Space Complexity of Euclidean Clustering

Xiaoyi Zhu
School of Data Science, Fudan University, Shanghai, China

Yuxiang Tian
School of Data Science, Fudan University, Shanghai, China

Lingxiao Huang[1]
State Key Laboratory of Novel Software Technology, Nanjing University, China

Zengfeng Huang[1]
School of Data Science, Fudan University, Shanghai, China

Abstract

The (k, z)-Clustering problem in Euclidean space $\mathbb{R}^d$ has been extensively studied. Given the scale of data involved, compression methods for the Euclidean (k, z)-Clustering problem, such as data compression and dimension reduction, have received significant attention in the literature. However, the space complexity of the clustering problem, specifically, the number of bits required to compress the cost function within a multiplicative error ε, remains unclear in existing literature.

This paper initiates the study of space complexity for Euclidean (k, z)-Clustering and offers both upper and lower bounds. Our space bounds are nearly tight when k is constant, indicating that storing a coreset, a well-known data compression approach, serves as the optimal compression scheme. Furthermore, our lower bound result for (k, z)-Clustering establishes a tight space bound of $\Theta(nd)$ for terminal embedding, where n represents the dataset size. Our technical approach leverages new geometric insights for principal angles and discrepancy methods, which may hold independent interest.

2012 ACM Subject Classification Theory of computation → Computational geometry; Theory of computation → Facility location and clustering; Theory of computation → Data compression

Keywords and phrases Space complexity, Euclidean clustering, coreset, terminal embedding

Digital Object Identifier 10.4230/LIPIcs.SoCG.2024.82

Related Version *Full Version*: `https://arxiv.org/abs/2403.02971`

Funding This work is supported by National Natural Science Foundation of China under Grant 62276066 and U2241212.

1 Introduction

Clustering problems are fundamental in theoretical computer science and machine learning with various applications [3, 13, 36]. An important class of clustering is called Euclidean (k, z)-Clustering where, given a dataset $P \subset \mathbb{R}^d$ of n points and a $k \geq 1$, the goal is to find a center set $C \subset \mathbb{R}^d$ of k points that minimizes the cost $\mathrm{cost}_z(P, C) := \sum_{p \in P} d(p, C)^z$. Here, $d^z(p, C) = \min_{c \in C} d^z(p, c)$ is the z-th power Euclidean distance of p to C. Well-known examples of (k, z)-Clustering include k-Median (when $z = 1$) and k-Means (when $z = 2$).

In many real-world scenarios, the dataset P is large and the dimension d is high, and it is desirable to compress P to reduce storage and computational requirements in order to solve the underlying clustering problem efficiently. Previous studies have proposed two approaches: data compression and dimension reduction. On one hand, coresets have been proposed as a solution to data compression [24] – a coreset is a small representative

[1] Corresponding Authors

subset S that approximately preserves the clustering cost for all possible center sets. Recent research has focused on developing efficient coresets [40, 16, 14, 15, 28], showing that the coreset size remains independent of both the size n of dataset and the dimension d. On the other hand, dimension reduction methods have also proven to be effective for (k,z)-Clustering, including techniques like Johnson-Lindenstrauss (JL) [37, 10] and terminal embedding [39, 29]. Specifically, terminal embedding (Definition 4), which projects a dataset P to a low-dimensional space while approximately preserving all pairwise distances between P and $\mathbb{R}^d$, is the key for removing the size dependence on d for coreset [29, 16].

While the importance of compression for clustering has been widely acknowledged, the literature currently lacks clarity regarding the space complexity of the clustering problem itself. Specifically, one may want to know how many bits are required to compress the cost function. Space complexity, a fundamental factor in theoretical computer science, serves as a measure of the complexity of the cost function. Previous research has investigated the space complexity for various other problems, including approximate nearest neighbor [30], inner products [2], Euclidean metric compression [31], and graph cuts [9].

To investigate the space complexity of the (k,z)-Clustering problem, one initial approach is to utilize a coreset S, which yields a space requirement of $\tilde{O}(|S| \cdot d)$ using standard quantization methods (see Theorem 2 in the paper). Here the d factor arises from preserving all coordinates of each point in the coreset S. One might wonder if it is possible to combine the benefits of coreset construction and dimension reduction to eliminate the dependence on the dimension d in terms of space requirements. This leads to a natural question: "Is it possible to obtain an $|S| \cdot o(d)$ bound? Additionally, is coreset the most efficient compression scheme for the (k,z)-Clustering problem?" Perhaps surprisingly, we show that $\tilde{\Omega}(|S| \cdot d)$ is necessary for interesting parameter regimes (see Theorem 3 in the paper). This means a quantized coreset is optimal and dimensionality reduction does not help for space complexity. The proof of the lower bound for space complexity is our main contribution, which encounters more technical challenges. Unlike upper bounds, existing lower bounds for coresets do not directly translate into lower bounds for space complexity since compression approaches can go beyond simply storing a subset of points as a coreset. Overall, the study of space complexity is intricately connected to the optimality of coresets and poses technical difficulties.

1.1 Problem Definition and Our Results

In this paper, we initiate the study of the space complexity for the Euclidean (k,z)-Clustering problem. We first formally define the notion of space complexity. Assume that $P \subseteq [\Delta]^d$ for some integer $\Delta \geq 1$, i.e., every $p \in P$ is a grid point in $[\Delta]^d = \{1, 2, \ldots, \Delta\}^d$. This assumption is standard in the literature, e.g., for clustering [8, 26], facility location [18], minimum spanning tree [23], and the max-cut problem [11], and necessary for analyzing the space complexity. [1] Let $\mathcal{C}$ denote the collection of all k-center sets in $\mathbb{R}^d$, i.e., $\mathcal{C} := \left\{C \subset \mathbb{R}^d : |C| = k\right\}$. An ε-sketch for P is a data structure $\mathcal{O}$ that given any center set $C \in \mathcal{C}$, returns a value $\mathcal{O}(C) \in (1 \pm \varepsilon) \cdot \mathrm{cost}_z(P, C)$ which recovers the value $\mathrm{cost}_z(P, C)$ up to a multiplicative error of ε. We give the following notion.

[1] We need such an assumption to ensure that the precision of every coordinate of $p \in P$ is bounded. Otherwise, when P contains a unique point $p \in \mathbb{R}^d$, we need to maintain all coordinates of p such that the information of $\mathrm{cost}_z(P, \{p\}) = 0$ is preserved. Then if the precision of p can be arbitrarily large, the space complexity is unlimited.

▶ **Definition 1** (Space complexity for Euclidean (k,z)-Clustering). *Given a dataset $P \subseteq [\Delta]^d$, integers $n, k \geq 1$, constant $z \geq 1$ and an error parameter $\varepsilon \in (0,1)$, we define $sc(P, \Delta, k, z, d, \varepsilon)$ to be the minimum possible number of bits of an ε-sketch for P. Moreover, we define $sc(n, \Delta, k, z, d, \varepsilon) := \sup_{P \subseteq [\Delta]^d : |P| \leq n} sc(P, \Delta, k, z, d, \varepsilon)$ to be the space complexity function, i.e., the maximum cardinality $sc(P, \Delta, k, z, d, \varepsilon)$ over all datasets $P \subseteq [\Delta]^d$ of size at most n.*

Space upper bounds. Our first contribution is to provide upper bounds for the space complexity $sc(n, \Delta, k, z, d, \varepsilon)$. We apply the idea of storing an ε-coreset and have theorem 2. Here, an ε-coreset for (k,z)-Clustering is a weighted subset $S \subseteq P$ together with a weight function $w : S \to \mathbb{R}_{\geq 0}$ such that for every $C \in \mathcal{C}$, $\sum_{p \in S} w(p) \cdot d^z(p, C) \in (1 \pm \varepsilon) \cdot \mathrm{cost}_z(P, C)$.

▶ **Theorem 2** (Space upper bounds). *Suppose for any dataset $P \subseteq [\Delta]^d$ of size n, there exists an ε-coreset of P for (k,z)-Clustering of size at most $\Gamma(n) \geq 1$. We have the following space upper bounds:*

- *When $n \leq k$, $sc(n, \Delta, k, z, d, \varepsilon) \leq O(nd \log \Delta)$;*
- *When $n > k$, $sc(n, \Delta, k, z, d, \varepsilon) \leq O(kd \log \Delta + \Gamma(n)(d \log 1/\varepsilon + d \log\log \Delta + \log\log n))$.*

The proof of theorem 2 can be found in Section 2, which relies on novel geometric findings; see Section 1.2 for discussion. Fully storing a coreset S requires $\Gamma(n) \cdot d \log \Delta$ bits for points and $\Gamma(n) \cdot \log n$ for its weight function w. To further reduce the storage space, we provide a quantification scheme for the weight function w and points in S (Algorithm 1). Ignoring the logarithmic term, we have $sc(n, \Delta, k, z, d, \varepsilon) \leq \tilde{O}(\Gamma(n) \cdot d)$. [2] Combining with the recent breakthroughs that shows $\Gamma(n) = \tilde{O}(\min\left\{k^{\frac{2z+2}{z+2}} \varepsilon^{-2}, k\varepsilon^{-z-2}\right\})$ [16, 14, 15, 28], when $n > k$,

$$sc(n, \Delta, k, z, d, \varepsilon) \leq \tilde{O}(d \cdot \min\left\{\frac{k^{\frac{2z+2}{z+2}}}{\varepsilon^2}, \frac{k}{\varepsilon^{z+2}}\right\}). \tag{1}$$

Space lower bounds. Our main contribution is to provide the following lower bounds for the space complexity $sc(n, \Delta, k, z, d, \varepsilon)$.

▶ **Theorem 3** (Space lower bounds). *We have the following space lower bounds:*

- *When $n \leq k$, $sc(n, \Delta, k, z, d, \varepsilon) \geq \Omega(nd \log \Delta)$;*
- *When $n > k \geq 2$ and $\Delta = \Omega\left(\frac{k^{\frac{1}{d}}\sqrt{d}}{\varepsilon}\right)$, $sc(n, \Delta, k, z, d, \varepsilon) \geq \Omega\left(kd \log \Delta + kd \min\left\{\frac{1}{\varepsilon^2}, \frac{d}{\log d}, \frac{n}{k}\right\} + k \log\log \frac{n}{k}\right)$.*

The proof of this theorem can be found in Section 3. Compared to Theorem 2, our lower bound for space complexity is tight when $n \leq k$. For the case when $n > k$, the key term in our lower bound is $\Omega(kd \min\left\{\frac{1}{\varepsilon^2}, \frac{d}{\log d}, \frac{n}{k}\right\})$. Comparing this with Inequality (1), we can conclude that the optimal space complexity $sc(n, \Delta, k, z, d, \varepsilon) \approx \frac{d}{\varepsilon^2}$ when $k = O(1)$, $n \geq \Omega(\frac{1}{\varepsilon^2})$ and $d \geq \Omega(\frac{1}{\varepsilon^2 \log 1/\varepsilon})$. As a corollary, we can affirm that the coreset method is indeed the optimal compression method when the size and dimension of the dataset P are large and the number of centers k is constant. It would be interesting to further investigate whether the coreset method remains optimal for large k. Another corollary of Theorem 2 is a lower bound $\Omega(\frac{k}{\varepsilon^2})$ for the coreset size $\Gamma(n)$. This bound matches the previous result

[2] In this paper, $\tilde{O}(\cdot)$ may hide a factor of $2^{O(z)}$ and the logarithmic term of the input parameters $n, \Delta, k, d, 1/\varepsilon$.

in [14], and it has been recently improved to $\Omega(\frac{k}{\varepsilon^{-z-2}})$ when $\varepsilon = \Omega(k^{\frac{1}{z+2}})$ [28]. Since the technical approach is different, our methods for space lower bounds may also be useful to further improve the coreset lower bounds.

It is worth noting that d still appears in our lower bound results, which implies that exploiting dimension reduction techniques does not necessarily lead to a reduction in storage space. Although this may seem counter-intuitive, it is reasonable since we still need to maintain the mapping from the original space to the embedded space (which is also the space consumed by the dimensionality reduction itself), and the storage of this mapping could also be relatively large. Moreover, we can utilize this fact to lower bound the space cost of these dimension reduction methods from our results; see the following applications.

Application: Tight space lower bound for terminal embedding. Our Theorem 3 also yields an interesting by-product: a nearly tight lower bound for the space complexity of terminal embedding, which is a well-known dimension reduction method recently introduced by [21, 39]. It is a pre-processing step to map input data from one metric space $(\mathcal{X}, d_{\mathcal{X}})$ to the target metric space $(\mathcal{Y}, d_{\mathcal{Y}})$. The definition of it is given as follows.

▶ **Definition 4** (Terminal embedding). *Let $\varepsilon \in (0,1)$ and $X \subseteq \mathbb{R}^d$ be a collection of n points. A mapping $f : \mathbb{R}^d \to \mathbb{R}^m$ is called an ε-terminal embedding of X if for any $p \in X$ and $q \in \mathbb{R}^d$, $d(p,q) \leq d(f(p), f(q)) \leq (1+\varepsilon) \cdot d(p,q)$.*

This result is obtained by another natural idea for sketch construction: maintaining a terminal embedding function f for a coreset S and the projection $f(S)$ of a coreset S, in which the storage space for $f(S)$ can be independent on the dimension d. As a consequence of Theorem 3, the preservation of the terminal embedding function f must incur a large space cost; summarized as follows.

▶ **Theorem 5** (Informal; see the full version). *Let $\varepsilon \in (0,1)$ and assume $d = \Omega\left(\frac{\log n \log(n/\varepsilon)}{\varepsilon^2}\right)$. An ε-terminal embedding, that projects a given dataset $P \subset \mathbb{R}^d$ of size n to a target dimension $O(\frac{\log n}{\varepsilon^2})$, requires space at least $\Omega(nd)$.*

The bound $\Omega(nd)$ is not surprising since terminal embedding can be used to approximately recover the original dataset. In the case when $d \geq \Omega(\frac{\log n \log(n/\varepsilon)}{\varepsilon^2})$, our result improves upon the previous lower bound of $\Omega(\frac{n \log n}{\varepsilon^2})$ from [2]. [3] We replace their factor of $\frac{\log n}{\varepsilon^2}$ with d. Furthermore, our lower bound of $\Omega(nd)$ matches the prior upper bound of $\tilde{O}(nd)$ for terminal embedding [12], making it nearly tight.

1.2 Technical Overview

We mainly focus on analyzing the technical idea behind our main contribution Theorem 3. In general, our approach involves using a clever counting argument to establish lower bounds on space. We do this by creating a large family of datasets $\mathcal{P}$ where, for any pair P and Q from this family, there exists a center set C that separates their cost function by a significant margin, denoted as $\mathrm{cost}_z(P, C) \notin (1 \pm O(\varepsilon))\,\mathrm{cost}_z(Q, C)$. This difference in cost implies that P and Q can not share the same sketch, which leads to a lower bound on space of $\log(|\mathcal{P}|)$ (Lemma 8). Hence, we focus on how to construct such a family $\mathcal{P}$.

[3] Although the paper does not directly study terminal embedding, their bound for preserving inner products (Theorem 1.1 in [2]) implies a lower bound of $\Omega(\frac{n \log n}{\varepsilon^2})$ for terminal embedding.

We discuss the most technical bound, which is $\Omega\left(kd\min\left\{\frac{1}{\varepsilon^2},\frac{d}{\log d},\frac{n}{k}\right\}\right)$, when $n>k\geq 2$ and $\Delta=\Omega\left(\frac{k^{\frac{1}{d}}\sqrt{d}}{\varepsilon}\right)$. The proofs for other bounds are pretty standard. For brevity, we will explain the technical idea for the case of $z=k=2$ (2-Means). The extension to general z and k is straightforward, by analyzing Taylor expansions for $(1+x)^z$ and making $\Omega(k)$ copies of datasets in $\mathcal{P}$ (In the full version).

Our construction of $\mathcal{P}$ relies on a fundamental geometric concept known as *principal angles* (Definition 9). The Cosine of these angles, when given the orthonormal bases $P=\{p_i : i\in[n]\}$ and $Q=\{q_i : i\in[n]\}$ of two distinct subspaces in $\mathbb{R}^d$, uniquely correspond to the singular values of $P^\top Q$ (Lemma 10). This correspondence essentially measures how orthogonal the two subspaces are to each other. With principal angles in mind, we outline the two main components of our proof. Assuming that $d>n$, the first component (Lemma 11) demonstrates that if the largest $O(n)$ principal angles between two orthonormal bases P and Q are sufficiently large, there exists a center set $C=\{c,-c\}\in\mathcal{C}$ with $\|c\|_2=1$ such that $\mathrm{cost}_2(P,\{c,-c\})-\mathrm{cost}_2(Q,\{c,-c\})\geq\Omega\left(\sqrt{n}\right)$. This induced error of $\Omega\left(\sqrt{n}\right)$ from C achieves the desired scale of $\varepsilon\cdot\mathrm{cost}_z(P,C)=O(\varepsilon n)$ when $n=O\left(\frac{1}{\varepsilon^2}\right)$. The second component (Lemma 12) states that when $n=O\left(\frac{d}{\log d}\right)$, there exists a large family $\mathcal{P}$ of orthonormal bases (for different n-dimensional subspaces) with size $\exp(nd)$ such that most principal angles of any two different orthonormal bases in the family are sufficiently large. The space lower bound $\Omega\left(d\min\left\{\frac{1}{\varepsilon^2},\frac{d}{\log d},n\right\}\right)$ directly follows from these two lemmas.

Next, we delve into the technical insights behind Lemmas 11 and 12.

Lemma 11: Reduction from principal angles to cost difference. Recall that we aim to show the existence of a center set $C=\{c,-c\}$ that incurs a large cost difference between two orthonormal bases $P=\{p_i : i\in[n]\}$ and $Q=\{q_i : i\in[n]\}$. By the formulation of C, we note that $\mathrm{cost}_2(P,C)-\mathrm{cost}_2(Q,C)=2\left(\sum_{i=1}^n|\langle q_i,c\rangle|-|\langle p_i,c\rangle|\right)$. Hence, we focus on showing the existence of a unit vector $c\in\mathbb{R}^d$ such that

$$\sum_{i=1}^{n}|\langle q_i,c\rangle|-|\langle p_i,c\rangle|\geq\Omega(\sqrt{n}). \tag{2}$$

Intuitively, our goal is to increase the magnitude of the first term $\sum_{i=1}^n|\langle q_i,c\rangle|$ while decreasing the magnitude of the second term $\sum_{i=1}^n|\langle p_i,c\rangle|$. One initial approach is to choose $c=\frac{1}{\sqrt{n}}Q\lambda=\frac{1}{\sqrt{n}}\sum_{i\in[n]}\lambda_i q_i$, where $\lambda\in\{-1,+1\}^n$ is a full coloring on $[n]$. By this selection, center c lies on the subspace spanned by Q and maximizes the first term $\sum_{i=1}^n|\langle q_i,c\rangle|$ to be $\sqrt{n}$. Moreover, the second term becomes $\sum_{i=1}^n|\langle p_i,c\rangle|=\frac{1}{\sqrt{n}}\|P^\top Q\lambda\|_1\leq\sqrt{n}\|P^\top Q\lambda\|_\infty$. Ideally, if we can find a coloring $\lambda\in\{-1,+1\}^n$ such that $\|P^\top Q\lambda\|_\infty\leq 0.5$, we can achieve the desired cost difference in Inequality (2). However, the existence of such λ appears to be non-trivial. For instance, if we randomly select a coloring λ from $\{-1,+1\}^n$, the expected value of $\|P^\top Q\lambda\|_\infty$ can be as large as $O(\log n)$ [41].

Motivated by the discrepancy literature (e.g., [41, 19]), we enhance the previous idea by allowing $\lambda\in\{-1,0,+1\}^n$ to be a partial coloring with $\|\lambda\|_1=0.75n$. With this modification, we still have $\sum_{i=1}^n|\langle q_i,c\rangle|=0.75\sqrt{n}$, and it suffices to bound $\|P^\top Q\lambda\|_\infty\leq 0.5$ to achieve Inequality (2). Spencer et al. [41] have shown the existence of such a partial coloring λ with $\|P^\top Q\lambda\|_\infty\leq\alpha$ for some constant $\alpha\geq 1$ and orthonormal bases P and Q. However, we require a stricter bound of $\alpha\leq 0.5$, which calls for new ideas. Fortunately, we discover that large principal angles indicate the existence of a partial coloring λ such that $\|P^\top Q\lambda\|_\infty\leq 0.5$ (see Lemmas 13 and 14). Geometrically, large principal angles imply that the two subspaces spanned by P and Q are nearly orthogonal to each other in many directions. As a consequence, most inner products $\langle p_i, q_j \rangle$ are close to 0, indicating that the majority of the row norms $\|(P^\top Q)_i\|_2$ are smaller than 0.1 (Lemma 13). These small row sums enable us to find a partial coloring λ that further reduces the bound for $\|P^\top Q\lambda\|_\infty$ to 0.5 (Lemma 14), employing similar approaches as in [41]. In summary, we have completed the proof of Lemma 11.

Lemma 12: Construction of $\mathcal{P}$. Our construction is inspired by a geometric observation made in Absil et al. [1], which states that the largest principal angle between the orthonormal bases P and Q of two n-dimensional subspaces, independently drawn from the uniform distribution on the Grassmann manifold of n-planes in $\mathbb{R}^d$, is at least $\Omega(1)$ with high probability. We extend this result and prove that even the largest $O(n)$ principal angles between P and Q are at least $\Omega(1)$ (Lemma 15). This extension relies on a more careful integral calculation for the density function of principal angles. Moreover, this extension leads to an enhanced geometric observation: on average, these two orthonormal bases P and Q are distinct with respect to principal angles, which could be of independent research interest. Then using standard probabilistic arguments, we can randomly select a family $\mathcal{P}$ of $\exp(\Omega(nd))$ orthonormal bases, ensuring that the largest $O(n)$ principal angles between any pair P and Q from $\mathcal{P}$ are consistently large.

1.3 Other Related Work

Coreset construction for clustering. There are a series of works towards closing the upper and lower bounds of coreset size for (k, z)-Clustering in high dimensional Euclidean spaces [22, 7, 16, 14, 15, 28]. The current best upper bound is $\tilde{O}(\min\{\frac{k^{\frac{2z+2}{z+2}}}{\varepsilon^2}, \frac{k}{\varepsilon^{z+2}}\})$ by [16, 14, 15, 28]. On the other hand, Huang and Vishnoi [29] proved a size lower bound $\Omega(k \min\{2^{z/20}, d\})$ and Cohen-addad et al. [14] showed bound $\Omega(k\varepsilon^{-2})$. Very recently, Huang and Li [28] gave a size lower bound of $\Omega(k\varepsilon^{-z-2})$ for $\varepsilon = \Omega(k^{-1/(z+2)})$, which matches the size upper bound and is nearly tight. There have also been studies for the coreset size when the dimension is small, see e.g. [24, 27]. In addition to offline settings, coresets have also been studied in the stream setting [24, 6, 17], distributed setting [4] and dynamic setting [25].

Dimension reduction. Dimension reduction is an important technique for data compression, including techniques like Johnson-Lindenstrauss (JL) [37, 10] and terminal embedding [39, 29]. The target dimension of any embedding satisfying the JL lemma is shown to be $\Theta(\varepsilon^{-2}\log n)$ [32, 2, 35], where n is the size of the data set. The space complexity of JL is shown to be $O(\log d + \log(1/\delta)(\log\log(1/\delta) + \log(1/\varepsilon)))$ random bits [34], where ε and δ are error and fail probability respectively. In the context of (k, z)-Clustering, Makarychev et al. [37] gives a nearly optimal target dimension $O(\log(k/\varepsilon)/\varepsilon^2)$ by applying JL. Their reduction ensures that the cost of the optimal clustering is preserved within a factor of $(1+\varepsilon)$ instead of preserving the clustering cost for all center sets. For terminal embedding, Narayanan and Nelson [39] provided an optimal terminal embedding with target dimension $O(\varepsilon^{-2}\log n)$. For the space complexity, the best-known construction of terminal embedding costs $\tilde{O}(nd)$ bits [12].

2 Proof of Theorem 2: Space Upper Bounds

For the first part when $n \le k$, we simply store all data points. Since $P \subseteq [\Delta]^d$, the storage space for each coordinate is at most $\log\Delta$, which results in the space upper bound $O(nd\log\Delta)$. Next, we focus on the second part when $n > k$. The main idea is to construct a sketch to store a coreset using space as small as possible. We need the following lemma for preparation.

▶ **Lemma 6** (Relaxed triangle inequality (Lemma 10 of [14])). *Let a, b, c be an arbitrary set of points in a metric space with distance function d, and let z be a positive integer. Then for any $\varepsilon > 0, d(a,b)^z \leq (1+\varepsilon)^{z-1} d(a,c)^z + \left(\frac{1+\varepsilon}{\varepsilon}\right)^{z-1} d(b,c)^z$ and $|d(a,b)^z - d(a,c)^z| \leq \varepsilon \cdot d(a,c)^z + \left(\frac{z+\varepsilon}{\varepsilon}\right)^{z-1} d(b,c)^z$.*

Let $P \subseteq [\Delta]^d$ be a dataset of size $n > k$, (S, w) be an $\frac{\varepsilon}{5}$-coreset of P for (k,z)-Clustering, and $C^\star$ be an $O(1)$-approximation of optimal center set, that is, a center set satisfying $\mathrm{cost}_z(P, C^\star) \leq O(1) \cdot \min_{C \in \mathcal{C}} \mathrm{cost}_z(P, C)$. We argue that by rounding each $c_l^\star \in C^\star$ to the nearest point in P, i.e. $p_{c^\star} := \mathrm{argmin}_{p \in P} d(c^\star, p)$, it remains the property of $O(1)$-approximation. To this end, by Lemma 6, for any $p \in P$, $|d(p,c^\star)^z - d(p, p_{c^\star})^z| \leq d(p,c^\star)^z + (1+z)^{z-1} d(p_{c^\star}, c^\star)^z \leq (1 + (1+z)^{z-1}) d(p, c^\star)^z$. As z is constant, the claim is proved. Without of loss of generality, we assume $C^\star \subseteq P$ and $\mathrm{cost}_z(P, C^\star) \leq 2 \min_{C \in \mathcal{C}} \mathrm{cost}_z(P, C)$.

The compression scheme is summarized in Algorithm 1. Intuitively, we need to compress the weight $w(p)$ and each coordinate $i \in [d]$ of $p \in P$. For $w(p)$, we either safely ignore too small weight, i.e. $w(p) \leq \frac{\varepsilon}{4|S|}$, or remain its highest position by γ_p and the first $\lceil \log 4/\varepsilon \rceil$ digits by $w_{p,\varepsilon}$. For each p, we denote $c_p^\star$ to be the closest center of p in $C^\star$ and $c_{p,i}^\star$ to be the i'th coordinate of $c^\star$. Then compress each coordinate $p_i - c_{p,i}^\star$ by a similar idea as for weight. In Algorithm 1 we abuse the notation $c_l^\star$ to denote the l'th point in $C^\star$, and $c_{l,i}^\star$ similarly.

Algorithm 1 A compression scheme based on coreset.

Input: Error parameter $\varepsilon \in (0,1)$, an $\frac{\varepsilon}{4}$-coreset $S \subseteq P$ of size $|S| \leq \Gamma(n)$ together with a weight function $w : S \to \mathbb{R}_{\geq 0}$, an 1-approximate center set $C^\star \subseteq P$ of P for (k,z)-Clustering
Output: A sketch T of P for (k,z)-Clustering
Partition S into $S_l := \left\{ p \in S | c_l = \mathrm{argmin}_{c_l^\star \in C^\star} d(p, c_l^\star) \right\}, l = 1, \cdots, k$;
for $c_l^\star \in C^\star$ **do**
  for $p \in S_l$ **do**
    if $w(p) \leq \frac{\varepsilon}{4|S|}$ **then** $(w_{p,\varepsilon}, \gamma_p) \leftarrow (0,0)$;
    else $w_{p,\varepsilon} \leftarrow \frac{w(p)}{2^{\lfloor \log w(p) \rfloor}}$, rounding to $\lceil \log 4/\varepsilon \rceil$ decimal places; $\gamma_p \leftarrow \lfloor \log w(p) \rfloor$;
    end if
    $w_p \leftarrow (w_{p,\varepsilon}, \gamma_p)$;
    for each coordinate $i \in [d]$ **do**
      if $p_i - c_{l,i}^\star = 0$ **then** $(\tau_{i,\varepsilon}, \gamma_{p,i}) \leftarrow (0,0)$;
      else $\tau_{p,i} \leftarrow \frac{p_i - c_{l,i}^\star}{2^{\lfloor \log(p_i - c_{l,i}^\star) \rfloor}}$, rounding to $\lceil \log 4z/\varepsilon \rceil$ decimal places;
        $\gamma_{p,i} \leftarrow \lfloor \log(p_i - c_{l,i}^\star) \rfloor$;
      end if
      $\tau_{p,i} \leftarrow (\tau_{p,i}, \gamma_{p,i})$;
    end for
    $T_l \leftarrow \cup_{p \in S_l} (\{\tau_{p,i}\}_{i=1}^d, w_p)$;
  end for
end for
return $T \leftarrow \cup_{l \in [k]} (c_l^\star, T_l)$

Since S is a coreset, we will make use of the following lemma.

▶ **Lemma 7** (Sum of weights). *We have that $\sum_{p \in S} w(p) \in (1 \pm \varepsilon) n$.*

Proof. Consider a center set $C \in \mathcal{C}$ far away from P, e.g. $C = +\infty^d$. then $\frac{\sum_{p\in P} d(p,C)^z}{\sum_{p\in S} w(p)d(p,C)^z} \approx \frac{n}{\sum_{p\in S} w(p)}$. By the coreset definition we have $\frac{\mathrm{cost}_z(P,C)}{\sum_{p\in S} w(p)d(p,C)^z} = \frac{\sum_{p\in P} d(p,C)^z}{\sum_{p\in S} w(p)d(p,C)^z} \in 1\pm\varepsilon$. ◀

Now we are ready to prove the second part of Theorem 2.

Proof of Theorem 2 (second part).

Correctness analysis. We first show Algorithm 1 outputs an ε-sketch of P for (k,z)-Clustering. We use T to obtain $\widehat{w}_p = w_{p,\varepsilon} \cdot 2^{\gamma_p}$ and $\widehat{p} = (\hat{\tau}_1, \cdots, \hat{\tau}_d) + c^\star_p$ with $\hat{\tau}_i = \tau_{p,i} \cdot 2^{\gamma_{p,i}}$ for $i \in [d]$. Given a center set $C \in \mathcal{C}$, we approximate (k,z)-Clustering function by the value: $\sum_{p\in S} \widehat{w}_p \cdot d(\widehat{p}, C)^z$. We claim that for each $p \in S$, $\widehat{w}_p \in (1 \pm \frac{\varepsilon}{4})w(p)$ when $w(p) > \frac{\varepsilon}{4|S|}$. This is because $\frac{w(p)}{2} \le 2^{\gamma_p} \le w(p)$ and $|w_{p,\varepsilon} - \frac{w(p)}{2^{\gamma_p}}| \le \frac{\varepsilon}{4}$, which implies that $w_{p,\varepsilon} \in (1 \pm \frac{\varepsilon}{4})\frac{w(p)}{2^{\gamma_p}}$ and $\widehat{w}_p \in (1 \pm \frac{\varepsilon}{4})w(p)$. When $w(p) \le \frac{\varepsilon}{4|S|}$, we have $w(p)d(p,C)^z \le \frac{\varepsilon}{4|S|}d(p,C)^z \le \frac{\varepsilon}{4|S|}\sum_{p\in P} d(p,C)^z$, which means this quantity is too small to affect (k,z)-Clustering function and we could set such $w(p)$ to zero.

We next analysis $\widehat{p}$. By Lemma 6, for any $c \in C$, $|d(p,c)^z - d(\widehat{p},c)^z| \le \frac{\varepsilon}{4}d(p,c)^z + (1 + \frac{4z}{\varepsilon})^{z-1}d(p,\widehat{p})^z$. By construction, $d(p,\widehat{p}) = d(p - c^\star_p, \hat{\tau}) = \sqrt{\sum_{i=1}^d (p_i - c^\star_{p,i} - \hat{\tau}_i)^2}$. Similarly for weight, $p_i - c^\star_{p,i} - \hat{\tau}_i \le \frac{\varepsilon}{4z}(p_i - c^\star_{p,i})$, and thus $d(p,\widehat{p}) \le \frac{\varepsilon}{4z}d(p,c^\star_p)$. Therefore,

$$\begin{aligned}
\sum_{p\in S} \widehat{w}_p \cdot d(\widehat{p}, C)^z \in& \sum_{p\in S} w(p)\left(1 \pm \frac{\varepsilon}{4}\right)\left[d(p,c_p)^z \pm \left(\frac{\varepsilon}{4}d(p,c_p)^z + \left(1+\frac{4z}{\varepsilon}\right)^{z-1} d(p,\widehat{p})^z\right)\right] \\
&\in \left(1 \pm \frac{\varepsilon}{4}\right)\sum_{p\in S} w(p)\left[\left(1 \pm \frac{\varepsilon}{4}\right)d(p,c_p)^z \pm \frac{\varepsilon}{4z}\left(1+\frac{\varepsilon}{4z}\right)^{z-1} d(p,c^\star_p)^z\right] \\
&\in \left(1 \pm \frac{\varepsilon}{4}\right)\sum_{p\in S} w(p)\left[\left(1 \pm \frac{\varepsilon}{4}\right)d(p,c_p)^z \pm \frac{\varepsilon}{8}d(p,c^\star_p)^z\right] \\
&\in \left(1 \pm \frac{\varepsilon}{4}\right)\sum_{p\in S} w(p)\left(1 \pm \frac{\varepsilon}{2}\right)d(p,c_p)^z \\
&\in \left(1 \pm \frac{\varepsilon}{4}\right)\left(1 \pm \frac{\varepsilon}{2}\right)\left(1 \pm \frac{\varepsilon}{5}\right)\sum_{p\in P} d(p,c_p)^z \in (1\pm\varepsilon)\sum_{p\in P} d(p,c_p)^z,
\end{aligned}$$

where the third line follows from $\ln(1 + \frac{\varepsilon}{4z}) \le \frac{\varepsilon}{4z}$, the fourth line follows from $C^\star$ is a 1-approximation of an optimal center set, and the penultimate line follows from the construction of coreset. Therefore we construct an ε-sketch for P.

Space complexity analysis. The storage for k grid points $C^\star$ is $O(kd\log\Delta)$. We store each weight by a set $(w_{p,\varepsilon}, \gamma_p)$, where the first number is up to $O(\lceil \log 4/\varepsilon \rceil)$ decimal places, and representing the integer number $\gamma_p = \lfloor \log w(p) \rfloor$ requires $O(\log\max\{\log\frac{4|S|}{\varepsilon}, \log n\})$ bits by Lemma 7. Similarly, the storage for each τ_p is $O(d\log 4z/\varepsilon + d\log\log\Delta)$ bits. Combining them and notice that $|S| \le n$, we obtain the final bound

$$\begin{aligned}
sc(P,\Delta,k,z,d,\varepsilon) \le& O\left(kd\log\Delta + |S|\left(\log\frac{4}{\varepsilon} + \log\max\left\{\log\frac{4|S|}{\varepsilon}, \log n\right\}\right.\right. \\
&\left.\left. + \ d\log\frac{4z}{\varepsilon} + d\log\log\Delta\right)\right) \\
=& O\left(kd\log\Delta + \Gamma(n)(d\log 1/\varepsilon + d\log\log\Delta + \log\log n)\right),
\end{aligned}$$

where we ignore the dependence on z. ◀

3 Proof of Theorem 3: Space Lower Bounds

In this section, we prove the space lower bounds. The high-level idea is to construct a large family of datasets such that any two of them can not use the same sketch.

▶ **Lemma 8** (A family of datasets leads to space lower bounds). *Suppose there exists a family $\mathcal{P}$ of datasets of size $n \geq 1$ such that for any two datasets $P, Q \in \mathcal{P}$, there exists a center set $C \in \mathcal{C}$ with $\mathrm{cost}_z(P, C) \notin (1 \pm 3\varepsilon)\, \mathrm{cost}_z(Q, C)$. Then we have $sc(n, \Delta, k, z, d, \varepsilon) \geq \Omega(\log |\mathcal{P}|)$.*

Proof. We prove this by contradiction. Assume that $sc(n, \Delta, k, z, d, \varepsilon) = o(\log |\mathcal{P}|)$, we must be able to find two datasets P and Q such that they correspond to the same ε-sketch $\mathcal{O}$. Since $\mathcal{O}$ is an ε-sketch for both P and Q, we have for every center set $C \in \mathcal{C}$,

$$\mathcal{O}(C) \in (1 \pm \varepsilon) \cdot \mathrm{cost}_z(P, C),\ \ \mathcal{O}(C) \in (1 \pm \varepsilon) \cdot \mathrm{cost}_z(Q, C),$$
$$\mathrm{cost}_z(P, C) \leq \frac{1}{1-\varepsilon}\mathcal{O}(C) \leq \frac{1+\varepsilon}{1-\varepsilon}\,\mathrm{cost}_z(Q, C) \leq (1 + 3\varepsilon)\,\mathrm{cost}_z(Q, C),$$
$$\mathrm{cost}_z(P, C) \geq \frac{1}{1+\varepsilon}\mathcal{O}(C) \geq \frac{1-\varepsilon}{1+\varepsilon}\,\mathrm{cost}_z(Q, C) \geq (1 - 3\varepsilon)\,\mathrm{cost}_z(Q, C).$$

This contradicts with our assumption that $\mathrm{cost}_z(P, C) \notin (1 \pm 3\varepsilon)\,\mathrm{cost}_z(Q, C)$. ◀

3.1 Proof of Theorem 3

We first prove the lower bound $\Omega\left(nd \log \Delta\right)$ when $n \leq k$.

Proof of Theorem 3 (first part). We construct a family $\mathcal{P}$ as follows: for each dataset $P \in \mathcal{P}$, we choose n different grid points in $[\Delta]^d$. Note that for each single grid point, there are Δ^d choices. Therefore, the size $|\mathcal{P}|$ is $\binom{\Delta^d}{n}$, which implies that $\log |\mathcal{P}| = \Omega(nd \log \Delta)$.

For any two datasets $P, Q \in \mathcal{P}$, there must be a single grid point p such that $p \in Q \setminus P$. Let $C = P \cup \{\underbrace{c, \cdots, c}_{k-n \text{ points}}\}$, where $c \in \mathbb{R}^d$ is an arbitrary point with $c \neq p$. We have

$$\mathrm{cost}_z\left(Q, C\right) \geq d\left(p, C\right)^z > 0,\ \ \mathrm{cost}_z\left(P, C\right) = 0 \notin (1 \pm 3\varepsilon)\,\mathrm{cost}_z\left(Q, C\right).$$

Using Lemma 8, we obtain the space lower bound $\Omega\left(nd \log \Delta\right)$. ◀

We then consider the second part of Theorem 3 when $n > k$. We first prove the lower bound $\Omega\left(kd \min\left\{\frac{1}{\varepsilon^2}, \frac{d}{\log d}, \frac{n}{k}\right\}\right)$. Recall that we have $n > 2$ and $\Delta = \Omega(\frac{k^{\frac{1}{d}}\sqrt{d}}{\varepsilon})$. For preparation, we introduce the notion of principal angles.

▶ **Definition 9** (Principal angles). *Suppose $n \leq d$. Given two n-dimensional subspaces $\mathcal{X}$ and $\mathcal{Y}$ of $\mathbb{R}^d$, the principal angles (or canonical angles) $\theta_1(\mathcal{X}, \mathcal{Y}), \cdots, \theta_n(\mathcal{X}, \mathcal{Y}) \in \left[0, \frac{\pi}{2}\right]$ between them are defined recursively by the following equations: $\forall i \in [n]$, $\cos\left(\theta_i(\mathcal{X}, \mathcal{Y})\right) = \sup_{x \in \mathcal{X}, y \in \mathcal{Y}} |x^T y| = |x_l^T y_l|$, subject to $x \perp x_1, \cdots, x_{i-1}, \|x\|_2 = 1, y \perp y_1, \cdots, y_{i-1}, \|y\|_2 = 1$.*

The notion of principal angles between subspaces was first introduced by Jordan [33] and has many important applications in statistics and numerical analysis [20, 42]. Intuitively, small principal angles indicate that the two subspaces are nearly parallel in many directions, while large principal angles imply that the two subspaces span many directions that are nearly orthogonal to each other. For example, when $\mathcal{X} \perp \mathcal{Y}$, all principal angles $\theta_i(\mathcal{X}, \mathcal{Y}) = \frac{\pi}{2}$. The following lemma shows a relation between principal angles and singular value decomposition.

▶ **Lemma 10** (Property of principal angles (Theorem 1 in [5])). *Given two n-dimensional subspaces $\mathcal{X}$ and $\mathcal{Y}$ of $\mathbb{R}^d$, let the columns of matrices $X \in \mathbb{R}^{d\times n}$ and $Y \in \mathbb{R}^{d\times n}$ form orthonormal bases for the subspaces $\mathcal{X}$ and $\mathcal{Y}$ respectively. Denote $1 \geq \sigma_1 \geq \cdots \geq \sigma_n$ to be the singular values of the inner product matrix X^TY. We have $\sigma_i = \cos(\theta_i(\mathcal{X},\mathcal{Y})), \forall i \in [n]$.*

Lemma 10 implies that the principal angles are symmetric, i.e. $\theta(\mathcal{X},\mathcal{Y}) = \theta(\mathcal{Y},\mathcal{X})$. It also implies that the principal angles are orthogonally invariant, i.e. $\theta(\mathcal{X}A,\mathcal{Y}B) = \theta(\mathcal{X},\mathcal{Y})$ for any orthogonal matrix $A, B \in \mathbb{R}^{n\times n}$, since $\sigma\left(A^TX^TYB\right) = \sigma\left(X^TY\right)$. Hence, Lemma 10 holds for any choice of orthonormal bases X and Y of corresponding subspaces.

The idea is still to construct a large family $\mathcal{P}$ of datasets to apply Lemma 8. For ease of analysis, we first do not require the construction of datasets $P \subseteq [\Delta]^d$ and ensure that every P consists of orthonormal bases of some subspaces. At the end of the proof, we will show how to round and scale these datasets P into $[\Delta]^d$.

For any two datasets $P, Q \in \mathcal{P}$, let their inner product matrix be

$$U := P^TQ = [U_1, \cdots, U_n]^T = \begin{bmatrix} p_1^Tq_1 & \cdots & p_1^Tq_n \\ \vdots & \ddots & \vdots \\ p_n^Tq_1 & \cdots & p_n^Tq_n \end{bmatrix} = (U_{ij})_{i,j\in[n]}.$$

Slightly abuse of notation, we use $\theta_i(P,Q)$ to represent the i-th least principal angles of the two subspaces spanned by P and Q. The following lemma shows that large principal angles between P and Q imply a large cost difference on some center set $\{c, -c\}$.

▶ **Lemma 11** (Principal angles to cost difference). *Let P, Q be datasets of n orthonormal bases (with $100 \leq n \leq \frac{d}{2}$) satisfying that $\theta_{\frac{1}{32}10^{-6}\cdot n}(P,Q) \geq \arccos\left(\frac{10^{-3}}{4\sqrt{2}}\right)$. There exists a unit vector $c \in \mathbb{R}^d$ such that $\mathrm{cost}_2(P,\{c,-c\}) - \mathrm{cost}_2(Q,\{c,-c\}) \geq \frac{1}{2}\sqrt{n}$.*

Proof. Proof of Lemma 11 can be found in section 3.2. ◀

Applying $n = O\left(\frac{1}{\varepsilon^2}\right)$ to the above lemma leads to our desired cost difference $\Omega(\varepsilon n) = \Omega(\varepsilon \mathrm{cost}_2(P,\{c,-c\}))$. We then show it is possible to construct a large family $\mathcal{P}$ such that the principal angles between any two datasets in $\mathcal{P}$ are sufficiently large.

▶ **Lemma 12** (Construction of a large family of datasets). *When $n = O\left(\frac{d}{\log d}\right)$, there is a family $\mathcal{P}$ of size $\exp\left(\frac{1}{256}10^{-6}\log\left(\frac{1}{1-\frac{1}{32}10^{-6}}\right)\cdot nd\right)$ such that for any two dataset $P, Q \in \mathcal{P}$, we have $\theta_{\frac{1}{32}10^{-6}\cdot n}(P,Q) \geq \arccos\left(\frac{10^{-3}}{4\sqrt{2}}\right)$.*

Proof. Proof of Lemma 12 can be found in section 3.3. ◀

Combining Lemma 11 and 12, we are ready to prove the second part of Theorem 3.

Proof of Theorem 3 (second part). The lower bound of $\Omega(kd\log\Delta)$ is trivial since $sc(n,\Delta,k,z,d,\varepsilon)$ is non-decreasing with n. Then by the first part of Theorem 3, we have $sc(n,\Delta,k,z,d,\varepsilon) \geq sc(k,\Delta,k,z,d,\varepsilon) \geq \Omega(kd\log\Delta)$.

Next, we prove the lower bound of $\Omega\left(kd\min\left\{\frac{1}{\varepsilon^2}, \frac{d}{\log d}, \frac{n}{k}\right\}\right)$. We mainly prove the case of $k = z = 2$. The extensions to general k and z can be found in the full version. Denote

$$\tilde{n} = \min\left\{\Theta\left(\frac{1}{484\varepsilon^2}\right), \Theta\left(\frac{d}{\log d}\right), n\right\} \geq 100,$$

where the first term $\Theta\left(\frac{1}{484\varepsilon^2}\right)$ is for achieving a large cost difference by Lemma 11 and the second term $\Theta\left(\frac{d}{\log d}\right)$ is to satisfy the condition of Lemma 12. Since $sc(n, \Delta, 2, 2, d, \varepsilon)$ is non-decreasing with n, it suffices to prove a lower bound for $sc(\tilde{n}, \Delta, 2, 2, d, \varepsilon)$.

Lemma 12 shows that we can find a family $\mathcal{P}$ of size $\exp\left(\frac{1}{256}10^{-6}\log\left(\frac{1}{1-\frac{1}{32}10^{-6}}\right)\cdot \tilde{n}d\right)$ such that for any two dataset in this set P and Q, we have $\theta_{\frac{1}{32}10^{-6}\cdot n}(P,Q) \geq \arccos\left(\frac{10^{-3}}{4\sqrt{2}}\right)$. Lemma 11 shows that such a condition allows us to find a unit-norm vector c such that $\mathrm{cost}_2(P,\{c,-c\}) - \mathrm{cost}_2(Q,\{c,-c\}) \geq \frac{1}{2}\sqrt{\tilde{n}}$. By our choice of $\tilde{n}$, we have $\frac{1}{2}\sqrt{\tilde{n}} \geq 11\varepsilon\tilde{n} \geq 5\varepsilon \cdot \mathrm{cost}_2(P,\{c,-c\})$ and $\tilde{n} = O(\frac{d}{\log d})$. Hence, there exists a family $\mathcal{P}$ of size $\exp(\Theta(\tilde{n}d)) = \exp\left(\Theta\left(d\min\left\{\frac{1}{\varepsilon^2}, \frac{d}{\log d}, n\right\}\right)\right)$ such that for any two datasets $P, Q \in \mathcal{P}$, there exists a unit vector $c \in \mathbb{R}^d$ with $\mathrm{cost}_2(P,\{c,-c\}) - \mathrm{cost}_2(Q,\{c,-c\}) \geq 11\varepsilon\tilde{n}$.

We then round and scale every dataset $P \in \mathcal{P}$ to $[\gamma]^d$, where $\gamma = \lceil\frac{10\sqrt{d}}{\varepsilon}\rceil$. The extra term $k^{\frac{1}{d}}$ in Δ will show up when we extend the result to general $k \geq 2$ (In the full version). Without loss of generality, we assume that γ is an odd integer. Otherwise, we let $\gamma = \lceil\frac{10\sqrt{d}}{\varepsilon}\rceil + 1$.

Denote $\mathbf{1} = (1, \cdots, 1)$. For a dataset $P = (p_1, \cdots, p_{\tilde{n}}) \in \mathcal{P}$, we will construct $\tilde{\mathcal{P}}$ to be our final family as follows: For each of dataset $P \in \mathcal{P}$, we shift the origin to $\lceil\frac{\gamma}{2}\rceil \cdot \mathbf{1}$, scale it by a factor of $\frac{\gamma}{2}$ and finally perform an upward rounding on each dimension to put every point on the grid: $\tilde{P} = \left(\lceil\frac{\gamma}{2}p_1\rceil, \cdots, \lceil\frac{\gamma}{2}p_{\tilde{n}}\rceil\right) + \lceil\frac{\gamma}{2}\rceil\cdot\mathbf{1} = (\tilde{p}_1, \cdots, \tilde{p}_{\tilde{n}})$. We will then show that this set fulfills the requirement of Lemma 8. For ease of explanation, we also define $\hat{P}$ to be the dataset without rounding: $\hat{P} = \left(\frac{\gamma}{2}p_1, \cdots, \frac{\gamma}{2}p_{\tilde{n}}\right) + \lceil\frac{\gamma}{2}\rceil\cdot\mathbf{1} = (\hat{p}_1, \cdots, \hat{p}_{\tilde{n}})$. Moreover, let $\bar{c} = \frac{\gamma}{2}c + \lceil\frac{\gamma}{2}\rceil\cdot\mathbf{1}$. We must have that for the scaling dataset,

$$\begin{aligned}
&\mathrm{cost}_2\left(\hat{P},\{\bar{c},-\bar{c}\}\right) = \frac{\gamma^2}{4}\mathrm{cost}_2\left(P,\{c,-c\}\right) \leq \frac{\gamma^2\tilde{n}}{2},\\
&\mathrm{cost}_2\left(\hat{P},\{\bar{c},-\bar{c}\}\right) - \mathrm{cost}_2\left(\hat{Q},\{\bar{c},-\bar{c}\}\right)\\
&= \frac{\gamma^2}{4}\left(\mathrm{cost}_2(P,\{c,-c\}) - \mathrm{cost}_2(Q,\{c,-c\})\right) \geq \frac{11\gamma^2\varepsilon\tilde{n}}{4}.
\end{aligned}$$

On the other hand, for the rounding dataset, we have

$$\left|\|\hat{p}_i - \bar{c}\|_2^2 - \|\tilde{p}_i - \bar{c}\|_2^2\right| \leq 2\left\|\hat{p}_i - \tilde{p}_i\right\|_2\left\|\hat{p}_i - \bar{c}\right\|_2 + \left\|\hat{p}_i - \tilde{p}_i\right\|_2^2 \leq 2\gamma\sqrt{d} + d \leq \frac{\gamma^2\varepsilon}{4}.$$

The case for $-\bar{c}$ and other datasets is similar. Therefore, we have that for any dataset

$$\begin{aligned}
&\left|\mathrm{cost}_2(\hat{P},\{\bar{c},-\bar{c}\}) - \mathrm{cost}_2(\tilde{P},\{\bar{c},-\bar{c}\})\right| \leq \frac{\gamma^2\varepsilon\tilde{n}}{4},\\
&\mathrm{cost}_2\left(\tilde{P},\{\bar{c},-\bar{c}\}\right) \leq \mathrm{cost}_2\left(\hat{P},\{\bar{c},-\bar{c}\}\right) + \frac{\gamma^2\varepsilon\tilde{n}}{4} \leq \frac{\gamma^2\tilde{n}}{2} + \frac{\gamma^2\varepsilon\tilde{n}}{4} \leq \frac{3\gamma^2\tilde{n}}{4}.
\end{aligned}$$

We now have a rounded family $\tilde{\mathcal{P}}$ such that all points are on the grid and we can find $\bar{c}$ that

$$\begin{aligned}
&\mathrm{cost}_2\left(\tilde{P},\{\bar{c},-\bar{c}\}\right) - \mathrm{cost}_2\left(\tilde{Q},\{\bar{c},-\bar{c}\}\right)\\
&\geq \mathrm{cost}_2\left(\hat{P},\{\bar{c},-\bar{c}\}\right) - \mathrm{cost}_2\left(\hat{Q},\{\bar{c},-\bar{c}\}\right) - \frac{\gamma^2\varepsilon\tilde{n}}{2} \geq \frac{9\gamma^2\varepsilon\tilde{n}}{4}.
\end{aligned}$$

Since the cost function value is upper bounded by $\frac{3\gamma^2\tilde{n}}{4}$, we must have that $\mathrm{cost}_2(\tilde{P},\{c,-c\}) \notin (1\pm 3\varepsilon)\,\mathrm{cost}_2(\tilde{Q},\{c,-c\})$. Moreover, we can find the origin being $\lceil\frac{\gamma}{2}\rceil\cdot\mathbf{1}$ such that all the center points and data points have a distance to it less than $\frac{\gamma}{2} + \sqrt{d} \leq \gamma$. By Lemma 8,

$$sc(n,\gamma,2,2,d,\varepsilon) \geq \Omega\left(\log|\tilde{\mathcal{P}}|\right) = \Omega\left(\log|\mathcal{P}|\right) \geq \Omega\left(d\min\left\{\frac{1}{\varepsilon^2}, \frac{d}{\log d}, n\right\}\right).$$

The extension to any constant $z \geq 1$ can be found in the full version, which relies on the analysis of the Taylor expansion of function $f_z(x) = (1+x)^z$. The extension to general $k \geq 2$ can be found in the full version, whose main idea is to let every dataset consist of $\Theta(k)$ datasets from $\mathcal{P}$ and set the positions of their center points in $[\Delta]^d$ "remote" from each other.

Finally, we prove the lower bound $\Omega\left(k \log\log \frac{n}{k}\right)$. We again construct a large family $\mathcal{P}$ of datasets. For preparation, we find arbitrary $\frac{k}{2}$ points, denoted as $p_1, \cdots, p_{\frac{k}{2}}$, such that the distance between every two points is at least 10. This is available since $\Delta^d = \Omega(k)$. Every dataset $P \in \mathcal{P}$ is constructed as follow: for each $i \in \left[\frac{k}{2}\right]$, we select $m_i \in [\log \frac{n}{k}]$ and put 2^{m_i} points at $p_i + e_1$ and $\frac{2n}{k} - 2^{m_i}$ points at p_i. Therefore, the total number of possible datasets is $|\mathcal{P}| = \prod_{i=1}^{k} m_i = O\left(\left(\log \frac{n}{k}\right)^{\frac{k}{2}}\right)$. We then consider for any two different datasets $P, Q \in \mathcal{P}$, there must exist l such that P and Q have different assignments for p_l and $p_l + e_1$. Without loss of generality,assume that P put 2^i at $p_l + e_1$ while Q put 2^j for $i < j$. Choosing center set $C = \left\{p_1, p_1 + e_1, \cdots, p_l, p_l + 2e_1, \cdots, p_{\frac{k}{2}}, p_{\frac{k}{2}} + e_1\right\}$, we must have that $\mathrm{cost}_z(P, C) = 2^i \leq \frac{1}{2} 2^j \notin \left(1 \pm \frac{1}{2}\right) \mathrm{cost}_z(Q, C)$, which satisfies our requirement of $\mathcal{P}$. Lemma 8 provides us with a lower bound of $\Omega\left(\log |\mathcal{P}|\right) \geq \Omega\left(k \log\log \frac{n}{k}\right)$. ◀

3.2 Proof of Lemma 11: Principal Angles to Cost Difference

Let $P = \left\{p_i \in \mathbb{R}^d\right\}_{i \in [n]}$ and $Q = \left\{q_i \in \mathbb{R}^d\right\}_{i \in [n]}$ be two orthonormal bases. We have

$$\mathrm{cost}_2(P, \{c, -c\}) = \sum_{i=1}^{n} \left(\|p_i\|_2^2 + \|c\|_2^2 - 2\left|\langle p_i, c\rangle\right|\right) = 2n - 2\sum_{i=1}^{n} \left|\langle p_i, c\rangle\right| \leq 2n, \forall c \in \mathbb{R}^d,$$

$$\mathrm{cost}_2(P, \{c, -c\}) - \mathrm{cost}_2(Q, \{c, -c\}) = 2\left(\sum_{i=1}^{n} \left|\langle q_i, c\rangle\right| - \left|\langle p_i, c\rangle\right|\right).$$

Then it suffices to show the existence of a unit vector c such that $\sum_{i=1}^{n} |\langle q_i, c\rangle| - \sum_{i=1}^{n} |\langle p_i, c\rangle| \geq \frac{1}{4}\sqrt{n}$. To this end, we first show that large principal angles imply that most rows of the inner product matrix U have a small ℓ_2^2-norm.

▶ **Lemma 13** (Principal angles to row norms). *Let P, Q be datasets of n orthonormal bases following the condition that $\theta_{\frac{1}{32}10^{-6} \cdot n}(P, Q) \geq \arccos\left(\frac{10^{-3}}{4\sqrt{2}}\right)$. There exists a set K of size larger than $\left(1 - \frac{10^{-4}}{16}\right) n$ and having property that $\|U_i\|_2^2 := \sum_{j=1}^{n} (U_{ij})^2 = \begin{cases} \leq 10^{-2}, i \in K \\ \leq 1, i \notin K \end{cases}$.*

Proof. Proof of Lemma 13 can be found in the full version. ◀

Next, we show that a partial coloring can be found for the rows of U. We use a technique similar to [41].

▶ **Lemma 14** (Row norms to partial coloring). *Considering that we have a matrix $U_{n \times n}$ (with $n \geq 100$) such that we can find a set K of size larger than $\left(1 - \frac{10^{-4}}{16}\right) n$ and having the property that $\sum_{j=1}^{n} U_{ij}^2 = \begin{cases} \leq 10^{-2}, i \in K \\ \leq 1, i \notin K \end{cases}$. we can find a partial coloring $\{\lambda_j\}_{[n]} \in \{-1, 0, 1\}_{[n]}$ such that $|j : \lambda_j = 0| \leq \frac{1}{4} n$ and $\left|\sum_{j=1}^{n} \lambda_j U_{ij}\right| \leq \frac{1}{2}, \forall j \in [n]$.*

Proof. Proof of Lemma 14 can be found in the full version. ◀

By Lemmas 13 and 14, we are ready to prove Lemma 11.

Proof of Lemma 11. By Lemmas 13 and 14, there exists a partial coloring $\{\lambda_j\}_{[n]} \in \{-1,0,1\}_{[n]}$ such that $|j : \lambda_j = 0| \le \frac{1}{4}n$ and $\left|\sum_{j=1}^n \lambda_j U_{ij}\right| \le \frac{1}{2}, \forall i \in [n]$. Let $\tilde{c} = \sum_{j=1}^n \frac{\lambda_j}{\sqrt{n}} q_j$.

$$\mathrm{cost}_2(P,\{\tilde{c},-\tilde{c}\}) - \mathrm{cost}_2(Q,\{\tilde{c},-\tilde{c}\}) = 2\sum_{i=1}^n |\langle q_i,\tilde{c}\rangle| - \sum_{i=1}^n |\langle p_i,\tilde{c}\rangle|$$

$$= 2\sum_{j=1}^n \left|\frac{\lambda_j}{\sqrt{n}}\right| - \sum_{i=1}^n \left|\langle p_i, \sum_{j=1}^n \frac{\lambda_j}{\sqrt{n}} q_j\rangle\right| = 2\sum_{j=1}^n \left|\frac{\lambda_j}{\sqrt{n}}\right| - \sum_{i=1}^n \left|\sum_{j=1}^n \frac{\lambda_j}{\sqrt{n}} U_{ij}\right|.$$

Due to our choice of $\{\lambda_j\}_{[n]}$, we would have that

$$\sum_{j=1}^n \left|\frac{\lambda_j}{\sqrt{n}}\right| = \left(1 - \frac{|j:\lambda_j = 0|}{n}\right)\sqrt{n} \ge \frac{3}{4}\sqrt{n},$$

$$\sum_{i=1}^n \left|\sum_{j=1}^n \frac{\gamma_j}{\sqrt{n}} U_{ij}\right| = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left|\sum_{j=1}^n \lambda_j U_{ij}\right| \le \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{1}{2} = \frac{1}{2}\sqrt{n}.$$

Combining together, we would have that $\mathrm{cost}_2(P,\{\tilde{c},-\tilde{c}\}) - \mathrm{cost}_2(Q,\{\tilde{c},-\tilde{c}\}) \ge 2(\frac{3}{4}\sqrt{n} - \frac{1}{2}\sqrt{n}) = \frac{1}{2}\sqrt{n}$. To complete proof, note that $\|\tilde{c}\|_2 = \sqrt{\sum_{k=1}^n \frac{\lambda_j^2}{n}} \le \sqrt{\sum_{j=1}^n \frac{1}{n}} = 1$. Since $d > 2n$, we can always find a vector $\hat{c}$ such that $\hat{c} \perp P, Q$ and $\|\hat{c}\|_2^2 = 1 - \|\tilde{c}\|_2^2 \ge 0$. Let $c = \tilde{c} + \hat{c}$, we should have that c is of unit norm and

$$\mathrm{cost}_2(P,\{c,-c\}) - \mathrm{cost}_2(Q,\{c,-c\}) = \mathrm{cost}_2(P,\{\tilde{c},-\tilde{c}\}) - \mathrm{cost}_2(Q,\{\tilde{c},-\tilde{c}\}) \ge \frac{1}{2}\sqrt{n}. \quad ◀$$

3.3 Proof of Lemma 12: Construction of A Large Family $\mathcal{P}$

We first show that the principal angles are likely to be large between random subspaces.

▶ **Lemma 15** (Large principal angles between random subspaces). *Let $\mathcal{X}, \mathcal{Y}$ be two subspaces chosen from the uniform distribution on the Grassmann manifold of n-planes in $\mathbb{R}^d$ $\left(n = O\left(\frac{d}{\log d}\right)\right)$ endowed with its canonical metric. We have*

$$\Pr\left(\theta_{\frac{1}{32}10^{-6}\cdot n}(\mathcal{X},\mathcal{Y}) < \arccos\left(\frac{10^{-3}}{4\sqrt{2}}\right)\right) < \exp\left(-\frac{1}{128}10^{-6}\log\left(\frac{1}{1-\frac{1}{32}10^{-6}}\right)\cdot nd\right).$$

Proof. Random Matrix Theory e.g. [38] shows that with high probability the Hilbert-Schmidt norm of a random matrix is small. However, they only consider the square matrices and their results may not be strong enough to have a bound of $\exp(-O(nd))$. Therefore, we take advantage of the techniques in [1] which they use to calculate the distribution of the largest canonical angle. We generalized their results to bound the value of $\theta_{O(n)}$ with high probability. Proof of Lemma 15 can be found in the full version. ◀

With Lemma 15, we are ready to show the existence of a large enough family $\mathcal{P}$.

Proof of Lemma 12. We will generate $\exp\left(\frac{1}{256}10^{-6}\log\left(\frac{1}{1-\frac{1}{32}10^{-6}}\right)\cdot nd\right)$ sub-spaces, each of which is chosen from uniform distribution on the Grassmann manifold of n-planes in $\mathbb{R}^d$. We then let our dataset be arbitrary orthonormal bases of the subspace. Notice that by Lemma 10, the choice of the orthonormal bases will not affect the value of the principal angles. With the result of lemma 15, we have that for any two datasets P and Q,

$$\Pr\left(\theta_{\frac{1}{32}10^{-6}\cdot n}(P,Q) < \arccos\left(\frac{10^{-3}}{4\sqrt{2}}\right)\right) < \exp\left(-\frac{1}{128}10^{-6}\log\left(\frac{1}{1-\frac{1}{32}10^{-6}}\right)\cdot nd\right).$$

We then consider that

$$\begin{aligned}
&\Pr\left(\forall P \neq Q \in \mathcal{P}, \theta_{\frac{1}{32}10^{-6}\cdot n}(P,Q) \geq \arccos\left(\frac{10^{-3}}{4\sqrt{2}}\right)\right)\\
&= 1 - \Pr\left(\exists P \neq Q \in \mathcal{P}, \theta_{\frac{1}{32}10^{-6}\cdot n}(P,Q) < \arccos\left(\frac{10^{-3}}{4\sqrt{2}}\right)\right)\\
&\geq 1 - \sum_{i\neq j} \Pr\left(\theta_{\frac{1}{32}10^{-6}\cdot n}(P,Q) < \arccos\left(\frac{10^{-3}}{4\sqrt{2}}\right)\right)\\
&> 1 - \left(\exp\left(\frac{1}{256}10^{-6}\log\left(\frac{1}{1-\frac{1}{32}10^{-6}}\right)\cdot nd\right)\right)^2\\
&\qquad \cdot \exp\left(-\frac{1}{128}10^{-6}\log\left(\frac{1}{1-\frac{1}{32}10^{-6}}\right)\cdot nd\right) = 0.
\end{aligned}$$

Therefore, there is a positive probability for us to find enough datasets that fulfill our requirement, which shows the existence of the family. ◀

4 Conclusions and Future Work

In this study, we initiate the exploration of space complexity for the Euclidean (k,z)-Clustering problem, presenting both upper and lower bounds. Our findings suggest that a coreset serves as the optimal compression scheme when k is constant. Furthermore, the space lower bounds for (k,z)-Clustering directly imply a tight space lower bound for terminal embedding when $d \geq \Omega(\frac{\log n \log(n/\varepsilon)}{\varepsilon^2})$. The techniques we employ for establishing these lower bounds contribute to a deeper geometric understanding of principal angles, which may be of independent research interest.

Our work opens up several interesting research directions. One immediate challenge is to further narrow the gap between the upper and lower bounds of the space complexity for Euclidean (k,z)-Clustering. Additionally, it would be valuable to investigate whether a coreset remains optimal for compression when k is large.

References

1. P-A Absil, Alan Edelman, and Plamen Koev. On the largest principal angle between random subspaces. *Linear Algebra and its applications*, 414(1):288–294, 2006.
2. Noga Alon and Bo'az Klartag. Optimal compression of approximate inner products and dimension reduction. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 639–650. IEEE, 2017.
3. David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2007.
4. Maria-Florina F Balcan, Steven Ehrlich, and Yingyu Liang. Distributed k-means and k-median clustering on general topologies. *Advances in neural information processing systems*, 26, 2013.
5. Ake Björck and Gene H Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of computation*, 27(123):579–594, 1973.
6. Vladimir Braverman, Dan Feldman, Harry Lang, and Daniela Rus. Streaming coreset constructions for m-estimators. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2019)*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019.

7 Vladimir Braverman, Dan Feldman, Harry Lang, Adiel Statman, and Samson Zhou. New frameworks for offline and streaming coreset constructions. *arXiv preprint*, 2016. arXiv:1612.00889.

8 Vladimir Braverman, Gereon Frahling, Harry Lang, Christian Sohler, and Lin F. Yang. Clustering high dimensional dynamic data streams. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 576–585. PMLR, 2017.

9 Charles Carlson, Alexandra Kolla, Nikhil Srivastava, and Luca Trevisan. Optimal lower bounds for sketching graph cuts. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2565–2569. SIAM, 2019.

10 Moses Charikar and Erik Waingarten. The Johnson-Lindenstrauss lemma for clustering and subspace approximation: From coresets to dimension reduction. *arXiv preprint*, 2022. arXiv:2205.00371.

11 Xiaoyu Chen, Shaofeng H.-C. Jiang, and Robert Krauthgamer. Streaming Euclidean max-cut: Dimension vs data reduction. In *STOC*, pages 170–182. ACM, 2023.

12 Yeshwanth Cherapanamjeri and Jelani Nelson. Terminal embeddings in sublinear time. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1209–1216. IEEE, 2022.

13 Adam Coates and Andrew Y Ng. Learning feature representations with k-means. In *Neural Networks: Tricks of the Trade: Second Edition*, pages 561–580. Springer, 2012.

14 Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, and Chris Schwiegelshohn. Towards optimal lower bounds for k-median and k-means coresets. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1038–1051, 2022.

15 Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, Chris Schwiegelshohn, and Omar Ali Sheikh-Omar. Improved coresets for Euclidean k-Means. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, 2022.

16 Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. A new coreset framework for clustering. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 169–182, 2021.

17 Vincent Cohen-Addad, David P Woodruff, and Samson Zhou. Streaming euclidean k-median and k-means with o (log n) space. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 883–908. IEEE, 2023.

18 Artur Czumaj, Christiane Lammersen, Morteza Monemizadeh, and Christian Sohler. $(1+\varepsilon)$-approximation for facility location in data streams. In *SODA*, pages 1710–1728. SIAM, 2013.

19 Daniel Dadush, Haotian Jiang, and Victor Reis. A new framework for matrix discrepancy: partial coloring bounds via mirror descent. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 649–658, 2022.

20 G Dahlquist, B Sjöberg, and P Svensson. Comparison of the method of averages with the method of least squares. *Mathematics of Computation*, 22(104):833–845, 1968.

21 Michael Elkin, Arnold Filtser, and Ofer Neiman. Terminal embeddings. *Theoretical Computer Science*, 697:1–36, 2017.

22 Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 569–578, 2011.

23 Gereon Frahling, Piotr Indyk, and Christian Sohler. Sampling in dynamic data streams and applications. *Int. J. Comput. Geom. Appl.*, 18(1/2):3–28, 2008.

24 Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-medians clustering. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 291–300, 2004.

25 Monika Henzinger and Sagar Kale. Fully-dynamic coresets. *arXiv preprint*, 2020. arXiv:2004.14891.

26 Wei Hu, Zhao Song, Lin F. Yang, and Peilin Zhong. Nearly optimal dynamic k-Means clustering for high-dimensional data. *arXiv: Data Structures and Algorithms*, 2018. URL: https://api.semanticscholar.org/CorpusID:127972547.

27 Lingxiao Huang, Ruiyuan Huang, Zengfeng Huang, and Xuan Wu. On coresets for clustering in small dimensional Euclidean spaces. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 13891–13915. PMLR, 2023.

28 Lingxiao Huang, Jian Li, and Xuan Wu. On optimal coreset construction for Euclidean (k, z)-clustering, 2022. arXiv:2211.11923.

29 Lingxiao Huang and Nisheeth K Vishnoi. Coresets for clustering in Euclidean spaces: importance sampling is nearly optimal. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1416–1429, 2020.

30 Piotr Indyk and Tal Wagner. Approximate nearest neighbors in limited space. In *Conference On Learning Theory*, pages 2012–2036. PMLR, 2018.

31 Piotr Indyk and Tal Wagner. Optimal (Euclidean) metric compression. *SIAM Journal on Computing*, 51(3):467–491, 2022.

32 William B Johnson, Joram Lindenstrauss, and Gideon Schechtman. Extensions of lipschitz maps into Banach spaces. *Israel Journal of Mathematics*, 54(2):129–138, 1986.

33 Camille Jordan. Essai sur la géométrie à n dimensions. *Bulletin de la Société mathématique de France*, 3:103–174, 1875.

34 Daniel Kane, Raghu Meka, and Jelani Nelson. Almost optimal explicit Johnson-Lindenstrauss families. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*, pages 628–639. Springer, 2011.

35 Kasper Green Larsen and Jelani Nelson. Optimality of the Johnson-Lindenstrauss lemma. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 633–638. IEEE, 2017.

36 Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

37 Konstantin Makarychev, Yury Makarychev, and Ilya Razenshteyn. Performance of Johnson-Lindenstrauss transform for k-means and k-medians clustering. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1027–1038, 2019.

38 Elizabeth S. Meckes. *The Random Matrix Theory of the Classical Compact Groups*. Cambridge Tracts in Mathematics. Cambridge University Press, 2019.

39 Shyam Narayanan and Jelani Nelson. Optimal terminal dimensionality reduction in Euclidean space. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1064–1069, 2019.

40 Christian Sohler and David P. Woodruff. Strong coresets for k-Median and subspace approximation: Goodbye dimension. In *FOCS*, pages 802–813. IEEE Computer Society, 2018.

41 Joel Spencer. Six standard deviations suffice. *Transactions of the American mathematical society*, 289(2):679–706, 1985.

42 James M Varah. Computing invariant subspaces of a general matrix when the eigensystem is poorly conditioned. *Mathematics of Computation*, 24(109):137–149, 1970.