completed in $O(nd)$ time. \square

Theorem 4. *Let P be a set of n points in the Euclidean \mathbb{R}^d , and let z be a nonnegative integer smaller than n . Suppose that there is a set $T \subset P$ of ℓ centers that is a solution to the ℓ -center clustering problem for P with z outliers in the Euclidean space, where the maximum distance from a non-outlier to its closest center in T is at most r . Let $\epsilon \in (0, \frac{1}{2})$. In time $\tilde{O}(n^2(\epsilon^{-2} + \ell))$, one can construct a conservative solution to the ℓ -center clustering problem for P with z outliers in the Euclidean space such that the maximum distance from a non-outlier to its closest center is at most $(3 + \epsilon)r$ w.h.p.*

Proof. Let us run $DIMREDCENOUT(\ell, P, \delta, z)$, where δ is a fraction of ϵ to be specified later. Note that $r = \|p_i - p_j\|_2$ for some $i, j \in [n]$. Recall also that both W'_{ij} and $W'_{ij}(1 + 2\delta)$ are considered in the binary search in Step 5 of $DIMREDCENOUT(\ell, P, \delta, z)$. If $W'_{ij} \leq r$ then $r \leq (1 + 2\delta)W'_{ij} \leq (1 + 2\delta)r$ by Lemma 2. Otherwise, we have $r \leq W'_{ij} \leq (1 + \delta)r$ by Corollary 1. We infer that in the binary search, a value r' between r and $(1 + 2\delta)r$ is considered. It follows from Lemma 8 that $DIMREDCENOUT(\ell, P, \delta, z)$ produces a conservative solution, where the maximum distance from a non-outlier to its closest center is at most $(3 + \delta)(1 + 2\delta)r$. This is at most $(3 + \epsilon)r$ when δ is set to $\frac{\epsilon}{8}$.

$DIMREDCENOUT(\ell, P, \epsilon/8, z)$ can be implemented in time $\tilde{O}(n^2(\epsilon^{-2} + \ell))$ by Lemma 10. \square

Since a solution to an instance of the ℓ -center clustering problem with outliers can be easily transformed to a conservative one by at most doubling the distances from non-outlier points to their closest centers, we obtain the following corollary.

Corollary 3. *Let P be a set of n points in the Euclidean \mathbb{R}^d , and let z be a nonnegative integer smaller than n . For an arbitrary $\epsilon \in (0, \frac{1}{2})$, the ℓ -center clustering problem for P with z outliers in the Euclidean space admits, w.h.p., a $(6 + \epsilon)$ -approximation in time $\tilde{O}(n^2(\epsilon^{-2} + \ell))$.*

By slightly modifying the procedures $DIMREDCENOUT$ and $GREEDY$, we can obtain analogous theorem and corollary for ℓ -center clustering with outliers in Hamming spaces. Simply, in the modified procedures, we use as the approximate distance between points v and u the squared distance $\|f(v) - f(u)\|_2$ instead of $\|f(v) - f(u)\|_2$. The asymptotic running times of the modified procedures are the same as those of the original ones. The proof of the approximation guarantee is analogous to that in the Euclidean case. Instead of Corollary 1, we use Corollary 2 and instead of Lemma 2, its Hamming equivalent. See Appendix A for details.

Theorem 5. *Let P be a set of n points in the Hamming space $\{0, 1\}^d$, and let z be a nonnegative integer smaller than n . Suppose that there is a set $T \subset P$ of ℓ centers that is a solution to the ℓ -center clustering problem for P with z outliers in the Hamming space $\{0, 1\}^d$, where the maximum Hamming distance from a non-outlier to its closest center in T is at most r . Let $\epsilon \in (0, \frac{1}{2})$. In time $\tilde{O}(n^2(\epsilon^{-2} + \ell))$, one can construct a conservative solution to the ℓ -center clustering problem for P with z outliers in the Hamming space $\{0, 1\}^d$ such that the maximum Hamming distance from a non-outlier to its closest center is at most $(3 + \epsilon)r$ w.h.p.*

Corollary 4. *Let P be a set of n points in the Hamming space $\{0, 1\}^d$, and let z be a nonnegative integer smaller than n . For an arbitrary $\epsilon \in (0, \frac{1}{2})$, the ℓ -center clustering problem for P with z outliers in the Hamming space $\{0, 1\}^d$ admits, w.h.p., a $(6 + \epsilon)$ -approximation in time $\tilde{O}(n^2(\epsilon^{-2} + \ell))$.*

7. Final Remarks

Relatively recently, Jiang et al. [6] showed that for some problems, including the ℓ -center clustering problem in Euclidean spaces, there is a randomized dimension reduction even to a sublogarithmic-size subspace. Such a reduction preserves the optimal value of the solution within a factor related to the size of the subspace. Following [6], their main result can be stated informally as follows.

Fact 4 ([6]). For every d, α, ℓ, n , where $\ell \leq n$, there is a random linear map $g : \mathbb{R}^d \rightarrow \mathbb{R}^t$, where $t = O(\frac{\log n}{\alpha^2} + \log \ell)$, such that for every set $P \subset \mathbb{R}^d$ of n points, with high probability, g preserves the value of an optimal solution to the ℓ -center clustering problem within the $O(\alpha)$ factor.

We could replace the variant of JL randomized dimension reduction from [11] in the general method presented in Section 3 with the enhanced randomized dimension reduction from [6]. Roughly speaking, this would decrease the asymptotic running time by $O(\alpha^2)$ at the cost of increasing the approximation guarantee by $O(\alpha)$. As the authors of [6] note, when one is interested in $O(1)$ -approximation and hence $\alpha = O(1)$, then $\frac{\log n}{\alpha^2} = \Omega(\log n)$, and their method does not yield any improvement in the asymptotic running time. On the other hand, the randomized dimension reduction from [6] can be advantageous when, for instance, algorithms of exponential dependence on the dimension are applied in the target subspace.

Author Contributions: All authors contributed to various parts of the manuscript in ideas, proofs, and writing, and reviewed it before submission. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. $O(1)$ -Approximation for ℓ -Center Clustering with Outliers in Hamming Spaces

Our Hamming versions of the procedures *DIMREDCENOUT* and *GREEDY* are as follows.

procedure *HAMDIMREDCENOUT*(ℓ, P, ϵ, z)
Input: Positive integers ℓ, z , a set P of points $p_1, \dots, p_n \in \{0, 1\}^d$, where $n > \ell, z$, a real $\epsilon \in (0, \frac{1}{2})$.
Output: A set T of ℓ centers for a subset of P of cardinality $n - z$.

1. Set n to the number of input points and k to $O(\log n / \epsilon^2)$.
2. Generate a random $d \times k$ matrix R with entries in $\{-1, 1\}$, defining the function $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ by $f(x) = \frac{1}{\sqrt{k}}xR$ (see Fact 1).
3. Compute the values of the function f for each point $p_i \in P$, i.e., for $i = 1, \dots, n$, compute $\frac{1}{\sqrt{k}}p_iR$.
4. For $i, j \in [n]$, compute $W''_{ij} = (\|f(p_i) - f(p_j)\|_2)^2$ and set W'' to the matrix (W''_{ij}) .
5. (a) Compute and sort the set $B' = \{W''_{ij} | i, j \in [n]\} \cup \{W''_{ij}(1 + 2\epsilon) | i, j \in [n]\}$.
 (b) By binary search find the smallest r in the sorted B' such that *HAMGREEDY*($\ell, W'', \epsilon, z, r$) returns *YES*.
6. Set T to the set of centers returned by the successful call of *HAMGREEDY* with the smallest r .

procedure *HAMGREEDY*($\ell, W'', \epsilon, z, r$)
Input: The input parameters $\ell, P, \epsilon, z, W'' = (W''_{ij})$ are specified as in *HAMDIMREDCENOUT*(ℓ, P, ϵ, z); r is a positive real number.
Output: YES if there is an ℓ -center clustering of an $(n - z)$ -point subset of P such that the maximum Hamming distance from a point in the subset to its nearest center does not exceed $3(1 + \epsilon)r$ otherwise NO.

1. For $i \in [n]$, compute the set G'_i of points $p_j \in P$ s.t. $W''_{ij} \leq r(1 + \epsilon)$ and the set E'_i of points $p_j \in P$ s.t. $W''_{ij} \leq 3r(1 + \epsilon)$. Set S to $[n]$.
2. ℓ times iterate the following block:
 - (a) Select a set G'_j of the largest cardinality among (the not yet selected) sets $G'_i, i \in S$. Select also E'_j . Set S to $S \setminus \{j\}$.
 - (b) Mark the points in E'_j and remove the newly marked points from all (the not yet selected) sets G'_i, E'_i , for $i \in S$.
3. If the total number of marked points is at least $n - z$, i.e., $|\bigcup_{i=1}^{\ell} E'_i| \geq n - z$, then return YES along with $E'_i, i \in [\ell]$ else output NO.

Lemma A1. *If there is a set T of ℓ centers in $P \subset \{0, 1\}^d$, such that for at least $n - z$ points $p \in P$, $\min_{t \in T} \text{ham}(t, p) \leq r$ then *HAMGREEDY*($\ell, P, W', \epsilon, z, r$) returns YES w.h.p.*

Proof. We may assume w.l.o.g. that p_1, \dots, p_ℓ are the centers in the sets G'_i , and their extensions $E'_i, i \in [\ell]$, produced by *HAMGREEDY*($\ell, P, W'', \epsilon, z, r$). Let $T = \{t_1, \dots, t_\ell\}$ and for $i \in [\ell]$, let $T_i = \{p \in P | \text{ham}(t, p_i) \leq r\}$. As in the Euclidean case, to prove that the lemma is sufficient to show by induction on $i \in [\ell]$, that one can order the sets T_i so that $E'_1 \cup E'_2 \dots E'_i$ cover at least as many points in P as $T_1 \cup T_2 \dots T_i$ w.h.p. The proof is obtained by assigning to each point in the latter set union a distinct point in the former set union.

By the inductive hypothesis, we may assume that the sets T_1, T_2, \dots, T_{i-1} and the assignment of a distinct point in $E'_1 \cup E'_2 \dots E'_{i-1}$ to each point in $T_1 \cup T_2 \dots T_{i-1}$ have been determined. Suppose that the set G'_i intersects some remaining set T_j . Then, E'_i covers all points in T_j not covered by $E'_1 \cup E'_2 \dots E'_{i-1}$ w.h.p. by Corollary 2. Therefore, we can assign to each point in T_j not covered by $E'_1 \cup E'_2 \dots E'_{i-1}$ the point itself. Otherwise, by the greedy choice of G'_i and Corollary 2, G'_i and consequently also E'_i contain at least as many points outside $E'_1 \cup E'_2 \dots E'_{i-1}$ as T_j . Therefore, we can assign to each point in T_j not covered by $E'_1 \cup E'_2 \dots E'_{i-1}$ a distinct point in E'_i . Hence, we can further rearrange the order of the remaining sets $T_q, i \leq q \leq \ell$, so T_j becomes T_i . Importantly, no point can be doubly assigned since in the i -th iteration of the *HAMGREEDY* procedure, the updated sets E'_i are disjoint from the sets $E'_q, q < i$. □

Lemma A2. *Assume the notation from Lemma A1. *HAMGREEDY*($\ell, P, W'', \epsilon, z, r$) can be implemented in $\tilde{O}(n^2\ell)$ time.*

Proof. is analogous to that of Lemma 9. □

Lemma A3. **DIMREDCENOUT*(ℓ, P, ϵ, z) can be implemented in time $\tilde{O}(n^2(\epsilon^{-2} + \ell))$.*

Proof. is analogous to that of Lemma 10. □

We also need a Hamming equivalent of Lemma 2.

Lemma A4. Assume the notation from Fact 1. Suppose that $\epsilon \in (0, 1/2)$ and $P \subset \{0, 1\}^d \subset \mathbb{R}^d$. Then, for all $v, u \in P$, the following inequalities hold w.h.p.:

$$(1 - \epsilon)(\|f(v) - f(u)\|_2)^2 \leq \text{ham}(v, u),$$

$$\text{ham}(v, u) \leq (1 + 2\epsilon)(\|f(v) - f(u)\|_2)^2.$$

Proof. By the right-hand inequality in Corollary 2, we obtain for any $v, u \in P$, and $\epsilon \in (0, 1/2)$, $\frac{(\|f(v) - f(u)\|_2)^2}{1 + \epsilon} \leq \text{ham}(v, u)$. It follows that $(\|f(v) - f(u)\|_2)^2 - \frac{\epsilon}{1 + \epsilon}(\|f(v) - f(u)\|_2)^2 \leq \text{ham}(v, u)$. Consequently, we obtain the first inequality in this lemma.

Similarly, by the left-hand inequality in Corollary 2, we infer that for any $v, u \in P$, and $\epsilon \in (0, 1/2)$, $\text{ham}(v, u) \leq \frac{(\|f(v) - f(u)\|_2)^2}{1 - \epsilon}$. This yields $\text{ham}(v, u) \leq (\|f(v) - f(u)\|_2)^2 + \frac{\epsilon}{1 - \epsilon}(\|f(v) - f(u)\|_2)^2$. Since $\epsilon \in (0, 1/2)$, the second inequality in this lemma follows. \square

Proof of Theorem 5. Let us run $HAMDIMREDCENOUT(\ell, P, \delta, z)$, where δ is set to $\frac{\epsilon}{8}$. Note that $r = \text{ham}(p_i, p_j) = (\|p_i - p_j\|_2)^2$ for some $i, j \in [n]$. Recall that W''_{ij} and $W''_{ij}(1 + 2\delta)$ are considered in the binary search in Step 5 of $HAMDIMREDCENOUT(\ell, P, \delta, z)$. If $W''_{ij} \leq r$ then $r \leq (1 + 2\delta)W''_{ij} \leq (1 + 2\delta)r$ by Lemma A4. Otherwise, we have $r \leq W''_{ij} \leq (1 + \delta)r$ by Corollary 2. We infer that in the binary search, a value r' between r and $(1 + 2\delta)r$ is considered. It follows from Lemma A1 that $HAMDIMREDCENOUT(\ell, P, \delta, z)$ produces a conservative solution, where the maximum distance from a non-outlier to its closest center is at most $(3 + \delta)(1 + 2\delta)r$. This is at most $(3 + \epsilon)r$ by $\delta = \frac{\epsilon}{8}$.

$HAMDIMREDCENOUT(\ell, P, \epsilon/8, z)$ can be implemented in time $\tilde{O}(n^2(\epsilon^{-2} + \ell))$ by Lemma A3. \square

References

- González, T. Clustering to minimize the maximum intercluster distance. *Theor. Comput. Sci.* **1985**, *38*, 293–306. [CrossRef]
- Feder, T.; Greene, D. Optimal algorithms for approximate clustering. In Proceedings of the ACM Symposium on Theory of Computing (STOC 1988), Chicago, IL, USA, 2–4 May 1988; pp. 434–444.
- Har-Peled, S.; Mendel, M. Fast construction of nets in low-dimensional metrics and their applications. *SIAM J. Comput.* **2006**, *35*, 1148–1184. [CrossRef]
- Eppstein, D.; Har-Peled, S.; Sidiropoulos, A. Approximate greedy clustering and distance selection for graph metrics. *J. Comput. Geom.* **2020**, *11*, 629–652.
- Filtser, A.; Jiang, S.; Li, Y.; Naredla, A.; Psarros, I.; Yang, Q.; Zhang, Q. Faster Approximation Algorithms for k -Center via Data Reduction. *arXiv* **2025**. [CrossRef]
- Jiang, S.; Krauthgamer, R.; Sapir, S. Moderate Dimension Reduction for k -Center Clustering. In Proceedings of the International Symposium on Computational Geometry (SoCG 2024), Athens, Greece, 11–14 June 2024; LIPIcs; Volume 293, pp. 64:1–64:16.
- Ebbens, M.; Funk, N.; Höockendorff, J.; Sohler, C.; Weil, V. A Subquadratic Time Approximation Algorithm for Individually Fair k -Center. 2024. In Proceedings of the 28th International Conference on Artificial Intelligence and Statistics, Phuket, Thailand, 3–5 May 2025; Volume 258, pp. 2287–2295.
- Arslan, A.N.; Chidri, A. A Clustering-Based Matrix Multiplication Algorithm. In Proceedings of the 2011 International Conference on Scientific Computing (CSC 2011), Las Vegas, NV, USA, 18–21 July 2011; pp. 303–307.
- Jansson, J.; Kowaluk, M.; Lingas, A.; Persson, M. Multiplication of 0-1 matrices via clustering. In *Proceedings of the Frontiers of Algorithmics—The 19th International Joint Conference (IJTCS-FAW 2025)*; Lecture Notes in Computer Science; Springer Nature: Berlin/Heidelberg, Germany, 2025; Volume 15828, pp. 92–102.
- Charikar, M.; Khuller, S.; Mount, D.; Narasimhan, G. Algorithms for facility location problems with outliers. In Proceedings of the 12th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2001), Washington, DC, USA, 7–9 January 2001; pp. 642–651.
- Achlioptas, D. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.* **2003**, *66*, 671–687. [CrossRef]
- Johnson, W.; Lindenstrauss, J. Extensions of Lipschitz mappings into a Hilbert space. In *Proceedings of the Conference in Modern Analysis and Probability*; Contemporary Mathematics; American Mathematical Society: Providence, RI, USA, 1984; Volume 26, pp. 189–206.

13. Hochbaum, D.; Shmoys, D. A Best Possible Heuristic for the k-Center Problem. *Math. Oper. Res.* **1985**, *10*, 180–184. [[CrossRef](#)]
14. Harris, D.; Pensyl, T.; Srinivasan, A.; Trinh, K. A Lottery Model for Center-type Problems with Outliers. In Proceedings of the International Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX 2017), Berkeley, CA, 16–18 August 2017; LIPIcs; Volume 81, pp. 10:1–10:19.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.