Dimension Reduction for Clustering: The Curious Case of Discrete Centers

Shaofeng H.-C. Jiang
Peking University, Beijing, China

Robert Krauthgamer
The Harry Weinrebe Professorial Chair of Computer Science, Weizmann Institute of Science, Rehovot, Israel

Shay Sapir
Weizmann Institute of Science, Rehovot, Israel

Sandeep Silwal
University of Wisconsin-Madison, WI, USA

Di Yue
Peking University, Beijing, China

Abstract

The Johnson-Lindenstrauss transform is a fundamental method for dimension reduction in Euclidean spaces, that can map any dataset of n points into dimension $O(\log n)$ with low distortion of their distances. This dimension bound is tight in general, but one can bypass it for specific problems. Indeed, tremendous progress has been made for clustering problems, especially in the *continuous* setting where centers can be picked from the ambient space $\mathbb{R}^d$. Most notably, for k-median and k-means, the dimension bound was improved to $O(\log k)$ [Makarychev, Makarychev and Razenshteyn, STOC 2019].

We explore dimension reduction for clustering in the *discrete* setting, where centers can only be picked from the dataset, and present two results that are both parameterized by the doubling dimension of the dataset, denoted as ddim. The first result shows that dimension $O_\varepsilon(\text{ddim} + \log k + \log\log n)$ suffices, and is moreover tight, to guarantee that the cost is preserved within factor $1 \pm \varepsilon$ for every set of centers. Our second result eliminates the $\log\log n$ term in the dimension through a relaxation of the guarantee (namely, preserving the cost only for all approximately-optimal sets of centers), which maintains its usefulness for downstream applications.

Overall, we achieve strong dimension reduction in the discrete setting, and find that it differs from the continuous setting not only in the dimension bound, which depends on the doubling dimension, but also in the guarantees beyond preserving the optimal value, such as which clusterings are preserved.

2012 ACM Subject Classification Theory of computation → Random projections and metric embeddings

Keywords and phrases dimension reduction, clustering, k-median, k-means, doubling dimension

Digital Object Identifier 10.4230/LIPIcs.ITCS.2026.82

Related Version *Full Version*: `https://arxiv.org/abs/2509.07444` [23]

Funding *Shaofeng H.-C. Jiang*: Work partially supported by a national key R&D program of China No. 2021YFA1000900.
Robert Krauthgamer: Work partially supported by the Israel Science Foundation grant #1336/23.

Acknowledgements The authors would like to thank Lingxiao Huang and Yury Makarychev for insightful discussions during the initial stage of this work.

1 Introduction

Oblivious dimension reduction, in the spirit of the Johnson and Lindenstrauss (JL) Lemma [24], is a fundamental technique for many Euclidean optimization problems over large, high-dimensional datasets. It has a strong guarantee: there is a random linear map $\pi : \mathbb{R}^d \to \mathbb{R}^t$, for a suitable target dimension $t = O(\varepsilon^{-2} \log n)$, such that for every n-point dataset $P \subset \mathbb{R}^d$, with high probability, π preserves all pairwise distances in P within factor $1 \pm \varepsilon$:

$$\forall x, y \in P, \qquad \|\pi(x) - \pi(y)\| \in (1 \pm \varepsilon)\|x - y\|,$$

where throughout $\|\cdot\|$ is the Euclidean norm. This guarantee is extremely powerful, particularly for algorithms: to solve a Euclidean problem on input P, one can apply the map π, solve the same problem on $\pi(P)$, which is often more efficient since $\pi(P)$ lies in low dimension, and "lift" the solution back to the original dimension (as discussed further in Section 1.2).

However, many problems require computational resources that grow exponentially with the dimension (the curse of dimensionality), and hence even dimension $t = O(\varepsilon^{-2} \log n)$ might be too large. Unfortunately, this dimension bound is tight in general, i.e., for preserving all pairwise distances [29], but interestingly one may bypass it for *specific* optimization problems, by showing that the optimal value/solution is preserved even when the dimension is reduced beyond the JL Lemma, say to dimension $t = O(\varepsilon^{-2})$, which is completely independent of n. This raises an important question:

> For which problems does dimension $o(\varepsilon^{-2} \log n)$ suffice for oblivious dimension reduction?

Prior work has revealed an affirmative answer for several key problems, as we discuss below. This paper studies this question for fundamental clustering problems, captured by (k, z)-clustering, which includes the famous k-means and k-median problems as its special cases. In (k, z)-clustering, the input is a dataset $P \subset \mathbb{R}^d$, and the goal is to find a set of centers C of size $|C| \leq k$ that minimizes

$$\mathrm{cost}^z(P, C) := \sum_{p \in P} \mathrm{dist}^z(p, C), \quad \text{where} \quad \mathrm{dist}^z(p, C) := \min_{c \in C} \|p - c\|^z.$$

We can distinguish two variants, differing in their space of potential centers. In the *continuous* variant, C is a subset of $\mathbb{R}^d$ (the centers lie in the ambient space), and in the *discrete* variant, also called sometimes *k-medoids*, C is a subset of P (or maybe of a larger set given as input). A key feature of the discrete version, is that $\pi : P \to \pi(P)$ is invertible, hence each potential center in $\pi(P)$ corresponds to a unique potential center in P (in contrast, a potential center in the ambient space $\mathbb{R}^t$ has many preimages in $\mathbb{R}^d$). Thus, in the discrete version, a set of centers computed for the dataset $\pi(P)$ can be mapped back to the higher dimension and serve as centers for the dataset P. See Section 1.3 for a discussion on practical applications of the discrete variant.

The continuous variant is a success story of the "beyond JL" program. A series of papers [3, 6, 2, 32] has culminated showing that target dimension $t = O(\varepsilon^{-2} \log \frac{k}{\varepsilon})$, which is independent of n, suffices to preserve all the solutions within factor $1 \pm \varepsilon$. Curiously, Charikar and Waingarten [4] observed that the discrete variant *behaves very differently*: certain instances require $t = \Omega(\log n)$, even for $k = 1$ (when using the standard Gaussian-based map π). Counterintuitively, restricting the centers to be data points makes dimension reduction significantly harder!

To bypass this limitation, we consider the doubling dimension, which was identified in previous work as a natural parameter that is very effective in achieving "beyond JL" bounds [20, 33, 22, 18, 14]. Formally, the *doubling dimension* of P, denoted $\mathrm{ddim}(P)$, is the smallest positive number such that every ball in the finite metric P can be covered by $2^{\mathrm{ddim}(P)}$ balls of half the radius. For several problems, including nearest neighbor [20], facility location [33, 18], and maximum matching [14], target dimension $t = O(\varepsilon^{-2} \log \frac{1}{\varepsilon} \cdot \mathrm{ddim}(P))$ suffices. Note that restricting the doubling dimension does not immediately imply a better dimension reduction of the JL flavor, as there are datasets $P \subset \mathbb{R}^d$ with $\mathrm{ddim}(P) = O(1)$ where no linear map can approximately preserve all pairwise distances (see e.g., [20, Remark 4.1]).

1.1 Main results

We present the first dimension reduction results for discrete (k, z)-clustering, along with matching lower bounds. Our first result (Theorem 1.1) provides a strong approximation guarantee, but requires a $\log \log n$ term in the target dimension, which we show is necessary. Our main result (Theorem 1.2) avoids this $\log \log n$ term, through a relaxation of the guarantee that maintains its algorithmic usefulness, e.g., it still implies that the optimal value is preserved up to factor $1 \pm \varepsilon$.

In all our results, the random linear map π is given by a matrix $G \in \mathbb{R}^{t \times d}$ of iid Gaussians $N(0, \frac{1}{t})$, which we refer to as a *Gaussian JL map*. This is nowadays a standard JL map [19, 13], and our results may extend to other JL maps, similarly to prior work in this context. We denote the optimal value of discrete (k, z)-clustering by

$$\mathrm{opt}^z(P) = \min_{C \subset P, |C| = k} \mathrm{cost}^z(P, C),$$

however for sake of exposition, we omit z and focus on $z = 1$ or $z = 2$, which are discrete k-median and k-means. We use the notation $\tilde{O}(f)$ to hide factors that are logarithmic in f, although below it only hides a $\log \frac{1}{\varepsilon}$ factor.

▶ **Theorem 1.1** (Informal version of Theorem 3.1). *For suitable* $t = \tilde{O}(\varepsilon^{-2}(\mathrm{ddim}(P) + \log k + \log \log n))$*, with probability at least* $2/3$*,*

1. $\mathrm{opt}(G(P)) \le (1 + \varepsilon)\,\mathrm{opt}(P)$*, and*
2. *for all* $C \subseteq P, |C| \le k$*, we have* $\mathrm{cost}(G(P), G(C)) \ge (1 - \varepsilon)\,\mathrm{cost}(P, C)$*.*

This theorem has immediate algorithmic applications. First, it implies that the optimal value is preserved, i.e., $\mathrm{opt}(G(P)) \in (1 \pm \varepsilon)\,\mathrm{opt}(P)$. Second, for every $C \subset P$ and $\beta > 1$, if the set of centers $G(C)$ is a β-approximate solution for the instance $G(P)$, then C is a $(1 + O(\varepsilon))\beta$-approximate solution for the instance P. Therefore, the theorem fit into the general paradigm of using oblivious linear maps – apply the mapping, solve the problem in low dimension, and lift the centers back to the higher dimension.

It is interesting to compare our result with the continuous variant of (k, z)-clustering. On the one hand, to preserve the optimal value in the continuous variant, we know from [32] that target dimension $O(\varepsilon^{-2} \log \frac{k}{\varepsilon})$ suffices, independently of $\mathrm{ddim}(P)$. On the other hand, Theorem 1.1 further provides a "for all centers" guarantee, which is not attainable in the continuous version (by any linear map), by simply considering centers in the kernel of the linear map (see Theorem 6.1). We examine and discuss these guarantees more carefully in Section 1.2.

Matching lower bounds. The results in Theorem 1.1 are nearly tight for Gaussian JL maps, and likely for all oblivious linear maps. It is known that achieving $\mathrm{opt}(G(P)) \in (1 \pm \varepsilon)\,\mathrm{opt}(P)$ requires target dimension $t = \Omega(\log k)$, even for a dataset P of doubling dimension $O(1)$ [33], and another known lower bound is that $t = \Omega(\mathrm{ddim}(P))$, even for $k = O(1)$ [4]. It is easy to tighten these bounds with respect to the dependence on ε. We complete the picture, and show in Theorem 6.2 the multiplicative approximation of Theorem 1.1 requires dimension $t = \Omega(\varepsilon^{-2} \log\log n)$, even for $k = 1$ and a dataset P of doubling dimension $O(1)$.

To get some intuition about the discrete variant, we briefly recall the hard instance of [4], taking $z = 1$ for simplicity. Consider $k = 2$, and let P be the first n standard basis vectors, thus $\mathrm{ddim}(P) = \log n$. The pairwise distances all equal $\sqrt{2}$, hence $\mathrm{opt}(P) = \sqrt{2} \cdot (n-2)$. The standard basis vectors form a well-known hard instance for the JL Lemma, hence, when using target dimension $t = o(\varepsilon^{-2} \log n)$, with high probability, there exists $j_1 \in [\frac{n}{2}]$ such that $\|Ge_{j_1}\| < 1 - 10\varepsilon$. Similarly, let $j_2 > \frac{n}{2}$ be such an index for the last $\frac{n}{2}$ standard basis vectors. Let Ge_{j_1}, Ge_{j_2} be the two centers for $G(P)$, and assign the first $\frac{n}{2}$ basis vectors to Ge_{j_2} and the last $\frac{n}{2}$ vectors to Ge_{j_1}. Now a simple argument using the independence between the two halves shows that $\mathrm{opt}(G(P)) \le (1-\varepsilon)\sqrt{2} \cdot n \le (1 - \varepsilon/2)\,\mathrm{opt}(P)$ with probability $2/3$.

A relaxed guarantee. Our main result avoids the $\log\log n$ term in Theorem 1.1 by slightly relaxing the guarantee, while keeping it useful for downstream applications.

▶ **Theorem 1.2** (Informal version of Theorem 5.1). *For suitable $t = \tilde{O}(\varepsilon^{-2}(\mathrm{ddim}(P) + \log k))$, with probability at least $2/3$,*

1. $\mathrm{opt}(G(P)) \le (1+\varepsilon)\,\mathrm{opt}(P)$*, and*
2. *for all $C \subseteq P, |C| \le k$, we have* $\mathrm{cost}(G(P), G(C)) \ge \min\{(1-\varepsilon)\,\mathrm{cost}(P, C), 100\,\mathrm{opt}(P)\}$.

This theorem implies that the optimal value is preserved, i.e., $\mathrm{opt}(G(P)) \in (1 \pm \varepsilon)\,\mathrm{opt}(P)$. Let us further examine which solutions are preserved under this guarantee: For all $C \subset P$ and $1 < \beta < \frac{100}{1+\varepsilon}$, if the set of centers $G(C)$ is a β-approximate solution for the instance $G(P)$, then C is a $(1 + O(\varepsilon))\beta$-approximate solution for the instance P. Recall that for Theorem 1.1, we had a similar claim, but without the restriction $\beta < \frac{100}{1+\varepsilon}$. The constant 100 here is arbitrary, and can be changed to any $\alpha > 2$, at the cost of increasing the target dimension by an additive $O(\varepsilon^{-2} \log\log \alpha)$ term.

1.2 Various notions for preserving solutions

We study several definitions for dimension reduction for k-clustering. All these definitions require (perhaps implicitly) that $\mathrm{opt}(G(P)) \le (1+\varepsilon)\,\mathrm{opt}(P)$, i.e., that the optimal value has bounded expansion. This direction is often easy because it suffices to analyze one optimal solution for P. In the other direction, one may naively require that $\mathrm{opt}(G(P)) \ge (1-\varepsilon)\,\mathrm{opt}(P)$, however this is rather weak, as it does not guarantee that *solutions* are preserved. Moreover, even requiring that an optimal solution for $G(P)$ is a near-optimal solution for P is quite limited, because a near-optimal solution for $G(P)$, say one found by a $(1+\varepsilon)$-approximation algorithm, may be lifted to a poor solution for P. In fact, such a phenomenon was observed for minimum spanning tree (MST) when using target dimension $t = o(\log n)$: an optimal MST of $G(P)$ is a $(1+\varepsilon)$-approximate MST of P, however a $(1+\varepsilon)$-approximate MST of $G(P)$ may have large cost for P [33]. Ideally, we want the cost of *every solution* to have bounded contraction, as it allows to lift any solution for $G(P)$ to a solution for P, and we thus consider several different notions for the set of solutions, as follows. For simplicity, we present these for $z = 1$ in the discrete setting, but they extend naturally to all $z \ge 1$ and to the continuous setting.

1. **Partitions.** A solution is a partition $\mathcal{P} = (P_1, \ldots, P_k)$ of P. Its cost is defined as $\text{cost}(\mathcal{P}) := \sum_{i=1}^{k} \min_{c \in P_i} \sum_{p \in P_i} \|p - c\|$.
2. **Centers.** A solution is a set of centers $C = (c_1, \ldots, c_k) \subseteq P$. Its cost is defined as $\text{cost}(P, C) := \sum_{p \in P} \text{dist}(p, C)$.
3. **Centers and partitions.** A solution is a partition $\mathcal{P} = (P_1, \ldots, P_k)$ of P and a set of centers $C = (c_1, \ldots, c_k) \subseteq P$. Its cost is defined as $\text{cost}(\mathcal{P}, C) := \sum_{i=1}^{k} \sum_{p \in P_i} \|p - c_i\|$.

These definitions are fairly natural, and were used in prior work on dimension reduction, e.g., partition-based solutions were used in [32] for k-means and k-median, and center-based solutions were used in [22] for k-center. It was observed in [4] that not all "for all" guarantees are the same; in particular, "for all centers" and "for all partitions" are incomparable. However,"for all centers and partitions" is clearly stronger than both.

Next, we define contraction for solutions, capturing the two notions in Theorems 1.1 and 1.2. The notion in Theorem 1.1 is simply of *multiplicative contraction*: A solution S has $(1-\varepsilon)$-contraction if $\text{cost}(G(S)) \geq (1-\varepsilon)\,\text{cost}(S)$. The notion in Theorem 1.2 is new, at least in the context of dimension reduction, and goes as follows.

▶ **Definition 1.3** (Relaxed Contraction). *A solution S has α-relaxed $(1-\varepsilon)$-contraction (for $\alpha > 1$, $\varepsilon > 0$) if* $\text{cost}(G(S)) \geq \min\{\alpha\,\text{opt}(P), (1-\varepsilon)\,\text{cost}(S)\}$.

Using these definitions, we can restate Theorem 1.1 as having $(1-\varepsilon)$-contraction for all centers, and restate Theorem 1.2 as achieving 100-relaxed $(1-\varepsilon)$-contraction for all centers. In fact, we can strengthen Theorem 1.1 to assert $(1-\varepsilon)$-contraction for all *centers and partitions*.

▶ **Theorem 1.4** (Strengthened Theorem 1.1, informal). *For suitable $t = \tilde{O}(\varepsilon^{-2}(\text{ddim}(P) + \log k + \log\log n))$, with probability at least 2/3, for all partitions $\mathcal{P} = (P_1, \ldots, P_k)$ of P and sets of centers $C = (c_1, \ldots, c_k) \subseteq P$,*

$$\text{cost}(G(\mathcal{P}), G(C)) \geq (1-\varepsilon)\,\text{cost}(\mathcal{P}, C).$$

This strengthening is not attainable for Theorem 1.2, as dimension $\Omega(\varepsilon^{-2}\log\log n)$ is needed to get a "for all centers and partitions" guarantee, even for relaxed contraction (see Theorem 6.3). However, we do not know if a "for all partitions" guarantee is possible without the $\log\log n$ term. If it is possible, then a curious phenomenon will occur: we get a "for all partitions" and a "for all centers" guarantees, but not a "for all centers and partitions" guarantee. All our results are summarized in Table 1.

Candidate centers. We consider also a more general variant of k-clustering, where the candidate centers are part of the input (given either explicitly or implicitly): Given a dataset P and candidate-centers set Q, the goal is to find $C \subseteq Q$ of size $|C| \leq k$ that minimizes $\sum_{p \in P} \text{dist}^z(p, C)$. When $Q = \mathbb{R}^d$ or $Q = P$, we obtain the continuous and discrete variants, respectively.

We observe a slightly different phenomenon in terms of the attainable contraction: to get $(1-\varepsilon)$-contraction, one needs target dimension $\Theta(\varepsilon^{-2}\log|Q|)$, and the lower bound holds even when both P and Q are doubling and $k = 1$. We can still obtain claims analogous to Theorems 1.2 and 1.4, albeit with relaxed contraction: a "for all partitions and centers" using dimension $t = \tilde{O}(\varepsilon^{-2}(\text{ddim}(P \cup Q) + \log k + \log\log n))$, and a "for all centers" for the same target dimension but without the $\log\log n$ term. See Table 1 for references.

Table 1 Summary of our results for dimension reduction for k-clustering. The notions of "for all" centers and/or partitions, and of multiplicative/relaxed contraction are as explained in Section 1.2. Some lower bounds apply even for preserving the optimal value; for clarity, it is noted in the table they hold "even for value". In the setting of candidate centers, the size of the candidate set is denoted by s. Suppressing $\log \frac{1}{\varepsilon}$ terms and the dependence on α for α-relaxed contraction.

Problem	Target dimension	∀ partitions	∀ centers	contraction	Reference
Continuous	$O(\varepsilon^{-2} \log k)$	yes	no	multiplicative	[32]
	$\Omega(\varepsilon^{-2} \log k)$	no	no	even for value	[33]
	$> d - 1$	no	yes	even for relaxed	Thm 6.1
Discrete	$O(\varepsilon^{-2}(\mathrm{ddim} + \log k + \log\log n))$	yes	yes	multiplicative	Thm 3.1
	$O(\varepsilon^{-2}(\mathrm{ddim} + \log k))$	no	yes	relaxed	Thm 5.1
	?	yes	no	any	OPEN
	$\Omega(\varepsilon^{-2} \log\log n)$	yes	yes	even for relaxed	Thm 6.3
	$\Omega(\varepsilon^{-2} \log\log n)$	no	yes	multiplicative	Thm 6.2
	$\Omega(\varepsilon^{-2} \log k)$	no	no	even for value	[33]
	$\Omega(\varepsilon^{-2} \mathrm{ddim})$	no	no	even for value	[4]
Candidate centers	$O(\varepsilon^{-2} \log s)$	yes	yes	multiplicative	Thm 4.1
	$O(\varepsilon^{-2}(\mathrm{ddim} + \log k + \log\log n))$	yes	yes	relaxed	Thm 4.2
	$O(\varepsilon^{-2}(\mathrm{ddim} + \log k))$	no	yes	relaxed	Thm 5.1
	$\Omega(\varepsilon^{-2} \log s)$	no	yes	multiplicative	Thm 6.4

1.3 Other related work

Besides the aforementioned results for "beyond JL" for clustering problems, there are also several improved bounds for other classes of problems such as Max-Cut [28, 27, 5], numerical linear algebra [31, 37, 7], and other applications [1, 15, 21].

The discrete k-median problem in Euclidean space was originally shown to be NP-hard by Papadimitriou, even for the case of $d = 2$ [34]. In terms of hardness of approximation, the current state of the art is that one cannot approximate the discrete k-means or k-median problem beyond 1.07 and 1.17, respectively, assuming P $\neq$ NP [11, 12]. As for upper bounds, the best approximation factors known in polynomial time are $2 + \varepsilon$ for any fixed $\varepsilon > 0$ for discrete Euclidean k-median [10] and 5.912 for discrete Euclidean k-means [8]. There are also algorithms that achieve $1 + \varepsilon$ approximation (again in the discrete case) in time that is doubly exponential in the doubling dimension, see [9] for a thorough discussion.

The discrete variant that we study may also be preferred over the continuous version in certain applications. First, it is thought to be less sensitive to outliers in practice than the continuous version [35, 25]. Second, in applications where cluster centers are used as data summarization, interpretability might require the centers to be part of the dataset. For example, in applications based on machine-learning embeddings of objects such as text [38], an arbitrary vector in the embedding space might not represent any actual object. A similar issue arises for structured data such as sparse data or images, e.g., the "average image" is visually random noise [30, 36] or the average of sparse vectors is not necessarily sparse. A discrete center, however, represents an actual underlying object, and thus preserves the underlying properties of the input points.

1.4 Technical overview

Since the dimension-error tradeoff behaves differently between the discrete and continuous settings, it is not surprising that our results for the discrete setting require new techniques. To simplify the discussion, we focus on the k-medoids ($z = 1$) case, and an alternative guarantee that only preserves the optimal *value*, i.e.,

$$\mathrm{opt}(G(P)) \in (1 \pm \varepsilon)\,\mathrm{opt}(P), \tag{1}$$

with target dimension bound $t = \tilde{O}(\varepsilon^{-2}(\mathrm{ddim}(P) + \log k))$ which is the same as that in Theorem 1.2. While this is a weaker guarantee than both Theorem 1.1 and Theorem 1.2, it already introduces major technical challenges, and the techniques for this claim covers most of our new ideas.

We begin our discussion with the case $k = 1$. We first argue that even for this case, a natural framework based on extension theorems (which has been used in previous works on dimension reduction for clustering) fails in our discrete case.

Failure of extension theorems in the discrete setting. To prove (1) (and possibly more general claims), a natural framework based on *extension theorems* have been widely used in dimension reduction for clustering. Specifically, given an arbitrary center v in the target space (e.g., v is the optimal 1-median center of $G(P)$), one can define an "inverse image" u in the original space such that $\mathrm{cost}(P, u) \le (1+\varepsilon)\,\mathrm{cost}(G(P), v)$, and this directly implies $\mathrm{opt}(G(P)) \ge \frac{1}{1+\varepsilon}\,\mathrm{opt}(P)$. The key step of defining "inverse image" is precisely what an extension theorem does. This framework is widely used in prior works such as [32, 22, 4], in the spirit of the classic Kirszbraun extension theorem [26] or the robust one-point extension theorem [32, Theorem 5.2]. However, such extension theorems are only known to work in the continuous setting, which require to pick the inverse image $u \in \mathbb{R}^d$ from the entire $\mathbb{R}^d$ and cannot be restricted only to the data points $u \in P$.[1]

Our techniques. We start with $k = 1$ case (a detailed discussion can be found in Section 1.4.1). In this case, we first obtain a target dimension bound with an $O(\log\log n)$ factor, by utilizing the existence of a small movement-based coreset. A coreset is a small accurate proxy of the dataset, and the movement-based coreset additionally requires the existence of a "local" mapping such that each data point can be mapped to a nearby coreset point. The dimension reduction simply preserves the pairwise distance on the coreset, and (1) is argued via the local mapping. A conceptually similar coreset-to-dimension-reduction idea has been employed in [4] for continuous k-median, and one main difference is that we also utilize the locality property (movement-based) of the coreset, see a more detailed comparison in Section 1.4.1.

Then, to remove the $O(\log\log n)$ factor, we consider a weaker guarantee as in Theorem 1.2, where we prove the $(1+\varepsilon)$ relative error only for near-optimal solutions, and for the other solutions we have a flat $100\,\mathrm{opt}(P)$ error. This relaxed guarantee is strong enough for (1) (and many other applications), which may be of independent interest to further studies. Our analysis is crucially built on this small vs large cost case, albeit we also need to consider the middle ground of the mix of the two.

Finally, we discuss the generalization to $k > 1$ in Section 1.4.2, which introduces several nontrivial technical complications from $k = 1$.

1.4.1 The $k = 1$ case

The easy side of (1) is the upper bound $\mathrm{opt}(G(P)) \le (1+\varepsilon)\,\mathrm{opt}(P)$, even for the general k case. The reason is that it suffices to preserve the cost w.r.t. an optimal center set C^*, and since C^* is a fixed solution, even a target dimension $t = O(\varepsilon^{-2}\log(1/\varepsilon))$ will be sufficient. This is a standard argument also observed in prior works. The lower bound

[1] We note that the Kirszbraun theorem may be adapted to work for the discrete case when the target dimension $t = O(\log n)$, but this dimension bound is too large to be useful.

$\mathrm{opt}(G(P)) \geq (1-\varepsilon)\,\mathrm{opt}(P)$ is the major challenge. To prove this inequality, we want to preserve the clustering cost w.r.t. the optimal center set of $G(P)$, denoted by C. Since C is a random set that depends on G, preserving its cost is almost the same as preserving the cost of *all* center sets, which is exactly the guarantee 2 of Theorems 1.1 and 1.2.

To introduce our new techniques, we first establish a weaker target dimension bound of $O(\varepsilon^{-2}(\mathrm{ddim} + \log\log n))$, and this part contains main ideas for proving Theorem 1.1. We then overview the key steps to eliminate the extra $\log\log n$ term, which also reflects how we prove Theorem 1.2.

The $O(\log\log n)$ bound: from coreset to dimension reduction. To prove (1), we use an approach inspired by the movement-based coreset construction in Euclidean spaces [17]. Roughly speaking, a movement-based coreset[2] is a subset $S \subseteq P$, such that there exists a mapping $\sigma\colon P \to S$ satisfying $\sum_{p\in P} \|p-\sigma(p)\| \leq O(\varepsilon)\,\mathrm{opt}(P)$. Our framework is summarized as follows: we first construct a movement-based coreset S to compress the dataset P. Next, we apply the standard JL lemma to preserve pairwise distances in the coreset S within $(1\pm\varepsilon)$, which requires $O(\varepsilon^{-2}\log|S|)$ target dimensions. After this step, the optimal value of S is already preserved, namely, $\mathrm{opt}(G(S)) \in (1\pm\varepsilon)\,\mathrm{opt}(S)$. Finally, it suffices to show that the cost of snapping data points to their nearest neighbor in S (i.e., $\sum_{p\in P}\|p-S(p)\|$ and $\sum_{p\in P}\|Gp-GS(p)\|$) is negligible in both original and target spaces.

The construction of the coreset is essentially the same as that in [17], except that [17] also assigns weight to the coreset points and here we only need the point set itself. We review the construction. This construction is based on a sequence of *nets*, a standard tool for discretizing metrics. Formally, a ρ-net of a point set P is a subset $N \subseteq P$, such that 1) the interpoint distances in N are at least ρ, and 2) every point in P has a point in N within distance ρ. (See the more detailed definition in Definition 2.3). Denote $c^* \in P$ as an optimal discrete 1-median center. We construct nets on a sequence of balls centered at c^* with geometrically decreasing radii. Denote $r_0 := \mathrm{opt}(P)$ and $r_\ell := r_0/2^\ell$ for $\ell = 1, 2, \ldots, \log n$. Construct the level ℓ net N_ℓ as an εr_ℓ-net on the ball $B(c^*, r_\ell)$, and denote $N := \bigcup_{\ell=0}^{\log n} N_\ell$ to be the union of all $\log n$ levels of nets.

By the standard packing property of doubling metrics, each net has size $|N_\ell| \leq O(\varepsilon^{-O(\mathrm{ddim})})$, thus $|N| \leq O(\varepsilon^{-O(\mathrm{ddim})}\log n)$, which implies a target dimension $t = O(\varepsilon^{-2}(\mathrm{ddim}\log\varepsilon^{-1} + \log\log n))$. On the other hand, let $G(c) \in G(P)$ be an optimal discrete 1-median center of $G(P)$. Then the total cost of snapping c and all data points to the nearest neighbor in N (i.e., $\sum_{p\in P}(\|p-N(p)\| + \|c-N(c)\|)$) can be bounded by $O(\varepsilon)(\mathrm{opt}(P) + \mathrm{cost}(P,c))$ in the original space. Based on results in [20], we further show that this snapping cost in the target space (i.e., $\sum_{p\in P}(\|Gp-GN(p)\| + \|Gc-GN(c)\|)$) can increase by at most a constant factor.

Finally, we note that the above analysis can be applied to obtain the "for all centers" guarantee in Theorem 1.1, or even the stronger "for all centers and partitions" guarantee in Theorem 1.4.

Comparison to [4]. The locality property (movement-based) of the coreset is crucial to our analysis. To see this, let us investigate the proof in [4], which also uses coresets for dimension reduction but only works for the continuous setting. Specifically, it shows the existence of a coreset S of P which with high probability satisfies: (i) S is a coreset of P; (ii) the pairwise distances within S are preserved by G within $(1\pm\varepsilon)$-factor; and (iii) $G(S)$

[2] This definition is tailored to our need and may be slightly different to that in the literature.

is a coreset of $G(P)$. To see why this implies the desired bound $\mathrm{opt}(G(P)) \geq (1-\varepsilon)\,\mathrm{opt}(P)$, let $v \in \mathbb{R}^m$ be an optimal center for continuous 1-median of $G(P)$. By property (iii), $\mathrm{opt}(G(P)) \geq (1-\varepsilon)\,\mathrm{cost}(G(S), v)$. Property (ii) and the Kirszbraun Extension Theorem [26] imply, as explained in [4], the existence of $u \in \mathbb{R}^d$ (intuitively, a pre-image of v) such that $\mathrm{cost}(G(S), v) \geq (1-\varepsilon)\,\mathrm{cost}(S, u)$. Finally, by property (i) we can bound $\mathrm{cost}(S, u) \geq (1-\varepsilon)\,\mathrm{cost}(P, u) \geq (1-\varepsilon)\,\mathrm{opt}(P)$.

This elegant argument for the continuous 1-median breaks for discrete variant at the step of finding u that is a pre-image of an optimal center $v \in G(P)$. This issue may seem surprising, because $u := G^{-1}(v) \in P$ is a natural candidate. However, it is not clear whether $\mathrm{cost}(G(S), v) \geq (1-\varepsilon)\,\mathrm{cost}(S, u)$ holds, because property (ii) only preserves pairwise distances within S, and it is possible that $v \notin G(S)$. Another option is to define u via the Kirszbraun Extension Theorem, but then u is not guaranteed to be in P, i.e., might be an infeasible solution, hence we can no longer argue that $\mathrm{cost}(P, u) \geq \mathrm{opt}(P)$. Therefore, in the discrete setting we need to utilize more structures of the coreset, and the locality properties of movement-based coresets are key for resolving the issue.

Removing the $\log\log n$ term via relaxed guarantee Let us first recall the cause of the $\log\log n$ term. We apply the JL Lemma to N, which is a union of $\log n$ nets, each of size $\varepsilon^{-O(\mathrm{ddim})}$. The $\log\log n$ thus comes from a union bound over all $\log n$ levels. To bypass this union bound, we use two technical ideas. First, we avoid touching cross-level pairs and only apply the union bound for each N_ℓ separately. This requires us to always snap p and c to the same level of net when handling each $p \in P$. Second, for a single level, we analyze its maximum distance distortion which is a random variable, and bound the expectation. We remark that some levels will be distorted significantly, but the average distortion is $(1+O(\varepsilon))$. Similar ideas have been used by prior works (e.g., [14]).

Consider the following two extremes. First, suppose c is the closest point to c^*, say, $\forall p \in P, \|c-c^*\| \leq \|p-c^*\|$. For every $p \in P$, we can snap p to its nearest neighbor in net N_p. Observe that c can also be covered by N_p. The cost of snapping p and c can both be bounded by $O(\varepsilon)\cdot\|p-c^*\|$, and we show that *on average*, the cost of snapping Gp and Gc is bounded by $O(\varepsilon)\cdot\|p-c^*\|$ as well, which adds up to $O(\varepsilon)\,\mathrm{opt}(P)$. The other extreme is that c is very far from c^*, i.e, $\|c-c^*\| > \mathrm{opt}(P)/10$. In this case, we can no longer snap c to the same net as p (like the previous case). We show that in this case, $\mathrm{cost}(G(P), Gc) \geq 100\,\mathrm{opt}(P)$.

If c does not fall into any of the above two extremes, our analysis is a combination of them. Indeed, we show the *relaxed* "for all centers" guarantee,

$$\forall c \in P, \qquad \mathrm{cost}(G(P), Gc) \geq \min\{(1-\varepsilon)\,\mathrm{cost}(P, c), 100\,\mathrm{opt}(P)\}. \tag{2}$$

Note that this is exactly the same as the guarantee 2 of Theorem 1.2, and that the two terms in the min correspond to the aforementioned two extremes, respectively. Specifically, we first specify a level ℓ and its corresponding radius r_ℓ. If $\|c-c^*\| > r_\ell$, then we fall into the second extreme and show that $\mathrm{cost}(G(P), Gc) \geq 100\,\mathrm{opt}(P)$. Otherwise, $\|c-c^*\| \leq r_\ell$, then we handle each $p \in P$ differently, depending on the distance $\|p-c^*\|$. If $\|p-c^*\| \geq r_\ell$, then we use the same argument as the first extreme – snapping both p and c to N_p, bounding the snapping cost, and analyzing the additive contraction. If $\|p-c^*\| < r_\ell$, then we snap both p and c to N_ℓ. Since ℓ is a fixed level, a union bound over N_ℓ is affordable and we obtain $\mathrm{cost}(G(P), Gc) \geq (1-\varepsilon)\,\mathrm{cost}(P, c)$ in this case.

1.4.2 Generalization to $k > 1$

Instead of directly generalizing (2), we first show a weaker guarantee: for target dimension $t = O(\varepsilon^{-2} \operatorname{ddim} \log k)$,

$$\forall C \subseteq P, |C| = k, \qquad \sum_{p \in P} \|Gp - GC(p)\| \geq \min\{(1-\varepsilon)\operatorname{cost}(P, C), 100 \operatorname{opt}(P)\}, \tag{3}$$

where $C(p)$ is the center in C closest to p. Note that (3) is weaker than what we desire in Theorem 1.2, for the following two reasons. First, the target dimension is worse than the $O(\varepsilon^{-2}(\operatorname{ddim} + \log k))$ in Theorem 1.2. Second, the left hand side of (3) can be much larger than $\operatorname{cost}(G(P), G(C))$, since the image of $C(p)$ under G (i.e., $GC(p)$) is not necessarily the nearest neighbor of Gp in $G(C)$. Nonetheless, the proof of (3) already captures most of our key ideas. In the end of this section, we briefly discuss how we obtain a sharper target dimension bound as well as a stronger guarantee.

Suppose $C^* \subseteq P$ is an optimal solution, which induces a clustering $\mathcal{C}^* = \{S_1^*, S_2^*, \ldots, S_k^*\}$. Our general proof framework is the same as the $k = 1$ case – considering the "distance" between C and C^*, if C is "far from" C^*, then we show $\operatorname{cost}(G(P), G(C)) \geq 100 \operatorname{opt}(P)$; otherwise we show $\operatorname{cost}(G(P), G(C)) \geq (1-\varepsilon)\operatorname{cost}(P, C)$.

However, an immediate issue is how to define that C and C^* are far from or close to each other. For each $i \in [k]$, we specify a "threshold level" of cluster S_i^*, denoted by ℓ_i. We say C is "far from" C^* if there exists $i \in [k]$, such that $\operatorname{dist}(c_i^*, C) > 10 r_{\ell_i}$. In this case, the cost of connecting $B(c_i^*, r_{\ell_i})$ to C is already high. We further prove that $\operatorname{cost}(G(P), G(C)) \geq 100 \operatorname{opt}(P)$, by careful analysis of the randomness of G.

Now suppose C is "close to" C^*, i.e., $\forall i \in [k], \operatorname{dist}(c_i^*, C) \leq 10 r_{\ell_i}$. Our key observation is that for every $p \in S_i^*$, $C(p)$ should also be close to c_i^*, i.e.,

$$\forall p \in S_i^*, \qquad \|C(p) - c_i^*\| \leq O(\max\{\|p - c_i^*\|, r_{\ell_i}\}). \tag{4}$$

As a natural generalization of the $k = 1$ case, we lower bound $\|Gp - GC(p)\|$ for $p \in S_i^*$ differently, depending on the distances $\|C(p) - c_i^*\|$. If $\|C(p) - c_i^*\| \geq r_{\ell_i}$, then we snap both p and $C(p)$ to the (enlarged) net N_p. (We can do this since (4) holds.) Otherwise, we snap both p and $C(p)$ to the (enlarged) net N_{ℓ_i}. The snapping cost and the distance contraction are bounded similarly to the $k = 1$ case. This simply introduces an extra $\log k$ factor in the target dimension.

Decoupling ddim from $\log k$. So far, we only obtain an $O_\varepsilon(\operatorname{ddim} \log k)$ bound, instead of $O_\varepsilon(\operatorname{ddim} + \log k)$. This is due to error accumulation: Recall we handle each (optimal) cluster S_i^* separately, each of which incurs an $O(\varepsilon)\operatorname{opt}(P)$ additive error; hence, we have to rescale ε by a $1/k$ factor to compensate the accumulated error of k clusters, resulting in an $O(\varepsilon^{-2} \operatorname{ddim} \log k)$ target dimension (naïvely, that results in $\tilde{O}(\varepsilon^{-2} k^2 \operatorname{ddim})$ target dimension, but this is avoided by an easy adaptation).

To decouple these two factors, we need more delicate analysis for the error. For "far" points $p \in S_i^*$ with $\|C(p) - c_i^*\| \geq r_{\ell_i}$, the snapping and distortion error is $O(\varepsilon)\|p - c_i^*\|$ in expectation, which adds up to $O(\varepsilon)\operatorname{opt}(P)$ and does not incur any error accumulation. However, the error accumulation happens for "close" points p with $\|C(p) - c_i^*\| < r_{\ell_i}$, where the snapping cost within a single cluster S_i^*, namely $\sum_{p \in S_i^*} \|p - N_\ell(p)\|$, is already $O(\varepsilon)\operatorname{opt}(P)$, which accumulates to $O(k\varepsilon)\operatorname{opt}(P)$.

To reduce the error accumulation, we further divide the close points (i.e., $\|C(p) - c_i^*\| < r_{\ell_i}$) into two ranges, namely, the *close range* $\|C(p) - c_i^*\| < r_{\ell_i}/k$ and the *middle range* $\|C(p) - c_i^*\| \in [r_{\ell_i}/k, r_{\ell_i}]$, and handle these two ranges differently. The cost of points in the

close range can be bounded by $O(\varepsilon/k)\operatorname{opt}(P)$, which adds up to $O(\varepsilon)\operatorname{opt}(P)$. For points in the middle range, we handle them in a point-by-point manner, at the cost of $\operatorname{poly}(k)e^{-\Omega(\varepsilon^2 t)}$ per point. Since there are at most $k \cdot O(\log k)$ levels in the middle range, a union bound over all net points at these levels will be affordable.

Handling nearest neighbor assignment in the target space. Recall that (4) conerns the cost $\|Gp - GC(p)\|$, which is the cost in the target space with respect to the nearest neighbor assignment in the *original* space. However, what we really need is the nearest neighbor assignment in the *target* space. To capture such misalignment in the original and target spaces, we define a mapping f to be the assignment in the target space, i.e., $f(p)$ is the center in C realizing $\operatorname{dist}(Gp, G(C))$, so that $\operatorname{cost}(G(P), G(C)) = \sum_{p\in P}\|Gp - Gf(p)\|$, and $f(p) = C(p)$ does not hold in general. We attempt to modify the previous analysis to lower bound each $\|Gp - Gf(p)\|$ instead of $\|Gp - GC(p)\|$.

To lower bound this distance, we attempt to replace every $C(p)$ with $f(p)$ in our previous proof. The analysis becomes problematic, as our structural observation (4) no longer holds if we change $C(p)$ to $f(p)$, and this turns out to be the only place where our analysis does not go through. To resolve this issue, let us focus on the bad scenario where $f(p)$ is sufficiently far from c_i^*, i.e., $\|f(p) - c_i^*\| \gg \max\{\|p - c_i^*\|, r_{\ell_i}\}$. This implies $f(p)$ is also far from p. We further show that $\|Gp - Gf(p)\| \gg \|p - c_i^*\|$ by careful analysis of G's randomness. On the other hand, we have $\|p - C(p)\| \le O(\|p - c_i^*\|)$ by (4). Therefore, we can directly lower bound $\|Gp - Gf(p)\|$ by $\|p - C(p)\|$ in this case.

2 Preliminaries

Consider a point set $P \subset \mathbb{R}^d$. For every $x \in \mathbb{R}^d$, denote by $P(x)$ the point in P closest to x and $\operatorname{dist}(x, P) := \|x - P(x)\|$ (recall that throughout $\|\cdot\|$ is the Euclidean norm). Denote $\operatorname{diam}(P) := \max\{\operatorname{dist}(p, q)\colon p, q \in P\}$ as the *diameter* of P. For $x \in \mathbb{R}^d$ and $r > 0$, denote by $B(x, r) := \{y \in \mathbb{R}^d\colon |x - y\| \le r\}$ the *ball* centered at x with radius r. Recall that for $k \in \mathbb{N}$ and $z \ge 1$, the (k, z)-clustering cost of P w.r.t. center set $C \subset \mathbb{R}^d, |C| \le k$ is $\operatorname{cost}_k^z(P, C) := \sum_{p\in P}\operatorname{dist}(p, C)^z$. The optimal discrete (k, z)-clustering cost of P w.r.t. a candidate center set $Q \subset \mathbb{R}^d$ is denoted by $\operatorname{opt}_k^z(P, Q) := \min_{C\subseteq Q, |C|\le k}\operatorname{cost}_k^z(P, C)$, and by $\operatorname{opt}(P, Q)$ for short when k, z are clear from the context. Denote $\operatorname{opt}(P) := \operatorname{opt}(P, P)$ and $\operatorname{opt-cont}(P) := \operatorname{opt}(P, \mathbb{R}^d)$ for simplicity.

We use the following generalized triangle inequalities.

▶ **Lemma 2.1** (Generalized triangle inequalities [32]). *Let (X, dist) be a metric space. Then for every $z \ge 1$, $\varepsilon \in (0, 1)$ and $p, q, r \in X$,*

$$\operatorname{dist}(p, q)^z \ge (1 - z\varepsilon)\operatorname{dist}(p, r)^z - \varepsilon^{-z}\operatorname{dist}(q, r)^z.$$

$$\operatorname{dist}(p, q)^z \le (1 + \varepsilon)^{z-1}\operatorname{dist}(p, r)^z + \left(\frac{1 + \varepsilon}{\varepsilon}\right)^{z-1}\operatorname{dist}(q, r)^z.$$

2.1 Doubling dimension and nets

▶ **Definition 2.2** (Doubling dimension [16]). *The* doubling dimension *of a set $P \subseteq \mathbb{R}^d$, denoted* $\operatorname{ddim}(P)$*, is the minimum $m > 0$, such that $\forall r > 0$, every ball in P with radius r can be covered by at most 2^m balls of radius $r/2$.*

Our proof uses ρ-nets for doubling sets, whose definition and key properties are described here.

▶ **Definition 2.3** (ρ-net). *Let $P \subseteq \mathbb{R}^d$ and $\rho > 0$. A subset $N \subseteq P$ is called a* ρ-packing *of P if $\forall u, v \in N,\ \|u - v\| > \rho$. The subset N is called a* ρ-covering *of P if $\forall x \in P$, there exists $u \in N$ such that $x \in B(u, \rho)$. The subset N is called a* ρ-net *of P if N is both a ρ-packing and ρ-covering of P.*

▶ **Lemma 2.4** (Packing property [16]). *Let $P \subseteq \mathbb{R}^d$ and $N \subseteq P$ be a ρ-packing of P. Then $|N| \leq (\operatorname{diam}(P)/\rho)^{O(\operatorname{ddim}(P))}$.*

2.2 Dimension reduction

For simplicity, we only consider random linear maps defined by a matrix of iid Gaussians, which are known to satisfy the JL Lemma [19, 13].

▶ **Definition 2.5.** *A Gaussian JL map is a $t \times d$ matrix with i.i.d. entries drawn from $N(0, \frac{1}{t})$.*

Recall the following concentration bound [20, Eq. (7)] (see also [33, Eq. (5)]), from which one can deduce the JL lemma.

▶ **Lemma 2.6** ([20, Eq. (7)]). *Let $x \in \mathbb{R}^d, \varepsilon > 0$ and a Gaussian JL map $G \in \mathbb{R}^{t \times d}$. We have*

$$\Pr(\|Gx\| \notin (1 \pm \varepsilon)\|x\|) \leq \exp(-\varepsilon^2 t/8).$$

The following two lemmas regard Gaussian JL maps when applied to doubling sets.

▶ **Lemma 2.7** ([20, Lemma 4.2]). *There exist universal constants $A_1, A_2 > 0$ such that for every subset $P \subset B(\vec{0}, 1)$ of the Euclidean unit ball in $\mathbb{R}^d$, $t > A_1 \cdot \operatorname{ddim}(P) + 1, D \geq 10$, and a Gaussian JL map $G \in \mathbb{R}^{t \times d}$,*

$$\Pr(\exists x \in P, \|Gx\| > D) \leq e^{-A_2 t D^2}.$$

▶ **Lemma 2.8** ([18, Lemma 3.21]). *There exists universal constants $A_1, A_2, L > 1$, such that for every $P \subset \mathbb{R}^d \setminus B(\vec{0}, 1), \varepsilon > 0, t > A_1 \operatorname{ddim}(P)$, and a Gaussian JL map $G \in \mathbb{R}^{t \times d}$,*

$$\Pr(\exists x \in P, \|Gx\| < \tfrac{1}{L}) \leq e^{-A_2 t}.$$

3 The first upper bound

We prove Theorem 1.1 (a.k.a Theorem 1.4) in this section, formally stated below.

▶ **Theorem 3.1.** *Let $\varepsilon > 0$, $z \geq 1$ and $d, \operatorname{ddim}, k \in \mathbb{N}$ and a Gaussian JL map $G \in \mathbb{R}^{t \times d}$ with suitable $t = O(z^2 \varepsilon^{-2}(\operatorname{ddim} \log(z/\varepsilon) + \log k + \log\log n))$. For every set $P \subseteq \mathbb{R}^d$ with $\operatorname{ddim}(P) \leq \operatorname{ddim}$, with probability at least $2/3$,*

1. $\operatorname{opt}_k^z(G(P)) \leq (1 + \varepsilon) \operatorname{opt}_k^z(P)$, *and*
2. *for all centers $C = (c_1, \ldots, c_k) \subseteq P$ and all partitions $\mathcal{P} = (S_1, \ldots, S_k)$ of P,*

$$\operatorname{cost}_k^z(G(\mathcal{P}), G(C)) \geq (1 - \varepsilon) \operatorname{cost}_k^z(\mathcal{P}, C),$$

where $\operatorname{cost}_k^z(\mathcal{P}, C) = \sum_{i=1}^k \sum_{p \in S_i} \|p - c_i\|^z$.

We use the following lemma to bound the clustering cost of a fixed set of centers and partition of P. The proof can be found in the full version [23].

▶ **Lemma 3.2.** *Let $\varepsilon > 0$, $z \geq 1$ and $d, k \in \mathbb{N}$ and a Gaussian JL map $G \in \mathbb{R}^{t\times d}$ with suitable $t = O(z^2\varepsilon^{-2}\log\varepsilon^{-1})$. For every set $P \subseteq \mathbb{R}^d$, every set of centers $(c_1, \ldots, c_k) \subset \mathbb{R}^d$ and every partition $\mathcal{P} = (S_1, \ldots, S_k)$ of P, with probability at least 9/10,*

$$\mathrm{cost}_k^z(G(\mathcal{P}), G(C)) \leq (1+\varepsilon)\,\mathrm{cost}_k^z(\mathcal{P}, C).$$

Proof of Theorem 3.1. Consider an optimal discrete k-median of P. Denote by $C^* = \{c_1^*, \ldots, c_k^*\} \subseteq P$ and by $S_1^*, \ldots, S_k^*$ the centers and clusters (respectively) in that solution. Applying Lemma 3.2 to the optimal center set C^* and the partition $\mathcal{P}^* = (S_1^*, \ldots, S_k^*)$, we have that with probability at least 9/10,

$$\mathrm{opt}(G(P)) \leq \mathrm{cost}(G(\mathcal{P}^*), G(C^*)) \leq (1+\varepsilon)\,\mathrm{cost}(\mathcal{P}^*, C^*) = (1+\varepsilon)\,\mathrm{opt}(P),$$

concluding the first part of the theorem.

Denote by r_0 the largest radius of any cluster S_i^*. Pick a suitable $m = O(\log n)$ such that $2^m = n^{10}$. For $i \in [0, m]$ and $j \in [k]$, set $r_i = r_0/2^i$, and $P_{ij} = S_j \cap B(c_j, r_i)$, i.e., for every cluster, we have a sequence of geometrically decreasing balls. Additionally, let N_i be an $\varepsilon^3 r_i$-net of $\cup_j P_{ij}$. By Lemma 2.4, $|N_i| \leq k\varepsilon^{-O(\mathrm{ddim}(P))}$.

For each $x, y \in \cup_{i\in[0,m]} N_i$, by Lemma 2.6,

$$\Pr(\|Gx - Gy\| > (1+\varepsilon)\|x-y\|) \leq \exp(-\varepsilon^2 t/8) \leq \frac{\varepsilon^{\Omega(\mathrm{ddim}(P))}}{k^2 m^2}.$$

Thus, by a union bound, w.p. at least 9/10,

$$\forall x, y \in \cup_{i\in[0,m]} N_i, \qquad \|Gx - Gy\| \leq (1+\varepsilon)\|x-y\|. \tag{5}$$

Furthermore, for each $i \in [0, m], y \in N_i$, by Lemma 2.7,

$$\Pr(\exists p \in P \cap B(y, \varepsilon^3 r_i) \text{ s.t. } \|G(p-y)\| > 10\varepsilon^3 r_i) \leq \exp(-\Omega(t)).$$

By a union bound, w.p. at least 9/10,

$$\forall i \in [0, m], y \in N_i, p \in P \cap B(y, \varepsilon^3 r_i), \qquad \|G(p-y)\| \leq 10\varepsilon^3 r_i. \tag{6}$$

By another union bound, Equations (5) and (6) hold with probability at least 2/3.

We are now ready to prove the second part of the theorem. Let $C = \{c_1, \ldots, c_k\} \subseteq P$ and let a partition $\mathcal{P} = (S_1, \ldots, S_k)$ of P. For every $p \in P$ we denote by u_p the nearest net-point to p in the level such that $P_i \setminus P_{i+1}$ contains p, and the radius of that level is denoted r_p. Denote by $f(p)$ the center in C assigned to p according to the partition $\mathcal{P}$. Recall that $C^*(p)$ is a point in C^* that is nearest to p. Observe that

$$\sum_{p\in P} r_p^z \leq n \cdot \left(\frac{r_0}{n^{10}}\right)^z + \sum_{j=1}^{k}\sum_{i=0}^{m-1} \sum_{p\in P_{i,j}\setminus P_{i+1,j}} (2\|p - c_j^*\|)^z = O(2^z)\cdot \mathrm{opt}(P),$$

and

$$\begin{aligned}
(\tfrac{1}{2} r_{f(p)})^z &\leq \|f(p) - C^*(f(p))\|^z && \text{by definition} \\
&\leq \|f(p) - C^*(p)\|^z && C^*(f(p)) \text{ is nearest to } f(p) \text{ from } C^* \\
&\leq 2^{z-1}\|p - f(p)\|^z + 2^{z-1}\|p - C^*(p)\|^z && \text{by Lemma 2.1.}
\end{aligned} \tag{7}$$

Therefore,

$$\begin{aligned}
&\operatorname{cost}(G(\mathcal{P}), G(C)) \\
&\equiv \sum_{p\in P} \|Gp - Gf(p)\|^z \\
&\geq \sum_{p\in P} (1-z\varepsilon)\|Gu_p - Gu_{f(p)}\|^z - \varepsilon^{-z}\|Gp - Gu_p\|^z - \varepsilon^{-z}\|Gf(p) - Gu_{f(p)}\|^z \\
&\geq \sum_{p\in P} (1-z\varepsilon)(1-\varepsilon)^z\|u_p - u_{f(p)}\|^z - \varepsilon^{-z}(10\varepsilon^3 r_p)^z - \varepsilon^{-z}(10\varepsilon^3 r_{f(p)})^z \\
&\geq \sum_{p\in P} (1-z\varepsilon)^2(1-\varepsilon)^z\|p - f(p)\|^z - O(\varepsilon)^z r_p^z - O(\varepsilon)^z r_{f(p)}^z \\
&\geq \sum_{p\in P} (1-3z\varepsilon)\|p - f(p)\|^z - O(\varepsilon)^z r_p^z - O(\varepsilon)^z 2^{2z-1}(\|p - f(p)\|^z + \|p - C^*(p)\|^z) \\
&\geq (1 - O(z\varepsilon))\operatorname{cost}(\mathcal{P}, C) - O(\varepsilon)\cdot\operatorname{opt}(P),
\end{aligned}$$

where the first and the third inequalities follow from the generalized triangle inequality (Lemma 2.1), the second inequality follows from (5) and (6) and the fourth inequality follows from (7). Rescaling $\varepsilon \to \varepsilon/z$ concludes the proof. ◀

4 General candidate centers

We now consider a generalization of Theorem 3.1, to the setting where the centers are from a (possibly different than the input) candidate set Q. Unfortunately, to obtain multiplicative contraction in this setting, we have to pay $\Theta(\varepsilon^{-2}\log|Q|)$ in the target dimension. We state the upper bound below, and the matching lower bound is provided in Theorem 6.4. The proofs appear in the full version [23].

▶ **Theorem 4.1.** *Let $\varepsilon > 0$, $z \geq 1$ and $d, k, s \in \mathbb{N}$ and a Gaussian JL map $G \in \mathbb{R}^{t\times d}$ with suitable $t = O(z^2\varepsilon^{-2}(\log s + z\log(z/\varepsilon)))$. For every set $P \subseteq \mathbb{R}^d$ and every candidate center set $Q \subseteq \mathbb{R}^d$ with $|Q| = s \geq k$, with probability at least $2/3$,*

1. $\operatorname{opt}_k^z(G(P), G(Q)) \leq (1+\varepsilon)\operatorname{opt}_k^z(P, Q)$, *and*
2. *for every $C = (c_1, \ldots, c_k) \subseteq Q$ and every partition $\mathcal{P} = (S_1, \ldots, S_k)$ of P,*

$$\operatorname{cost}_k^z(G(\mathcal{P}), G(C)) \geq (1-\varepsilon)\operatorname{cost}_k^z(\mathcal{P}, C),$$

where $\operatorname{cost}_k^z(\mathcal{P}, C) = \sum_{i=1}^k \sum_{p\in S_i} \|p - c_i\|^z$.

To bypass the $O(\varepsilon^{-2}\log|Q|)$ barrier in the target dimension, we consider relaxed contraction, and prove the following.

▶ **Theorem 4.2.** *Let $\varepsilon > 0$, $z \geq 1$ and $d, \operatorname{ddim}, k \in \mathbb{N}$ and a Gaussian JL map $G \in \mathbb{R}^{t\times d}$ with suitable $t = O(z^2\varepsilon^{-2}(\operatorname{ddim}\log(z/\varepsilon) + \log k + \log\log\alpha + \log\log n))$. For every n-point set $P \subseteq \mathbb{R}^d$ and every candidate center set $Q \subseteq \mathbb{R}^d$ with $\operatorname{ddim}(P\cup Q) \leq \operatorname{ddim}$, with probability at least $2/3$,*

1. $\operatorname{opt}_k^z(G(P), G(Q)) \leq (1+\varepsilon)\operatorname{opt}_k^z(P, Q)$, *and*
2. *for every $C = (c_1, \ldots, c_k) \subseteq Q$ and every partition $\mathcal{P} = (S_1, \ldots, S_k)$ of P,*

$$\operatorname{cost}_k^z(G(\mathcal{P}), G(C)) \geq \min\{\alpha\cdot\operatorname{opt}_k^z(P, Q), (1-\varepsilon)\operatorname{cost}_k^z(\mathcal{P}, C)\},$$

where $\operatorname{cost}_k^z(\mathcal{P}, C) = \sum_{i=1}^k \sum_{p\in S_i} \|p - c_i\|^z$.

5 Improved upper bound: Removing the $\log\log n$ term

We prove Theorem 1.2 in this section. In fact, we prove the following for the more general candidate centers setting, and get Theorem 1.2 by setting $Q = P$.

▶ **Theorem 5.1.** *Let $0 < \varepsilon < \frac{1}{2}$, $z \geq 1$, $\alpha > 2$ and $d, \mathrm{ddim} \in \mathbb{N}$ and a Gaussian JL map $G \in \mathbb{R}^{t\times d}$ with suitable $t = O(z^2\varepsilon^{-2}(\mathrm{ddim}\log(z/\varepsilon) + z\log(z/\varepsilon) + \log k + \log\log\alpha))$, the following holds. For every $P, Q \subseteq \mathbb{R}^d$ with $\mathrm{ddim}(P\cup Q) \leq \mathrm{ddim}$, with probability at least $2/3$,*

1. $\mathrm{opt}_k^z(G(P), G(Q)) \leq (1+\varepsilon)\,\mathrm{opt}_k^z(P,Q)$, *and*
2. $\forall C \subseteq Q, |C| = k$,

$$\mathrm{cost}_k^z(G(P), G(C)) \geq \min\{\alpha\cdot\mathrm{opt}_k^z(P,Q), (1-\varepsilon)\,\mathrm{cost}_k^z(P,C)\}.$$

Consider an optimal discrete k-median of P w.r.t. candidate center set Q. Denote by $C^* = \{c_1^*, \ldots, c_k^*\} \subseteq Q$ and by $S_1^*, \ldots, S_k^*$ the centers and clusters (respectively) in that solution. Denote $r_0 := \mathrm{opt}(P,Q)^{1/z}$. For $\ell \in \mathbb{N}$ and $i \in [k]$, set $r_\ell = r_0/2^\ell$, and $P_\ell^i = S_i^* \cap B(c_i^*, r_\ell)$, i.e., for every cluster, we have a sequence of geometrically decreasing balls. Additionally, let N_ℓ^i be an $\varepsilon^3 r_\ell$-net of $(P\cup Q)\cap B(c_i^*, r_{\ell-\log\varepsilon^{-1}})$. Let $N_\ell := \bigcup_i N_\ell^i$.

For $p \in P\cup Q$, recall $C^*(p)$ is the closest center to p in C^*. Let $j_p \in \mathbb{N}$ be the level satisfying $r_{j_p+1} \leq \|p - C^*(p)\| \leq r_{j_p}$. Denote $r_p := r_{j_p}$ for simplicity. We have the following claim.

▶ **Lemma 5.2.** $\sum_{p\in P} r_p^z \leq 2^z\,\mathrm{opt}_k^z(P,Q)$.

For $C \subseteq Q$ and $p \in P$, recall we denote by $C(p)$ the point closest to p in C. We have the following lemma that upper bounds the distance from $C(p)$ to $C^*(p)$ (and also the distance from $C(p)$ to p).

▶ **Lemma 5.3.** *Let $C \subseteq Q$. Then for every $i \in [k]$ and $p \in S_i^*$, it holds that $\|C(p) - c_i^*\| \leq 4\max\{r_p, \|c_i^* - C(c_i^*)\|\}$.*

Proof.

$$\begin{aligned}
\|C(p) - c_i^*\| &\leq \|C(p) - p\| + \|p - c_i^*\| && \text{by triangle inequality}\\
&\leq \|C(c_i^*) - p\| + \|p - c_i^*\| && C(p) \text{ is the point closest to } p \text{ in } C\\
&\leq \|C(c_i^*) - c_i^*\| + \|c_i^* - p\| + \|p - c_i^*\| && \text{by triangle inequality}\\
&\leq 4\max\{r_p, \|c_i^* - C(c_i^*)\|\}. &&
\end{aligned}$$

◀

Proof of Theorem 5.1. The first guarantee is the same as Theorem 3.1, so we omit its proof and focus on the second guarantee. For a generic solution $C \subseteq Q, |C| = k$, denote $C = \{c_1, c_2, \ldots, c_k\}$. Denote $f(p) := G^{-1}(GC(Gp))$, i.e., $f(p)$ is a center in C realizing $\mathrm{dist}(Gp, G(C))$. For $j \in [k]$, denote $S_j := \{p \in P\colon f(p) = c_j\}$ as the cluster induced by c_j.

For every $i \in [k]$, define the "threshold level" of cluster i as

$$\ell_i := \max\{\ell\colon |P_\ell^i|\cdot r_\ell^z > \alpha\,\mathrm{opt}(P,Q)\}. \tag{8}$$

We also define the i-th "buffer" as $I_i := [\ell_i - \log(2000L^2), \ell_i + \log(\alpha k)]$, where L is the (sufficiently large) constant in Lemma 2.8.

For $0 \le \ell \le m$, denote random variable β_ℓ to be the minimum real, such that $\forall u, v \in N_\ell,\ \|Gu - Gv\| \ge (1 - \varepsilon - \beta_\ell \varepsilon)\|u - v\|$. Denote random variable γ_ℓ to be the minimum real, such that $\forall u \in N_\ell, v \in B(u, \varepsilon^3 r_\ell),\ \|Gu - Gv\| \le \gamma_\ell \varepsilon^3 r_\ell$. For $p \in P \cup Q$, write $\beta_p := \beta_{j_p}$ and $\gamma_p := \gamma_{j_p}$ for simplicity.

In the following lemma, we define our good events and bound their success probability. The proof can be found in the full version [23].

▶ **Lemma 5.4.** *With probability at least* 0.99*, the following events happen simultaneously.*

(a) $\sum_{p\in P} \beta_p r_p^z \le e^{-\Omega(\varepsilon^2 t)} \cdot \mathrm{opt}(P,Q)$*, and* $\sum_{p\in P} \gamma_p^z r_p^z \le 10^z \cdot O(\mathrm{opt}(P,Q))$.

(b) $\forall i \in [k], \forall \ell \in I_i,\ \forall u \in N_\ell, v \in B(u, \varepsilon^3 r_\ell),\ \|Gu - Gv\| \le 10\varepsilon^3 r_\ell$.

(c) $\forall i \in [k], \forall \ell \in I_i$*, every net point* $u \in N_\ell$ *satisfies that* $\forall P' \subseteq P$*,*

$$\sum_{p\in P'} \|Gp - Gu\|^z \ge (1-\varepsilon)^{3z} \sum_{p\in P'} \|p - u\|^z - \frac{\varepsilon}{k^2}\,\mathrm{opt}(P,Q).$$

(d) $\forall i \in [k], \forall y \in B(c_i^*, 40L \cdot r_{\ell_i}),\ \|Gy - Gc_i^*\| \le 400L \cdot r_{\ell_i}$.

(e) $\forall i \in [k], \forall y \in (P \cup Q) \setminus B(c_i^*, 2000L^2 \cdot r_{\ell_i}),\ \|Gy - Gc_i^*\| > 2000L \cdot r_{\ell_i}$.

(f) *For* $p \in P$*, denote by random variable* $\xi_p := \min_{y:\ \|y-p\| > 9L\cdot r_{\ell_i}} \|Gy - Gp\|$*. Then* $\forall i \in [k]$*,*

$$\sum_{p\in P^i_{\ell_i}} \xi_p^z > \alpha\,\mathrm{opt}(P,Q).$$

(g) *For* $p \in P$*, denote* $\eta_p := \min_{y:\ \|y-p\| > 9Lr_p} \|Gy - Gp\|$*. Then* $\forall i \in [k]$*,*

$$\sum_{p\in S_i^*} \max\{0, (9r_p)^z - \eta_p^z\} \le e^{-\Omega(t)} \cdot \sum_{p\in S_i^*} r_p^z.$$

The proof proceeds by a careful case analysis.

Case 1, one cluster with no cover: $\max_{1\le i\le k}\{\|c_i^* - C(c_i^*)\| - 10L \cdot r_{\ell_i}\} > 0$**.** Then there exists $i \in [k]$, such that $\|c_i^* - C(c_i^*)\| > 10L \cdot r_{\ell_i}$. Intuitively, this means all points in C are far away from c_i^*. Write

$$\mathrm{cost}(G(P), G(C)) \ge \mathrm{cost}(G(P^i_{\ell_i}), G(C)) = \sum_{p\in P^i_{\ell_i}} \|Gp - Gf(p)\|^z. \tag{9}$$

Note that for every $p \in P^i_{\ell_i}$,

$$\begin{aligned}
\|p - f(p)\| &\ge \|p - C(p)\| \\
&\ge \|C(p) - c_i^*\| - \|p - c_i^*\| \\
&\ge \|c_i^* - C(c_i^*)\| - \|p - c_i^*\| \\
&> 10L \cdot r_{\ell_i} - r_p \\
&\ge 9L \cdot r_{\ell_i}.
\end{aligned}$$

Therefore, $\|Gp - Gf(p)\| \ge \xi_p$. Combining with (9) yields

$$\mathrm{cost}(G(P), G(C)) \ge \sum_{p\in P^i_{\ell_i}} \|Gp - Gf(p)\|^z \ge \sum_{p\in P^i_{\ell_i}} \xi_p^z > \alpha\,\mathrm{opt}(P,Q),$$

where the last inequality follows from event f.

Case 2, $\max_{1\le i\le k}\{\|c_i^*-C(c_i^*)\|-10L\cdot r_{\ell_i}\}\le 0$. Then for every $i\in[k]$, $\|c_i^*-C(c_i^*)\|\le 10L\cdot r_{\ell_i}$, which intuitively means every center in C^* has a nearby neighbor in C.

Comparing "fake" centers to optimal centers. Let $i\in[k]$. For every $p\in S_i^*$, we consider the distance of p's "fake" center $f(p)$ (recall, $Gf(p)$ realizes $\mathrm{dist}(Gp,G(C))$) from p's optimal center c_i^*. There are three ranges we consider for $\|f(p)-c_i^*\|$.

Define $R_i:=\{p\in S_i^*:\, r_{\ell_i}/(\alpha k)\le\|f(p)-c_i^*\|\le 2000L^2\cdot r_{\ell_i}\}$, and denote $R:=\bigcup_{i=1}^k R_i$ (called "the middle range"). Moreover, define $T_i:=\{p\in S_i^*:\,\|f(p)-c_i^*\|\le r_{\ell_i}/(\alpha k)\}$, and denote $T:=\bigcup_{i=1}^k T_i$ (called "the close range").

Case 2.1, the middle range $p\in R$. Let us first lower bound $\|Gp-Gf(p)\|$ for $p\in R$. Assume $C^*(p)=c_i^*$ and $f(p)=c_j$, where $i,j\in[k]$. Since $p\in R_i$, we can assume $r_{\ell+1}<\|c_j-c_i^*\|\le r_\ell$ for some level $\ell\in I_i$. Let $u_{i,j}$ be the net point in N_ℓ closest to c_j. Then

$$\begin{aligned}
\|Gp-Gf(p)\|^z &\ge (1-z\varepsilon)\|Gp-Gu_{i,j}\|^z-\varepsilon^{-z}\|Gc_j-Gu_{i,j}\|^z\\
&\ge (1-z\varepsilon)\|Gp-Gu_{i,j}\|^z-\varepsilon^{-z}(10\varepsilon^3 r_\ell)^z\\
&\ge (1-z\varepsilon)\|Gp-Gu_{i,j}\|^z-O(\varepsilon)^{2z}\|c_j-c_i^*\|^z\\
&\ge (1-z\varepsilon)\|Gp-Gu_{i,j}\|^z-O(\varepsilon)^{2z}\|p-c_j\|^z-O(\varepsilon)^{2z}\|p-c_i^*\|^z\\
&= (1-z\varepsilon)\|Gp-Gu_{i,j}\|^z-O(\varepsilon)^{2z}\|p-f(p)\|^z-O(\varepsilon)^{2z}\|p-C^*(p)\|^z,
\end{aligned}$$

where the first and the last inequalities follow from Lemma 2.1 and the second inequality follows from event b. Summing over $p\in R$, we have

$$\begin{aligned}
&\sum_{p\in R}\|Gp-Gf(p)\|^z\\
=&\sum_{i=1}^k\sum_{j=1}^k\sum_{p\in R_i\cap S_j}\|Gp-Gc_j\|^z\\
\ge&\sum_{i=1}^k\sum_{j=1}^k\sum_{p\in R_i\cap S_j}\Big((1-z\varepsilon)\|Gp-Gu_{i,j}\|^z-O(\varepsilon)^{2z}\|p-f(p)\|^z-O(\varepsilon)^{2z}\|p-C^*(p)\|^z\Big)\\
\ge&\sum_{i=1}^k\sum_{j=1}^k\sum_{p\in R_i\cap S_j}(1-z\varepsilon)\|Gp-Gu_{i,j}\|^z-O(\varepsilon)^{2z}\sum_{p\in R}\|p-f(p)\|^z-O(\varepsilon)^{2z}\,\mathrm{opt}(P,Q)
\end{aligned}$$

Applying event c to net point $u_{i,j}$ and subset $R_i\cap S_j$, we have

$$\begin{aligned}
&\ge\sum_{i=1}^k\sum_{j=1}^k\left((1-O(z\varepsilon))\sum_{p\in R_i\cap S_j}\|p-u_{i,j}\|^z-\frac{\varepsilon}{k^2}\,\mathrm{opt}(P,Q)\right)\\
&\quad-O(\varepsilon)\sum_{p\in R}\|p-f(p)\|^z-O(\varepsilon)\,\mathrm{opt}(P,Q)\\
&\ge(1-O(z\varepsilon))\sum_{p\in R}\|p-f(p)\|^z-O(\varepsilon)\,\mathrm{opt}(P,Q)\\
&\ge(1-O(z\varepsilon))\sum_{p\in R}\|p-C(p)\|^z-O(\varepsilon)\,\mathrm{opt}(P,Q).
\end{aligned}\tag{10}$$

Case 2.2, the close range $p \in T$. This is somewhat of a special case of Case 2.1. Assume $C^*(p) = c_i^*$ and $f(p) = c_j$, where $i, j \in [k]$. Since $p \in R_i$, we have $\|c_j - c_i^*\| \le r_\ell$ for $\ell = \ell_i + \log(\alpha k)$. Let $u_{i,j}$ be the net point in N_ℓ closest to c_j. We have,

$$\begin{aligned}\|Gp - Gf(p)\|^z &\ge (1 - z\varepsilon)\|Gp - Gu_{i,j}\|^z - \varepsilon^{-z}\|Gc_j - Gu_{i,j}\|^z \\ &\ge (1 - z\varepsilon)\|Gp - Gu_{i,j}\|^z - O(\varepsilon)^{2z} r_\ell^z.\end{aligned}$$

If $p \notin B(c_i^*, r_{\ell_i+1})$, then

$$r_\ell \le \tfrac{1}{2k}\|p - c_i^*\| \le \tfrac{1}{2k}(\|p - c_j\| + \|c_j - c_i^*\|) \le \tfrac{1}{2k}(\|p - c_j\| + r_\ell).$$

Rearranging, we obtain $r_\ell \le \|p - c_j\|$. Summing over $p \in T$, we have

$$\begin{aligned}&\sum_{p\in T} \|Gp - Gf(p)\|^z \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{p\in T_i\cap S_j} \|Gp - Gc_j\|^z \\ &\ge \sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{p\in T_i\cap S_j} \Big((1 - z\varepsilon)\|Gp - Gu_{i,j}\|^z \\ &\qquad - O(\varepsilon)^{2z}\|p - c_j\|^z\Big) - O(\varepsilon)^{2z}\Big|P \cap B(c_i^*, r_{\ell_i+1})\Big| \cdot r_\ell^z \\ &\ge \sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{p\in T_i\cap S_j} \Big((1 - z\varepsilon)\|Gp - Gu_{i,j}\|^z - O(\varepsilon)^{2z}\|p - c_j\|^z\Big) - O(\varepsilon\,\mathrm{opt})\end{aligned}$$

Applying event c to net point $u_{i,j}$ and subset $T_i \cap S_j$, we have

$$\begin{aligned}&\ge \sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{p\in T_i\cap S_j} \Big((1 - O(z\varepsilon))\|p - u_{i,j}\|^z - \frac{\varepsilon}{k^2}\,\mathrm{opt} - O(\varepsilon)^{2z}\|p - c_j\|^z\Big) - O(\varepsilon\,\mathrm{opt}) \\ &\ge \sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{p\in T_i\cap S_j} \Big((1 - O(z\varepsilon))\|p - c_j\|^z - \frac{\varepsilon}{k^2}\,\mathrm{opt}\Big) - O(\varepsilon\,\mathrm{opt}) \\ &\ge (1 - O(z\varepsilon))\sum_{p\in T}\|p - C(p)\|^z - O(\varepsilon)\,\mathrm{opt}(P, Q). \qquad (11)\end{aligned}$$

Case 2.3, the far range $p \notin R \cup T$. We now consider points $p \in S_i^* \setminus (R \cup T)$, i.e., $\|f(p) - c_i^*\| \ge 2000L^2 r_{\ell_i}$. Suppose $f(p) = c_j$. By e, $\|Gc_j - Gc_i^*\| \ge 2000Lr_{\ell_i}$.

▷ Claim 5.5. In this case, $r_p \ge 10Lr_{\ell_i}$.

Proof. Assume by contradiction that $r_p < 10Lr_{\ell_i}$. By Lemma 5.3, $\|C(p) - c_i^*\| \le 4\max\{r_p, \|c_i^* - C(c_i^*)\|\} \le 40Lr_{\ell_i}$. Thus by d, $Gp, GC(p) \in B(Gc_i^*, 400Lr_{\ell_i})$. Therefore,

$$\|Gc_j - Gc_i^*\| \le \|Gc_j - Gp\| + \|Gp - Gc_i^*\| \le \|GC(p) - Gp\| + \|Gp - Gc_i^*\| \le 800Lr_{\ell_i},$$

contradiction. ◁

Therefore, by Lemma 5.3, $\|C(p) - c_i^*\| \le 4r_p$ and hence

$$\|p - C(p)\| \le \|p - c_i^*\| + \|C(p) - c_i^*\| \le 5r_p. \qquad (12)$$

On a high level, as can be seen by the claim, we have that both $f(p)$ and p are far from c_i^*. We split into cases depending on which of p or $f(p)$ is farther from c_i^* (up to a constant), as follows.

Case 2.3.1, $p \in S_i^* \setminus (R \cup T)$, and $\|f(p) - c_i^*\| > 10Lr_p$. By triangle inequality,

$$\|p - f(p)\| \geq \|f(p) - c_i^*\| - \|p - c_i^*\| \geq 9L \cdot r_p.$$

By the definition of η_p, we have $\|Gp - Gf(p)\| \geq \eta_p$. Therefore,

$$\begin{aligned}
&\sum_{\substack{p \in S_i^* \setminus (R \cup T) \\ \|f(p) - c_i^*\| > 10Lr_p}} \|Gp - Gf(p)\|^z \\
\geq &\sum_{\substack{p \in S_i^* \setminus (R \cup T) \\ \|f(p) - c_i^*\| > 10Lr_p}} \eta_p^z && \text{since } \|p - f(p)\| \geq 9Lr_p \\
\geq &\sum_{\substack{p \in S_i^* \setminus (R \cup T) \\ \|f(p) - c_i^*\| > 10Lr_p}} (9r_p)^z - e^{-\Omega(t)} \cdot \sum_{p \in S_i^*} r_p^z && \text{by event g} \\
\geq &\sum_{\substack{p \in S_i^* \setminus (R \cup T) \\ \|f(p) - c_i^*\| > 10Lr_p}} (5r_p)^z - e^{-\Omega(t)} \cdot \sum_{p \in S_i^*} r_p^z \\
\geq &\sum_{\substack{p \in S_i^* \setminus (R \cup T) \\ \|f(p) - c_i^*\| > 10Lr_p}} \|p - C(p)\|^z - e^{-\Omega(t)} \cdot \sum_{p \in S_i^*} r_p^z && \text{by (12)} \qquad (13)
\end{aligned}$$

Case 2.3.2, $p \in S_i^* \setminus (R \cup T)$, and $\|f(p) - c_i^*\| \leq 10Lr_p$. Denote u_p and $u_{f(p)}$ to be the net points in N_{j_p} that are closest to p and $f(p)$, respectively. Then

$$\begin{aligned}
&\|Gp - Gf(p)\|^z \\
&\geq (1 - 2z\varepsilon)\|Gu_p - Gu_{f(p)}\|^z - \varepsilon^{-z}\|Gp - Gu_p\|^z - \varepsilon^{-z}\|Gf(p) - Gu_{f(p)}\|^z \\
&\geq (1 - 2z\varepsilon)(1 - \varepsilon - \beta_p\varepsilon)^z\|u_p - u_{f(p)}\|^z - 2\varepsilon^{-z}(\gamma_p\varepsilon^3 r_p)^z \\
&\geq (1 - 3z\varepsilon - \beta_p z\varepsilon)\|u_p - u_{f(p)}\|^z - O(\varepsilon)^{2z}\gamma_p^z r_p^z \\
&\geq (1 - 3z\varepsilon - \beta_p z\varepsilon)\|p - f(p)\|^z - O(\varepsilon)^{2z} r_p^z - O(\varepsilon)^{2z}\gamma_p^z r_p^z
\end{aligned}$$

Since $\|p - f(p)\| \leq \|p - c_i^*\| + \|f(p) - c_i^*\| \leq r_p + 10Lr_p \leq 20Lr_p$, we have

$$\geq (1 - 3z\varepsilon)\|p - f(p)\|^z - \beta_p z\varepsilon \cdot (20L)^z r_p^z - O(\varepsilon)^{2z} r_p^z - O(\varepsilon)^{2z}\gamma_p^z r_p^z.$$

Therefore,

$$\begin{aligned}
&\sum_{\substack{p \in S_i^* \setminus (R \cup T) \\ \|f(p) - c_i^*\| \leq 10Lr_p}} \|Gp - Gf(p)\|^z \\
\geq (1 - 3z\varepsilon) &\sum_{\substack{p \in S_i^* \setminus (R \cup T) \\ \|f(p) - c_i^*\| \leq 10Lr_p}} \|p - f(p)\|^z - z\varepsilon(20L)^z \sum_{p \in S_i^*} \beta_p r_p^z - O(\varepsilon)^{2z} \sum_{p \in S_i^*} (1 + \gamma_p^z) r_p^z. \qquad (14)
\end{aligned}$$

Wrap Up. Combining (14) and (13), we have

$$\begin{aligned}
\sum_{p \in S_i^* \setminus (R \cup T)} \|Gp - Gf(p)\|^z \geq (1 - 3z\varepsilon) &\sum_{p \in S_i^* \setminus (R \cup T)} \|p - C(p)\|^z \\
&- z\varepsilon(20L)^z \sum_{p \in S_i^*} \beta_p r_p^z - O(\varepsilon)^{2z} \sum_{p \in S_i^*} (1 + \gamma_p^z) r_p^z.
\end{aligned}$$

Summing over $i \in [k]$ yields

$$\begin{aligned}
&\sum_{p\in P\setminus(R\cup T)} \|Gp - Gf(p)\|^z \\
&\geq (1-3z\varepsilon) \sum_{p\in P\setminus(R\cup T)} \|p - C(p)\|^z - z\varepsilon(20L)^z \sum_{p\in P} \beta_p r_p^z - O(\varepsilon)^{2z} \sum_{p\in P} (1+\gamma_p^z) r_p^z \\
&\geq (1-3z\varepsilon) \sum_{p\in P\setminus(R\cup T)} \|p - C(p)\|^z - z\varepsilon(20L)^z e^{-\Omega(\varepsilon^2 t)} \cdot \operatorname{opt}(P,Q) - O(\varepsilon)^{2z} \cdot \operatorname{opt}(P,Q) \\
&\geq (1-3z\varepsilon) \sum_{p\in P\setminus(R\cup T)} \|p - C(p)\|^z - O(\varepsilon) \cdot \operatorname{opt}(P,Q), \qquad (15)
\end{aligned}$$

where the second last inequality follows from event a and Lemma 5.2. Finally, we combine (10),(11) and (15) and obtain

$$\operatorname{cost}(G(P), G(C)) \geq (1 - O(z\varepsilon)) \operatorname{cost}(P,C) - O(\varepsilon) \cdot \operatorname{opt}(P,Q) \geq (1 - O(z\varepsilon)) \operatorname{cost}(P,C).$$

Rescaling $\varepsilon \to \varepsilon/z$ concludes the proof. ◀

6 Lower bounds

In this section, we state our lower bounds. For simplicity, we do not try to optimize the dependence on z. All lower bounds are presented for $z = 1$. All the proofs are provided in the full version [23]. Denote by 0_d the origin of $\mathbb{R}^d$. For ease of presentation, we allow P to be a multi-set.

▶ **Theorem 6.1** (Continuous, for all centers). *Let $n, d \in \mathbb{N}$, and $P = \{0_d\}^n$. Let $G \in \mathbb{R}^{(d-1)\times d}$ be any linear map. Then, there exists $c \in \mathbb{R}^d$ such that $\sum_{p\in P} \|Gp - Gc\| = 0$ and $\sum_{p\in P} \|p - c\| = n$.*

The next theorem shows that in order to bound the (multiplicative) contraction for all centers, we need either dimension $\Omega(\log\log n)$, or to relax the definition of contraction (as is done in Theorem 5.1).

▶ **Theorem 6.2** (Discrete, for all centers). *Let $n, d \in \mathbb{N}$ and $\varepsilon \in (0, \frac{1}{2})$. There exists $P \subset \mathbb{R}^d$ of size $|P| = n$ and $\operatorname{ddim}(P) = \Theta(1)$, such that if G is a Gaussian JL map onto dimension $t \leq a\varepsilon^{-2}\log\log n$ for a sufficiently small constant $a > 0$, then with probability at least $2/3$, there exists $c \in P$ such that $\sum_{p\in P} \|Gp - Gc\| \leq (1-\varepsilon)\sum_{p\in P} \|p - c\|$.*

Next, we show that dimension $\Omega(\log\log n)$ is necessary, even for the relaxed notion of contraction, for preserving all partitions and centers.

▶ **Theorem 6.3** (Discrete, for all partitions and centers). *Let $n, d \in \mathbb{N}$ and $\varepsilon \in (0, \frac{1}{2})$. There exists $P \subset \mathbb{R}^d$ of size $|P| = n$ and $\operatorname{ddim}(P) = \Theta(1)$, such that if G is a Gaussian JL map onto dimension $\frac{1}{a}\varepsilon^{-2} \leq t \leq a\varepsilon^{-2}\log\log n$ for a sufficiently small constant $a > 0$, then with probability at least $2/3$, there exists $(c_1, c_2) \subset P$ and a partition (P_1, P_2) of P such that*

$$\sum_{i\in\{1,2\}} \sum_{p\in P_i} \|Gp - Gc_i\| < \min\left\{(1-\varepsilon) \sum_{i\in\{1,2\}} \sum_{p\in P_i} \|p - c_i\|, 100\operatorname{opt}(P)\right\}.$$

▶ **Theorem 6.4** (Discrete, for all centers, with candidate center set). *Let $n, s, d \in \mathbb{N}$ and $\varepsilon \in (0, \frac{1}{2})$. There exists $P, Q \subset \mathbb{R}^d$ of sizes $|P| = n, |Q| = s$, and $\operatorname{ddim}(P \cup Q) = O(1)$, such that if G is a Gaussian JL map onto dimension $t \leq a\varepsilon^{-2}\log s$ for a sufficiently small constant $a > 0$, then with probability at least $2/3$, there exists $c \in Q$ such that $\sum_{p\in P} \|Gp - Gc\| \leq (1-\varepsilon)\sum_{p\in P} \|p - c\|$.*

References

1 Yair Bartal, Ben Recht, and Leonard J Schulman. Dimensionality reduction: beyond the johnson-lindenstrauss bound. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 868–887. SIAM, 2011. doi:10.1137/1.9781611973082.68.

2 Luca Becchetti, Marc Bury, Vincent Cohen-Addad, Fabrizio Grandoni, and Chris Schwiegelshohn. Oblivious dimension reduction for k-means: beyond subspaces and the Johnson-Lindenstrauss Lemma. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC*, pages 1039–1050, 2019. doi:10.1145/3313276.3316318.

3 Christos Boutsidis, Anastasios Zouzias, and Petros Drineas. Random projections for k-means clustering. In *24th Annual Conference on Neural Information Processing Systems, NeurIPS*, pages 298–306. Curran Associates, Inc., 2010. URL: https://proceedings.neurips.cc/paper/2010/hash/73278a4a86960eeb576a8fd4c9ec6997-Abstract.html.

4 Moses Charikar and Erik Waingarten. The Johnson-Lindenstrauss Lemma for clustering and subspace approximation: From coresets to dimension reduction. In *SODA*, pages 3172–3209. SIAM, 2025. doi:10.1137/1.9781611978322.102.

5 Xiaoyu Chen, Shaofeng H.-C. Jiang, and Robert Krauthgamer. Streaming Euclidean Max-Cut: Dimension vs data reduction. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing, STOC*, pages 170–182, 2023. doi:10.1145/3564246.3585170.

6 Michael B. Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC*, pages 163–172, 2015. doi:10.1145/2746539.2746569.

7 Michael B. Cohen, Jelani Nelson, and David P. Woodruff. Optimal approximate matrix product in terms of stable rank. In *43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016)*, volume 55 of *LIPIcs*, pages 11:1–11:14. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2016. doi:10.4230/LIPIcs.ICALP.2016.11.

8 Vincent Cohen-Addad, Hossein Esfandiari, Vahab Mirrokni, and Shyam Narayanan. Improved approximations for Euclidean k-means and k-median, via nested quasi-independent sets. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1621–1628, 2022. doi:10.1145/3519935.3520011.

9 Vincent Cohen-Addad, Andreas Emil Feldmann, and David Saulpic. Near-linear time approximation schemes for clustering in doubling metrics. *Journal of the ACM*, 68(6):1–34, 2021. doi:10.1145/3477541.

10 Vincent Cohen-Addad, Fabrizio Grandoni, Euiwoong Lee, Chris Schwiegelshohn, and Ola Svensson. A $(2+\varepsilon)$-approximation algorithm for metric k-median. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 615–624, 2025. doi:10.1145/3717823.3718299.

11 Vincent Cohen-Addad and C. S. Karthik. Inapproximability of clustering in Lp metrics. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 519–539. IEEE, 2019. doi:10.1109/FOCS.2019.00040.

12 Vincent Cohen-Addad, C. S. Karthik, and Euiwoong Lee. Johnson coverage hypothesis: Inapproximability of k-means and k-median in ℓ_p-metrics. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1493–1530. SIAM, 2022. doi:10.1137/1.9781611977073.63.

13 Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, 2003. doi:10.1002/rsa.10073.

14 Jie Gao, Rajesh Jayaram, Benedikt Kolbe, Shay Sapir, Chris Schwiegelshohn, Sandeep Silwal, and Erik Waingarten. Randomized dimensionality reduction for Euclidean maximization and diversity measures. In *Forty-second International Conference on Machine Learning*, 2025. URL: https://openreview.net/forum?id=Rcivp36KzO.

15 Lee-Ad Gottlieb and Robert Krauthgamer. A nonlinear approach to dimension reduction. *Discrete & Computational Geometry*, 54(2):291–315, 2015. doi:10.1007/s00454-015-9707-9.

16 Anupam Gupta, Robert Krauthgamer, and James R. Lee. Bounded geometries, fractals, and low-distortion embeddings. In *44th Symposium on Foundations of Computer Science, FOCS*, pages 534–543. IEEE Computer Society, 2003. doi:10.1109/SFCS.2003.1238226.

17 Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *STOC*, pages 291–300. ACM, 2004. doi:10.1145/1007352.1007400.

18 Lingxiao Huang, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Di Yue. Near-optimal dimension reduction for facility location. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing, STOC*, pages 665–676, 2025. doi:10.1145/3717823.3718214.

19 Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing, STOC*, pages 604–613, 1998. doi:10.1145/276698.276876.

20 Piotr Indyk and Assaf Naor. Nearest-neighbor-preserving embeddings. *ACM Trans. Algorithms*, 3(3):31, 2007. doi:10.1145/1273340.1273347.

21 Zachary Izzo, Sandeep Silwal, and Samson Zhou. Dimensionality reduction for Wasserstein barycenter. *Advances in neural information processing systems*, 34:15582–15594, 2021. URL: https://proceedings.neurips.cc/paper/2021/hash/8346db44a721fa863ca38180638bad3d-Abstract.html.

22 Shaofeng H.-C. Jiang, Robert Krauthgamer, and Shay Sapir. Moderate dimension reduction for k-center clustering. In *40th International Symposium on Computational Geometry (SoCG 2024)*, volume 293 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 64:1–64:16. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024. doi:10.4230/LIPIcs.SoCG.2024.64.

23 Shaofeng H. C. Jiang, Robert Krauthgamer, Shay Sapir, Sandeep Silwal, and Di Yue. Dimension reduction for clustering: The curious case of discrete centers, 2025. Full version. doi:10.48550/arXiv.2509.07444.

24 William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into Hilbert space. *Contemporary mathematics*, 26:189–206, 1984. doi:10.1090/conm/026/737400.

25 Leonard Kaufman and Peter J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009. doi:10.1002/9780470316801.

26 M. Kirszbraun. Über die zusammenziehende und lipschitzsche transformationen. *Fundamenta Mathematicae*, 22(1):77–108, 1934. doi:10.4064/fm-22-1-77-108.

27 Christiane Lammersen. *Approximation Techniques for Facility Location and Their Applications in Metric Embeddings*. PhD thesis, Technische Universität Dortmund, 2010. doi:10.17877/DE290R-8506.

28 Christiane Lammersen, Anastasios Sidiropoulos, and Christian Sohler. Streaming embeddings with slack. In *11th International Symposium on Algorithms and Data Structures, WADS*, volume 5664 of *Lecture Notes in Computer Science*, pages 483–494. Springer, 2009. doi:10.1007/978-3-642-03367-4_42.

29 Kasper Green Larsen and Jelani Nelson. Optimality of the Johnson-Lindenstrauss Lemma. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS*, pages 633–638, 2017. doi:10.1109/FOCS.2017.64.

30 Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of massive data sets*. Cambridge University Press, 3rd edition, 2020. doi:10.1017/9781108684163.

31 Michael W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011. doi:10.1561/2200000035.

32 Konstantin Makarychev, Yury Makarychev, and Ilya P. Razenshteyn. Performance of Johnson-Lindenstrauss transform for k-means and k-medians clustering. In *Proceedings of the 51st Annual Symposium on Theory of Computing, STOC*, pages 1027–1038, 2019. doi:10.1145/3313276.3316350.

33 Shyam Narayanan, Sandeep Silwal, Piotr Indyk, and Or Zamir. Randomized dimensionality reduction for facility location and single-linkage clustering. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 7948–7957. PMLR, 2021. URL: http://proceedings.mlr.press/v139/narayanan21b.html.

34 Christos H. Papadimitriou. Worst-case and probabilistic analysis of a geometric location problem. *SIAM Journal on Computing*, 10(3):542–557, 1981. doi:10.1137/0210040.

35 Hae-Sang Park and Chi-Hyuck Jun. A simple and fast algorithm for K-medoids clustering. *Expert Syst. Appl.*, 36(2):3336–3341, 2009. doi:10.1016/J.ESWA.2008.01.039.

36 Mo Tiwari, Martin J Zhang, James Mayclin, Sebastian Thrun, Chris Piech, and Ilan Shomorony. Banditpam: Almost linear time k-medoids clustering via multi-armed bandits. *Advances in Neural Information Processing Systems*, 33:10211–10222, 2020.

37 David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1—2):1–157, 2014. doi:10.1561/0400000060.

38 Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 478–487. PMLR, 2016. URL: https://proceedings.mlr.press/v48/xieb16.html.