

# The Johnson-Lindenstrauss Lemma for Clustering and Subspace Approximation: From Coresets to Dimension Reduction\*

Moses Charikar<sup>†</sup>

Erik Waingarten<sup>‡</sup>

## Abstract

We study the effect of Johnson-Lindenstrauss transforms in various projective clustering problems, generalizing results which only applied to center-based clustering [40]. We ask the general question: for a Euclidean optimization problem and an accuracy parameter  $\varepsilon \in (0, 1)$ , what is the smallest target dimension  $t \in \mathbb{N}$  such that a Johnson-Lindenstrauss transform  $\Pi: \mathbb{R}^d \rightarrow \mathbb{R}^t$  preserves the cost of the optimal solution up to a  $(1 + \varepsilon)$ -factor. We give a new technique which uses coreset constructions to analyze the effect of the Johnson-Lindenstrauss transform. Our technique, in addition applying to center-based clustering, improves on (or is the first to address) other Euclidean optimization problems, including:

- For  $(k, z)$ -subspace approximation: we show that  $t = \tilde{O}(zk^2/\varepsilon^3)$  suffices, whereas the prior best bound, of  $O(k/\varepsilon^2)$ , only applied to the case  $z = 2$  [13].
- For  $(k, z)$ -flat approximation: we show  $t = \tilde{O}(zk^2/\varepsilon^3)$  suffices, completely removing the dependence on  $n$  from the prior bound  $\tilde{O}(zk^2 \log n/\varepsilon^3)$  of [36].
- For  $(k, z)$ -line approximation: we show  $t = O((k \log \log n + z + \log(1/\varepsilon))/\varepsilon^3)$  suffices, and ours is the first to give any dimension reduction result.

## 1 Introduction

The Johnson-Lindenstrauss lemma [35] concerns dimensionality reduction for high-dimensional Euclidean spaces. It states that, for any set of  $n$  points  $x_1, \dots, x_n$  in  $\mathbb{R}^d$  and any  $\varepsilon \in (0, 1)$ , there exists a map  $\Pi: \mathbb{R}^d \rightarrow \mathbb{R}^t$ , with  $t = O(\log n/\varepsilon^2)$  such that, for any  $i, j \in [n]$ ,

$$(1.1) \quad \frac{1}{1 + \varepsilon} \cdot \|x_i - x_j\|_2 \leq \|\Pi(x_i) - \Pi(x_j)\|_2 \leq (1 + \varepsilon) \cdot \|x_i - x_j\|_2.$$

From a computational perspective, the lemma has been extremely influential in designing algorithms for high-dimensional geometric problems, partly because proofs show that a *random* linear map, oblivious to the data, suffices. Proofs specify a distribution  $\mathcal{J}_{d,t}$  supported on linear maps  $\mathbb{R}^d \rightarrow \mathbb{R}^t$  which is independent of  $x_1, \dots, x_n$ , for example, given by a  $t \times d$  matrix of i.i.d  $\mathcal{N}(0, 1/t)$  entries [32, 18], and show that a draw  $\Pi \sim \mathcal{J}_{d,t}$  satisfies (1.1) with probability at least 0.9.

In this paper, we study the Johnson-Lindenstrauss transforms for projective clustering problems, building on a line-of-work which gave dimension reduction results for center-based clustering [7, 13, 6, 40]. Our goal is to reduce the dimensionality of the input of a more general projective clustering problem from  $d$  to  $t$ , with  $t \ll d$ , without affecting the cost of the optimal solution significantly. We map  $d$ -dimensional points to  $t$ -dimensional points via a random linear map  $\Pi$  such that the optimal cost in  $t$ -dimensions is within a  $(1 + \varepsilon)$ -factor of the optimal cost in the original  $d$ -dimensional space. We study this for a variety of problems, each of which is specified by a set of candidate solutions  $\mathcal{C}_d$  and a cost function. By varying the family of candidate solutions  $\mathcal{C}_d$  and the cost functions considered, one obtains center-based clustering problems (like  $k$ -means and  $k$ -median), as well as subspace approximation problems (like principal components analysis), and beyond (like clustering with subspaces). The key question we address here is:

\*The full version of the paper can be accessed at <https://arxiv.org/abs/2205.00371>

<sup>†</sup>Stanford University.

<sup>‡</sup>University of Pennsylvania. Part of this work was done while Erik Waingarten was a postdoc at Stanford University, supported by an NSF postdoctoral fellowship and by Moses Charikar's Simons Investigator Award. Erik Waingarten is currently supported by NSF CCF-2337993.

**Main Question:** For a projective clustering problem, how small can  $t$  be as a function of the dataset size  $n$  and the accuracy parameter  $\varepsilon$ , such that the cost of the optimization is preserved up to  $(1 \pm \varepsilon)$ -factor with probability at least 0.9 over  $\Pi \sim \mathcal{J}_{d,t}$ ?

Our results fit into a line of prior work on the power of Johnson-Lindenstrauss maps beyond the original discovery of [35]. These have been investigated before for various problems and in various contexts, including nearest neighbor search [32, 27, 3], numerical linear algebra [43, 39, 50], prioritized and terminal embeddings [21, 38, 42, 12], and clustering and facility location problems [7, 13, 36, 6, 40, 41, 33, 11, 34].

**1.1 Our Contribution** In the  $(k, j)$ -projective clustering problem with  $\ell_z$  cost ([29, 1, 20, 19, 49, 36, 25, 47]), the goal is to cluster a dataset  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ , where each cluster is approximated by an affine  $j$ -dimensional subspace. Namely, we define an objective function for  $(k, j)$ -projective clustering problems with  $\ell_z$  cost on a dataset  $X$ , which aims to minimize

$$\min_{c \in \mathcal{C}_d} \text{cost}_z(X, c),$$

where the candidate solutions  $\mathcal{C}_d$  consist of all  $k$ -tuples of  $j$ -dimensional affine subspaces. The cost function  $\text{cost}_z(X, \cdot)$  maps each candidate solution  $c \in \mathcal{C}_d$  to a cost in  $\mathbb{R}_{\geq 0}$  given by the  $\ell_z$ -norm of the vector of distances between each dataset point  $x \in X$  to its nearest point on one of the  $k$  subspaces. Intuitively, each point of the dataset  $x_i \in X$  “pays” for the Euclidean distance to the nearest subspace in  $c \in \mathcal{C}_d$ , and the total cost is the  $\ell_z$ -norm of the  $n$  payments, one for each point.

There has been significant prior work which showed surprising results for the special case of  $(k, z)$ -clustering (like  $k$ -means and  $k$ -median, which corresponds to  $(k, 0)$ -projective clustering with  $\ell_z$ -cost) as well as for low-rank approximation (which corresponds to  $(1, k)$ -projective clustering for non-affine subspaces with  $\ell_2$ -cost) [7, 13, 6, 40]. It is important to note that the techniques in prior works are specifically tailored to the Euclidean optimization problem at hand. For example, the results of [40], which apply for  $(k, 0)$ -clustering with  $\ell_z$ -norm, rely on using center points as the approximation and do not generalize to affine subspaces beyond dimension 0. The other result of [13] for low-rank approximation uses the specific algebraic properties of the  $\ell_2$ -norm which characterize the optimal low-rank approximation but do not hold generally. In summary, these prior show that for  $(k, z)$ -clustering and low-rank approximation, even though many pairwise distances among dataset points become highly distorted, the cost of the optimization need not be significantly distorted.

**Our Results.** We show that  $(k, 0)$ -clustering with  $\ell_z$ -norm and low-rank approximation are not isolated incidents, but rather, part of a more general phenomenon. Our main conceptual contribution is the following: we use algorithms for constructing *coresets* (via the sensitivity sampling framework of [23]) to obtain bounds on dimension reduction. Then, the specific bounds that we obtain for the various problems depend on the sizes of the coresets that the algorithms can produce. We can instantiate our framework to new upper bounds for the following problems:

- **$(k, z)$ -Subspace Approximation.** This problem is a restricted  $(1, k)$ -projective clustering problems with  $\ell_z$ -cost. We seek to minimize over a  $k$ -dimensional subspace  $S$  of  $\mathbb{R}^d$  the  $\ell_z$ -norm of the  $n$ -dimensional vector where the coordinate  $i \in [n]$  encodes the distance between  $x_i$  and the closest point in  $S$ .<sup>1</sup>
- **$(k, z)$ -Flat Approximation.** This problem is exactly the  $(1, k)$ -projective clustering problem with  $\ell_z$  cost. It is similar to  $(k, z)$ -subspace approximation, except we optimize over all affine subspaces.
- **$(k, z)$ -Line Approximation.** This problem corresponds to  $(k, 1)$ -projective clustering with  $\ell_z$ -cost. The optimization minimizes over an arbitrary set  $L$  of  $k$  of 1-dimensional affine subspaces  $l_1, \dots, l_k$  (i.e.,  $k$  lines in  $\mathbb{R}^d$ ), the  $\ell_z$ -norm of the  $n$ -dimensional vector where the coordinate  $i \in [n]$  encodes the distance between  $x_i$  and the closest point on any line in  $L$ .

Concretely, our quantitative results are summarized by the following theorem.

<sup>1</sup>This is a restricted version of projective clustering because subspaces are not affine and required to go through the origin.

Target Dimension for Johnson-Lindenstrauss Transforms		
Problem	New Result	Prior Best
$(k, z)$ -subspace	$\tilde{O}(zk^2/\varepsilon^3)$	$O(k/\varepsilon^2)$ , only $z = 2$ [13]
$(k, z)$ -flat	$\tilde{O}(zk^2/\varepsilon^3)$	$\tilde{O}(zk^2 \log n/\varepsilon^3)$ [36]
$(k, z)$ -line	$O((k \log \log n + z + \log(1/\varepsilon))/\varepsilon^3)$	None

Figure 1: Comparison to Prior Bounds

**THEOREM 1.1. (MAIN RESULT—INFORMAL)** *Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be any dataset, and  $\mathcal{J}_{d,t}$  denote a  $t \times d$  matrix of i.i.d. entries from  $\mathcal{N}(0, 1/t)$ . Let  $\mathcal{C}_d$  and  $\mathcal{C}_t$  be candidate solutions for a projective clustering problem in  $\mathbb{R}^d$  and  $\mathbb{R}^t$ , respectively. For any  $\varepsilon \in (0, 1)$ , we have*

$$(1.2) \quad \Pr_{\Pi \sim \mathcal{J}_{d,t}} \left[ \left| \frac{\min_{c' \in \mathcal{C}_t} \text{cost}(\Pi(X), c')}{\min_{c \in \mathcal{C}_d} \text{cost}(X, c)} - 1 \right| \leq \varepsilon \right] \geq 0.9$$

whenever:

- $(k, z)$ -subspace and  $(k, z)$ -flat approximation:  $\mathcal{C}_d$  and  $\mathcal{C}_t$  are all  $k$ -dimensional subspaces of  $\mathbb{R}^d$  and  $\mathbb{R}^t$ , respectively; the cost measures the  $\ell_z$ -norm of distances between points to the subspace; and,  $t = \tilde{O}(zk^2/\varepsilon^3)$ . Similarly, the same bound on  $t$  holds for  $\mathcal{C}_d$  and  $\mathcal{C}_t$  varying over all affine  $k$ -dimensional subspaces.
- $(k, z)$ -line approximation:  $\mathcal{C}_d$  and  $\mathcal{C}_t$  are all  $k$ -tuples of lines in  $\mathbb{R}^d$  and  $\mathbb{R}^t$ , respectively; the cost measures the  $\ell_z$ -norm of distances between points and the nearest line; and,  $t = O((k \log \log n + z + \log(1/\varepsilon))/\varepsilon^3)$ .

In all cases, the bound that we obtain is directly related to the size of the best coresets from the sensitivity sampling framework, and all of our proofs follow the same format.<sup>2</sup> Our proofs are not entirely black-box applications of coresets, since we use the specific instantiations of the sensitivity sampling framework. We believe that any improvement on the size of the best coresets will likely lead to quantitative improvements on the dimension reduction bounds. However, improving our current bounds by either better coresets or via a different argument altogether (for example, improving on the cubic dependence on  $\varepsilon$ ) seems to require significantly new ideas.

**REMARK 1.1. (HIGH PROBABILITY BOUNDS)** *Theorems are stated with respect to probability 0.9 and can be made arbitrarily close to 1 (by increasing the constant factor in the dimensionality  $t$ ). For high-probability guarantees, i.e., those which hold with probability  $1 - \delta$  for any  $\delta > 0$ , a straight-forward adaptation would incur extra factors of  $O(\log(1/\delta))$ . As we will see, one factor would come from ensuring randomized coreset constructions hold with probability  $1 - \delta$ , and another factor to ensure the dimension reduction preserves distances (or subspaces spanned) among coreset points with probability  $1 - \delta$ .*

**A Subtlety in “For-All” versus “Optimal” Guarantees.** Our results focus on applying the Johnson-Lindenstrauss transform and preserving the optimal cost, i.e., that the minimizing solution in the original and the dimension reduced space have approximately the same cost. A stronger guarantee which one may hope for, a so-called “for-all” guarantee, asks that after applying the Johnson-Lindenstrauss transform, every solution has its cost approximately preserved before and after dimension reduction. We do not achieve “for all” guarantees, like those appearing in [40]. However, we emphasize that various subtleties arise in what is meant by “a solution.” In particular, prior work on dimension reduction and coresets parametrize to solutions differently—the coreset literature considers optimal centers, and [40] to partitions of inputs.

Consider the 1-medoid problem, a constrained version of the 1-means problem. The 1-medoid cost of a dataset  $X$  is the minimum over centers  $c$  chosen from the dataset  $X$ , of the sum of squares of distances from each dataset point  $x$  to  $c$ . The subtlety is: one can apply a Johnson-Lindenstrauss transform to  $t = O(\log(1/\varepsilon)/\varepsilon^2)$  dimensions and preserve the 1-means cost—for 1-medoid, we show that one cannot.

<sup>2</sup>The reason we did not generalize the results to  $(k, j)$ -projective clustering with  $\ell_z$ -cost with  $j > 1$  is that these problems do not admit small coresets [26]. Researchers have studied “integer  $(k, j)$ -projective clustering” where one restricts the input points to have bounded integer coordinates, where small coresets do exist [20]. However, using this approach for dimension reduction would incur additional additive errors, so we have chosen not to pursue this route.

**THEOREM 1.2.** (JOHNSON-LINDENSTRAUSS FOR MEDOID—INFORMAL (SEE THEOREM [A.1](#))) *For any  $t = o(\log n)$ , applying a Johnson-Lindenstrauss transform to dimension  $t$  decreases the cost of the 1-medoid problem by a factor approaching 2.*

Theorem [1.2](#) does not contradict the “for-all” guarantee of [\[40\]](#) because there, a candidate solution for  $(k, z)$ -clustering refers to a partition of  $X$  into  $k$  parts and not a set of centers. For  $k = 1$ , there are many possible centers but only one partition. For  $(k, z)$ -subspace and -flat approximation, an analogous issue arises. Consider the 1-column subset selection problem, a constrained version of the subspace approximation problem, where one must choose a 1-dimensional subspace going through a dataset point to approximate the entire dataset. The 1-column subset selection cost of a dataset is the minimum over 1-dimensional subspaces spanned by a dataset point of  $X$ , of the sum of squares of distances from each dataset point  $x$  to the projection onto the subspace. Similarly to Theorem [1.2](#) a Johnson-Lindenstrauss transform does not preserve the cost of 1-column subset selection.

**THEOREM 1.3.** (JOHNSON-LINDENSTRAUSS FOR COLUMN SUBSET SELECTION—INFORMAL (SEE THEOREM [A.2](#))) *For any  $t = o(\log n)$ , applying a Johnson-Lindenstrauss transform to dimension  $t$  decreases the cost of the 1-column subset selection problem by a factor approaching  $3/2$ .*

The above theorem does not contradict the “for-all” guarantee of [\[13\]](#) for similar reasons (which, in addition, crucially rely on having  $z = 2$ , and which we elaborate on in Appendix [A](#)). For  $(k, z)$ -line approximation, however, there is an interesting open problem: is it true that after applying a Johnson-Lindenstrauss transform to  $t = \text{poly}(k \log \log n / \varepsilon)$ , for all partitions of  $X$  into  $k$  parts, the cost of optimally approximating each part with a line has its cost preserved.

## 1.2 Related Work

**Dimension Reduction.** Our paper continues a line of work initiated by Boutsidis, Zouzias, and Drineas [\[7\]](#), who first studied the effect of a Johnson-Lindenstrauss transform for  $k$ -means clustering, and showed that  $t = O(k/\varepsilon^2)$  sufficed for a  $(2 + \varepsilon)$ -approximation. The bound was improved to  $(1 + \varepsilon)$ -approximation with  $t = O(k/\varepsilon^2)$  by Cohen, Elder, Musco, Musco, Persu [\[13\]](#), who also showed that  $t = O(\log k/\varepsilon^2)$  gave a  $(9 + \varepsilon)$ -approximation. Becchetti, Bury, Cohen-Addad, Grandoni, Schwiegelshohn [\[6\]](#) showed that  $t = O((\log k + \log \log n) \log(1/\varepsilon)/\varepsilon^6)$  sufficed for preserving the costs of all  $k$ -mean clusterings. Makarychev, Makarychev, and Razenshteyn [\[40\]](#) improved and generalized the above bounds for all  $(k, z)$ -clustering. They showed that  $t = O((\log k + z \log(1/\varepsilon) + z^2)/\varepsilon^2)$  preserved costs to  $(1 \pm \varepsilon)$ -factor.

For subspace approximation problems, [\[13\]](#) showed that  $t = O(k/\varepsilon^2)$  preserves the cost of  $(k, 2)$ -subspace approximation to  $(1 + \varepsilon)$ -factor. In addition, [\[36\]](#) showed that  $O(zk^2 \log n/\varepsilon^3)$  preserved the cost of the  $(k, z)$ -flat approximation to  $(1 + \varepsilon)$ -factor.

**Coresets.** Coresets are a well-studied technique for reducing the size of a dataset, while approximately preserving a particular desired property. Since its formalization in Agarwal, Har-Peled, and Varadarajan [\[2\]](#), coresets have played a foundational role in computational geometry, and found widespread application in clustering, numerical linear algebra, and machine learning (see the recent survey [\[22\]](#)). Indeed, even for clustering problems in Euclidean spaces, there is a long line of work [\[4, 28, 2, 10, 37, 23, 49, 48, 24, 9, 46, 31, 5, 16, 14, 15, 8, 17, 30\]](#) exploring the best coreset constructions.

Most relevant to our work is the sensitivity sampling framework of Feldman and Langberg [\[23\]](#), which gives algorithms for constructing coresets for the projective clustering problems we study. In light of the results of [\[23\]](#), as well as the classical formulation of the Johnson-Lindenstrauss lemma [\[35\]](#), it may seem natural to apply coreset algorithms and dimensionality reduction concurrently. However, this is not without a few technical challenges. As we will see in the next subsection, it is not necessarily the case that coreset algorithms and random projections “commute.” Put succinctly, the random projection  $\Pi$  of a coreset of  $X$  may not be a coreset of the random projection  $\Pi(X)$ . Indeed, proving such a statement constitutes the bulk of the technical work.

**1.3 Organization** The following section (Section [2](#)) overviews the high-level plan, since all our results follow the same technique. To highlight the technique, the first technical section considers the case of  $(k, z)$ -clustering (Section [4](#)), where the technique of arguing via coresets shows to obtain  $t = O((\log(k) + z \log(1/\varepsilon))/\varepsilon^2)$ . The remaining sections cover the technical material for  $(k, z)$ -subspace approximation (in Section [5](#)),  $(k, z)$ -flat approximation (in Section [6](#)), and finally  $(k, z)$ -line approximation (in Section [7](#)).

## 2 Overview of Techniques

In this subsection, we give a high-level overview of the techniques employed. As it will turn out, all results in this paper follow from one general technique, which we instantiate for the various problem instances.

We give an abstract instantiation of the approach. We will be concerned with geometric optimization problems of the following sort:

- For each  $d \in \mathbb{N}$ , we specify a class of candidate solutions given by a set  $\mathcal{C}_d$ . For example, in center-based clustering,  $\mathcal{C}_d$  may be given by a tuple of  $k$  points in  $\mathbb{R}^d$ , corresponding to  $k$  centers for a center-based clustering. In subspace approximation, the set  $\mathcal{C}_d$  may denote the set of all  $k$ -dimensional subspaces of  $\mathbb{R}^d$ .
- There will be a cost function  $f_d: \mathbb{R}^d \times \mathcal{C}_d \rightarrow \mathbb{R}_{\geq 0}$  which, takes a point  $x \in \mathbb{R}^d$  and a potential solution  $c \in \mathcal{C}_d$ , and outputs the cost of  $x$  on  $c$ . Continuing on the example on center-based clustering,  $f_d$  may denote the distance from a dataset point  $x \in \mathbb{R}^d$  to its nearest point in  $c$ . In subspace approximation,  $f_d$  may denote the distance from a dataset point to the orthogonal projection of that point onto the subspace  $c$ . For a parameter  $z \in \mathbb{N}$ , we will denote the cost of using  $c$  for a dataset  $X \subset \mathbb{R}^d$  by

$$\text{cost}_{d,z}(X, c) = \left( \sum_{x \in X} f_d(x, c)^z \right)^{1/z}.$$

For simplicity in the notation, we will drop the subscripts from the functions  $f$  and cost when they are clear from context.

We let  $\mathcal{J}_{d,t}$  denote a distribution over linear maps  $\Pi: \mathbb{R}^d \rightarrow \mathbb{R}^t$  which will satisfy some “Johnson-Lindenstrauss” guarantees (we will specify in the preliminaries the properties we will need). For concreteness, we will think of  $\Pi \sim \mathcal{J}_{d,t}$  given by matrix multiplication by a  $t \times d$  matrix of i.i.d draws of  $\mathcal{N}(0, 1/t)$ . We ask, for a particular bound on the dataset size  $n \in \mathbb{N}$ , a geometric optimization problem (specified by  $\{\mathcal{C}_d\}_{d \in \mathbb{N}}$ ,  $f_d$  and  $z$ ), and a parameter  $\varepsilon \in (0, 1)$ , what is the smallest  $t \in \mathbb{N}$  such that with probability at least 0.9 over the draw of  $\Pi \sim \mathcal{J}_{d,t}$ ,

$$(2.3) \quad \frac{1}{1 + \varepsilon} \cdot \min_{c \in \mathcal{C}_d} \text{cost}(X, c) \leq \min_{c \in \mathcal{C}_t} \text{cost}(\Pi(X), c) \leq (1 + \varepsilon) \cdot \min_{c \in \mathcal{C}_d} \text{cost}(X, c).$$

The right-most inequality in (2.3) claims that the cost after applying  $\Pi$  does not increase significantly, i.e.,  $\min_{c \in \mathcal{C}_t} \text{cost}(\Pi(X), c) \leq (1 + \varepsilon) \min_{c \in \mathcal{C}_d} \text{cost}(X, c)$ . This direction is easy to prove for the following reason. For a dataset  $X \subset \mathbb{R}^d$ , we consider the solution  $c^* \in \mathcal{C}_d$  minimizing  $\text{cost}(X, c^*)$ . We sample  $\Pi \sim \mathcal{J}_{d,t}$  and we find a candidate solution  $c^{**} \in \mathcal{C}_t$  which exhibits an upper bound on  $\min_{c \in \mathcal{C}_t} \text{cost}(\Pi(X), c) \leq \text{cost}(\Pi(X), c^{**})$ . For example, in the center-based clustering,  $c^* \in \mathcal{C}_d$  is a set of  $k$  centers in  $\mathbb{R}^d$ , and we may consider  $c^{**} \in \mathcal{C}_t$  as the  $k$  centers from  $c^*$  after applying  $\Pi$ . The fact that  $\text{cost}(\Pi(X), c^{**}) \leq (1 + \varepsilon) \text{cost}(X, c^*)$  with high probability over  $\Pi \sim \mathcal{J}_{d,t}$  will follow straight-forwardly from properties of  $\mathcal{J}_{d,t}$ . Importantly, the optimal solution  $c^*$  does not depend on  $\Pi \sim \mathcal{J}_{d,t}$ . In fact, while we expect  $\Pi: \mathbb{R}^d \rightarrow \mathbb{R}^t$  to distort some distances substantially, we can pick  $c^{**} \in \mathcal{C}_t$  so that too many distortions on these points is unlikely.

However, the same reasoning does not apply to the left-most inequality in (2.3). This is because the solution  $c^{**} \in \mathcal{C}_t$  which minimizes  $\min_{c \in \mathcal{C}_t} \text{cost}(\Pi(X), c)$  depends on  $\Pi$ . Indeed, we would expect  $c^{**} \in \mathcal{C}_t$  to take advantage of distortions in  $\Pi$  in order to decrease the cost of the optimal solution. We proceed by the following high level plan. We identify a sequence of important events defined over the draw of  $\Pi \sim \mathcal{J}_{d,t}$  which occur with probability at least 0.9. The special property is that if  $\Pi$  satisfies these events, we can identify, from  $c^{**} \in \mathcal{C}_t$  minimizing  $\text{cost}(\Pi(X), c^{**})$ , a candidate solution  $c^* \in \mathcal{C}_d$  which exhibits an upper bound  $\text{cost}(X, c^*) \leq (1 + \varepsilon) \text{cost}(\Pi(X), c^{**})$ .

We now specify how exactly we define, for an optimal  $c^{**} \in \mathcal{C}_t$  (depending on  $\Pi$ ), a candidate solution  $c^* \in \mathcal{C}_d$  whose cost is not much higher than  $\text{cost}(\Pi(X), c^{**})$ . For that, we will use the notion of coresets. Before the formal definition, we note there is a natural extension of cost for weighted datasets. In particular, if  $S \subset \mathbb{R}^d$  is a set of points and  $w: S \rightarrow \mathbb{R}_{\geq 0}$  is a set of weights for  $S$ , then we will use  $\text{cost}((S, w), c)$  as  $1/z$ -th power of the sum over all  $x \in S$  of  $w(x) \cdot f_d(x, c)^z$ .

**DEFINITION 2.1.** ((WEAK)<sup>3</sup> CORESETS, SEE ALSO [22]) For  $d \in \mathbb{N}$ , let  $\mathcal{C}_d$  denote a class of candidate solutions and  $f: \mathbb{R}^d \times \mathcal{C}_d \rightarrow \mathbb{R}_{\geq 0}$  specify the cost of a point to a solution. For a dataset  $X \subset \mathbb{R}^d$  and a parameter  $\varepsilon \in (0, 1)$ , a (weak)  $\varepsilon$ -coreset for  $X$  is a weighted set of points  $S \subset \mathbb{R}^d$  and  $w: S \rightarrow \mathbb{R}_{\geq 0}$  which satisfy

$$\frac{1}{1 + \varepsilon} \cdot \min_{c \in \mathcal{C}_d} \text{cost}(X, c) \leq \min_{c \in \mathcal{C}_d} \text{cost}((S, w), c) \leq (1 + \varepsilon) \cdot \min_{c \in \mathcal{C}_d} \text{cost}(X, c).$$



It will be crucial for us that these problems admit small coresets. More specifically, for the problems considered in this paper, there exists algorithms which can produce small-size coresets from a dataset. In what follows, **ALG** is a randomized algorithm which receives as input a dataset  $X \subset \mathbb{R}^d$  and outputs a weighted subset of points  $(\mathbf{S}, \mathbf{w})$  which is a weak  $\varepsilon$ -coreset for  $X$  with high probability. Computationally, the benefit of using coresets is that the sets  $\mathbf{S}$  tend to be much smaller than  $X$ , so that one may compute on  $(\mathbf{S}, \mathbf{w})$  and obtain an approximately optimal solution for  $X$ . For us, the benefit will come in defining the important events. At a high level, since  $\mathbf{S}$  is small, the important events defined with respect to  $\Pi$  will only worry about distortions within the subset (or subspace spanned by)  $\mathbf{S}$ .

Consider the following approach:

1. We begin with the original dataset  $X \subset \mathbb{R}^d$ , and we consider the solution  $c \in \mathcal{C}_d$  which minimizes  $\text{cost}(X, c)$ . The goal is to show that  $\text{cost}(X, c)$  cannot be much larger than  $\text{cost}(\Pi(X), c^{**})$ , where  $c^{**} \in \mathcal{C}_t$  minimizes  $\text{cost}(\Pi(X), c^{**})$ .
2. Instead of considering the entire dataset  $X$ , we execute **ALG**( $X$ ) and consider the weak  $\varepsilon$ -coreset  $(\mathbf{S}, \mathbf{w})$  that we obtain. If we can identify a candidate solution  $c^* \in \mathcal{C}_d$  whose cost  $\text{cost}((\mathbf{S}, \mathbf{w}), c^*) \leq (1 + \varepsilon)\text{cost}(\Pi(X), c^{**})$ , we would be done. Indeed,  $\min_{c \in \mathcal{C}_d} \text{cost}((\mathbf{S}, \mathbf{w}), c) \leq \text{cost}((\mathbf{S}, \mathbf{w}), c^*)$ , and the fact  $(\mathbf{S}, \mathbf{w})$  is a weak  $\varepsilon$ -coreset implies  $\min_{c \in \mathcal{C}_d} \text{cost}(X, c) \leq (1 + \varepsilon) \min_{c \in \mathcal{C}_d} \text{cost}((\mathbf{S}, \mathbf{w}), c)$ .
3. Moving to a coreset  $(\mathbf{S}, \mathbf{w})$  allows one to relate  $\text{cost}((\mathbf{S}, \mathbf{w}), c^*)$  and  $\text{cost}((\Pi(\mathbf{S}), \mathbf{w}), c^{**})$  by considering the performance of  $\Pi$  on  $\mathbf{S}$ . The benefit is that the important events, defined over the draw of  $\Pi \sim \mathcal{J}_{d,t}$ , set  $t$  as a function of  $|\mathbf{S}|$ , instead of  $|X|$ . A useful example to consider is requiring  $\Pi^{-1}$  be  $(1 + \varepsilon)$ -Lipschitz on the entire subspace spanned by  $\mathbf{S}$ , which requires  $t = \Theta(|\mathbf{S}|/\varepsilon^2)$ . For the problems considered here, a nearly optimal  $c^{**} \in \mathcal{C}_t$  for  $(\mathbf{S}, \mathbf{w})$  will be in the subspace spanned by  $\mathbf{S}$ , so we may identify  $c^* \in \mathcal{C}_d$  whose cost on  $(\mathbf{S}, \mathbf{w})$  is not much higher than the  $\text{cost}((\Pi(\mathbf{S}), \mathbf{w}), c^{**})$  by evaluating  $\Pi^{-1}(c^{**})$  since  $c^{**}$  lies inside  $\text{span}(\mathbf{S})$ <sup>4</sup>.
4. The remaining step is showing  $\text{cost}((\Pi(\mathbf{S}), \mathbf{w}), c^{**}) \leq (1 + \varepsilon)\text{cost}(\Pi(X), c^{**})$ . In particular, one would like to claim  $(\Pi(\mathbf{S}), \mathbf{w})$  is a weak  $\varepsilon$ -coreset for  $\Pi(X)$  and use the right-most inequality in Definition 2.1. However, it is not clear this is so. The problem is that the algorithm **ALG** depends on the  $d$ -dimensional representation of  $X \subset \mathbb{R}^d$ , and  $(\Pi(\mathbf{S}), \mathbf{w})$  may not be a valid output for **ALG**( $\Pi(X)$ ). As we show, this does work for (some) coreset algorithms built on the sensitivity sampling framework (see, [23, 9])<sup>5</sup>.

**2.0.1 Sensitivity Sampling for Step 4** In the remainder of this section, we briefly overview the sensitivity sampling framework, and the components required to make Step 4 go through. At a high level, coreset algorithms in the sensitivity sampling framework proceed in the following way. Given a dataset  $X \subset \mathbb{R}^d$ , the algorithm computes a sensitivity sampling distribution  $\tilde{\sigma}$  supported on  $X$ . The requirement is that, for each potential solution  $c \in \mathcal{C}_d$ , sampling from  $\tilde{\sigma}$  gives a low-variance estimator for  $\text{cost}_{d,z}(X, c)^z$ . In particular, we let  $\tilde{\sigma}(x)$  be the probability of sampling  $x \in X$ . Then, for any distribution  $\tilde{\sigma}$  and any  $c \in \mathcal{C}_d$ ,

$$(2.4) \quad \mathbf{E}_{\mathbf{x} \sim \tilde{\sigma}} \left[ \frac{1}{\tilde{\sigma}(\mathbf{x})} \cdot \frac{f_d(\mathbf{x}, c)^z}{\text{cost}_{d,z}(X, c)^z} \right] = 1.$$

Equation 2.4 implies that, for any  $m \in \mathbb{N}$ , if  $\mathbf{S}$  is  $m$  i.i.d samples from  $\tilde{\sigma}$  and  $\mathbf{w}(x) = 1/(m\tilde{\sigma}(x))$ , the expectation of  $\text{cost}_{d,z}((\mathbf{S}, \mathbf{w}), c)^z$  is  $\text{cost}_{d,z}(X, c)^z$ . In addition, the algorithm designs  $\tilde{\sigma}$  so that, for a parameter  $T > 0$ ,

$$(2.5) \quad \sup_{c \in \mathcal{C}_d} \mathbf{E}_{\mathbf{x} \sim \tilde{\sigma}} \left[ \left( \frac{1}{\tilde{\sigma}(\mathbf{x})} \cdot \frac{f_d(\mathbf{x}, c)^z}{\text{cost}_{d,z}(X, c)^z} \right)^2 \right] \leq T.$$

If we set  $m \geq T/\varepsilon^2$ , (2.5) and Chebyshev's inequality implies  $\text{cost}_{d,z}((\mathbf{S}, \mathbf{w}), c)^z \approx_{1 \pm \varepsilon} \text{cost}_{d,z}(X, c)^z$  for each  $c \in \mathcal{C}_d$  with a high constant probability, and the remaining work is in increasing  $m$  by a large enough factor to “union

<sup>4</sup>While the above results in bounds for  $t$  which are already meaningful, we will exploit other geometric aspects of the problems considered to get bounds on  $t$  which are logarithmic in the coreset size. For center-based clustering, [40] showed that one may apply Kirzbraun's theorem. For subspace approximation, we use the geometric lemmas of [44].

<sup>5</sup>We will not prove that with high probability over  $\Pi$  and **ALG**( $X$ ),  $(\Pi(\mathbf{S}), \mathbf{w})$  is a weak  $\varepsilon$ -coreset for  $\Pi(X)$ . Rather, all we need is that the right-most inequality in Definition 2.1 holds for  $(\Pi(\mathbf{S}), \mathbf{w})$  and  $\Pi(X)$ , which is what we show.

bound” over all  $c \in \mathcal{C}_d$ . There is a canonical way of ensuring  $\tilde{\sigma}$  and  $T$  satisfy (2.5): we first define  $\sigma: X \rightarrow \mathbb{R}_{\geq 0}$ , known as a “sensitivity function”, which sets for each  $x \in X$ ,

$$(2.6) \quad \sigma(x) \geq \sup_{c \in \mathcal{C}_d} \frac{f_d(x, c)^z}{\text{cost}_{d,z}(X, c)^z}, \quad \text{and} \quad T = \sum_{x \in X} \sigma(x),$$

which is known as the “total sensitivity.” Then, the distribution is given by letting  $\tilde{\sigma}(x) = \sigma(x)/T$ .

We now show how to incorporate the map  $\mathbf{\Pi} \sim \mathcal{J}_{d,t}$ , to argue Step 4. Recall that we let  $\mathbf{S}$  denote  $m$  i.i.d draws from  $\tilde{\sigma}$  and the weights be  $\mathbf{w}(x) = 1/(m\tilde{\sigma}(x))$ . We want to argue that, with high constant probability over the draw of  $(\mathbf{S}, \mathbf{w})$  and  $\mathbf{\Pi} \sim \mathcal{J}_{d,t}$ , we have

$$(2.7) \quad \text{cost}_{t,z}((\mathbf{\Pi}(\mathbf{S}), \mathbf{w}), c^{**}) \leq (1 + \varepsilon) \cdot \text{cost}_{t,z}(\mathbf{\Pi}(X), c^{**}).$$

First, note that the analogous version of (2.4) for  $\text{cost}_{t,z}(\mathbf{\Pi}(X), c)$  continues to hold. In particular, for any map  $\mathbf{\Pi}$  in the support of  $\mathcal{J}_{d,t}$  and  $c^{**} \in \mathcal{C}_t$  minimizing  $\text{cost}_{t,z}(\mathbf{\Pi}(X), c)$ ,

$$(2.8) \quad \mathbf{E}_{\mathbf{x} \sim \tilde{\sigma}} \left[ \frac{1}{\tilde{\sigma}(\mathbf{x})} \cdot \frac{f_t(\mathbf{\Pi}(\mathbf{x}), c^{**})^z}{\text{cost}_{t,z}(\mathbf{\Pi}(X), c^{**})^z} \right] = 1.$$

Hence, it remains to define  $\tilde{\sigma}$  satisfying (2.6) which also satisfies one additional requirement. With high probability over  $\mathbf{\Pi} \sim \mathcal{J}_{d,t}$ , we should have

$$(2.9) \quad \mathbf{E}_{\mathbf{x} \sim \tilde{\sigma}} \left[ \left( \frac{1}{\tilde{\sigma}(\mathbf{x})} \cdot \frac{f_t(\mathbf{\Pi}(\mathbf{x}), c^{**})^z}{\text{cost}_{t,z}(\mathbf{\Pi}(X), c^{**})^z} \right)^2 \right] \lesssim T.$$

The above translates to saying, for most  $\mathbf{\Pi} \sim \mathcal{J}_{d,t}$  the variance of  $\text{cost}((\mathbf{\Pi}(\mathbf{S}), \mathbf{w}), c^{**})$ , when  $m = O(T/\varepsilon^2)$ , is small. Once that is established, we may apply Chebyshev’s inequality and conclude (2.7) with high constant probability.<sup>6</sup>

**2.0.2 The Circularity and How to Break It** One final technical hurdle arises. While one may define the sensitivity function  $\sigma(x)$  to be exactly  $\sup_{c \in \mathcal{C}_d} f_d(x, c)^z / \text{cost}_{d,z}(X, c)^z$  and automatically satisfy (2.6), it becomes challenging to argue that (2.9) holds. In the end, the complexity we seek to optimize is the total sensitivity  $T$ , so there is flexibility in defining  $\sigma$  while showing (2.9) holds. In fact, sensitivity functions  $\sigma$  which are computationally simple tend to be known, since an algorithm using coresets must quickly compute  $\sigma(x)$  for every  $x \in X$ .

The sensitivity functions  $\sigma$  used in the literature (for instance, in [23, 49]) are defined with respect to an approximately optimal  $c \in \mathcal{C}_d$  (or bi-criteria approximation) for  $\text{cost}_{d,z}(X, c)$ . Furthermore, the arguments used to show these function satisfy (2.6), which we will also employ for (2.9), crucially utilize the approximation guarantee on  $c \in \mathcal{C}_d$ . The apparent circularity appears in approximation algorithms and also shows up in the analysis here:

- For  $X \subset \mathbb{R}^d$ , we identify the optimal  $c \in \mathcal{C}_d$  minimizing  $\text{cost}_{d,z}(X, c)$ , and use  $c$  to define  $\sigma: X \rightarrow \mathbb{R}_{\geq 0}$ . The fact that  $c \in \mathcal{C}_d$  is optimal (and therefore approximately optimal) allows us to use known arguments (in particular, those in [49, 48]) to establish (2.6) and give an upper bound on  $T$ .
- We use the proof of the “easy” direction to identify a solution  $c' \in \mathcal{C}_t$  with  $\text{cost}_{t,z}(\mathbf{\Pi}(X), c') \leq (1 + \varepsilon) \text{cost}_{d,z}(X, c)$  (recall this was used to establish the right-most inequality in (2.3)). From the analytical perspective, it is useful to think of  $\sigma': \mathbf{\Pi}(X) \rightarrow \mathbb{R}_{\geq 0}$  as the function one would get from defining a sensitivity function like in the previous step with  $c'$  instead of  $c$ . If we could show  $c' \in \mathcal{C}_t$  was approximately optimal for  $\mathbf{\Pi}(X)$ , we could use [49, 48] again to argue (2.9). The circularity is the following. Showing  $c' \in \mathcal{C}_t$  is approximately optimal means showing an upper bound on  $\min_{c \in \mathcal{C}_t} \text{cost}_{t,z}(\mathbf{\Pi}(X), c)$  in terms of  $\text{cost}_{t,z}(\mathbf{\Pi}(X), c')$ . Since, we picked  $\text{cost}_{t,z}(\mathbf{\Pi}(X), c')$  to be at most  $(1 + \varepsilon) \min_{c \in \mathcal{C}_d} \text{cost}_{d,z}(X, c)$ , this is exactly what we sought to prove.

<sup>6</sup>Since Steps 1–4 only argued about the optimal  $c^{**} \in \mathcal{C}_t$ , there is no need to “union bound” over all  $c \in \mathcal{C}_t$  in our arguments.

If “approximately optimal” above required  $c' \in \mathcal{C}_t$  be a  $(1 + \varepsilon)$ -approximation to the optimal  $\Pi(X)$ , we would have a complete circularity and be unable to proceed. However, similarly to the case of approximation algorithms, it suffices to have a poor approximation. Suppose we showed  $c' \in \mathcal{C}_t$  was a  $C$ -approximation, then, increasing  $m$  by a factor depending on  $C$  (which would affect the resulting dimensionality  $t$ ) would account for this increase and drive the variance back down to  $\varepsilon^2$ . Moreover, showing  $c'$  is a  $O(1)$ -approximation with probability at least 0.99 over  $\Pi \sim \mathcal{J}_{d,t}$ , given Steps 1–4 is straight-forward. Instead of showing the stronger bound that  $\text{cost}_{t,z}((\Pi(\mathbf{S}), \mathbf{w}), c^{**}) \leq (1 + \varepsilon) \text{cost}_{t,z}(\Pi(X), c^{**})$ , we show that  $\text{cost}_{t,z}((\Pi(\mathbf{S}), \mathbf{w}), c^{**}) \leq O(1) \cdot \text{cost}_{t,z}(\Pi(X), c^{**})$ . The latter (loose) bound is a consequence of applying Markov’s inequality to (2.8).

In summary, we overcome the circularity by going through Steps 1–4 twice. In the first time, we show a weaker  $O(1)$ -approximation. Specifically, we show that  $\Pi \sim \mathcal{J}_{d,t}$  preserves the cost of  $\min_{c \in \mathcal{C}_t} \text{cost}_{d,z}(\Pi(X), c)$  up to factor  $O(1)$ . The first time around, we won’t upper bound the variance in (2.9), and we simply use Markov’s inequality to (2.8) in order to prove a (loose) bound on Step 4. Once we’ve established the  $O(1)$ -factor approximation, we are guaranteed that  $c' \in \mathcal{C}_t$  is a  $O(1)$ -approximation to  $\min_{c \in \mathcal{C}_t} \text{cost}_{t,z}(\Pi(X), c)$ . This means that, actually, the sensitivity sampling distribution  $\tilde{\sigma}$  we had considered (when viewed as a sensitivity sampling distribution for  $\Pi(X)$ ) gives estimators with bounded variance, as in (2.9). In particular, going through Steps 1–4 once again implies that  $c' \in \mathcal{C}_t$  was actually the desired  $1 \pm \varepsilon$ -approximation.

### 3 Preliminaries

We specify the properties we use from the distribution  $\mathcal{J}_{d,t}$ . We will refer to these as “Johnson-Lindenstrauss” distributions. Throughout the proof, we will often refer to  $\mathcal{J}_{d,t}$  as given by a  $t \times d$  matrix of i.i.d draws from  $\mathcal{N}(0, 1/t)$ . The goal of specifying the useful properties is to use other “Johnson-Lindenstrauss”-like distributions. The first property we need is that  $\Pi: \mathbb{R}^d \rightarrow \mathbb{R}^t$  is a linear map, and that any  $x, y \in \mathbb{R}^d$  satisfies

$$\mathbf{E}_{\Pi \sim \mathcal{J}_{d,t}} \left[ \frac{\|\Pi(x) - \Pi(y)\|_2^2}{\|x - y\|_2^2} \right] = 1.$$

We use the standard property of  $\mathcal{J}_{d,t}$ , that  $\Pi$  preserves distances with high probability, i.e., for any  $x, y \in \mathbb{R}^d$ ,

$$\Pr_{\Pi \sim \mathcal{J}_{d,t}} \left[ \left| \frac{\|\Pi(x) - \Pi(y)\|_2^2}{\|x - y\|_2^2} - 1 \right| \geq \varepsilon \right] \leq e^{-\Omega(\varepsilon^2 t)}.$$

More generally, we use the conclusion of the following lemma. We give a proof when  $\mathcal{J}_{d,t}$  is a  $t \times d$  i.i.d entries of  $\mathcal{N}(0, 1/t)$ .

**LEMMA 3.1.** *Let  $\mathcal{J}_{d,t}$  denote a Johnson-Lindenstrauss distribution over maps  $\mathbb{R}^d \rightarrow \mathbb{R}^t$  given by a matrix multiplication on the left by a  $t \times d$  matrix of i.i.d draws of  $\mathcal{N}(0, 1/t)$ . If  $t \gtrsim z/\varepsilon^2$ , then for any  $x, y \in \mathbb{R}^d$ ,*

$$\mathbf{E}_{\Pi \sim \mathcal{J}_{d,t}} \left[ \left( \frac{\|\Pi(x) - \Pi(y)\|_2^z}{\|x - y\|_2^z} - 1 \right)^+ \right] \leq \frac{(1 + \varepsilon)^z - 1}{100} \quad \text{7}$$

*Proof.* We note that by the 2-stability of the Gaussian distribution, we have  $\|\Pi(x) - \Pi(y)\|_2^2$  is equivalently distributed as  $\|\mathbf{g}\|_2^2 \cdot \|x - y\|_2^2$ , where  $\mathbf{g} \sim \mathcal{N}(0, I_t/t)$ . Therefore, we have that for any  $\lambda > 0$  which we will optimize shortly to be a small constant times  $\varepsilon$ ,

$$\mathbf{E}_{\Pi \sim \mathcal{J}_{d,t}} \left[ \left( \frac{\|\Pi(x) - \Pi(y)\|_2^z}{\|x - y\|_2^z} - 1 \right)^+ \right] \leq (1 + \lambda)^z - 1 + \mathbf{E}_{\mathbf{g} \sim \mathcal{N}(0, I_t/t)} \left[ (\|\mathbf{g}\|_2^z - (1 + \lambda)^z)^+ \right].$$

Furthermore,

$$\mathbf{E}_{\mathbf{g} \sim \mathcal{N}(0, I_t/t)} \left[ (\|\mathbf{g}\|_2^z - (1 + \lambda)^z)^+ \right] = \int_{u: (1+\lambda)^z}^{\infty} \Pr[\|\mathbf{g}\|_2^z \geq u] du = \frac{z}{2} \int_{v: (1+\lambda)^2}^{\infty} \Pr[\|\mathbf{g}\|_2^2 \geq v] \cdot v^{z/2-1} dv.$$

<sup>7</sup>We use the notation  $(t)^+ = \max\{t, 0\}$ .



We will upper bound the probability that  $\|\mathbf{g}\|_2^2$  exceeds  $v$  by the Chernoff-Hoeffding method. In particular, recall that  $\|\mathbf{g}\|_2^2$  when  $\mathcal{N}(0, I_t/t)$  is distributed as a  $\chi^2$ -random variable with  $t$  degrees of freedom, rescaled by  $1/t$ , such that the moment generating function of  $\|\mathbf{g}\|_2^2$  has the following closed form solution whenever  $\alpha < t/2$ :

$$\log \left( \mathbf{E}_{\mathbf{g} \sim \mathcal{N}(0, I_t/t)} [\exp(\alpha \|\mathbf{g}\|_2^2)] \right) = -\frac{t}{2} \log \left( 1 - \frac{2\alpha}{t} \right) \leq \alpha + \frac{2\alpha^2}{t}$$

In particular, for any  $\alpha < t/2$ , we may upper bound

$$\begin{aligned} \frac{z}{2} \int_{v:(1+\lambda)^2}^{\infty} \Pr [\|\mathbf{g}\|_2^2 \geq v] v^{z/2-1} &\leq \frac{z}{2} \cdot \exp \left( \alpha + \frac{2\alpha^2}{t} \right) \int_{v:(1+\lambda)^2}^{\infty} \exp \left( - \left( \alpha - \frac{z}{2} - 1 \right) v \right) dv \\ &\leq \frac{z}{2\alpha - z - 2} \cdot \exp \left( \frac{2\alpha^2}{t} - \alpha\lambda + (1+\lambda)(z/2 + 1) \right) < \frac{\lambda}{100}, \end{aligned}$$

by letting setting  $\alpha = t\lambda/10$  whenever  $t \gtrsim z/\lambda^2$ . Setting  $\lambda$  to be a small constant of  $\varepsilon$  gives the desired guarantees.  $\square$

**DEFINITION 3.1. (SUBSPACE EMBEDDINGS)** *Let  $d \in \mathbb{N}$  and  $A \subset \mathbb{R}^d$  denote a subspace of  $\mathbb{R}^d$ . For  $\varepsilon > 0$ , a map  $f: \mathbb{R}^d \rightarrow \mathbb{R}^t$  is an  $\varepsilon$ -subspace embedding of  $A$  if, for any  $x \in A$ ,*

$$\frac{1}{1+\varepsilon} \cdot \|f(x)\|_2 \leq \|x\|_2 \leq (1+\varepsilon) \cdot \|f(x)\|_2.$$

**LEMMA 3.2.** *Let  $d \in \mathbb{N}$  and  $A \subset \mathbb{R}^d$  be a subspace of dimension at most  $k$ . For  $\varepsilon, \delta \in (0, 1/2)$ , let  $\mathcal{J}_{d,t}$  denote a Johnson-Lindenstrauss distribution over maps  $\mathbb{R}^d \rightarrow \mathbb{R}^t$  given by a matrix multiplication on the left by a  $t \times d$  matrix of i.i.d draws of  $\mathcal{N}(0, 1/t)$ . If  $t \sim (k + \log(1/\delta))/\varepsilon^2$ , then  $\mathbf{\Pi} \sim \mathcal{J}_{d,t}$  is an  $\varepsilon$ -subspace embedding of  $A$  with probability at least  $1 - \delta$ .*

#### 4 Center-based $(k, z)$ -clustering

In the  $(k, z)$ -clustering problems, for any set  $C \subset \mathbb{R}^d$  of  $k$  points, and point  $x \in \mathbb{R}^d$ , we write

$$\text{cost}_z^z(x, C) = \min_{c \in C} \|x - c\|_2^z,$$

and for a subset  $X \subset \mathbb{R}^d$ ,

$$\text{cost}_z^z(X, C) = \sum_{x \in X} \text{cost}_z^z(x, C) = \sum_{x \in X} \min_{c \in C} \|x - c\|_2^z.$$

We extend the above notation to weighted subsets, where for a subset  $S \subset \mathbb{R}^d$  with (non-negative) weights  $w: S \rightarrow \mathbb{R}_{\geq 0}$ , we write  $\text{cost}_z^z((S, w), C) = \sum_{x \in S} w(x) \min_{c \in C} \|x - c\|_2^z$ . The main result of this section is the following theorem.

**THEOREM 4.1. (JOHNSON-LINDENSTRAUSS FOR CENTER-BASED CLUSTERING)** *Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be any set of points, and let  $C \subset \mathbb{R}^d$  denote the optimal  $(k, z)$ -clustering of  $X$ . For any  $\varepsilon \in (0, 1/2)$ , suppose we let  $\mathcal{J}_{d,t}$  be the distribution over Johnson-Lindenstrauss maps where*

$$t \gtrsim \frac{\log k + z \log(1/\varepsilon)}{\varepsilon^2},$$

*Then, with probability at least 0.9 over the draw of  $\mathbf{\Pi} \sim \mathcal{J}_{d,t}$ ,*

$$\frac{1}{1+\varepsilon} \cdot \text{cost}_z(X, C) \leq \min_{\substack{C' \subset \mathbb{R}^t \\ |C'|=k}} \text{cost}_z(\mathbf{\Pi}(X), C') \leq (1+\varepsilon) \text{cost}_z(X, C).$$

There are two directions to showing dimension reduction: (1) the optimal clustering in the *reduced space* is not too expensive, and (2) the optimal clustering in the reduced space is not too cheap. We note that (1) is simple because we can exhibit a clustering in the reduced space whose cost is not too high; however, (2) is much trickier, since we need to rule out a too-good-to-be-true clustering in the reduced space.

#### 4.1 Easy Direction: Optimum Cost Does Not Increase

LEMMA 4.1. Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be any set of points and let  $C \subset \mathbb{R}^d$  of size  $k$  be the centers of the optimal  $(k, z)$ -clustering of  $X$ . We let  $\mathcal{J}_{d,t}$  be the distribution over Johnson-Lindenstrauss maps. If  $t \gtrsim z/\varepsilon^2$ , then with probability at least 0.99 over the draw of  $\Pi \sim \mathcal{J}_{d,t}$ ,

$$\left( \sum_{x \in X} \|\Pi(x) - \Pi(c(x))\|_2^z \right)^{1/z} \leq (1 + \varepsilon) \text{cost}_z(X, C),$$

and hence,

$$\min_{\substack{C' \subset \mathbb{R}^t \\ |C'|=k}} \text{cost}_z(\Pi(X), C') \leq (1 + \varepsilon) \min_{\substack{C \subset \mathbb{R}^d \\ |C|=k}} \text{cost}_z(X, C).$$

*Proof.* For  $x \in X$ , let  $c(x) \in C$  denote the closest point from  $C$  to  $x$ . We compute the expected positive deviation from assigning  $\Pi(x)$  to  $\Pi(c(x))$ , and note that the  $\text{cost}_z(\Pi(X), \Pi(C))$  can only be lower. Hence, if we can show

$$(4.10) \quad \frac{1}{\text{cost}_z^z(X, C)} \mathbf{E}_{\Pi \sim \mathcal{J}_{d,t}} \left[ \sum_{x \in X} (\|\Pi(x) - \Pi(c(x))\|_2^z - \|x - c(x)\|_2^z)^+ \right] \lesssim \frac{(1 + \varepsilon)^z - 1}{100},$$

then by Markov's inequality, we will obtain  $\text{cost}_z^z(\Pi(X), \Pi(C)) \leq (1 + \varepsilon)^z \text{cost}_z^z(X, C)$ , and obtain the desired bound when raising to power  $1/z$ . This last part follows from Lemma 3.1  $\square$

**4.2 Hard Direction: Optimum Cost Does Not Decrease** We now show that after applying a Johnson-Lindenstrauss map, the cost of the optimal clustering in the dimension-reduced space is not too cheap. This section will be significantly more difficult, and will draw on the following preliminaries.

##### 4.2.1 Preliminaries

DEFINITION 4.1. (WEAK AND STRONG CORESETS) Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be a set of points. A (weak)  $\varepsilon$ -coreset of  $X$  for  $(k, z)$ -clustering is a subset  $S \subset \mathbb{R}^d$  of points with weights  $w: S \rightarrow \mathbb{R}_{\geq 0}$  such that,

$$\frac{1}{1 + \varepsilon} \cdot \min_{\substack{C \subset \mathbb{R}^d \\ |C| \leq k}} \text{cost}_z(X, C) \leq \min_{\substack{C \subset \mathbb{R}^d \\ |C| \leq k}} \text{cost}_z((S, w), C) \leq (1 + \varepsilon) \cdot \min_{\substack{C \subset \mathbb{R}^d \\ |C| \leq k}} \text{cost}_z(X, C).$$

The coreset  $(S, w)$  is a strong  $\varepsilon$ -coreset if, for all  $C = \{c_1, \dots, c_k\} \subset \mathbb{R}^d$ , we have

$$\frac{1}{1 + \varepsilon} \cdot \text{cost}_z(X, C) \leq \text{cost}_z((S, w), C) \leq (1 + \varepsilon) \cdot \text{cost}_z(X, C).$$

Notice that Definition 4.1 gives an approach to finding an approximately optimal  $(k, z)$ -clustering. Given  $X \subset \mathbb{R}^d$ , we find a (weak)  $\varepsilon$ -coreset  $(S, w)$  and find the optimal clustering  $C \subset \mathbb{R}^d$  with respect to the coreset  $(S, w)$ . Then, we can deduce that a clustering which is optimal for the coreset is also approximately optimal for the original point set. A common and useful framework for building coresets is by utilizing the “sensitivity” sampling framework.

DEFINITION 4.2. (SENSITIVITIES) Let  $n, d \in \mathbb{N}$ , and consider any set of points  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ , as well as  $k \in \mathbb{N}$ ,  $z \geq 1$ . A sensitivity function  $\sigma: X \rightarrow \mathbb{R}_{\geq 0}$  for  $(k, z)$ -clustering  $X$  in  $\mathbb{R}^d$  is a function satisfying, that for all  $x \in X$ ,

$$\sup_{\substack{C \subset \mathbb{R}^d \\ |C| \leq k}} \frac{\text{cost}_z^z(x, C)}{\text{cost}_z^z(X, C)} \leq \sigma(x).$$

The total sensitivity of the sensitivity function  $\sigma$  is given by

$$\mathfrak{S}_\sigma = \sum_{x \in X} \sigma(x).$$

For a sensitivity function, we let  $\tilde{\sigma}$  denote the sensitivity sampling distribution, supported on  $X$ , which samples  $x \in X$  with probability proportional to  $\sigma(x)$ .

The following lemma gives a particularly simple sensitivity sampling distribution, which will be useful for analyzing our dimension reduction procedure. The proof below will follow from two applications of the triangle inequality which we reproduce from Claim 5.6 in [31].

LEMMA 4.2. *Let  $n, d \in \mathbb{N}$  and consider a set of points  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ . Let  $C \subset \mathbb{R}^d$  of size  $k$  be optimal  $(k, z)$ -clustering of  $X$ , and let  $c: X \rightarrow C$  denote the function which sends  $x \in X$  to its closest point in  $C$ , and let  $X_c \subset X$  be the set of points where  $c(x) = c$ . Then, the function  $\sigma: X \rightarrow \mathbb{R}_{\geq 0}$  given by*

$$\sigma(x) = 2^{z-1} \cdot \frac{\|x - c(x)\|_2^z}{\text{cost}_z^z(X, C)} + \frac{2^{2z-1}}{|X_{c(x)}|}$$

*is a sensitivity function for  $(k, z)$ -clustering  $X$  in  $\mathbb{R}^d$ , satisfying*

$$\mathfrak{S}_\sigma = 2^{z-1} + 2^{2z-1} \cdot k$$

*Proof.* Consider any set  $C' \subset \mathbb{R}^d$  of  $k$  points, and let  $c': X \rightarrow C'$  be the function which sends each  $x \in X$  to its closest point in  $C'$ . Then, we have

$$(4.11) \quad \|x - c'(x)\|_2^z \leq (\|x - c(x)\|_2 + \|c(x) - c'(x)\|_2)^z \leq 2^{z-1} \|x - c(x)\|_2^z + 2^{z-1} \|c(x) - c'(x)\|_2^z$$

$$(4.12) \quad \leq 2^{z-1} \|x - c(x)\|_2^z + \frac{2^{z-1}}{|X_{c(x)}|} \sum_{x \in X} \|c(x) - c'(x)\|_2^z$$

$$(4.13) \quad \leq 2^{z-1} \|x - c(x)\|_2^z + \frac{2^{2(z-1)}}{|X_{c(x)}|} (\text{cost}_z^z(X, C) + \text{cost}_z^z(X, C')),$$

where we used the triangle inequality and Hölder inequality in (4.11), and added additional non-negative terms in (4.12), and we finally apply the triangle inequality and Hölder's inequality once more in (4.13). Hence, using the fact that  $C$  is the optimal clustering, we have

$$\frac{\text{cost}_z^z(x, C')}{\text{cost}_z^z(X, C')} \leq 2^{z-1} \cdot \frac{\|x - c(x)\|_2^z}{\text{cost}_z^z(X, C')} + \frac{2^{2(z-1)}}{|X_{c(x)}|} \left( \frac{\text{cost}_z^z(X, C)}{\text{cost}_z^z(X, C')} + 1 \right) \leq 2^{z-1} \cdot \frac{\|x - c(x)\|_2^z}{\text{cost}_z^z(X, C')} + \frac{2^{2z-1}}{|X_{c(x)}|}.$$

The bound on  $\mathfrak{S}_\sigma$  follows from summing over all  $x \in X$ , noting the fact that  $\sum_{x \in X} 1/|X_{c(x)}| = k$ .  $\square$

The main idea behind the sensitivity sampling framework for building coresets is to sample from a sensitivity sampling distribution enough times in order to build a coreset. For this work, it will be sufficient to consider the following theorem of [31], which shows that  $\text{poly}(k, 1/\varepsilon^z)$  draws from an appropriate sensitivity sampling distribution suffices to build strong  $\varepsilon$ -coresets for  $(k, z)$ -clustering in  $\mathbb{R}^d$ .

THEOREM 4.2. ( $\varepsilon$ -STRONG CORESETS FROM SENSITIVITY SAMPLING [31]) *For any subset  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  and  $\varepsilon \in (0, 1/2)$ . Let  $C \subset \mathbb{R}^d$  of size  $k$  be the optimal  $(k, z)$ -clustering of  $X$ , and let  $\tilde{\sigma}$  denote the sensitivity sampling distribution of Lemma 4.2.*

- *Let  $(\mathbf{S}, \mathbf{w})$  denote a random (multi-)set  $\mathbf{S} \subset X$  and  $\mathbf{w}: \mathbf{S} \rightarrow \mathbb{R}_{\geq 0}$  given by, for  $m = \text{poly}(k, 1/\varepsilon^z)$  iterations, sampling  $\mathbf{x} \sim \tilde{\sigma}$  i.i.d and letting  $\mathbf{w}(\mathbf{x}) = 1/(m\tilde{\sigma}(\mathbf{x}))$ .*
- *Then, with probability  $1 - o(1)$  over the draw of  $(\mathbf{S}, \mathbf{w})$ , it is an  $\varepsilon$ -strong coreset for  $X$ .*

THEOREM 4.3. (KIRSZBRAUN THEOREM) *Let  $Y \subset \mathbb{R}^{d_1}$  and  $\phi: Y \rightarrow \mathbb{R}^{d_2}$  be an  $L$ -Lipschitz map (with respect to Euclidean norms on  $\mathbb{R}^{d_1}$  and  $\mathbb{R}^{d_2}$ ). There exists a map  $\tilde{\phi}: \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2}$  which is  $L$ -Lipschitz extending  $\phi$ , i.e.,  $\phi(x) = \tilde{\phi}(x)$  for all  $x \in Y$ .*

**4.3 The Important Events** We now define the important events which will allow us to prove that the optimum  $(k, z)$ -clustering after dimension reduction does not decrease substantially. We first define the events, and then we prove that if the events are all satisfied, then we obtain our desired lower bound.

DEFINITION 4.3. (THE EVENTS) Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  and  $C \subset \mathbb{R}^d$  of size  $k$  be centers for an optimal  $(k, z)$ -clustering of  $X$ , and  $\tilde{\sigma}$  is the sensitivity sampling distribution of  $X$  with respect to  $C$  as in Lemma 4.2. We will consider the following experiment,

1. We generate a sample  $(\mathbf{S}, \mathbf{w})$  by sampling from  $\tilde{\sigma}$  for  $m = \text{poly}(k, 1/\varepsilon^z)$  i.i.d iterations  $\mathbf{x} \sim \tilde{\sigma}$  and set  $\mathbf{w}(\mathbf{x}) = 1/(m\tilde{\sigma}(\mathbf{x}))$ .
2. Furthermore, we sample  $\mathbf{\Pi} \sim \mathcal{J}_{d,t}$  which is a Johnson-Lindenstrauss map  $\mathbb{R}^d \rightarrow \mathbb{R}^t$ .
3. We let  $\mathbf{S}' = \mathbf{\Pi}(\mathbf{S}) \subset \mathbb{R}^t$  denote the image of  $\mathbf{\Pi}$  on  $\mathbf{S}$ .

The events are the following:

- $\mathbf{E}_1$  : The weighted (multi-)set  $(\mathbf{S}, \mathbf{w})$  is a weak  $\varepsilon$ -coreset for  $(k, z)$ -clustering  $X$  in  $\mathbb{R}^d$ .
- $\mathbf{E}_2$  : The map  $\mathbf{\Pi}: \mathbf{S} \rightarrow \mathbf{S}'$ , given by restricting  $\mathbf{\Pi}$  is  $(1 + \varepsilon)$ -bi-Lipschitz.
- $\mathbf{E}_3(\beta)$  : We let  $\mathbf{C}' \subset \mathbb{R}^t$  of size  $k$  be the optimal centers for  $(k, z)$ -clustering  $\mathbf{\Pi}(X)$  in  $\mathbb{R}^t$ . The weighted (multi-)set  $(\mathbf{\Pi}(\mathbf{S}), \mathbf{w})$  satisfies

$$\text{cost}_z^z((\mathbf{\Pi}(\mathbf{S}), \mathbf{w}), \mathbf{C}') \leq \beta \cdot \text{cost}_z^z(\mathbf{\Pi}(X), \mathbf{C}').$$

LEMMA 4.3. Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ , and suppose  $(\mathbf{S}, \mathbf{w})$  and  $\mathbf{\Pi}: \mathbb{R}^d \rightarrow \mathbb{R}^t$  satisfy events  $\mathbf{E}_1, \mathbf{E}_2$  and  $\mathbf{E}_3(\beta)$ , then,

$$\min_{\substack{\mathbf{C}' \subset \mathbb{R}^t \\ |\mathbf{C}'|=k}} \text{cost}_z(\mathbf{\Pi}(X), \mathbf{C}') \geq \frac{1}{\beta^{1/z}(1+\varepsilon)^{1+1/z}} \cdot \min_{\substack{\mathbf{C} \subset \mathbb{R}^d \\ |\mathbf{C}|=k}} \text{cost}_z(X, \mathbf{C}).$$

*Proof.* Let  $\mathbf{C}' \subset \mathbb{R}^t$  of size  $k$  denote the centers which give the optimal  $(k, z)$ -clustering of  $\mathbf{\Pi}(X)$  in  $\mathbb{R}^t$ . Then, by  $\mathbf{E}_3$ ,

$$\text{cost}_z^z(\mathbf{\Pi}(X), \mathbf{C}') \geq (1/\beta) \cdot \text{cost}_z^z((\mathbf{\Pi}(\mathbf{S}), \mathbf{w}), \mathbf{C}').$$

Now, we use Kirschbraun's Theorem, and extend  $\mathbf{\Pi}^{-1}: \mathbb{R}^t \rightarrow \mathbb{R}^d$  in a  $(1 + \varepsilon)$ -Lipschitz manner. Hence,

$$\text{cost}_z((\mathbf{\Pi}(\mathbf{S}), \mathbf{w}), \mathbf{C}') \geq \frac{1}{1 + \varepsilon} \cdot \text{cost}_z((\mathbf{S}, \mathbf{w}), \mathbf{\Pi}^{-1}(\mathbf{C}')) \geq \frac{1}{1 + \varepsilon} \cdot \min_{\substack{\mathbf{C}'' \subset \mathbb{R}^d \\ |\mathbf{C}''|=k}} \text{cost}_z((\mathbf{S}, \mathbf{w}), \mathbf{C}'').$$

Finally, using the fact that  $(\mathbf{S}, \mathbf{w})$  is an  $\varepsilon$ -weak coreset, we may conclude that

$$\text{cost}_z^z((\mathbf{S}, \mathbf{w}), \mathbf{C}'') \geq \frac{1}{1 + \varepsilon} \cdot \text{cost}_z^z(X, \mathbf{C}'').$$

□

We now turn to showing that an appropriate setting of parameters implies that the events occur often. For the first event, Theorem 4.2 from [31] implies event  $\mathbf{E}_1$  occurs with probability  $1 - o(1)$ . We state the usual guarantees of the Johnson-Lindenstrauss transform, which is what we need for event  $\mathbf{E}_2$  to hold.

LEMMA 4.4. Let  $S \subset \mathbb{R}^d$  be any set of  $m$  points, and  $\mathcal{J}_{d,t}$  denote the Johnson-Lindenstrauss map, with

$$t \gtrsim \frac{\log m}{\varepsilon^2}.$$

Then, with probability 0.99 over the draw of  $\mathbf{\Pi} \sim \mathcal{J}_{d,t}$ ,  $\mathbf{\Pi}: S \rightarrow \mathbf{\Pi}(S)$  is  $(1 \pm \varepsilon)$ -bi-Lipschitz, and hence,  $\mathbf{E}_2$  occurs with probability 0.99.

### 4.3.1 A Bad Approximation Guarantee

LEMMA 4.5. (WARM-UP LEMMA) Fix any  $\Pi \in \mathcal{J}_{d,t}$  and let  $C' \subset \mathbb{R}^t$  denote the optimal centers for  $(k, z)$ -clustering of  $\Pi(X)$ , then with probability at least 0.99 over the draw of  $(\mathbf{S}, \mathbf{w})$  as per Definition 4.3,

$$\sum_{x \in \mathbf{S}} \mathbf{w}(x) \cdot \min_{c \in C'} \|\Pi(x) - c\|_2^z \leq 100 \cdot \text{cost}_z^z(\Pi(X), C'),$$

with probability at least 0.99. In other words,  $\mathbf{E}_3(100)$  holds with probability at least 0.99.

*Proof.* For any  $x \in X$ , let  $c'(x) \in C'$  denote the point in  $C'$  closest to  $\Pi(x)$ . Then, we note

$$(4.14) \quad \sum_{x \in \mathbf{S}} \mathbf{w}(x) \cdot \|\Pi(x) - c'(x)\|_2^z = \mathbf{E}_{\mathbf{x} \sim \mathbf{S}} \left[ \frac{1}{\tilde{\sigma}(\mathbf{x})} \cdot \|\Pi(\mathbf{x}) - c'(\mathbf{x})\|_2^z \right],$$

so that in expectation over  $\mathbf{S}$ , we have

$$\mathbf{E}_{\mathbf{S}} \left[ \mathbf{E}_{\mathbf{x} \sim \mathbf{S}} \left[ \frac{1}{\tilde{\sigma}(\mathbf{x})} \cdot \|\Pi(\mathbf{x}) - c'(\mathbf{x})\|_2^z \right] \right] = \mathbf{E}_{\mathbf{x} \sim \tilde{\sigma}} \left[ \frac{1}{\tilde{\sigma}(\mathbf{x})} \cdot \|\Pi(\mathbf{x}) - c'(\mathbf{x})\|_2^z \right] = \text{cost}_z^z(\Pi(X), C').$$

By Markov's inequality, we obtain our desired bound.  $\square$

COROLLARY 4.1. Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be any set of points, and  $C \subset \mathbb{R}^d$  of size  $k$  be centers for optimally  $(k, z)$ -clustering  $X$ . For any  $\varepsilon \in (0, 1/2)$ , let  $\mathcal{J}_{d,t}$  be the Johnson-Lindenstrauss map with

$$t \gtrsim \frac{z \log(1/\varepsilon) + \log k}{\varepsilon^2},$$

Then, with probability at least 0.97 over the draw of  $\Pi \sim \mathcal{J}_{d,t}$ ,

$$(4.15) \quad \frac{1}{100^{1/z}(1+\varepsilon)^{1+1/z}} \cdot \text{cost}_z(X, C) \leq \min_{\substack{C' \subset \mathbb{R}^t \\ |C'|=k}} \text{cost}_z(\Pi(X), C').$$

*Proof.* We sample  $\Pi \sim \mathcal{J}_{d,t}$  and  $(\mathbf{S}, \mathbf{w})$  as per Definition 4.3. By Lemma 4.4, Lemma 4.5, Theorem 4.2, and a union bound, we have events  $\mathbf{E}_1$ ,  $\mathbf{E}_2$ , and  $\mathbf{E}_3(100)$  hold with probability at least 0.97. Hence, we obtain the desired result from applying Lemma 4.3.  $\square$

**4.3.2 Improving the Approximation** In what follows, we will improve upon the approximation of Corollary 4.1 significantly, to show that with large probability,  $\mathbf{E}_3(1+\varepsilon)$  holds. We let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  and denote  $C \subset \mathbb{R}^d$  of size  $k$  to be the optimal  $(k, z)$ -clustering of  $X$ . As before, we let  $c: X \rightarrow C$  map each  $x \in X$  to the point center in  $C$ , and  $\sigma: X \rightarrow \mathbb{R}_{\geq 0}$  be the sensitivities of  $X$  with respect to  $C$  as in Lemma 4.2, and  $\tilde{\sigma}$  be the sensitivity sampling distribution.

We define one more event,  $\mathbf{E}_4$ , with respect to the randomness of  $\Pi \sim \mathcal{J}_{d,t}$ . First, we let  $\mathbf{D}_x \in \mathbb{R}_{\geq 0}$  denote the random variable given by

$$(4.16) \quad \mathbf{D}_x \stackrel{\text{def}}{=} \frac{\|\Pi(x) - \Pi(c(x))\|_2}{\|x - c(x)\|_2}.$$

Notice that when  $\Pi$  consists of i.i.d  $\mathcal{N}(0, 1/t)$ , then  $t\mathbf{D}_x^2$  is distributed as  $\chi^2$ -random variable with  $t$  degrees of freedom. We say event  $\mathbf{E}_4$  holds whenever

$$(4.17) \quad \sum_{x \in X} \mathbf{D}_x^{2z} \cdot \sigma(x) \leq 100(k+1)2^z.$$

PROPOSITION 4.1. With probability at least 0.99 over the draw of  $\Pi \sim \mathcal{J}_{d,t}$ , event  $\mathbf{E}_4$  holds.



*Proof.* The proof will simply follow from computing the expectation of the left-hand side of (4.17), and applying Markov's inequality. In particular, for every  $x \in X$ , by the two-stability of the Gaussian,

$$\mathbf{E}_{\Pi \sim \mathcal{J}_{d,t}} [\mathbf{D}_x^{2z}] \leq 2^z.$$

The remaining part follows from the bound on  $\mathfrak{S}_\sigma$ .  $\square$

LEMMA 4.6. Let  $\Pi \in \mathcal{J}_{d,t}$  be a Johnson-Lindenstrauss map where, for  $\alpha > 1$ , the following events holds:

1. Guarantee from Lemma 4.1:  $\sum_{x \in X} \|\Pi(x) - \Pi(c(x))\|_2^z \leq \alpha \cdot \text{cost}_z^z(X, C)$ .
2. Guarantee from Corollary 4.1: letting  $C' \subset \mathbb{R}^t$  be the optimal  $(k, z)$ -clustering of  $\Pi(X)$ , then  $\text{cost}_z^z(\Pi(X), C') \geq (1/\alpha) \cdot \text{cost}_z^z(X, C)$ .
3. Event  $\mathbf{E}_4$  holds.

Then, if we let  $(\mathbf{S}, \mathbf{w})$  denote  $m = \text{poly}(k, 2^z, 1/\varepsilon, \alpha)$  i.i.d draws from  $\tilde{\sigma}$  and  $\mathbf{w}(x) = 1/(m\tilde{\sigma}(x))$ , with probability at least 0.99,

$$\text{cost}_z^z((\Pi(\mathbf{S}), \mathbf{w}), C') \leq (1 + \varepsilon) \cdot \text{cost}_z^z(\Pi(X), C').$$

*Proof.* The proof follows the same schema as Lemma 4.5. However, we give a bound on the variance of the estimator in order to improve upon the use of Markov's inequality. Specifically, we compute the variance of a rescaling of (4.14).

$$\begin{aligned} \text{Var}_{\mathbf{S}} \left[ \mathbf{E}_{\mathbf{x} \sim \mathbf{S}} \left[ \frac{1}{\tilde{\sigma}(\mathbf{x})} \cdot \frac{\|\Pi(\mathbf{x}) - c'(\mathbf{x})\|_2^z}{\text{cost}_z^z(\Pi(X), C')} \right] \right] &\leq \frac{1}{m} \mathbf{E}_{\mathbf{x} \sim \tilde{\sigma}} \left[ \left( \frac{1}{\tilde{\sigma}(\mathbf{x})} \cdot \frac{\|\Pi(\mathbf{x}) - c'(\mathbf{x})\|_2^z}{\text{cost}_z^z(\Pi(X), C')} \right)^2 \right] \\ (4.18) \quad &= \frac{\mathfrak{S}_\sigma}{m} \sum_{x \in X} \left( \frac{1}{\sigma(x)} \cdot \frac{\|\Pi(x) - c'(x)\|_2^{2z}}{\text{cost}_z^{2z}(\Pi(X), C')} \right). \end{aligned}$$

By the same argument as in the proof of Lemma 4.2 (applying the triangle inequality twice),

$$(4.19) \quad \frac{\|\Pi(x) - c'(x)\|_2^z}{\text{cost}_z^z(\Pi(X), C')} \leq 2^{z-1} \cdot \frac{\|\Pi(x) - \Pi(c(x))\|_2^z}{\text{cost}_z^z(\Pi(X), C')} + \frac{2^{2(z-1)}}{|X_{c(x)}|} \left( \frac{\text{cost}_z^z(\Pi(X), \Pi(C))}{\text{cost}_z^z(\Pi(X), C')} + 1 \right).$$

Recall that we have the lower bound (2) and the upper bound (1). Hence, along with the definition of  $D_x$  in (4.16) (we remove the bold-face as  $\Pi$  is fixed), we upper bound the left-hand side of (4.19) by

$$\alpha \cdot 2^{z-1} \cdot \frac{D_x^z \cdot \|x - c(x)\|_2^z}{\text{cost}_z^z(X, C)} + \frac{2^{2(z-1)}}{|X_{c(x)}|} (\alpha^2 + 1) \leq \alpha^2 \cdot 2^{100z} (D_x^z + 1) \cdot \sigma(x).$$

In particular, we may plug this in to (7.32) and use the definition of  $\sigma(x)$ . Specifically, one obtains the variance in (7.32) is upper bounded by

$$\begin{aligned} \frac{\mathfrak{S}_\sigma \cdot \alpha^4 \cdot 2^{200z}}{m} \sum_{x \in X} (D_x^z + 1)^2 \cdot \sigma(x) &\leq \frac{4\mathfrak{S}_\sigma^2 \cdot \alpha^4 \cdot 2^{200z}}{m} + \frac{4\mathfrak{S}_\sigma \cdot \alpha^4 \cdot 2^{200z}}{m} \sum_{x \in X} D_x^{2z} \cdot \sigma(x) \\ &\leq \frac{1000\mathfrak{S}_\sigma^2 \cdot \alpha^4 \cdot 2^{300z}}{m}, \end{aligned}$$

where in the final inequality, we used the fact that  $\mathbf{E}_4$  holds. Hence, letting  $m$  be a large enough polynomial in  $\text{poly}(k, 2^z, 1/\varepsilon, \alpha)$  implies the variance is smaller than  $o(\varepsilon^2)$ , so we can apply Chebyshev's inequality.  $\square$

COROLLARY 4.2. Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  by any set of points, and  $C \subset \mathbb{R}^d$  of size  $k$  be centers for optimally  $(k, z)$ -clustering  $X$ . For any  $\varepsilon \in (0, 1/2)$ , let  $\mathcal{J}_{d,t}$  be the Johnson-Lindenstrauss map with

$$t \gtrsim \frac{z \log(1/\varepsilon) + \log k}{\varepsilon^2}.$$

Then, with probability at least 0.92 over the draw of  $\Pi \sim \mathcal{J}_{d,t}$ ,

$$\frac{1}{(1+\varepsilon)^{1+2/z}} \cdot \text{cost}_z(X, C) \leq \min_{\substack{C' \subset \mathbb{R}^t \\ |C'| \leq k}} \text{cost}_z(\Pi(X), C').$$

*Proof.* We sample  $\Pi \sim \mathcal{J}_{d,t}$  and  $(\mathbf{S}, \mathbf{w})$  as per Definition 4.3. Note that by Theorem 4.2 and the setting of  $m$ , event  $\mathbf{E}_1$  holds with probability at least 0.99 over the draw of  $(\mathbf{S}, \mathbf{w})$ . By Lemma 4.1, Corollary 4.1, and Claim 4.1, and the setting of  $m$  and  $t$ , the conditions (1), (2) and (3) of Lemma 4.6 hold with probability at least 0.95, with  $\alpha$  being set to a large enough constant, and hence event  $\mathbf{E}_3(1+\varepsilon)$  holds with probability at least 0.94. Finally, event  $\mathbf{E}_2$  holds with probability 0.99 by Lemma 4.4 and taking a union bound and Lemma 4.3 gives the desired bound.  $\square$

## 5 Subspace Approximation

In the  $(k, z)$ -subspace approximation problem, we consider a subspace  $R \subset \mathbb{R}^d$  of dimension less than  $k$ , which we may encode by a collection of at most  $k$  orthonormal vectors  $r_1, \dots, r_k \in R$ . We let  $\rho_R: \mathbb{R}^d \rightarrow \mathbb{R}^d$  denote the map which sends each vector  $x \in \mathbb{R}^d$  to its closest point in  $R$ , and note that

$$\rho_R(x) = \operatorname{argmin}_{z \in R} \|x - z\|_2^2 = \sum_{i=1}^k \langle x, r_i \rangle \cdot r_i \in R,$$

For any subset  $X \subset \mathbb{R}^d$  and any  $k$ -dimensional subspace  $R$ , we let

$$\text{cost}_z^z(X, R) = \sum_{x \in X} \|x - \rho_R(x)\|_2^z.$$

In this section, we will show that we may compute the optimum  $k$ -subspace approximation after applying a Johnson-Lindenstrauss transform.

**THEOREM 5.1. (JOHNSON-LINDENSTRAUSS FOR  $k$ -SUBSPACE APPROXIMATION)** *Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be any set of points, and let  $R$  denote the optimal  $(k, z)$ -subspace approximation of  $X$ . For any  $\varepsilon \in (0, 1)$ , suppose we let  $\mathcal{J}_{d,t}$  be the distribution over Johnson-Lindenstrauss maps where*

$$t \gtrsim \frac{z \cdot k^2 \cdot \text{polylog}(k/\varepsilon)}{\varepsilon^3}.$$

*Then, with probability at least 0.9 over the draw of  $\Pi \sim \mathcal{J}_{d,t}$ ,*

$$\frac{1}{1+\varepsilon} \cdot \text{cost}_z(X, R) \leq \min_{\substack{R' \subset \mathbb{R}^t \\ \dim R' \leq k}} \text{cost}_z(\Pi(X), R') \leq (1+\varepsilon) \cdot \text{cost}_z(X, R).$$

The proof of the above theorem proceeds in two steps, and models the argument in the previous section. First, we show that the cost of the optimum does not increase substantially (the right-most inequality in the theorem). This is done in the next subsection. The second step is showing that the optimum does not decrease substantially (the left-most inequality in the theorem). The second step is done in the subsequent subsection.

**5.1 Easy Direction: Optimum Cost Does Not Increase** The first direction, which shows that the optimal  $(k, z)$ -subspace approximation does not increase follows similarly to Lemma 4.1

**LEMMA 5.1.** *Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be any set of points and let  $R \subset \mathbb{R}^d$  be optimal  $(k, z)$ -subspace approximation of  $X$ . For any  $\varepsilon \in (0, 1)$ , we let  $\mathcal{J}_{d,t}$  be the distribution over Johnson-Lindenstrauss maps. If  $t \gtrsim z/\varepsilon^2$ , then with probability at least 0.99 over the draw of  $\Pi \sim \mathcal{J}_{d,t}$ ,*

$$\sum_{x \in X} \|\Pi(x) - \Pi(\rho_R(x))\|_2^z \leq (1+\varepsilon) \cdot \text{cost}_z(X, R),$$

and hence,

$$\min_{\substack{R' \subset \mathbb{R}^t \\ \dim R' \leq k}} \text{cost}_z(\Pi(X), R') \leq (1+\varepsilon) \min_{\substack{R \subset \mathbb{R}^d \\ \dim R \leq k}} \text{cost}_z(X, R).$$

*Proof.* Note that  $\Pi$  is a linear map, so if we let  $r_1, \dots, r_k \in R$  denote  $k$  orthonormal unit vectors spanning  $R$ , then  $\Pi(r_1), \dots, \Pi(r_k) \in \mathbb{R}^t$  are  $k$  vectors spanning the subspace  $\Pi(R)$ . Furthermore, we may consider the  $k$ -dimensional subspace

$$\Pi(R) \stackrel{\text{def}}{=} \left\{ \sum_{i=1}^k \alpha_i \cdot \Pi(r_i) \in \mathbb{R}^t : \alpha_1, \dots, \alpha_k \in \mathbb{R} \right\}.$$

Notice that for any  $x \in \mathbb{R}^d$ , by linearity of  $\Pi$ ,

$$\Pi(\rho_R(x)) = \sum_{i=1}^k \langle x - \tau, r_i \rangle \cdot \Pi(r_i) \in \Pi(F),$$

which means that we may always upper bound

$$(5.20) \quad \min_{\substack{R' \subset \mathbb{R}^t \\ \dim R' \leq k}} \text{cost}_z(\Pi(X), R') \leq \left( \sum_{x \in X} \|\Pi(x) - \Pi(\rho_R(x))\|_2^z \right)^{1/z}.$$

It hence remains to upper bound the left-hand side of (5.20). We now use the fact that  $\Pi$  is drawn from a Johnson-Lindenstrauss distribution. Specifically, the lemma follows from applying Markov's inequality once we show

$$\frac{1}{\text{cost}_z^z(X, F)} \mathbf{E}_{\Pi \sim \mathcal{J}_{d,t}} \left[ \sum_{x \in X} (\|\Pi(x) - \Pi(\rho_F(x))\|_2^z - \|x - \rho_F(x)\|_2^z)^+ \right] \leq \frac{(1 + \varepsilon)^z - 1}{100},$$

which follows from Lemma 3.1  $\square$

## 5.2 Hard Direction: Optimum Cost Does Not Decrease

**5.2.1 Preliminaries** In the  $(k, z)$ -subspace approximation problem, there will be a difference between complexities of known strong coresets and weak coresets. Our argument will only use weak coresets, which is important for us, as strong coresets have a dependence on  $d$  (which we are trying to avoid).

**DEFINITION 5.1. (WEAK CORESETS FOR  $(k, z)$ -SUBSPACE APPROXIMATION)** Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be a set of points. A weak  $\varepsilon$ -coreset of  $X$  for  $(k, z)$ -subspace approximation is a weighted subset  $S \subset \mathbb{R}^d$  of points with weights  $w: S \rightarrow \mathbb{R}_{\geq 0}$  such that,

$$\frac{1}{1 + \varepsilon} \cdot \min_{\substack{R \subset \mathbb{R}^d \\ \dim R \leq k}} \text{cost}_z(X, R) \leq \min_{\substack{R \subset \mathbb{R}^d \\ \dim R \leq k}} \text{cost}_z((S, w), R) \leq (1 + \varepsilon) \cdot \min_{\substack{R \subset \mathbb{R}^d \\ \dim R \leq k}} \text{cost}_z(X, R).$$

Similarly to the case of  $(k, z)$ -clustering, algorithms for building weak coresets proceed by sampling according to the sensitivity framework. We proceed by defining sensitivity functions in the context of subspace approximation, and then state a lemma which gives a sensitivity function that we will use.

**DEFINITION 5.2. (SENSITIVITIES)** Let  $n, d \in \mathbb{N}$ , and consider any set of points  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ , as well as  $k \in \mathbb{N}$  and  $z \geq 1$ . A sensitivity function  $\sigma: X \rightarrow \mathbb{R}_{\geq 0}$  for  $(k, z)$ -subspace approximation in  $\mathbb{R}^d$  is a function satisfying that, for all  $x \in X$ ,

$$\sup_{\substack{R \subset \mathbb{R}^d \\ \dim R \leq k}} \frac{\|x - \rho_R(x)\|_2^z}{\text{cost}_z^z(X, R)} \leq \sigma(x).$$

The total sensitivity of the sensitivity function  $\sigma$  is given by

$$\mathfrak{S}_\sigma = \sum_{x \in X} \sigma(x).$$

For a sensitivity function, we let  $\tilde{\sigma}$  denote the sensitivity sampling distribution, supported on  $X$ , which samples  $x \in X$  with probability proportional to  $\sigma(x)$ .

We now state a specific sensitivity function that we will use. The proof will closely follow a method for bounding the total sensitivity of [49]. The resulting weak  $\varepsilon$ -coreset will have a worse dependence than the best-known coresets for this problem; however, the specific form of the sensitivity function will be especially useful for us. Specifically, the non-optimality of the sensitivity function will not significantly affect the final bound on dimension reduction.

LEMMA 5.2. (THEOREM 18 OF [49]) *Let  $n, d \in \mathbb{N}$ , and consider any set of points  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ , as well as  $k \in \mathbb{N}$  with  $k < d$ , and  $z \geq 1$ . Suppose  $R \subset \mathbb{R}^d$  is the optimal  $(k, z)$ -subspace approximation of  $X$  in  $\mathbb{R}^d$ . Then, the function  $\sigma: X \rightarrow \mathbb{R}_{\geq 0}$  given by*

$$\sigma(x) = 2^{z-1} \cdot \frac{\|x - \rho_R(x)\|_2^z}{\text{cost}_z^z(X, R)} + 2^{2z-1} \cdot \sup_{u \in \mathbb{R}^d} \frac{|\langle \rho_R(x), u \rangle|^z}{\sum_{x' \in X} |\langle \rho_R(x'), u \rangle|^z}$$

*is a sensitivity function for  $(k, z)$ -subspace approximation of  $X$  in  $\mathbb{R}^d$ , satisfying*

$$\mathfrak{S}_\sigma \leq 2^{z-1} + 2^{2z-1}(k+1)^{1+z}.$$

*Proof.* Consider any subspace  $R' \subset \mathbb{R}^d$  of dimension at most  $k$ . Then, for any  $x \in X$

$$\begin{aligned} \|x - \rho_{R'}(x)\|_2^z &\leq \|x - \rho_{R'}(\rho_R(x))\|_2^z \leq 2^{z-1} (\|x - \rho_R(x)\|_2^z + \|\rho_R(x) - \rho_{R'}(\rho_R(x))\|_2^z) \\ (5.21) \quad &\leq 2^{z-1} \left( \frac{\|x - \rho_R(x)\|_2^z}{\text{cost}_z^z(X, R)} \cdot \text{cost}_z^z(X, R) + \frac{\|\rho_R(x) - \rho_{R'}(\rho_R(x))\|_2^z}{\text{cost}_z^z(\rho_R(X), R')} \cdot \text{cost}_z^z(\rho_R(X), R') \right). \end{aligned}$$

Notice that  $\text{cost}_z^z(\rho_R(X), R') \leq 2^{z-1}(\text{cost}_z^z(X, R) + \text{cost}_z^z(X, R'))$  by the triangle inequality and Hölder's inequality, and that  $\text{cost}_z^z(X, R) \leq \text{cost}_z^z(X, R')$  since  $R$  is the optimal  $(k, z)$ -subspace approximation. Hence, dividing the left- and right-hand side of (5.21) we have

$$\frac{\|x - \rho_{R'}(x)\|_2^z}{\text{cost}_z^z(X, R')} \leq 2^{z-1} \cdot \frac{\|x - \rho_R(x)\|_2^z}{\text{cost}_z^z(X, R)} + 2^{2z-1} \cdot \frac{\|\rho_R(x) - \rho_{R'}(\rho_R(x))\|_2^z}{\text{cost}_z^z(\rho_R(X), R')}.$$

It remains to show that, for any set of points  $Y \subset \mathbb{R}^d$  (in particular, the set  $\{\rho_R(x) : x \in X\}$ ), and any  $y \in Y$ ,

$$\sup_{\substack{R' \subset \mathbb{R}^d \\ \dim R' \leq k}} \frac{\|y - \rho_{R'}(y)\|_2^z}{\text{cost}_z^z(Y, R')} \leq \sup_{\substack{H \subset \mathbb{R}^d \\ \dim H = d-1}} \frac{\|y - \rho_H(y)\|_2^z}{\text{cost}_z^z(Y, H)} = \sup_{u \in \mathbb{R}^d} \frac{|\langle y, u \rangle|^z}{\sum_{y' \in Y} |\langle y', u \rangle|^z}.$$

In particular, note that for any subspace  $R' \subset \mathbb{R}^d$  of dimension at most  $k$ , there exists a  $(d-1)$ -dimensional subspace  $H \subset \mathbb{R}^d$  containing  $R'$  given by all vectors orthogonal to  $y - \rho_{R'}(y)$ . In particular,  $\text{cost}_z^z(Y, H) \leq \text{cost}_z^z(Y, R')$  since  $R'$  is contained in  $H$ , and  $\|y - \rho_H(y)\|_2 = \|y - \rho_{R'}(y)\|_2$  by the definition of  $H$ . The bound on the total sensitivity then follows from Lemma 16 in [49], where we use the fact that  $\{\rho_R(x) : x \in X\}$  lies in a  $k$ -dimensional subspace.  $\square$

We will use the following geometric theorem of [44] in our proof. The theorem says that an approximately optimal  $(k, z)$ -subspace approximation lies in the span of a small set of points. We state the lemma for weighted point sets, even though [44] state it for unweighted points. We note that adding weights can be simulated by replicating points.

LEMMA 5.3. (THEOREM 3.1 [44]) *Let  $d, k \in \mathbb{N}$ , and consider a weighted set of points  $S \subset \mathbb{R}^d$  with weights  $w: S \rightarrow \mathbb{R}_{\geq 0}$ , as well as  $\varepsilon \in (0, 1)$  and  $z \geq 1$ . There exists a subset  $Q \subset S$  of  $O(k^2 \log(k/\varepsilon)/\varepsilon)$  and a  $k$ -dimensional subspace  $R' \subset \mathbb{R}^d$  within the span of  $Q$  satisfying*

$$\text{cost}_z((S, w), R') \leq (1 + \varepsilon) \min_{\substack{R \subset \mathbb{R}^d \\ \dim R \leq k}} \text{cost}_z((S, w), R).$$

Lemma 5.2 gives us an appropriate sensitivity function, and Lemma 5.3 limits the search of the subspace to just a small set of points. Similarly to the case of  $(k, z)$ -clustering, we can use this to construct weak  $\varepsilon$ -coresets for  $(k, z)$ -subspace approximation. The following theorem is Theorem 5.10 from [31]. We state the theorem with the sensitivity function of Lemma 5.2

**THEOREM 5.2.** (THEOREM 5.10 FROM [31]) For any subset  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  and  $\varepsilon \in (0, 1/2)$ , let  $\tilde{\sigma}$  denote the sensitivity sampling distribution from the sensitivity function of Lemma [5.2].

- Let  $(\mathbf{S}, \mathbf{w})$  denote the random (multi-)set  $\mathbf{S} \subset X$  and  $\mathbf{w}: \mathbf{S} \rightarrow \mathbb{R}_{\geq 0}$  given by, for

$$m = \text{poly}((k+1)^z, 1/\varepsilon)$$

iterations, sampling  $\mathbf{x} \sim \tilde{\sigma}$  i.i.d and letting  $\mathbf{w}(\mathbf{x}) = 1/(m\tilde{\sigma}(\mathbf{x}))$ .

- Then, with probability  $1 - o(1)$  over the draw of  $(\mathbf{S}, \mathbf{w})$ , it is an  $\varepsilon$ -weak coresets for  $(k, z)$ -subspace approximation of  $X$ .

**5.3 The Important Events** Similarly to the previous section, we define the important events, over the randomness in  $\mathbf{\Pi}$  such that, if these are satisfied, then the optimum of  $(k, z)$ -subspace approximation after dimension reduction does not decrease substantially. We first define the events, and then we prove that if the events are all satisfied, then we obtain our desired approximation.

**DEFINITION 5.3.** (THE EVENTS) Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ , and  $\tilde{\sigma}$  the sensitivity sampling distribution of  $X$  from Lemma [5.2]. We consider the following experiment,

1. We generate a sample  $(\mathbf{S}, \mathbf{w})$  by sampling from  $\tilde{\sigma}$  for  $m = \text{poly}(k^z, 1/\varepsilon)$  i.i.d iterations  $\mathbf{x} \sim \tilde{\sigma}$  and set  $\mathbf{w}(\mathbf{x}) = 1/(m\tilde{\sigma}(\mathbf{x}))$ .
2. Furthermore, we sample  $\mathbf{\Pi} \sim \mathcal{J}_{d,t}$ , which is a Johnson-Lindenstrauss map  $\mathbb{R}^d \rightarrow \mathbb{R}^t$ .
3. We let  $\mathbf{S}' = \mathbf{\Pi}(\mathbf{S}) \subset \mathbb{R}^t$  denote the image of  $\mathbf{\Pi}$  on  $\mathbf{S}$ .

The events are the following:

- $\mathbf{E}_1$  : The weighted (multi-)set  $(\mathbf{S}, \mathbf{w})$  is a weak  $\varepsilon$ -coresets for  $(k, z)$ -subspace approximation of  $X$  in  $\mathbb{R}^d$ .
- $\mathbf{E}_2$  : The map  $\mathbf{\Pi}: \mathbb{R}^d \rightarrow \mathbb{R}^t$  satisfies the following condition. For any choice of  $O(k^2 \log(k/\varepsilon)/\varepsilon)$  points of  $\mathbf{S}$ ,  $\mathbf{\Pi}$  is an  $\varepsilon$ -subspace embedding of the subspace spanned by these points.
- $\mathbf{E}_3(\beta)$  : Let  $\mathbf{R}' \subset \mathbb{R}^t$  denote the  $k$ -dimensional subspace for optimal  $(k, z)$ -subspace approximation of  $\mathbf{\Pi}(X)$  in  $\mathbb{R}^t$ . Then,

$$\text{cost}_z^z((\mathbf{\Pi}(\mathbf{S}), \mathbf{w}), \mathbf{R}') \leq \beta \cdot \text{cost}_z^z(\mathbf{\Pi}(X), \mathbf{R}').$$

**LEMMA 5.4.** Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ , and suppose  $(\mathbf{S}, \mathbf{w})$  and  $\mathbf{\Pi}: \mathbb{R}^d \rightarrow \mathbb{R}^t$  satisfy events  $\mathbf{E}_1$ ,  $\mathbf{E}_2$ , and  $\mathbf{E}_3(\beta)$ . Then,

$$\min_{\substack{\mathbf{R}' \subset \mathbb{R}^t \\ \dim \mathbf{R}' \leq k}} \text{cost}_z(\mathbf{\Pi}(X), \mathbf{R}') \geq \frac{1}{\beta^{1/z}(1+\varepsilon)^3} \cdot \min_{\substack{\mathbf{R} \subset \mathbb{R}^d \\ \dim \mathbf{R} \leq k}} \text{cost}_z(X, \mathbf{R}).$$

*Proof.* Consider a fixed  $\mathbf{\Pi}$  and  $(\mathbf{S}, \mathbf{w})$  satisfying the three events of Definition [5.3]. Let  $\mathbf{R}' \subset \mathbb{R}^t$  be the  $k$ -dimensional subspace which minimizes  $\text{cost}_z^z(\mathbf{\Pi}(X), \mathbf{R}')$ . Then, by event  $\mathbf{E}_3(\beta)$ , we have  $\text{cost}_z^z((\mathbf{\Pi}(\mathbf{S}), \mathbf{w}), \mathbf{R}') \leq \beta \cdot \text{cost}_z^z(\mathbf{\Pi}(X), \mathbf{R}')$ . Now, we apply Lemma [5.3] to  $(\mathbf{\Pi}(\mathbf{S}), \mathbf{w})$ , and we obtain a subset  $Q \subset \mathbf{S}$  of size  $O(k^2 \log(k/\varepsilon)/\varepsilon)$  for which there exists a  $k$ -dimensional subspace  $\mathbf{R}'' \subset \mathbb{R}^t$  within the span of  $\mathbf{\Pi}(Q)$  which satisfies

$$\left( \sum_{x \in \mathbf{S}} \mathbf{w}(x) \cdot \|\mathbf{\Pi}(x) - \rho_{\mathbf{R}''}(\mathbf{\Pi}(x))\|_2^z \right)^{1/z} = \text{cost}_z((\mathbf{\Pi}(\mathbf{S}), \mathbf{w}), \mathbf{R}'') \leq (1+\varepsilon) \cdot \text{cost}_z((\mathbf{\Pi}(\mathbf{S}), \mathbf{w}), \mathbf{R}').$$

Note that  $\mathbf{R}''$  is a  $k$ -dimensional subspace lying in the span of  $\mathbf{\Pi}(Q)$ . For any  $x \in \mathbf{S}$ , we will use the fact that  $\mathbf{E}_2$  is satisfied to say that  $\mathbf{\Pi}$  is an  $\varepsilon$ -subspace embedding of the subspace spanned by  $Q \cup \{x\}$ . This will enable us to find a subspace  $U \subset \mathbb{R}^d$  of dimension  $k$  whose cost of approximating  $(\mathbf{S}, \mathbf{w})$  is at most  $(1+\varepsilon) \cdot \text{cost}_z((\mathbf{\Pi}(\mathbf{S}), \mathbf{w}), \mathbf{R}'')$ .

Specifically, we write  $v_1, \dots, v_k \in \mathbb{R}^t$ , as orthogonal unit vectors which span  $\mathbf{R}''$ . Because  $\mathbf{R}''$  lies in the span of  $\mathbf{\Pi}(Q)$ , there are vectors  $u_1, \dots, u_k \in \mathbb{R}^d$  in the span of  $Q$  which satisfy

$$v_\ell = \mathbf{\Pi}(u_\ell) \in \mathbb{R}^t \quad \text{for} \quad u_\ell = \sum_{y \in Q} c_{\ell,y} y \in \mathbb{R}^d.$$



Hence, the subspace  $U$  given by the span of all vectors in  $U$  is a  $k$ -dimensional subspace lying in the span of  $Q$ . For  $x \in \mathbf{S}$ , we may write the coefficients  $\gamma_\ell(x) = \langle \mathbf{\Pi}(x), v_\ell \rangle$ , and we may express projection  $\rho_{R''}(\mathbf{\Pi}(x)) \in \mathbb{R}^t$  as

$$\rho_{R''}(\mathbf{\Pi}(x)) = \sum_{\ell=1}^k \gamma_\ell(x) \cdot v_\ell = \mathbf{\Pi} \left( \sum_{y \in Q} \left( \sum_{\ell=1}^k \gamma_\ell(x) c_{\ell,y} \right) \cdot y \right) = \mathbf{\Pi} \left( \sum_{\ell=1}^k \gamma_\ell(x) u_\ell \right),$$

which is a linear combination of  $Q$ . By event  $\mathbf{E}_2$ ,  $\mathbf{\Pi}$  is an  $\varepsilon$ -subspace embedding of the subspace spanned by  $Q \cup \{x\}$ , so

$$\|x - \rho_U(x)\|_2 \leq \left\| x - \sum_{\ell=1}^k \gamma_\ell(x) u_\ell \right\|_2 \leq (1 + \varepsilon) \|\mathbf{\Pi}(x) - \rho_{R''}(\mathbf{\Pi}(x))\|_2.$$

Combining the inequalities, we have

$$\text{cost}_z((\mathbf{S}, \mathbf{w}), U) \leq (1 + \varepsilon) \cdot \text{cost}_z((\mathbf{\Pi}(\mathbf{S}), \mathbf{w}), R''),$$

and finally, since  $(\mathbf{S}, \mathbf{w})$  is a  $\varepsilon$ -weak coreset, we obtain the desired inequality.  $\square$

We note that event  $\mathbf{E}_1$  will be satisfied with sufficiently high probability from Theorem 5.2. Furthermore, event  $\mathbf{E}_2$  is satisfied with sufficiently high probability from the following simple lemma. All that will remain is showing that event  $\mathbf{E}_3(\beta)$  is satisfied.

LEMMA 5.5. *Let  $S \subset \mathbb{R}^d$  be any set of  $m$  points and  $\ell \in \mathbb{N}$ , and let  $\mathcal{J}_{d,t}$  denote the Johnson-Lindenstrauss map, with*

$$t \gtrsim \frac{\ell \log m}{\varepsilon^2}.$$

*Then, with probability 0.99 over the draw of  $\mathbf{\Pi} \sim \mathcal{J}_{d,t}$ ,  $\mathbf{\Pi}$  is an  $\varepsilon$ -subspace embedding for all subspaces spanned by  $\ell$  vectors of  $S$ .*

*Proof.* There are at most  $\binom{m}{\ell}$  subspaces spanned by  $\ell$  vectors of  $S$ . If  $\mathbf{\Pi}$  is a subspace embedding for all of them, we obtain our desired conclusion. We use Lemma 3.2 with  $\delta$  to be a sufficiently small constant factor of  $1/m^\ell$  and union bound.  $\square$

### 5.3.1 A Bad Approximation Guarantee

LEMMA 5.6. (WARM-UP LEMMA) *Fix any  $\mathbf{\Pi} \in \mathcal{J}_{d,t}$  and let  $R' \subset \mathbb{R}^t$  denote the  $k$ -dimensional subspace for optimal  $(k, z)$ -subspace approximation of  $\mathbf{\Pi}(X)$  in  $\mathbb{R}^t$ . Then, with probability at least 0.99 over the draw of  $(\mathbf{S}, \mathbf{w})$  as per Definition 5.3,*

$$\sum_{x \in \mathbf{S}} \mathbf{w}(x) \cdot \|\mathbf{\Pi}(x) - \rho_{R'}(\mathbf{\Pi}(x))\|_2^z \leq 100 \cdot \text{cost}_z^z(\mathbf{\Pi}(X), R').$$

*In other words, event  $\mathbf{E}_3(100)$  holds with probability at least 0.99.*

*Proof.* Similarly to the proof of Lemma 4.5, we compute the expectation of the left-hand side of the inequality and use Markov's inequality.

$$(5.22) \quad \mathbf{E}_{\mathbf{S}} \left[ \mathbf{E}_{\mathbf{x} \sim \mathbf{S}} \left[ \frac{1}{\tilde{\sigma}(\mathbf{x})} \cdot \|\mathbf{\Pi}(x) - \rho_{R'}(\mathbf{\Pi}(x))\|_2^z \right] \right] = \mathbf{E}_{\mathbf{x} \sim \tilde{\sigma}} \left[ \frac{1}{\tilde{\sigma}(\mathbf{x})} \cdot \|\mathbf{\Pi}(x) - \rho_{R'}(\mathbf{\Pi}(x))\|_2^z \right] = \text{cost}_z^z(\mathbf{\Pi}(X), R').$$

$\square$

COROLLARY 5.1. *Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be any set of points. For any  $\varepsilon \in (0, 1/2)$ , let  $\mathcal{J}_{d,t}$  be the Johnson-Lindenstrauss map with*

$$t \gtrsim \frac{z \cdot k^2 \cdot \text{polylog}(k/\varepsilon)}{\varepsilon^3}.$$

*Then, with probability at least 0.97 over the draw of  $\mathbf{\Pi} \sim \mathcal{J}_{d,t}$ ,*

$$\frac{1}{100^{1/z}(1 + \varepsilon)^3} \cdot \text{cost}_z(X, R) \leq \min_{\substack{R' \subset \mathbb{R}^t \\ \dim R' \leq k}} \text{cost}_z(\mathbf{\Pi}(X), R').$$

*Proof.* We sample  $\Pi \sim \mathcal{J}_{d,t}$  and  $(\mathbf{S}, \mathbf{w})$  as per Definition 5.3. By Theorem 5.2 and Lemma 5.6, and a union bound, events  $\mathbf{E}_1$  and  $\mathbf{E}_3(100)$  hold with probability at least 0.98. Event  $\mathbf{E}_2$  occurs with probability at least 0.99 by apply Lemma 5.5 with  $m = \text{poly}(k^z, 1/\varepsilon)$  and  $\ell = O(k^2 \log(k/\varepsilon)/\varepsilon)$ . Hence, we apply Lemma 5.4  $\square$

**5.3.2 Improving the Approximation** We now improve on the approximation of Corollary 5.1 in a fashion similar to that of Subsection 4.3.2. We will show that with large probability,  $\mathbf{E}_3(1 + \varepsilon)$  holds. We let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  and  $R \subset \mathbb{R}^d$  be the subspace of dimension  $k$  for optimal  $(k, z)$ -subspace approximation of  $X$  in  $\mathbb{R}^d$ . As before, we let  $\sigma: X \rightarrow \mathbb{R}_{\geq 0}$  be the sensitivities of  $X$  with respect to  $R$  (as in Lemma 5.2), and  $\tilde{\sigma}$  be the sensitivity sampling distribution.

We define one more events,  $\mathbf{E}_4$  with respect to the randomness in  $\Pi \sim \mathcal{J}_{d,t}$ . Let  $\mathbf{D}_x \in \mathbb{R}_{\geq 0}$  denote the random variable given by

$$(5.23) \quad \mathbf{D}_x \stackrel{\text{def}}{=} \frac{\|\Pi(x) - \Pi(\rho_R(x))\|_2}{\|x - \rho_R(x)\|_2}.$$

We say event  $\mathbf{E}_4$  holds whenever

$$(5.24) \quad \sum_{x \in X} \mathbf{D}_x^{2z} \cdot \sigma(x) \leq 100 \cdot 2^z \cdot \mathfrak{S}_\sigma,$$

which holds with probability at least 0.99 over the draw of  $\Pi \sim \mathcal{J}_{d,t}$ , similarly to the proof of Claim 4.1 and Lemma 5.2

LEMMA 5.7. Let  $\Pi \in \mathcal{J}_{d,t}$  be a Johnson-Lindenstrauss map where, for  $\alpha > 1$ , the follows events hold:

1. Guarantee from Lemma 5.1:  $\sum_{x \in X} \|\Pi(x) - \Pi(\rho_R(x))\|_2^z \leq \alpha \cdot \text{cost}_z^z(X, R)$ .
2. Guarantee from Corollary 5.1: letting  $R' \subset \mathbb{R}^t$  be the optimal  $(k, z)$ -subspace approximation of  $\Pi(X)$ , then  $\text{cost}_z^z(X, R) \leq \alpha \text{cost}_z^z(\Pi(X), R')$ .
3. Event  $\mathbf{E}_4$  holds.

Then, if we let  $(\mathbf{S}, \mathbf{w})$  denote  $m = \text{poly}(k^z, 1/\varepsilon, \alpha)$  i.i.d draws from  $\tilde{\sigma}$  and  $\mathbf{w}(x) = 1/(m\tilde{\sigma}(x))$ , with probability at least 0.99,

$$\text{cost}_z^z((\Pi(\mathbf{S}), \mathbf{w}), R') \leq (1 + \varepsilon) \cdot \text{cost}_z^z(\Pi(X), R').$$

*Proof.* Again, the proof is similar to that of Lemma 4.6, where we bound the variance of the estimator to apply Chebyshev's inequality. In particular, we have

$$(5.25) \quad \text{Var}_{\mathbf{S}} \left[ \mathbf{E}_{x \sim \mathbf{S}} \left[ \frac{1}{\tilde{\sigma}(x)} \cdot \frac{\|\Pi(x) - \rho_{R'}(\Pi(x))\|_2^z}{\text{cost}_z^z(\Pi(X), R')} \right] \right] \leq \frac{\mathfrak{S}_\sigma}{m} \sum_{x \in X} \left( \frac{1}{\sigma(x)} \cdot \frac{\|\Pi(x) - \rho_{R'}(\Pi(x))\|_2^{2z}}{\text{cost}_z^{2z}(\Pi(X), R')} \right).$$

Similarly to the proof of Lemma 4.6, we will upper bound

$$\frac{\|\Pi(x) - \rho_{R'}(\Pi(x))\|_2^z}{\text{cost}_z^z(\Pi(X), R)}$$

as a function of  $\sigma(x)$  and  $D_x$  (given by (5.23) without boldface as  $\Pi$  is fixed). Toward this bound, we simplify the notation by letting  $y_x = \rho_R(x) \in \mathbb{R}^d$  and  $Y = \{y_x : x \in X\}$ . Then, since  $\Pi: \mathbb{R}^d \rightarrow \mathbb{R}^t$  is a linear map, for any  $x \in X$

$$(5.26) \quad \sup_{v \in \mathbb{R}^t} \frac{|\langle \Pi(y_x), v \rangle|^z}{\sum_{x' \in X} |\langle \Pi(y_{x'}), v \rangle|^z} \leq \sup_{u \in \mathbb{R}^d} \frac{|\langle y_x, u \rangle|^z}{\sum_{x' \in X} |\langle y_{x'}, u \rangle|^z}.$$

In particular, if we let  $M \in \mathbb{R}^{n \times d}$  be the matrix given by having the rows be points  $y_x \in Y$ , then writing  $\Pi \in \mathbb{R}^{d \times t}$ , we have  $M\Pi \in \mathbb{R}^{n \times t}$  is the matrix whose rows are  $\Pi(y_x)$ ; in particular, one may compare the left- and right-hand sides of (5.26) by letting  $u = \Pi v \in \mathbb{R}^d$ . Thus, we have

$$(5.27) \quad \begin{aligned} \|\Pi(x) - \rho_{R'}(\Pi(x))\|_2^z &\leq \|\Pi(x) - \rho_{R'}(\Pi(y_x))\|_2^z \leq 2^{z-1} (\|\Pi(x) - \Pi(y_x)\|_2^z + \|\Pi(y_x) - \rho_{R'}(\Pi(y_x))\|_2^z) \\ &\leq 2^{z-1} \left( D_x^z \cdot \frac{\|x - y_x\|_2^z}{\text{cost}_z^z(X, R)} \cdot \text{cost}_z^z(X, R) + \frac{\|\Pi(y_x) - \rho_{R'}(\Pi(y_x))\|_2^z}{\text{cost}_z^z(\Pi(Y), R')} \cdot \text{cost}_z^z(\Pi(Y), R') \right). \end{aligned}$$

We may now apply the triangle inequality, as well as (1) and (2), and we have

$$\begin{aligned} \text{cost}_z^z(\Pi(Y), R') &\leq 2^{z-1} \left( \text{cost}_z^z(\Pi(X), R') + \sum_{x \in X} \|\Pi(x) - \Pi(y_x)\|_2^z \right) \\ (5.28) \quad &\leq 2^{z-1} (\text{cost}_z^z(\Pi(X), R') + \alpha \cdot \text{cost}_z^z(X, R)) \leq 2^{z-1} (1 + \alpha^2) \cdot \text{cost}_z^z(\Pi(X), R'). \end{aligned}$$

Finally, we note that, similarly to the proof of Lemma 5.2

$$(5.29) \quad \frac{\|\Pi(y_x) - \rho_{R'}(\Pi(y_x))\|_2^z}{\text{cost}_z^z(\Pi(Y), R')} \leq \sup_{v \in \mathbb{R}^t} \frac{|\langle \Pi(y_x), v \rangle|^z}{\sum_{x' \in X} |\langle \Pi(y_{x'}), v \rangle|^z}.$$

Continuing to upper-bound the left-hand side of (5.27) by plugging in (5.26), (5.29) and (5.28),

$$\begin{aligned} \frac{\|\Pi(x) - \rho_{R'}(\Pi(x))\|_2^z}{\text{cost}_z^z(\Pi(X), R')} &\leq 2^{z-1} \left( D_x^z \alpha \cdot \frac{\|x - y_x\|_2^z}{\text{cost}_z^z(X, Y)} + 2^{z-1} (\alpha^2 + 1) \sup_{u \in \mathbb{R}^d} \frac{|\langle y_x, u \rangle|^z}{\sum_{x' \in X} |\langle y_{x'}, u \rangle|^z} \right) \\ &\leq D_x^z \cdot (\alpha^2 + 1) \cdot \sigma(x). \end{aligned}$$

In particular, the bound on the variance in (5.25) is at most

$$\frac{\mathfrak{S}_\sigma}{m} \cdot 2^{4z} (\alpha^2 + 1)^2 \sum_{x \in X} D_x^{2z} \sigma(x) \leq \frac{\mathfrak{S}_\sigma^2}{m} \cdot 2^{5z} (\alpha^2 + 1),$$

so letting  $m = \text{poly}(k^z, 1/\varepsilon, \alpha)$  gives the desired bound on the variance.  $\square$

**COROLLARY 5.2.** *Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be any set of points, and let  $R \subset \mathbb{R}^d$  be the subspace for optimal  $(k, z)$ -subspace approximation of  $X$ . For any  $\varepsilon \in (0, 1/2)$ , let  $\mathcal{J}_{d,t}$  be the Johnson-Lindenstrauss map with*

$$t \gtrsim \frac{z \cdot k^2 \cdot \text{polylog}(k/\varepsilon)}{\varepsilon^3}.$$

*Then, with probability at least 0.92 over the draw of  $\Pi \sim \mathcal{J}_{d,t}$ ,*

$$\frac{1}{(1 + \varepsilon)^{3+1/z}} \cdot \text{cost}_z(X, R) \leq \min_{\substack{R' \subset \mathbb{R}^t \\ \dim R' \leq k}} \text{cost}_z(\Pi(X), R').$$

*Proof.* We sample  $\Pi \sim \mathcal{J}_{d,t}$  and  $(\mathbf{S}, \mathbf{w})$  as per Definition 5.3. Again, Theorem 5.2 guarantees that  $\mathbf{E}_1$  occurs with probability at least 0.99 over the draw of  $(\mathbf{S}, \mathbf{w})$ . By Lemma 5.1 Corollary 5.1 and (5.24), the condition (1), (2), and (3) are satisfied with probability at least 0.95, so we may apply Lemma 5.7 and have  $\mathbf{E}_3(1 + \varepsilon)$  holds with probability at least 0.94. Finally, event  $\mathbf{E}_2$  holds with probability at least 0.99 by Lemma 5.5. Taking a union bound and applying Lemma 5.4 gives the desired bound.  $\square$

## 6 $k$ -Flat Approximation

In the  $(k, z)$ -flat approximation problem, we consider subspace  $R \subset \mathbb{R}^d$  of dimension less than  $k$ , which we may encode by a collection of at most  $k$  orthonormal vectors  $r_1, \dots, r_k \in R$ , as well as a translation vector  $\tau \in \mathbb{R}^d$ . The  $k$ -flat specified by  $R$  and  $\tau$  is given by the affine subspace

$$F = \{x + \tau \in \mathbb{R}^d : x \in R\}.$$

We let  $\rho_F: \mathbb{R}^d \rightarrow \mathbb{R}$  denote the map which sends each  $x \in \mathbb{R}^d$  to its closest point on  $F$ , and we note that

$$\rho_F(x) = \underset{y \in F}{\operatorname{argmin}} \|x - y\|_2^2 = \tau + \sum_{i=1}^r \langle x - \tau, r_i \rangle r_i$$

For any  $X \subset \mathbb{R}^d$ , we let

$$\text{cost}_z^z(X, F) \stackrel{\text{def}}{=} \sum_{x \in X} \|x - \rho_F(x)\|_2^z.$$

In this section, we show that we may find the optimal  $k$ -flat approximation after applying a Johnson-Lindenstrauss map. The proof will be almost exactly the same as the  $(k, z)$ -subspace approximation problem. Indeed, it only remains to incorporate a translation vector.

**THEOREM 6.1. (JOHNSON-LINDENSTRAUSS FOR  $k$ -FLAT APPROXIMATION)** *Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be any set of points, and let  $F \subset \mathbb{R}^d$  denote the optimal  $(k, z)$ -flat approximation of  $X$ . For any  $\varepsilon \in (0, 1)$ , suppose we let  $\mathcal{J}_{d,t}$  be a distribution over a Johnson-Lindenstrauss maps where*

$$t \gtrsim \frac{z \cdot k^2 \cdot \text{polylog}(k/\varepsilon)}{\varepsilon^3}.$$

*Then, with probability at least 0.9 over the draw of  $\Pi \sim \mathcal{J}_{d,t}$ ,*

$$\frac{1}{1+\varepsilon} \cdot \text{cost}_z(X, F) \leq \min_{\substack{F' \text{ } k\text{-flat} \\ \text{in } \mathbb{R}^t}} \text{cost}_z(\Pi(X), F') \leq (1+\varepsilon) \cdot \text{cost}_z(X, F).$$

## 6.1 Easy Direction: Optimum Cost Does Not Increase

**LEMMA 6.1.** *Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  by any set of points and let  $F \subset \mathbb{R}^d$  be the optimal  $(k, z)$ -flat approximation of  $X$ . We let  $\mathcal{J}_{d,t}$  be the distribution over Johnson-Lindenstrauss maps. If  $t \gtrsim z/\varepsilon^2$ , then with probability at least 0.99 over the draw of  $\Pi \sim \mathcal{J}_{d,t}$ ,*

$$\sum_{x \in X} \|\Pi(x) - \Pi(\rho_F(x))\|_2^z \leq (1+\varepsilon) \cdot \text{cost}_z^z(X, F),$$

and hence,

$$\min_{\substack{F' \text{ } k\text{-flat} \\ \text{in } \mathbb{R}^t}} \text{cost}_z(\Pi(X), F') \leq (1+\varepsilon) \cdot \text{cost}_z(X, F).$$

The proof follows in a similar fashion to Lemma 4.1 and Lemma 5.1. In particular, there is a natural definition of a  $k$ -flat  $\Pi(F) \subset \mathbb{R}^t$ , and the proof proceeds by upper bounding the expected dilation of  $\|\Pi(x) - \Pi(\rho_F(x))\|_2^z$ .

## 6.2 Hard Direction: Optimum Cost Does Not Decrease

**6.2.1 Preliminaries** The proof in this section follows similarly to that of  $(k, z)$ -subspace approximation.

**DEFINITION 6.1. (WEAK CORESETS FOR  $(k, z)$ -FLAT APPROXIMATION)** *Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be a set of points. A weak  $\varepsilon$ -coreset of  $X$  for  $(k, z)$ -flat approximation is a weighted subset  $S \subset \mathbb{R}^d$  of points with weights  $w: S \rightarrow \mathbb{R}_{\geq 0}$  such that,*

$$\frac{1}{1+\varepsilon} \cdot \min_{\substack{F \text{ } k\text{-flat} \\ \text{in } \mathbb{R}^d}} \text{cost}_z(X, F) \leq \min_{\substack{F \text{ } k\text{-flat} \\ \text{in } \mathbb{R}^d}} \text{cost}_z((S, w), F) \leq (1+\varepsilon) \cdot \min_{\substack{F \text{ } k\text{-flat} \\ \text{in } \mathbb{R}^d}} \text{cost}_z(X, F)$$

**DEFINITION 6.2. (SENSITIVITIES)** *Let  $n, d \in \mathbb{N}$ , and consider any set of points  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ , as well as  $k \in \mathbb{N}$  and  $z \geq 1$ . A sensitivity function  $\sigma: X \rightarrow \mathbb{R}_{\geq 0}$  for  $(k, z)$ -flat approximation in  $\mathbb{R}^d$  is a function satisfying that, for all  $x \in X$ ,*

$$\sup_{\substack{F \subset \mathbb{R}^d \\ k\text{-flat}}} \frac{\|x - \rho_F(x)\|_2^z}{\text{cost}_z^z(X, F)} \leq \sigma(x).$$

*The total sensitivity of the sensitivity function  $\sigma$  is given by*

$$\mathfrak{S}_\sigma = \sum_{x \in X} \sigma(x).$$

*For a sensitivity function, we let  $\tilde{\sigma}$  denote the sensitivity sampling distribution, supported on  $X$ , which samples  $x \in X$  with probability proportional to  $\sigma(x)$ .*

The sensitivity function we use here generalizes that of the previous section. In particular, the proof will follow similarly, and we will defer to the arguments in the previous section.

LEMMA 6.2. Let  $n, d \in \mathbb{N}$ , and consider any set of points  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ , as well as  $k \in \mathbb{N}$  with  $k < d$  and  $z \geq 1$ . Suppose  $F \subset \mathbb{R}^d$  is the optimal  $(k, z)$ -flat approximation of  $X$  in  $\mathbb{R}^d$ . Then, the function  $\sigma: X \rightarrow \mathbb{R}_{\geq 0}$  given by

$$\sigma(x) = 2^{z-1} \cdot \frac{\|x - \rho_F(x)\|_2^z}{\text{cost}_z^z(X, F)} + 2^{2z-1} \cdot \sup_{\substack{u \in \mathbb{R}^d \\ \phi \in \mathbb{R}}} \frac{|\langle \rho_F(x), u \rangle - \phi|^z}{\sum_{x' \in X} |\langle \rho_F(x'), u \rangle - \phi|^z}$$

is a sensitivity function for  $(k, z)$ -flat approximation of  $X$  in  $\mathbb{R}^d$ , satisfying

$$\mathfrak{S}_\sigma \leq 2^{z-1} + 2^{2z-1}(k+2)^{1+z}.$$

*Proof.* Consider any  $k$ -flat  $F' \subset \mathbb{R}^d$ , given by a subspace  $R \subset \mathbb{R}^d$  of dimension at most  $k$ , and a translation  $\tau \in \mathbb{R}^d$ . As in the proof of Lemma 5.2

$$\frac{\|x - \rho_{F'}(x)\|_2^z}{\text{cost}_z^z(X, F')} \leq 2^{z-1} \cdot \frac{\|x - \rho_F(x)\|_2^z}{\text{cost}_z^z(X, F)} + 2^{2z-1} \cdot \frac{\|\rho_F(x) - \rho_{F'}(\rho_F(x))\|_2^z}{\text{cost}_z^z(\rho_F(X), F')}.$$

We now have that for any  $Y \subset \mathbb{R}^d$ , and any  $y \in Y$ ,

$$\sup_{\substack{F' \subset \mathbb{R}^d \\ k\text{-flat}}} \frac{\|y - \rho_{F'}(y)\|_2^z}{\text{cost}_z^z(Y, F')} \leq \sup_{\tau \in \mathbb{R}^d} \sup_{u \in \mathbb{R}^d} \frac{|\langle y - \tau, u \rangle|^z}{\sum_{y' \in Y} |\langle y' - \tau, u \rangle|^z}.$$

Finally, for each  $y \in Y \subset \mathbb{R}^d$ , we may consider appending an additional coordinate and consider  $y^* \in \mathbb{R}^{d+1}$  where the  $d+1$ -th entry is 1. Then, by linearity

$$\sup_{\tau \in \mathbb{R}^d} \sup_{u \in \mathbb{R}^d} \frac{|\langle y - \tau, u \rangle|^z}{\sum_{y' \in Y} |\langle y' - \tau, u \rangle|^z} = \sup_{v \in \mathbb{R}^{d+1}} \frac{|\langle y^*, v \rangle|^z}{\sum_{y' \in Y} |\langle y'^*, v \rangle|^z},$$

and the bound on the total sensitivity follows from Lemma 5.2  $\square$

In the  $(k, z)$ -subspace approximation section, we used a lemma (Lemma 5.3) to narrow down the approximately optimal subspaces to those spanned by at most  $O(k^2 \log(k/\varepsilon)/\varepsilon)$  points. Here, we use a similar lemma in order to find an approximately optimal translation vector  $\tau \in \mathbb{R}^d$ , which is spanned by a small subset of points.

LEMMA 6.3. (LEMMA 3.3 [45]) Let  $d, k \in \mathbb{N}$ , and consider a weighted set of points  $S \subset \mathbb{R}^d$  with weights  $w: S \rightarrow \mathbb{R}_{\geq 0}$ , as well as  $\varepsilon \in (0, 1)$  and  $z \geq 1$ . Suppose  $F \subset \mathbb{R}^d$  is the optimal  $(k, z)$ -flat approximation of  $X$ , encoded by a  $k$ -dimensional subspace  $R \subset \mathbb{R}^d$  and translation vector  $\tau \in \mathbb{R}^d$ . There exists a subset  $Q \subset S$  of size  $O(\log(1/\varepsilon)/\varepsilon)$  and a point  $\tau' \in \text{conv}(Q)$  such that the  $k$ -flat

$$F' = \{\tau' + y \in \mathbb{R}^d : y \in R\}$$

satisfies

$$\text{cost}_z((S, w), F') \leq (1 + \varepsilon) \cdot \text{cost}_z((S, w), F).$$

THEOREM 6.2. ( $\varepsilon$ -WEAK CORESETS FOR  $k$ -FLATS VIA SENSITIVITY SAMPLING) For any subset  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  and  $\varepsilon \in (0, 1/2)$ , let  $\tilde{\sigma}$  denote the sensitivity sampling distribution.

- Let  $(\mathbf{S}, \mathbf{w})$  denote the random (multi-)set  $\mathbf{S} \subset X$  and  $\mathbf{w}: \mathbf{S} \rightarrow \mathbb{R}_{\geq 0}$  given by, for

$$m = \text{poly}((k+2)^z, 1/\varepsilon)$$

iterations, sampling  $\mathbf{x} \sim \tilde{\sigma}$  i.i.d and letting  $\mathbf{w}(\mathbf{x}) = 1/(m\tilde{\sigma}(\mathbf{x}))$ .

- Then, with probability  $1 - o(1)$  over the draw of  $(\mathbf{S}, \mathbf{w})$ , it is an  $\varepsilon$ -weak coresets for  $(k, z)$ -subspace approximation of  $X$ .



**6.3 The Important Events** The important events we consider mirror those the subspace approximation problem. The only event which would change is  $\mathbf{E}_2$ , where we require  $\Pi$  to be an  $\varepsilon$ -subspace embedding for all subsets of  $O(k^2 \log(k/\varepsilon)/\varepsilon) + O(\log(1/\varepsilon)/\varepsilon)$  points from  $\mathbf{S}$ . This will allow us to incorporate the translation  $\tau'$  from Lemma 6.3.

**DEFINITION 6.3. (THE EVENTS)** Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ , and  $\tilde{\sigma}$  the sensitivity sampling distribution of  $X$  of Lemma 6.2. We consider the following experiment,

1. We generate a sample  $(\mathbf{S}, \mathbf{w})$  by sampling from  $\tilde{\sigma}$  for  $m = \text{poly}(k^z, 1/\varepsilon)$  i.i.d iterations  $\mathbf{x} \sim \tilde{\sigma}$  and set  $\mathbf{w}(\mathbf{x}) = 1/(m\tilde{\sigma}(\mathbf{x}))$ .
2. Furthermore, we sample  $\Pi \sim \mathcal{J}_{d,t}$ , which is a Johnson-Lindenstrauss map  $\mathbb{R}^d \rightarrow \mathbb{R}^t$ .
3. We let  $\mathbf{S}' = \Pi(\mathbf{S}) \subset \mathbb{R}^t$  denote the image of  $\Pi$  on  $\mathbf{S}$ .

The events are the following:

- $\mathbf{E}_1$  : The weighted (multi-)set  $(\mathbf{S}, \mathbf{w})$  is a weak  $\varepsilon$ -coreset for  $(k, z)$ -flat approximation of  $X$  in  $\mathbb{R}^d$ .
- $\mathbf{E}_2$  : The map  $\Pi: \mathbb{R}^d \rightarrow \mathbb{R}^t$  satisfies the following condition. For any choice of  $O(k^2 \log(k/\varepsilon)/\varepsilon)$  points of  $\mathbf{S}$ ,  $\Pi$  is an  $\varepsilon$ -subspace embedding of the subspace spanned by these points.
- $\mathbf{E}_3(\beta)$  : Let  $\mathbf{F}' \subset \mathbb{R}^t$  denote the optimal  $(k, z)$ -flat approximation of  $\Pi(X)$  in  $\mathbb{R}^t$ . Then,

$$\text{cost}_z^z((\Pi(\mathbf{S}), \mathbf{w}), \mathbf{F}') \leq \beta \cdot \text{cost}_z^z(\Pi(X), \mathbf{F}').$$

**LEMMA 6.4.** Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ , and suppose  $(\mathbf{S}, \mathbf{w})$  and  $\Pi: \mathbb{R}^d \rightarrow \mathbb{R}^t$  satisfy events  $\mathbf{E}_1$ ,  $\mathbf{E}_2$ , and  $\mathbf{E}_3(\beta)$ . Then,

$$\min_{\substack{F' \text{ } k\text{-flat} \\ \text{in } \mathbb{R}^t}} \text{cost}_z(\Pi(X), F') \geq \frac{1}{\beta^{1/z}(1+\varepsilon)^4} \cdot \min_{\substack{F \text{ } k\text{-flat} \\ \text{in } \mathbb{R}^d}} \text{cost}_z(X, F).$$

*Proof.* Consider a fixed  $\Pi$  and  $(\mathbf{S}, \mathbf{w})$  satisfying the three events of Definition 5.3. Let  $F' \subset \mathbb{R}^t$  be the  $k$ -flat which minimizes  $\text{cost}_z^z(\Pi(X), F')$ . Suppose that  $F'$  is specified by a  $k$ -dimensional subspace  $R'$  and a translation  $\tau'$ . Then, by event  $\mathbf{E}_3(\beta)$ , we have  $\text{cost}_z^z((\Pi(\mathbf{S}), \mathbf{w}), F') \leq \beta \cdot \text{cost}_z^z(\Pi(X), F')$ . Now, we apply Lemma 6.3 to  $(\Pi(\mathbf{S}), \mathbf{w})$ , and we obtain a subset  $Q \subset \mathbf{S}$  of size  $O(\log(1/\varepsilon)/\varepsilon)$  for which there exists a translation vector  $\tau'' \in \mathbb{R}^t$  within the  $\text{conv}(\Pi(Q))$  such that  $k$ -flat  $F''$  given by  $\tau''$  and  $R'$  satisfies

$$\left( \sum_{x \in \mathbf{S}} \mathbf{w}(x) \cdot \|\Pi(x) - \rho_{F''}(\Pi(x))\|_2^z \right)^{1/z} = \text{cost}_z((\Pi(\mathbf{S}), \mathbf{w}), F'') \leq (1+\varepsilon) \cdot \text{cost}_z((\Pi(\mathbf{S}), \mathbf{w}), F').$$

Furthermore, by Lemma 5.3 to the weighted vectors  $(\Pi(\mathbf{S}) - \tau'', \mathbf{w})$ <sup>8</sup> there exists a subset  $Q' \subset \Pi(\mathbf{S}) - \tau''$  of size  $O(k^2 \log(k/\varepsilon)/\varepsilon)$  and a  $k$ -dimensional subspace  $R'' \subset \mathbb{R}^d$  within the span of  $Q'$  such that the  $k$ -flat  $F'''$  specified by  $R''$  and  $\tau''$  satisfies

$$\left( \sum_{x \in \mathbf{S}} \mathbf{w}(x) \cdot \|\Pi(x) - \rho_{F'''}(\Pi(x))\|_2^z \right)^{1/z} = \text{cost}_z((\Pi(\mathbf{S}), \mathbf{w}), F''') \leq (1+\varepsilon)^2 \cdot \text{cost}_z((\Pi(\mathbf{S}), \mathbf{w}), F').$$

Recall that (i)  $R''$  is a  $k$ -dimensional subspace lying in the span of  $\Pi(Q')$ , (ii)  $\tau'' \in \mathbb{R}^t$  is within  $\text{conv}(\Pi(Q))$ , and (iii) for any  $x \in \mathbf{S}$ ,  $\Pi$  is an  $\varepsilon$ -subspace embedding of the span of  $Q \cup Q' \cup \{x\}$ . Similarly to Lemma 5.4, we may find a  $k$ -flat  $U$  such that for every  $x \in \mathbf{S}$ ,

$$\|x - \rho_U(x)\|_2 \leq (1+\varepsilon) \|\Pi(x) - \rho_{F'''}(\Pi(x))\|_2,$$

and hence

$$\text{cost}_z((\mathbf{S}, \mathbf{w}), U) \leq (1+\varepsilon) \cdot \text{cost}_z((\Pi(\mathbf{S}), \mathbf{w}), F''').$$

Finally, since  $(\mathbf{S}, \mathbf{w})$  is a  $\varepsilon$ -weak coreset, we obtain the desired inequality.  $\square$

<sup>8</sup>Here, we are using the short-hand  $\Pi(\mathbf{S}) - \tau'' \stackrel{\text{def}}{=} \{\Pi(x) - \tau'' \in \mathbb{R}^t : x \in \mathbf{S}\}$ .

As in the previous section, events  $\mathbf{E}_1$  and  $\mathbf{E}_2$  hold with sufficiently high probability. All that remains is showing that  $\mathbf{E}_3(1 + \varepsilon)$  holds with sufficiently high probability. We proceed in a similar fashion, where we first show a loose approximation guarantee, and later improve on it.

LEMMA 6.5. Fix any  $\Pi \in \mathcal{J}_{d,t}$  and let  $F' \subset \mathbb{R}^t$  denote the  $k$ -flat for optimal  $(k, z)$ -flat approximation of  $\Pi(X)$  in  $\mathbb{R}^t$ . Then with probability at least 0.99 over the draw of  $(\mathbf{S}, \mathbf{w})$  as per Definition 6.3,

$$\sum_{x \in \mathbf{S}} \mathbf{w}(x) \cdot \|\Pi(x) - \rho_{F'}(\Pi(x))\|_2^z \leq 100 \cdot \text{cost}_z^z(\Pi(X), F').$$

In other words, event  $\mathbf{E}_3(100)$  holds with probability at least 0.99.

COROLLARY 6.1. Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be any set of points. For any  $\varepsilon \in (0, 1/2)$ , let  $\mathcal{J}_{d,t}$  be the Johnson-Lindenstrauss map with

$$t \gtrsim \frac{z \cdot k^2 \cdot \text{polylog}(k/\varepsilon)}{\varepsilon^3}.$$

Then, with probability at least 0.97 over the draw of  $\Pi \sim \mathcal{J}_{d,t}$ ,

$$\frac{1}{100^{1/z}(1 + \varepsilon)^4} \cdot \min_{\substack{F \subset \mathbb{R}^d \\ k\text{-flat}}} \text{cost}_z(X, F) \leq \min_{\substack{F' \subset \mathbb{R}^t \\ k\text{-flat}}} \text{cost}_z(\Pi(X), F').$$

**6.3.1 Improving the approximation** The improvement of the approximation, follows from upper bounding the variance, as in the  $(k, z)$ -clustering problem, and the  $(k, z)$ -subspace approximation problem. In particular, we show that  $\mathbf{E}_3(1 + \varepsilon)$  holds. Fix  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  and  $F \subset \mathbb{R}^d$  be the optimal  $(k, z)$ -flat approximation of  $X$  in  $\mathbb{R}^d$ . The sensitivity function  $\sigma: X \rightarrow \mathbb{R}_{\geq 0}$  specified in Lemma 6.2 specify the sensitivity sampling distribution  $\tilde{\sigma}$ .

We let  $\mathbf{E}_4$  denote the following event with respect to the randomness in  $\Pi \sim \mathcal{J}_{d,t}$ . For each  $x \in X$ , we let  $\mathbf{D}_x \in \mathbb{R}_{\geq 0}$  denote the random variable

$$\mathbf{D}_x \stackrel{\text{def}}{=} \frac{\|\Pi(x) - \Pi(\rho_F(x))\|_2}{\|x - \rho_F(x)\|_2},$$

and as in (5.23) and (5.24), event  $\mathbf{E}_4$ , which occurs with probability at least 0.99, whenever

$$\sum_{x \in X} \mathbf{D}_x^{2z} \cdot \sigma(x) \leq 100 \cdot 2^z \cdot \mathfrak{S}_\sigma.$$

LEMMA 6.6. Let  $\Pi \in \mathcal{J}_{d,t}$  be a Johnson-Lindenstrauss map where, for  $\alpha > 1$ , the following events hold:

1. Guarantee from Lemma 6.1:  $\sum_{x \in X} \|\Pi(x) - \Pi(\rho_F(x))\|_2^z \leq \alpha \cdot \text{cost}_z^z(X, F)$ .
2. Guarantee from Corollary 6.1: letting  $F' \subset \mathbb{R}^t$  be the optimal  $(k, z)$ -flat approximation of  $\Pi(X)$ , then  $\text{cost}_z^z(X, F) \leq \alpha \text{cost}_z^z(\Pi(X), F')$ .
3. Event  $\mathbf{E}_4$  holds.

Then, if we let  $(\mathbf{S}, \mathbf{w})$  denote  $m = \text{poly}(k^z, 1/\varepsilon, \alpha)$  i.i.d draws from  $\tilde{\sigma}$  and  $\mathbf{w}(x) = 1/(m\tilde{\sigma}(x))$ , with probability at least 0.99,

$$\text{cost}_z^z((\Pi(\mathbf{S}), \mathbf{w}), F') \leq (1 + \varepsilon) \cdot \text{cost}_z^z(\Pi(X), F').$$

*Proof.* We similarly bound the variance of

$$\mathbf{Vars} \left[ \mathbf{E}_{\mathbf{x} \sim \mathbf{S}} \left[ \frac{1}{\tilde{\sigma}(\mathbf{x})} \frac{\|\Pi(\mathbf{x}) - \rho_{F'}(\Pi(\mathbf{x}))\|_2^z}{\text{cost}_z^z(\Pi(X), F')} \right] \right] \leq \frac{\mathfrak{S}_\sigma}{m} \sum_{x \in X} \left( \frac{1}{\sigma(x)} \cdot \frac{\|\Pi(x) - \rho_{F'}(\Pi(x))\|_2^z}{\text{cost}_z^{2z}(\Pi(X), F')} \right).$$

It is not hard to show, as in the proof of Lemma 5.7, that writing  $y_x = \rho_F(x) \in \mathbb{R}^d$  and  $Y = \{y_x : x \in X\}$ , that

$$\frac{\|\Pi(x) - \rho_{F'}(\Pi(x))\|_2^z}{\text{cost}_z^z(\Pi(X), F')} \leq 2^{z-1} \alpha \mathbf{D}_x^z \cdot \frac{\|x - \rho_F(x)\|_2^z}{\text{cost}_z^z(X, F)} + 2^{2z-2} (1 + \alpha^2) \sup_{\substack{v \in \mathbb{R}^t \\ \mu \in \mathbb{R}}} \frac{|\langle \Pi(y_x), v \rangle - \mu|^z}{\sum_{x' \in X} |\langle \Pi(y_{x'}), v \rangle - \mu|^z},$$

and similarly to before, we have

$$\sup_{\substack{v \in \mathbb{R}^t \\ \mu \in \mathbb{R}}} \frac{|\langle \Pi(y_x), v \rangle - \mu|^z}{\sum_{x' \in X} |\langle \Pi(y_{x'}), v \rangle - \mu|^z} \leq \sup_{\substack{u \in \mathbb{R}^d \\ \phi \in \mathbb{R}}} \frac{|\langle y_x, u \rangle - \phi|^z}{\sum_{x' \in X} |\langle y_{x'}, u \rangle - \phi|^z}.$$

This implies that the variance is at most

$$2^{4z-2} \cdot \frac{\mathfrak{S}_\sigma \alpha^4}{m} \cdot \sum_{x \in X} \mathbf{D}_x^{2z} \cdot \sigma(x) \leq \varepsilon^2$$

by setting  $m = \text{poly}((k+2)^z, 1/\varepsilon, \alpha)$  to be large enough when  $\mathbf{E}_4$  holds, and we apply Chebyshev's inequality.  $\square$

**COROLLARY 6.2.** *Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be any set of points, and let  $F \subset \mathbb{R}^d$  be the optimal  $(k, z)$ -flat approximation of  $X$ . For any  $\varepsilon \in (0, 1/2)$ , let  $\mathcal{J}_{d,t}$  be the Johnson-Lindenstrauss map with*

$$t \gtrsim \frac{z \cdot k^2 \cdot \text{polylog}(k/\varepsilon)}{\varepsilon^3}.$$

*Then, with probability at least 0.92 over the draw of  $\Pi \sim \mathcal{J}_{d,t}$ ,*

$$\frac{1}{(1 + \varepsilon)^{4+1/z}} \cdot \text{cost}_z(X, F) \leq \min_{\substack{F' \subset \mathbb{R}^t \\ k\text{-flat}}} \text{cost}_z(\Pi(X), F').$$

## 7 $k$ -Line Approximation

In the  $(k, z)$ -line approximation problem, we consider a collection of  $k$  lines in  $\mathbb{R}^d$ . A line is encoded by a vector  $v \in \mathbb{R}^d$  and a unit vector  $u \in S^{d-1}$ , where we will write

$$\ell(v, u) = \{v + t \cdot u \in \mathbb{R}^d : t \in \mathbb{R}\}.$$

For a single line  $\ell$  encoded by  $v$  and  $u$ , we write  $\rho_\ell: \mathbb{R}^d \rightarrow \mathbb{R}$  as the orthogonal projection of a point onto  $\ell$ , i.e., the closest vector which lies on the line, where

$$\rho_\ell(x) = \underset{y \in \ell}{\text{argmin}} \|x - y\|_2^2 = v + \langle x - v, u \rangle u.$$

For any set of  $k$  lines,  $L = \{\ell_1, \dots, \ell_k\}$ , and a point  $x \in X$ , we write

$$\text{cost}_z^z(x, L) = \min_{\ell \in L} \|x - \rho_\ell(x)\|_2^z,$$

and for any dataset  $X \subset \mathbb{R}^d$  and set of lines  $L$ , we consider the map  $\ell: X \rightarrow L$  which sends a point  $x$  to its nearest line in  $L$ . Then, we write

$$\text{cost}_z^z(X, L) = \sum_{x \in X} \|x - \rho_{\ell(x)}(x)\|_2^z,$$

as the cost of representing the points in  $X$  according to the  $k$  lines in  $L$ . In this section, we show that we may find the optimal  $(k, z)$ -line approximation after applying a Johnson-Lindenstrauss map. Specifically, we prove:

**THEOREM 7.1. (JOHNSON-LINDENSTRAUSS FOR  $(k, z)$ -LINE APPROXIMATION)** *Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be any set of points, and let  $L = \{\ell_1, \dots, \ell_k\}$  denote a set of lines in  $\mathbb{R}^d$  for optimally  $(k, z)$ -line approximation of  $X$ . For any  $\varepsilon \in (0, 1)$ , suppose we let  $\mathcal{J}_{d,t}$  be a distribution over Johnson-Lindenstrauss maps where*

$$t \gtrsim \frac{k \log \log n + z + \log(1/\varepsilon)}{\varepsilon^3}.$$

*Then, with probability at least 0.9 over the draw of  $\Pi \sim \mathcal{J}_{d,t}$ ,*

$$\frac{1}{1+\varepsilon} \cdot \text{cost}_z(X, L) \leq \min_{\substack{L' \text{ } k \text{ lines} \\ \text{in } \mathbb{R}^t}} \text{cost}_z(\Pi(X), L') \leq (1+\varepsilon) \cdot \text{cost}_z(X, L).$$

## 7.1 Easy Direction: Optimum Cost Does Not Increase

**LEMMA 7.1.** *Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be any set of points and let  $L = \{\ell_1, \dots, \ell_k\}$  be a set of  $k$  lines in  $\mathbb{R}^d$  for optimal  $(k, z)$ -line approximation of  $X$ , and for each  $x \in X$ , let  $\ell(x) \in L$  be the line assigned to  $x$ . We let  $\mathcal{J}_{d,t}$  be the distribution over Johnson-Lindenstrauss maps. If  $t \gtrsim z/\varepsilon^2$ , then with probability at least 0.99 over the draw of  $\Pi \sim \mathcal{J}_{d,t}$ ,*

$$\left( \sum_{x \in X} \|\Pi(x) - \Pi(\rho_{\ell(x)}(x))\|_2^z \right)^{1/z} \leq (1+\varepsilon) \cdot \text{cost}_z(X, L),$$

and hence,

$$\min_{\substack{L' \text{ } k \text{ lines} \\ \text{in } \mathbb{R}^t}} \text{cost}_z(\Pi(X), L') \leq (1+\varepsilon) \cdot \text{cost}_z(X, L).$$

By now, there is a straight-forward way to prove the above lemma. For each set of  $k$  lines  $L = \{\ell_1, \dots, \ell_k\}$  in  $\mathbb{R}^d$ , there is an analogous definition of  $k$  lines  $\Pi(L)$  in  $\mathbb{R}^t$ . Hence, we use the two-stability of the Gaussian in previous sections to upper bound the cost  $\text{cost}_z(\Pi(X), \Pi(L))$ .

## 7.2 Hard Direction: Optimum Cost Does Not Decrease

**7.2.1 Preliminaries** At a high level, we proceed with the same argument as in previous sections: we consider a sensitivity function for  $(k, z)$ -line approximation of  $X$  in  $\mathbb{R}^d$ , and use it to build a weak coresets, as well as argue that sensitivity sampling is a low-variance estimator of the optimal  $(k, z)$ -line approximation in the projected space. The proof in this section will be significantly more complicated than the previous section. Defining the appropriate sensitivity functions, which will give a low-variance estimator in the projected space, is considerably more difficult than the expressions of Lemmas 4.2, 5.2, and 6.2. For this reason, we will proceed by assuming access to a sensitivity function which we will define later in the section.

**DEFINITION 7.1. (WEAK CORESETS FOR  $(k, z)$ -LINE APPROXIMATION)** *Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be a set of points. A weak  $\varepsilon$ -coreset of  $X$  for  $(k, z)$ -line approximation is a weighted subset  $S \subset \mathbb{R}^d$  of points with weights  $w: S \rightarrow \mathbb{R}_{\geq 0}$  such that*

$$\frac{1}{1+\varepsilon} \cdot \min_{\substack{L \text{ } k \text{ lines} \\ \text{in } \mathbb{R}^d}} \text{cost}_z(X, L) \leq \min_{\substack{L \text{ } k \text{ lines} \\ \text{in } \mathbb{R}^d}} \text{cost}_z((S, w), L) \leq (1+\varepsilon) \cdot \min_{\substack{L \text{ } k \text{ lines} \\ \text{in } \mathbb{R}^d}} \text{cost}_z(X, L)$$

**DEFINITION 7.2. (SENSITIVITIES)** *Let  $n, d \in \mathbb{N}$ , and consider any set of points  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ , as well as  $k \in \mathbb{N}$  and  $z \geq 1$ . A sensitivity function  $\sigma: X \rightarrow \mathbb{R}_{\geq 0}$  for  $(k, z)$ -line approximation in  $\mathbb{R}^d$  is a function which satisfies that, for all  $x \in X$ ,*

$$\sup_{\substack{L: \text{ } k \text{ lines} \\ \text{in } \mathbb{R}^d}} \frac{\|x - \rho_{\ell(x)}(x)\|_2^z}{\text{cost}_z^z(X, L)} \leq \sigma(x).$$

*The total sensitivity of the sensitivity function  $\sigma$  is given by*

$$\mathfrak{S}_\sigma = \sum_{x \in X} \sigma(x).$$

For a sensitivity function, we let  $\tilde{\sigma}$  denote the sensitivity sampling distribution, supported on  $X$ , which samples  $x \in X$  with probability proportional to  $\sigma(x)$ .

Similarly to before, we first give a lemma which narrows down the space of the optimal line approximations for a set of points. the following lemma is a re-formulations of Lemma 5.3 and Lemma 6.3 catered to the case of  $(k, z)$ -line approximation.

LEMMA 7.2. (THEOREM 3.1 AND LEMMA 3.3 OF [44]) *Let  $d \in \mathbb{N}$  and  $S \subset \mathbb{R}^d$  be any set of points with weights  $w: S \rightarrow \mathbb{R}_{\geq 0}$ ,  $\varepsilon \in (0, 1/2)$ , and  $z \geq 1$ . There exists a subset  $Q \subset S$  of size  $O(\log(1/\varepsilon)/\varepsilon)$  and a line  $\ell$  in  $\mathbb{R}^d$  within the span of  $Q$  such that*

$$\text{cost}_z((S, w), \{\ell\}) \leq (1 + \varepsilon) \min_{\substack{\ell' \text{ line} \\ \text{in } \mathbb{R}^d}} \text{cost}_z((S, w), \{\ell'\}).$$

LEMMA 7.3. (WEAK CORESETS FOR  $k$ -LINE APPROXIMATION [23, 49]) *For any subset  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  and  $\varepsilon \in (0, 1/2)$ , let  $\sigma$  denote a sensitivity function for  $(k, z)$ -line approximation of  $X$  with total sensitivity  $\mathfrak{S}_\sigma$  and let  $\tilde{\sigma}$  its sensitivity sampling distribution.*

- Let  $(\mathbf{S}, \mathbf{w})$  denote the random (multi-)set  $\mathbf{S} \subset X$  and  $\mathbf{w}: \mathbf{S} \rightarrow \mathbb{R}_{\geq 0}$  given by, for

$$m = \text{poly}(\mathfrak{S}_\sigma, k, 1/\varepsilon),$$

iterations, sampling  $\mathbf{x} \sim \tilde{\sigma}$  i.i.d and letting  $\mathbf{w}(\mathbf{x}) = 1/(m\tilde{\sigma}(\mathbf{x}))$ .

- Then, with probability  $1 - o(1)$  over the draw of  $(\mathbf{S}, \mathbf{w})$ , it is an  $\varepsilon$ -weak coreset for  $(k, z)$ -line approximation of  $X$ .

We note that [23] and [49] only give a strong coreset for  $(k, z)$ -line approximation of  $\text{poly}(\mathfrak{S}_\sigma, k, d, 1/\varepsilon)$ . For example, Theorem 13 in [49] giving the above bound follows from the fact that the “function dimension” (see Definition 3 of [49]) for  $(k, z)$ -line approximation is  $O(kd)$ . However, Lemma 7.2 implies that for any set of points, a line which approximates the points is within a span of  $O(\log(1/\varepsilon)/\varepsilon)$  points. This means that, for  $\varepsilon$ -weak coresets, it suffices to only consider  $k$  lines spanned by  $O(k \log(1/\varepsilon)/\varepsilon)$ , giving us a “function dimension” of  $O(k \log(1/\varepsilon)/\varepsilon)$ .

### 7.2.2 The Important Events

DEFINITION 7.3. (THE EVENTS) *Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ , and  $\sigma$  be a sensitivity function for  $(k, z)$ -line approximation of  $X$  in  $\mathbb{R}^d$ , with total sensitivity  $\mathfrak{S}_\sigma$  and sensitivity sampling distribution  $\tilde{\sigma}$ . We consider the following experiment,*

1. We generate a sample  $(\mathbf{S}, \mathbf{w})$  by sampling from  $\tilde{\sigma}$  for  $m = \text{poly}(\mathfrak{S}_\sigma, k, 1/\varepsilon)$  i.i.d iterations  $\mathbf{x} \sim \tilde{\sigma}$  and set  $\mathbf{w}(\mathbf{x}) = 1/(m\tilde{\sigma}(\mathbf{x}))$ .
2. Furthermore, we sample  $\mathbf{\Pi} \sim \mathcal{J}_{d,t}$ , which is a Johnson-Lindenstrauss map  $\mathbb{R}^d \rightarrow \mathbb{R}^t$ .
3. We let  $\mathbf{S}' = \mathbf{\Pi}(\mathbf{S}) \subset \mathbb{R}^t$  denote the image of  $\mathbf{\Pi}$  on  $\mathbf{S}$ .

The events are the following:

- $\mathbf{E}_1$  : The weighted (multi-)set  $(\mathbf{S}, \mathbf{w})$  is a weak  $\varepsilon$ -coreset for  $(k, z)$ -line approximation of  $X$  in  $\mathbb{R}^d$ .
- $\mathbf{E}_2$  : For any subset of  $O(\log(1/\varepsilon)/\varepsilon)$  points from  $\mathbf{S}$ , the map  $\mathbf{\Pi}: \mathbb{R}^d \rightarrow \mathbb{R}^t$  is an  $\varepsilon$ -subspace embedding for the subspace spanned by that subset.
- $\mathbf{E}_3(\beta)$  : Let  $\mathbf{L}' = \{\ell'_1, \dots, \ell'_k\}$  denote  $k$  lines in  $\mathbb{R}^t$  for optimal  $(k, z)$ -line approximation of  $\mathbf{\Pi}(X)$  in  $\mathbb{R}^t$ . Then,

$$\text{cost}_z^z((\mathbf{\Pi}(\mathbf{S}), \mathbf{w}), \mathbf{L}') \leq \beta \cdot \text{cost}_z^z(\mathbf{\Pi}(X), \mathbf{L}').$$



LEMMA 7.4. Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ , and suppose  $(\mathbf{S}, \mathbf{w})$  and  $\mathbf{\Pi}: \mathbb{R}^d \rightarrow \mathbb{R}^t$  satisfy events  $\mathbf{E}_1, \mathbf{E}_2$  and  $\mathbf{E}_3$ . Then,

$$\min_{\substack{L' \text{ } k \text{ lines} \\ \text{in } \mathbb{R}^t}} \text{cost}_z(\mathbf{\Pi}(X), L') \geq \frac{1}{\beta^{1/z}(1+\varepsilon)^3} \cdot \min_{\substack{L \text{ } k \text{ lines} \\ \text{in } \mathbb{R}^d}} \text{cost}_z(X, L).$$

*Proof.* Let  $\mathbf{\Pi} \sim \mathcal{J}_{d,t}$  and  $(\mathbf{S}, \mathbf{w})$  be sampled according to Definition 7.3, and suppose events  $\mathbf{E}_1, \mathbf{E}_2$  and  $\mathbf{E}_3$  all hold. Let  $\mathbf{L}' = \{\ell'_1, \dots, \ell'_k\}$  denote the set of  $k$  lines for optimal  $(k, z)$ -line approximation of  $\mathbf{\Pi}(X)$  in  $\mathbb{R}^t$ . Then, by event  $\mathbf{E}_3$ , we have  $\text{cost}_z^z((\mathbf{\Pi}(\mathbf{S}), \mathbf{w}), \mathbf{L}') \leq \beta \cdot \text{cost}_z^z(\mathbf{\Pi}(X), \mathbf{L}')$ . Consider the partition of  $\mathbf{S}$  into  $\mathbf{S}_1, \dots, \mathbf{S}_k$  induced by the lines in  $\mathbf{L}'$  closest to  $\mathbf{\Pi}(\mathbf{S})$ .

For each  $i \in [k]$ , we apply Lemma 7.2 to  $\mathbf{\Pi}(\mathbf{S}_i)$  with weights  $\mathbf{w}: \mathbf{S}_i \rightarrow \mathbb{R}_{\geq 0}$ . In particular, there exists subsets  $\mathbf{Q}_1 \subset \mathbf{S}_1, \dots, \mathbf{Q}_k \subset \mathbf{S}_k$  and  $k$  lines  $\mathbf{L}'' = \{\ell''_1, \dots, \ell''_k\}$  in  $\mathbb{R}^t$  such that each line  $\ell''_i$  lie in the span of  $\mathbf{Q}_i$ , and

$$\text{cost}_z((\mathbf{\Pi}(\mathbf{S}), \mathbf{w}), \mathbf{L}'') \leq (1 + \varepsilon) \cdot \text{cost}_z((\mathbf{\Pi}(\mathbf{S}), \mathbf{w}), \mathbf{L}').$$

Event  $\mathbf{E}_2$  implies that for each  $i \in [k]$  and each  $x \in \mathbf{S}$ ,  $\mathbf{\Pi}$  is an  $\varepsilon$ -subspace embedding for the subspace spanned by  $\mathbf{Q}_i \cup \{x\}$ . It is not hard to see, that there exists  $k$  lines  $\mathbf{H} = \{\mathbf{h}_1, \dots, \mathbf{h}_k\}$  in  $\mathbb{R}^d$  such that for all  $x \in \mathbf{S}$ ,

$$\|x - \rho_{\mathbf{h}_i}(x)\|_2 \leq (1 + \varepsilon) \cdot \|\mathbf{\Pi}(x) - \rho_{\ell''_i}(\mathbf{\Pi}(x))\|_2,$$

and therefore,

$$\text{cost}_z((\mathbf{S}, \mathbf{w}), \mathbf{H}) \leq (1 + \varepsilon) \cdot \text{cost}_z((\mathbf{\Pi}(\mathbf{S}), \mathbf{w}), \mathbf{L}'').$$

Lastly,  $(\mathbf{S}, \mathbf{w})$  is a  $\varepsilon$ -weak coreset for  $X$ , which means that

$$\min_{\substack{L \text{ } k \text{ lines} \\ \text{in } \mathbb{R}^d}} \text{cost}_z(X, L) \leq (1 + \varepsilon) \cdot \text{cost}_z((\mathbf{S}, \mathbf{w}), \mathbf{H}).$$

Combining all inequalities gives the desired lemma.  $\square$

By now, we note that it is straight-forward to prove the following corollary, which gives a dimension reduction bound which depends on the total sensitivity of a sensitivity function.

COROLLARY 7.1. Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be any set of points, and for  $k \in \mathbb{N}$  and  $z \geq 1$ , let  $\sigma: X \rightarrow \mathbb{R}_{\geq 0}$  be a sensitivity function for  $(k, z)$ -line approximation of  $X$  in  $\mathbb{R}^d$ . For any  $\varepsilon \in (0, 1/2)$ , let  $\mathcal{J}_{d,t}$  be the Johnson-Lindenstrauss map with

$$t \gtrsim \frac{\log(\mathfrak{S}_\sigma, k, 1/\varepsilon)}{\varepsilon^3}.$$

Then, with probability at least 0.97 over the draw of  $\mathbf{\Pi} \sim \mathcal{J}_{d,t}$ ,

$$\frac{1}{100^{1/z}(1+\varepsilon)^3} \min_{\substack{L \text{ } k \text{ lines} \\ \text{in } \mathbb{R}^d}} \text{cost}_z(X, L) \leq \min_{\substack{L' \text{ } k \text{ lines} \\ \text{in } \mathbb{R}^t}} \text{cost}_z(\mathbf{\Pi}(X), L')$$

**7.3 A Sensitivity Function for  $(k, z)$ -Line Approximation** We now describe a sensitivity function for  $(k, z)$ -line approximation of points in  $\mathbb{R}^d$ . Similarly to the previous section, we consider a set of points  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ , and we design a sensitivity function  $\sigma: X \rightarrow \mathbb{R}_{\geq 0}$  for  $(k, z)$ -line approximation of  $X$  in  $\mathbb{R}^d$ . The sensitivity function should satisfy two requirements. The first is that we have a good bound on the total sensitivity,  $\mathfrak{S}_\sigma$ , where the target dimension  $t$  will have logarithmic dependence on  $\mathfrak{S}_\sigma$  (for example, like in Corollary 7.1).

The second is that  $\mathbf{E}_3(1 + \varepsilon)$  will hold with sufficiently high probability over the draw of  $\mathbf{\Pi} \sim \mathcal{J}_{d,t}$ . In other words, we will proceed similarly to Lemmas 4.6, 5.7 and 6.6 and show that, for the optimal  $(k, z)$ -line approximation  $\mathbf{L}'$  of  $\mathbf{\Pi}(X)$  in  $\mathbb{R}^t$ , sampling according to the sensitivity sampling distribution gives a low-variance estimate for the cost of  $\mathbf{L}'$ .

**7.3.1 From Coresets for  $(k, \infty)$ -line approximation to Sensitivity Functions** Unfortunately, we do not know of a “clean” description of a sensitivity function for  $(k, z)$ -line approximation, as was the case in previous definitions. Certainly, one may define a sensitivity function to be  $\sigma(x) = \sup_L \text{cost}_z^z(x, L) / \text{cost}_z^z(X, L)$ , but then arguing that  $\mathbf{E}_3(1 + \varepsilon)$  holds with high probability becomes more complicated. The sensitivity function which we present follows the connection between sensitivity and  $\ell_\infty$ -coresets [48].

**DEFINITION 7.4. ( $c$ -CORESETS FOR  $(k, \infty)$ -LINE APPROXIMATION)** Let  $Y = \{y_1, \dots, y_n\} \subset \mathbb{R}^d$  be any subset of points, and  $c \geq 1$ . A subset  $A \subset Y$  is a  $c$ -coreset for  $(k, \infty)$ -line approximation if the following holds:

- Let  $L = \{\ell_1, \dots, \ell_k\}$  be any collection of  $k$  lines in  $\mathbb{R}^d$ , and  $r \in \mathbb{R}_{\geq 0}$  such that for all  $y \in A$ ,

$$\min_{\ell \in L} \|y - \rho_\ell(y)\|_2 \leq r.$$

- Then, for all  $x \in X$ ,

$$\min_{\ell \in L} \|y - \rho_\ell(y)\|_2 \leq cr.$$

Note that the  $(k, \infty)$ -line approximation is the problem of minimum enclosing cylinder: we are given a set of points  $Y$ , and want to find a set of  $k$  cylinders  $C_1, \dots, C_k \subset \mathbb{R}^d$  of smallest radius such that  $Y \subset \bigcup_{i=1}^k C_i$ . Thus, Definition 7.4, a set  $A \subset Y$  is a  $c$ -coreset for  $(k, \infty)$ -line approximation if, given any  $k$  cylinders which contain  $A$ , increasing the radii by a factor of  $c$  contains  $Y$ . The reason they will be relevant for defining a sensitivity function is the following simple lemma, whose main idea is from [48].

**LEMMA 7.5. (SENSITIVITIES FROM  $c$ -CORESETS FOR  $(k, \infty)$ -LINE APPROXIMATION (SEE LEMMA 3.1 IN [48]))** Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be any set of points and  $k \in \mathbb{N}$ ,  $z \geq 1$ . Let  $L = \{\ell_1, \dots, \ell_k\}$  be the  $k$  lines in  $\mathbb{R}^d$  for optimal  $(k, z)$ -line approximation of  $X$ , and let  $Y = \{y_x \in \mathbb{R}^d : x \in X\}$  where  $y_x = \rho_{\ell(x)}(x)$ . For  $c \geq 1$ , let the function  $\sigma : X \rightarrow \mathbb{R}_{\geq 0}$  be defined as follows:

- Let  $A_1, A_2, \dots, A_s$  denote a partition of  $Y$  where each  $A_i$  is a  $c$ -coreset for  $(k, \infty)$ -line approximation of  $Y \setminus \left(\bigcup_{i'=1}^{i-1} A_{i'}\right)$ .
- For each  $x \in X$ , where  $y_x \in A_i$  we let

$$\sigma(x) \stackrel{\text{def}}{=} 2^{z-1} \cdot \frac{\|x - y_x\|_2^z}{\text{cost}_z^z(X, L)} + 2^{2z-1} \cdot \frac{c}{i}.$$

Then,  $\sigma$  is a sensitivity function for  $(k, z)$ -line approximation, and the total sensitivity

$$\mathfrak{S}_\sigma = O\left(2^{2z} \cdot c \cdot \log n \cdot \max_{i \in [s]} |A_i|\right)$$

*Proof.* Suppose  $x \in X$  and  $y_x \in A_i$ . Consider any set  $L' = \{\ell'_1, \dots, \ell'_k\}$  of  $k$  lines in  $\mathbb{R}^d$ . The goal is to show that

$$(7.30) \quad \frac{\min_{\ell' \in L'} \|x - \rho_{\ell'}(x)\|_2^z}{\text{cost}_z^z(X, L')} \leq \sigma(x).$$

We will first use Hölder inequality and the triangle inequality, as well as the fact that  $y_x \in A_i$  in order to write the following:

$$(7.31) \quad \begin{aligned} \min_{\ell' \in L'} \|x - \rho_{\ell'}(x)\|_2^z &\leq 2^{z-1} \|x - y_x\|_2^z + 2^{z-1} \min_{\ell' \in L'} \|y_x - \rho_{\ell'}(y_x)\|_2^z \\ &\leq 2^{z-1} \|x - y_x\|_2^z + 2^{z-1} \left( \text{cost}_z^z(Y, L') \cdot \frac{c}{i} \right). \end{aligned}$$

The justification for (7.31) is the following: for every  $j \leq i$ ,  $A_j$  is a  $c$ -coreset for  $(k, \infty)$ -line approximation of a set of points which contains  $y_x$ . Therefore, if  $\min_{\ell' \in L'} \|y_x - \rho_{\ell'}(y_x)\|_2 = r$ , there must exist a point  $u \in A_j$  with  $\min_{\ell' \in L'} \|u - \rho_{\ell'}(u)\|_2 \geq r/c$ . Suppose otherwise: every  $u \in A_j$  satisfies  $\min_{\ell' \in L'} \|u - \rho_{\ell'}(u)\|_2 < r/c$ . Then, the

$k$  cylinders of radius  $r/c$  contain  $A_j$ , so increasing the radius by a factor of  $c$  contains  $y_x$ . However, this means  $\min_{\ell' \in L'} \|y_x - \rho_{\ell'}(y_x)\|_2 < c \cdot r/c < r$ , which is a contradiction.

Hence, we always have that  $y_x \in A_i$  satisfies

$$\min_{\ell' \in L'} \|y_x - \rho_{\ell'}(y_x)\|_2^z \leq \text{cost}_z^z(Y, L') \cdot \frac{c}{i}.$$

Continuing on upper-bounding (7.31), we now use the fact  $\text{cost}_z^z(Y, L') \leq 2^{z-1} \cdot \text{cost}_z^z(X, L) + 2^{z-1} \cdot \text{cost}_z^z(X, L') \leq 2^z \text{cost}_z^z(X, L')$  because  $\text{cost}_z^z(X, L)$  is the optimal  $(k, z)$ -line approximation. Therefore,

$$\min_{\ell' \in L'} \|x - \rho_{\ell'}(x)\|_2^z \leq 2^{z-1} \cdot \|x - y_x\|_2^z + 2^{2z-1} \cdot \frac{c}{i} \cdot \text{cost}_z^z(X, L'),$$

so dividing by  $\text{cost}_z^z(X, L')$  and noticing that  $\text{cost}_z^z(X, L) \leq \text{cost}_z^z(X, L')$  implies  $\sigma$  is a sensitivity function.

The bound on total sensitivity then follows from

$$\mathfrak{S}_\sigma = \sum_{x \in X} \sigma(x) = \sum_{i=1}^s \sum_{x \in A_{s'}} \sigma(x) = 2^{z-1} + 2^{2z-1} \sum_{i=1}^s |A_i| \cdot \frac{c}{s'} = O\left(2^{2z} \cdot c \cdot \max_{i \in [s]} |A_i| \cdot \log n\right),$$

since  $s \leq n$ .  $\square$

**7.3.2 A simple coresets for  $(k, \infty)$ -line approximation of one-dimensional instances** Suppose first, that a dataset  $Y = \{y_1, \dots, y_n\}$  lies on a line in  $\mathbb{R}^d$ , and let  $C_1, \dots, C_k$  be a collection of  $k$  cylinders. Then, the intersection of the cylinders with the line results in a union of  $k$  intervals on the line. If we increase the radius of each cylinder  $C_1, \dots, C_k$  by a factor of  $c$ , the lengths of the intervals are scaled by factor of  $c$  (while keeping center of interval fixed). We first show that, for any  $Y = \{y_1, \dots, y_n\}$  which lie on a line, there exists a small subset  $Q \subset Y$  such that: if  $I_1, \dots, I_k$  is any collection of  $k$  intervals which covers  $Q$ , then increasing the length of each interval by a factor of 3 (while keeping the center of the interval fixed) covers  $Y$ .

**LEMMA 7.6.** *There exists a large enough constant  $c_1 \in \mathbb{R}_{\geq 0}$  such that the following is true. Let  $Y = \{y_1, \dots, y_n\}$  be a set of points lying on a line in  $\mathbb{R}^d$ , and  $k \in \mathbb{N}$ . There exists a subset  $Q \subset Y$  which is a 3-coreset for  $(k, \infty)$ -line approximation of size at most  $(c_1 \log n)^k$ .*

*Proof.* The construction is recursive. Let  $\ell$  be the line containing  $Y$ , and after choosing an arbitrary direction on  $\ell$ , let  $y_1, \dots, y_n$  be the points in sorted order according to the chosen direction.

The set  $Q$  is initially empty, and we include  $Q \leftarrow \{y_1, y_{\lceil n/2 \rceil}, y_n\}$ . Suppose that  $\|y_1 - y_{\lceil n/2 \rceil}\|_2 \geq \|y_{\lceil n/2 \rceil} - y_n\|_2$  (the construction is symmetric, with  $y_1$  and  $y_n$  switched otherwise). We divide  $Y$  into two sets, the subsets  $Y_L = \{y_1, \dots, y_{\lceil n/2 \rceil}\}$  and  $Y_R = \{y_{\lceil n/2 \rceil}, \dots, y_n\}$ . Then, we perform three recursive calls: (i) we let  $Q_1$  be a 3-coreset for  $(k, \infty)$ -line approximation of  $Y_L$ , (ii) we let  $Q_2$  be a 3-coreset for  $(k-1, \infty)$ -line approximation of  $Y_L$ , and (iii) we let  $Q_3$  be a 3-coreset for  $(k-1, \infty)$ -line approximation of  $Y_R$ . We add  $Q_1, Q_2$ , and  $Q_3$  to  $Q$ .

The proof of correctness argues as follows. Let  $C_1, \dots, C_k$  be an arbitrary collection of  $k$  cylinders which covers  $Q$ . The goal is to show that increasing the radius of  $C_1, \dots, C_k$  by a factor of 3 covers  $Y$ . Let  $I_1, \dots, I_k$  be the intervals given by  $I_i = \ell \cap C_i$ . We let the indices  $u, v \in [k]$  be such that  $I_u$  is the first interval which contains  $y_1$ , and  $I_v$  the last interval which contains  $y_n$ . We note that  $I_1 \cup \dots \cup I_k$  covers  $Q$ . We must show that if we increase the length of each interval by a factor of 3, we cover  $Y$ . We consider three cases:

- Suppose there exists an index  $i^* \in [k]$  such that  $y_1$  and  $y_{\lceil n/2 \rceil}$  both lie in the interval  $I_{i^*}$ . Recall  $\|y_1 - y_{\lceil n/2 \rceil}\|_2 \geq \|y_{\lceil n/2 \rceil} - y_n\|_2$ , and all points are contained within  $y_1$  and  $y_n$ . Hence, when we increase the length of  $I_{i^*}$  by a factor of 3 while keeping center fixed,  $y_1$  and  $y_n$  lie in the same interval, and thus cover  $Y$ .
- Suppose  $y_1$  and  $y_{\lceil n/2 \rceil}$  lie in different intervals, but there exists  $i^*$  such that  $y_{\lceil n/2 \rceil}$  and  $y_n$  lie in the interval  $I_{i^*}$ . Then, since all points of  $Y_R$  between  $y_{\lceil n/2 \rceil}$  and  $y_n$ ,  $I_{i^*}$  covers  $Y_R$ . Since  $I_1, \dots, I_k$  covers  $Q_1$  and  $Q_1$  is a 3-coreset for  $(k, \infty)$ -line approximation of  $Y_L$ , increasing the length of each interval by a factor of 3 covers  $Y_L$ , and therefore all of  $Y$ .

- Suppose  $y_1, y_{\lceil n/2 \rceil}$  and  $y_n$  all lie in different intervals. Then, since  $y_1$  and  $y_{\lceil n/2 \rceil}$  are not on the same interval, the  $k-1$  intervals  $\bigcup_{i \in [k] \setminus \{u\}} I_i$  covers  $Q_2$ . Similarly,  $y_{\lceil n/2 \rceil}$  and  $y_n$  are not on the same interval, so the  $k-1$  intervals  $\bigcup_{i \in [k] \setminus \{v\}} I_i$  covers  $Q_3$ . Since  $Q_2$  is a 3-coreset for  $(k-1, \infty)$ -line approximation of  $Y_L$ , increasing the radius of each interval by a factor of 3 covers all of  $Y_L$ . In addition,  $Q_3$  is a 3-coreset for  $(k-1, \infty)$ -line approximation of  $Y_R$ , so increasing length of intervals by a factor of 3 covers  $Y_R$ .

This concludes the correctness of the coreset, and it remains to upper bound the size. Let  $f(k, n) \in \mathbb{N}$  be an upper bound on the coreset size of  $(k, \infty)$ -line approximation of a subset of size  $n$ . We have  $f(1, n) = 2$ , since any single interval which covers  $y_1$  and  $y_n$  covers everything in between them. By our recursive construction, we have

$$f(k, n) \leq 3 + f(k, n/2) + 2 \cdot f(k-1, n/2).$$

By a simple induction, one can show  $f(k, n)$  is at most  $(c_1 \log n)^k$  when  $k \geq 2$ , for large enough constant  $c_1$  and large enough  $n$ .  $\square$

### 7.3.3 The coreset for points on $k$ lines and the effect of dimension reduction

LEMMA 7.7. *There exists a large enough constant  $c_1 \in \mathbb{R}_{\geq 0}$  such that the following is true. Let  $Y = \{y_1, \dots, y_n\}$  be a set of points lying on  $k$  lines in  $\mathbb{R}^d$ . There exists a subset  $Q \subset Y$  which satisfies the following two requirements:*

1.  $Q$  is a 3-coreset for  $(k, \infty)$ -line approximation of  $Y$  size at most  $k(c_1 \log n)^k$ .
2. If  $\Pi: \mathbb{R}^d \rightarrow \mathbb{R}^t$  is a linear map, then  $\Pi(Q)$  is a 3-coreset for  $(k, \infty)$ -line approximation of  $\Pi(Y)$ .

*Proof.* Let  $Y_1, \dots, Y_k$  be the partition of  $Y$  into points lying on the lines  $\ell_1, \dots, \ell_k$  of  $\mathbb{R}^d$ , respectively. We may write each line  $\ell_i$  by two vectors  $u_i, v_i \in \mathbb{R}^d$ , and have

$$\ell_i = \{u_i + t \cdot v_i : t \in \mathbb{R}\}.$$

Let  $Q_i$  be the 3-coreset for  $(k, \infty)$ -line approximation of  $Y_i$  specified by Lemma 7.6. We let  $Q$  be the union of all  $Q_i$ . Item 1 follows from Lemma 7.6 since we are taking the union of  $k$  coresets.

We now argue Item 2. Since  $\Pi$  is a linear map, and every point in  $Y_i$  lies on the line  $\ell_i$ , there exists a map  $t: Y_i \rightarrow \mathbb{R}$  where each  $y \in Y_i$  satisfies

$$y = u_i + t(y) \cdot v_i \in \mathbb{R}^d \quad \text{and thus,} \quad \Pi(y) = \Pi(u_i) + t(y) \cdot \Pi(v_i) \in \mathbb{R}^t.$$

In other words,  $\Pi(Y_i)$  lies within a line in  $\mathbb{R}^t$ . We note that the relative order of points in  $\Pi(Y_i)$  remains the same, since for any two points  $y, y' \in Y_i$ ,

$$\|\Pi(y) - \Pi(y')\|_2 = |t(y) - t(y')| \cdot \|\Pi(v_i)\|_2, \quad \|y - y'\|_2 = |t(y) - t(y')| \cdot \|v_i\|_2.$$

We note that the construction of Lemma 7.6 only considers the order of points in  $Y_i$ , as well as the ratio of distances. Therefore, executing the construction of Lemma 7.6 on the points  $\Pi(Y_i)$  returns the set  $\Pi(Q_i)$ .  $\square$

COROLLARY 7.2. *Let  $Y = \{y_1, \dots, y_n\} \subset \mathbb{R}^d$  be a set of points lying on  $k$  lines in  $\mathbb{R}^d$ .*

- Let  $A_1, \dots, A_s$  denote a partition of  $Y$  where each  $A_i$  is a 3-coreset for  $(k, \infty)$ -line approximation of  $Y$  from Lemma 7.7 on the set  $Y \setminus \bigcup_{i'=1}^{i-1} A_{i'}$ .
- Let  $\Pi: \mathbb{R}^d \rightarrow \mathbb{R}^t$  be any linear map.

*For any set of  $k$  lines  $L' = \{\ell'_1, \dots, \ell'_k\}$  in  $\mathbb{R}^t$ , if  $y \in A_i$ , we have*

$$\|\Pi(y) - \rho_{L'}(\Pi(y))\|_2 \leq \text{cost}_z(\Pi(Y), L') \cdot \frac{3}{i}.$$

*Proof.* The proof follows from applying the same observation of Lemma 7.5 to  $\Pi(A_j)$ , which is a 3-coreset for  $(k, \infty)$ -line approximation by Lemma 7.7. Namely, for every  $j \leq i$ , the set  $\Pi(A_j)$  is a 3-coreset for  $(k, \infty)$ -line approximation of a set containing  $y$ . Thus, if  $\|y - \rho_{L'}(y)\|_2 = r$ , there must be a set of at least  $i$  points  $y' \in Y$  where  $\|y' - \rho_{L'}(y')\|_2 \geq r/3$ .  $\square$

**7.4 Improving the approximation** We now instantiate the sensitivity function of Lemma 7.7 and use Corollary 7.1 and Lemma 7.4 in order to improve on the approximation. Similarly to before, we show that event  $\mathbf{E}_3(1+\varepsilon)$  occurs with sufficiently high probability over the draw of  $\Pi$  and  $(\mathbf{S}, \mathbf{w})$  by giving an upper bound on the variance as in Lemma 4.6, Lemma 5.7, and Lemma 6.6

Fix  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ , and let  $L = \{\ell_1, \dots, \ell_k\}$  be the optimal  $(k, z)$ -line approximation of  $X$  in  $\mathbb{R}^d$ . For  $x \in X$ , we let  $y_x \in \mathbb{R}^d$  be given by  $y_x = \rho_{\ell(x)}(x)$ , and we denote the set  $Y = \{y_x : x \in X\}$ . The sensitivity function  $\sigma: X \rightarrow \mathbb{R}_{\geq 0}$  is specified by Lemma 7.5. Recall that we first let  $A_1, \dots, A_s$  denote a partition  $Y$ , where  $A_i$  is the 3-coreset for  $(k, \infty)$ -line approximation of  $Y$  from Lemma 7.7. For  $x \in X$  with  $y_x \in A_i$ , we have

$$\sigma(x) = 2^{z-1} \cdot \frac{\|x - y_x\|_2^z}{\text{cost}_z^z(X, L)} + 2^{2z-1} \cdot \frac{3}{i}.$$

We let  $\mathbf{E}_4$  denote the following event with respect to the randomness in  $\Pi \sim \mathcal{J}_{d,t}$ . For each  $x \in X$ , we let  $\mathbf{D}_x \in \mathbb{R}_{\geq 0}$  denote the random variable

$$\mathbf{D}_x = \frac{\|\Pi(x) - \Pi(\rho_{\ell(x)}(x))\|_2}{\|x - \rho_{\ell(x)}(x)\|_2},$$

and as in previous sections, event  $\mathbf{E}_4$ , which occurs with probability at least 0.99, whenever

$$\sum_{x \in X} \mathbf{D}_x^{2z} \cdot \sigma(x) \leq 100 \cdot 2^z \cdot \mathfrak{S}_\sigma.$$

LEMMA 7.8. Let  $\Pi \in \mathcal{J}_{d,t}$  be a Johnson-Lindenstrauss map where, for  $\alpha > 1$ , the following events hold:

- Guarantee from Lemma 7.1:  $\sum_{x \in X} \|\Pi(x) - \Pi(\rho_{\ell(x)}(x))\|_2^z \leq \alpha \cdot \text{cost}_z^z(X, L)$ .
- Guarantee from Corollary 7.1: letting  $L' = \{\ell'_1, \dots, \ell'_k\}$  be the optimal  $(k, z)$ -line approximation fo  $\Pi(X)$ , then  $\text{cost}_z^z(X, L) \leq \alpha \cdot \text{cost}_z^z(\Pi(X), L')$ .
- Event  $\mathbf{E}_4$  holds.

Then, if we let  $(\mathbf{S}, \mathbf{w})$  denote  $m = \text{poly}((\log n)^k, 1/\varepsilon, \alpha)$ , i.i.d draws from  $\tilde{\sigma}$  and  $\mathbf{w}(x) = 1/(m\tilde{\sigma}(x))$ , with probability at least 0.99,

$$\text{cost}_z^z((\Pi(\mathbf{S}), \mathbf{w}), L') \leq (1 + \varepsilon) \cdot \text{cost}_z^z(\Pi(X), L').$$

*Proof.* We bound the variance,

$$(7.32) \quad \text{Vars}_{\mathbf{x} \sim \mathbf{S}} \left[ \mathbf{E}_{\mathbf{x} \sim \mathbf{S}} \left[ \frac{1}{\tilde{\sigma}(\mathbf{x})} \cdot \frac{\|\Pi(\mathbf{x}) - \rho_{L'}(\Pi(\mathbf{x}))\|_2^z}{\text{cost}_z^z(\Pi(X), L')} \right] \right] \leq \frac{\mathfrak{S}_\sigma}{m} \sum_{x \in X} \left( \frac{1}{\sigma(x)} \cdot \frac{\|\Pi(x) - \rho_{L'}(\Pi(x))\|_2^{2z}}{\text{cost}_z^{2z}(\Pi(X), L')} \right).$$

We note that, as before, we will apply Hölder's inequality and the triangle inequality, followed by Corollary 7.2. Specifically, suppose  $x \in X$  with  $y_x \in A_i$ , then,

$$\begin{aligned} \|\Pi(x) - \rho_{L'}(\Pi(x))\|_2^z &\leq 2^{z-1} \cdot \mathbf{D}_x^z \cdot \|x - \rho_L(x)\|_2^z + 2^{z-1} \cdot \|\Pi(y_x) - \rho_{L'}(\Pi(y_x))\|_2^z \\ &\leq 2^{z-1} \cdot \mathbf{D}_x^z \cdot \|x - \rho_L(x)\|_2^z + 2^{z-1} \cdot \text{cost}_z^z(\Pi(Y), L') \cdot \frac{3}{i} \\ &\leq 2^{z-1} \cdot \mathbf{D}_x^z \cdot \|x - \rho_L(x)\|_2^z + \frac{2^{2z-2} \cdot 3}{i} (\text{cost}_z^z(\Pi(X), L') + \text{cost}_z^z(\Pi(X), \Pi(Y))). \end{aligned}$$

We note that from (7.8) and (7.8), we have  $\text{cost}_z^z(\Pi(X), \Pi(Y)) \leq \alpha \text{cost}_z^z(X, L) \leq \alpha^2 \text{cost}_z^z(\Pi(X), L')$ . So the above simplifies to

$$\begin{aligned} \frac{\|\Pi(x) - \rho_{L'}(\Pi(x))\|_2^z}{\text{cost}_z^z(\Pi(X), L')} &\leq 2^{z-1} \cdot \mathbf{D}_x^z \cdot \frac{\|x - \rho_L(x)\|_2^z}{\text{cost}_z^z(X, L)} + 2^{2z-2} \cdot \frac{3(1 + \alpha^2)}{i} \\ &\leq (\mathbf{D}_x^z + 1 + \alpha^2) \cdot \sigma(x). \end{aligned}$$

We now continue upper bounding (7.32), where the variance becomes less than

$$\frac{\mathfrak{S}_\sigma}{m} \sum_{x \in X} (\mathbf{D}_x^z + 1 + \alpha^2)^2 \cdot \sigma(x) \lesssim \frac{\mathfrak{S}_\sigma^2 \cdot \alpha^4}{m},$$

since event  $\mathbf{E}_4$  holds. Since  $\mathfrak{S}_\sigma \leq \text{poly}(2^{2z}, (\log n)^k)$ , we obtain our desired bound on the variance by letting  $m$  be a large enough polynomial of  $(\log n)^k$ ,  $\alpha$ , and  $1/\varepsilon$ .  $\square$

**COROLLARY 7.3.** *Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be any set of points, and let  $L = \{\ell_1, \dots, \ell_k\}$  be the optimal set of  $k$  lines for  $(k, z)$ -line approximation of  $X$ . For any  $\varepsilon \in (0, 1/2)$ , let  $\mathcal{J}_{d,t}$  be the Johnson-Lindenstrauss map with*

$$t \gtrsim \frac{k \log \log n + z + \log(1/\varepsilon)}{\varepsilon^3}.$$

*Then, with probability at least 0.92 over the draw of  $\mathbf{\Pi} \sim \mathcal{J}_{d,t}$ ,*

$$\min_{\substack{L' \text{ } k \text{ lines} \\ \text{in } \mathbb{R}^t}} \text{cost}_z(\mathbf{\Pi}(X), L') \geq \frac{1}{(1 + \varepsilon)^{3+1/z}} \cdot \text{cost}_z(X, L).$$

## A On preserving “all solutions” and comparisons to prior work

This section is meant for two things:

1. To help compare the guarantees of this work to that of prior works on  $(k, z)$ -clustering of [40] and  $(k, 2)$ -subspace approximation [13], expanding on the discussion in the introduction. In short, for  $(k, z)$ -clustering, the results of [40] are qualitatively stronger than the results obtained here. In  $(k, 2)$ -subspace approximation, the “for all” guarantees of [13] are for the qualitatively different problem of low-rank approximation. While the costs of low-rank approximation and  $(k, 2)$ -subspace approximation happen to agree at the optimum, the notion of a candidate solution is different.
2. To show that, for two related problems of “medoid” and “column subset selection,” one cannot apply the Johnson-Lindenstrauss transform to dimension  $o(\log n)$  while preserving the cost. The medoid problem is a center-based clustering problem, and column subset selection problem is a subspace approximation problem. The instances we will construct for these problems are very symmetric, such that uniform sampling will give small coresets. These give concrete examples ruling out a theorem which directly relates the size of coresets to the effect of the Johnson-Lindenstrauss transform.

**Center-Based Clustering** Consider the following (slight) modification to the center-based clustering problems known as the “medoid” problem.

**DEFINITION A.1.** (1-MEDOID PROBLEM) *Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be any set of points. The 1-medoid problem asks to optimize*

$$\min_{c \in X} \sum_{x \in X} \|x - c\|_2^2.$$

Notice the difference between 1-medoid and 1-mean: in 1-medoid the center is restricted to be from within the set of points  $X$ , whereas in 1-mean the center is arbitrary. Perhaps surprisingly, this modification has a dramatic effect on dimension reduction.

**THEOREM A.1.** *For large enough  $n, d \in \mathbb{N}$ , there exists a set of points  $X \subset \mathbb{R}^d$  (in particular, given by the  $n$ -basis vector  $\{e_1, \dots, e_n\} \subset \mathbb{R}^n$ ) such that, with high probability over the draw of  $\mathbf{\Pi} \sim \mathcal{J}_{d,t}$  where  $t = o(\log n)$ ,*

$$\frac{\min_{c \in X} \sum_{x \in X} \|x - c\|_2^2}{\min_{c' \in \mathbf{\Pi}(X)} \sum_{x \in X} \|\mathbf{\Pi}(x) - c'\|_2^2} \geq 2 - o(1).$$



Theorem [A.1](#) gives very strong lower bound for dimension reduction for  $k$ -medoid, showing that decreasing the dimension to any  $o(\log n)$  does not preserve (even the optimal) solutions within better-than factor 2. This is in stark contrast to the results on center-based clustering, where the 1-mean problem can preserve the solutions up to  $(1 \pm \varepsilon)$ -approximation without any dependence on  $n$  or  $d$ . The proof itself is also very straight-forward: each  $\mathbf{\Pi}(e_i)$  is an independent Gaussian vector in  $\mathbb{R}^t$ , and if  $t = o(\log n)$ , with high probability, there exists an index  $i \in [n]$  where  $\|\mathbf{\Pi}(e_i)\|_2^2 = o(1)$ . In a similar vein, with high probability  $\sum_{i=1}^n \|\mathbf{\Pi}(e_i)\|_2^2 \leq (1 + o(1))n$ . We take a union bound and set the center  $c' = \mathbf{\Pi}(e_i)$  for the index  $i$  where  $\|\mathbf{\Pi}(e_i)\|_2^2 = o(1)$ . By the pythagorean theorem, the cost of this 1-medoid solution is at most  $(1 + o(1))n$ . On the other hand, every 1-medoid solution in  $X$  has cost  $2(n - 1)$ .

We emphasize that Theorem [A.1](#) does not contradict [\[40, 6\]](#), even though it rules out that “all candidate centers” are preserved. The reason is that the notion of “candidate solution” is different. Informally, [\[40\]](#) shows that for any dataset  $X \subset \mathbb{R}^d$  of  $n$  vectors and any  $k \in \mathbb{N}, \varepsilon > 0$ , applying the Johnson-Lindenstrauss map  $\mathbf{\Pi} \sim \mathcal{J}_{d,t}$  with  $t = O(\log(k/\varepsilon)/\varepsilon^2)$  satisfies the following guarantee: for all partitions of  $X$  into  $k$  sets,  $(P_1, P_2, \dots, P_k)$ , the following is true:

$$\sum_{\ell=1}^k \min_{c'_\ell \in \mathbb{R}^t} \sum_{x \in P_\ell} \|\mathbf{\Pi}(x) - c'_\ell\|_2^2 \approx_{1 \pm \varepsilon} \sum_{\ell=1}^k \min_{c_\ell \in \mathbb{R}^d} \sum_{x \in P_\ell} \|x - c_\ell\|_2^2.$$

The “for all” quantifies over clusterings  $(P_1, \dots, P_k)$  is different (as seen from the 1-medoid example) from the “for all” over centers  $c_1, \dots, c_k$ .

**Subspace Approximation** The same subtlety appears in subspace approximation. Here, we can consider the 1-column subset selection problem, which at a high level, is the medoid version of subspace approximation. We want to approximate a set of points by their projections onto the subspace spanned by one of those points.

**DEFINITION A.2. (1-COLUMN SUBSET SELECTION)** *Let  $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$  be any set of points. The 1-column subset selection problem asks to optimize*

$$\min_{\substack{S = \text{span}(\{x_i\}) \\ x_i \in X}} \sum_{x \in X} \|x - \rho_S(x)\|_2^2$$

Again, notice the difference between 1-column subset selection and  $(k, 1)$ -subspace approximation: the subspace  $S$  is restricted to be in the span of a point from  $X$ . Given Theorem [A.1](#) it is not surprising that Johnson-Lindenstrauss cannot reduce the dimension of 1-column subset selection to  $o(\log n)$  without incurring high distortions.

**THEOREM A.2.** *For large enough  $n, d \in \mathbb{N}$ , there exists a set of points  $X \subset \mathbb{R}^d$  such that, with high probability over the draw of  $\mathbf{\Pi} \sim \mathcal{J}_{d,t}$  where  $t = o(\log n)$ ,*

$$\frac{\min_{\substack{S = \text{span}(x) \\ x \in X}} \sum_{x \in X} \|x - \rho_S(x)\|_2^2}{\min_{\substack{S' = \text{span}(\mathbf{\Pi}(x)) \\ x \in X}} \sum_{x \in X} \|\mathbf{\Pi}(x) - \rho_{S'}(\mathbf{\Pi}(x))\|_2^2} \geq 3/2 - o(1).$$

The proof is slightly more involved. The instance sets  $d = n + 1$ , and sets  $X = \{x_1, \dots, x_n\}$  where  $x_i = (e_{n+1} + e_i)/\sqrt{2}$ . For any subspace  $S$  spanned by any of the points  $x_i$ , via a straight-forward calculation, the distance between  $x_j$  and  $\rho_S(x_j)$  is  $\sqrt{3}/4$  when  $j \neq i$ , and therefore, the cost of 1-column subset selection in  $X$  is  $3/4 \cdot (n - 1)$ . We apply dimension reduction to  $t = o(\log n)$  and we think of  $\mathbf{g}_1, \dots, \mathbf{g}_{n+1} \in \mathbb{R}^t$  as the independent Gaussian vectors given by  $\mathbf{\Pi}(e_1), \dots, \mathbf{\Pi}(e_{n+1})$ . As in the 1-medoid case, there exists an index  $i \in [n]$  for which  $\|\mathbf{g}_i\|_2^2 = o(1)$ , and notice that when this occurs,  $\mathbf{\Pi}(x_i)$  is essentially  $\mathbf{g}_{n+1}/\sqrt{2}$  (because  $\|\mathbf{\Pi}(x_i) - \mathbf{g}_{n+1}/\sqrt{2}\|_2 = o(1)$ ). Letting  $S$  be the subspace spanned by  $\mathbf{\Pi}(x_i)$ , we get that the distance between the projection  $\|\mathbf{\Pi}(x_j) - \rho_S(\mathbf{\Pi}(x_j))\|_2^2$  is at most  $\|\mathbf{g}_j\|_2^2/2 + o(1)$ . This latter fact is because the subspace spanned by  $S$  is essentially spanned by  $\mathbf{g}_{n+1}$ . Therefore, the cost of the 1-column subset selection of  $\mathbf{\Pi}(X)$  is at most  $n/2(1 + o(1))$ .

As above, Theorem [A.2](#) does not contradict [13](#), even though it means that “all candidate *subspaces*” are preserved needs to be carefully considered. The notion of “candidate solutions” is different. In the matrix notation that [13](#) uses, the points in  $X$  are stacked into rows of an  $n \times d$  matrix (which we denote  $X$ ). A Johnson-Lindenstrauss map  $\Pi$  is represented by a  $d \times t$  matrix, and applying the map to every point in  $X$  corresponds to the operation  $X\Pi$  (which is now an  $n \times t$  matrix). [13](#) shows that if  $\Pi$  is sampled with  $t = O(k/\varepsilon^2)$ , the following occurs with high probability. For all rank- $k$  projection matrices  $P \in \mathbb{R}^{n \times n}$ , we have

$$\|X - PX\|_F^2 \approx_{1 \pm \varepsilon} \|X\Pi - PX\Pi\|_F^2.$$

Note that when we multiply the matrix  $X$  on the left-hand side by  $P$ , we are projecting the  $d$  columns of  $X$  to a  $k$ -dimensional subspace of  $\mathbb{R}^n$ . This is different from approximating all points in  $X$  with a  $k$ -dimensional subspace in  $\mathbb{R}^d$ , which would correspond to finding a rank- $k$  projection matrix  $S \in \mathbb{R}^{d \times d}$  and considering  $\|X - XS\|_F^2$ . In the matrix notation of [13](#), the dimension reduction result for  $(k, 2)$ -subspace approximation says that

$$(A.1) \quad \min_{\substack{S \in \mathbb{R}^{d \times d} \\ \text{rank-}k \\ \text{projection}}} \|X - XS\|_F^2 \approx_{1 \pm \varepsilon} \min_{\substack{S' \in \mathbb{R}^{t \times t} \\ \text{rank-}k \\ \text{projection}}} \|X\Pi - X\Pi S'\|_F^2.$$

At the optimal  $S \in \mathbb{R}^{d \times d}$  and the optimal  $P \in \mathbb{R}^{n \times n}$ , the costs coincide (a property which holds only for  $z = 2$ ). Thus, [13](#) implies [A.1](#), but it does not say that the cost of all subspaces of  $\mathbb{R}^d$  are preserved (as there is a type mismatch in the rank- $k$  projections on the left- and right-hand side of [A.1](#)).

## References

- [1] Panjak K Agarwal and Cecilia M Procopiuc. Approximation algorithms for projective clustering. *Journal of Algorithms*, 46(2):115–139, 2003.
- [2] Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. Geometric approximation via coresets. *Combinatorial and computational geometry*, 52(1):1–30, 2005.
- [3] Alexandr Andoni, Piotr Indyk, and Ilya Razenshteyn. Approximate nearest neighbor search in high dimensions. In *Proceedings of the International Congress of Mathematicians (ICM '2018)*, pages 3287–3318, 2018.
- [4] Mihai Badoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *Proceedings of the 34th ACM Symposium on the Theory of Computing (STOC '2002)*, 2002.
- [5] Daniel Baker, Vladimir Braverman, Lingxiao Huang, Shaofeng H-C Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering in graphs of bounded treewidth. In *Proceedings of the 37th International Conference on Machine Learning (ICML '2020)*, pages 569–579, 2020.
- [6] 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 51th ACM Symposium on the Theory of Computing (STOC '2019)*, 2019.
- [7] Christos Boutsidis, Anastasios Zouzias, and Petros Drineas. Random projections for  $k$ -means clustering. In *Proceedings of Advances in Neural Information Processing Systems 23 (NeurIPS '2010)*, 2010.
- [8] Vladimir Braverman, Vincent Cohen-Addad, H-C Shaofeng Jiang, Robert Krauthgamer, Chris Schwiegelshohn, Mads Bech Tofttrup, and Xuan Wu. The power of uniform sampling for coresets. In *Proceedings of the 63rd Annual IEEE Symposium on Foundations of Computer Science (FOCS '2022)*, 2022.
- [9] Vladimir Braverman, Dan Feldman, and Harry Lang. New frameworks for offline and streaming coreset constructions. *arXiv preprint arXiv:1612.00889*, 2016.
- [10] Ke Chen. On coresets for  $k$ -median and  $k$ -means clustering in metric and euclidean spaces and their applications. *SIAM Journal on Computing*, 39(3):923–947, 2009.
- [11] Xiaoyu Chen, Shaofeng H.-C. Jiang, and Robert Krauthgamer. Streaming euclidean max-cut: Dimension vs data reduction. In *Proceedings of the 55th ACM Symposium on the Theory of Computing (STOC '2023)*, pages 170–182, 2023.
- [12] Yeshwanth Cherapanamjeri and Jelani Nelson. Terminal embeddings in sublinear time. In *Proceedings of the 62nd Annual IEEE Symposium on Foundations of Computer Science (FOCS '2021)*, pages 1209–1216, 2021.
- [13] Michael B. Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Mădălina Persu. Dimensionality reduction for  $k$ -means clustering and low rank approximation. In *Proceedings of the 47th ACM Symposium on the Theory of Computing (STOC '2015)*, 2015.
- [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 ACM Symposium on the Theory of Computing (STOC '2022)*, 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 *Proceedings of Advances in Neural Information Processing Systems 35 (NeurIPS '2022)*, 2022.
- [16] Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. A new coreset framework for clustering. In *Proceedings of the 53rd ACM Symposium on the Theory of Computing (STOC '2021)*, pages 169–182, 2021.
- [17] Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. Deterministic clustering in high dimensional spaces: Sketches and approximation. In *Proceedings of the 64th Annual IEEE Symposium on Foundations of Computer Science (FOCS '2023)*, 2023.
- [18] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.
- [19] Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. *Theory of Computing*, 2(1):225–247, 2006.
- [20] Michael Edwards and Kasturi Varadarajan. No coreset, no cry: Ii. In *Proceedings of the 25th International Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS '2005)*, 2005.
- [21] Michael Elkin, Arnold Filter, and Ofer Neiman. Terminal embeddings. *Theoretical Computer Science*, 697:1–36, 2017.
- [22] Dan Feldman. Introduction to core-sets: an updated survey. *arXiv preprint arXiv:2011.09384*, 2020.
- [23] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd ACM Symposium on the Theory of Computing (STOC '2011)*, pages 569–578, 2011.
- [24] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: constant-size coresets for  $k$ -means, pca and projective clustering. In *Proceedings of the 24th ACM-SIAM Symposium on Discrete Algorithms (SODA '2013)*, pages 1434–1453, 2013.
- [25] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: constant-size coresets for  $k$ -means, pca, and projective clustering. *SIAM Journal on Computing*, 49(3):601–657, 2020.
- [26] Sarel Har-Peled. No, coreset, no cry. In *Proceedings of the 24th International Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS '2004)*, 2004.
- [27] Sarel Har-Peled, Piotr Indyk, and Rajeev Motwani. Approximate nearest neighbor: Towards removing the curse of dimensionality. *Theory of Computing*, 8(1):321–350, 2012.
- [28] Sarel Har-Peled and Soham Mazumdar. On coresets for  $k$ -means and  $k$ -median clustering. In *Proceedings of the 36th ACM Symposium on the Theory of Computing (STOC '2004)*, pages 291–300, 2004.
- [29] Sarel Har-Peled and Kasturi Varadarajan. Projective clustering in high-dimensions using core-sets. In *Proceedings of the 18th ACM Symposium on Computational Geometry (SoCG '2002)*, 2002.
- [30] Lingxiao Huang, Jian Li, and Xuan Wu. On optimal coreset construction for euclidean  $(k, z)$ -clustering. In *Proceedings of the 56th ACM Symposium on the Theory of Computing (STOC '2024)*, 2024.
- [31] Lingxiao Huang and Nisheeth K. Vishnoi. Coresets for clustering in euclidean spaces: importance sampling is nearly optimal. In *Proceedings of the 52nd ACM Symposium on the Theory of Computing (STOC '2020)*, pages 1416–1429, 2020.
- [32] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the 30th ACM Symposium on the Theory of Computing (STOC '1998)*, pages 604–613, 1998.
- [33] Zachary Izzo, Sandeep Silwal, and Samson Zhou. Dimensionality reduction for wasserstein barycenter. In *Proceedings of Advances in Neural Information Processing Systems 34 (NeurIPS '2021)*, 2021.
- [34] Shaofeng H.-C. Jiang, Robert Krauthgamer, and Shay Sapir. Moderate dimension reduction for  $k$ -center clustering. In *Proceedings of the 40th International Symposium on Computational Geometry (SoCG '2024)*, 2024.
- [35] William Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemporary Mathematics*, pages 189–206, 1984.
- [36] Michael Kerber and Sharath Raghvendra. Approximation and streaming algorithms for projective clustering via random projections. In *Proceedings of the 27th Canadian Conference on Computational Geometry (CCCG '2015)*, 2015.
- [37] Michael Langberg and Leonard J. Schulman. Universal epsilon-approximators for integrals. In *Proceedings of the 21st ACM-SIAM Symposium on Discrete Algorithms (SODA '2010)*, 2010.
- [38] Sepideh Mahabadi, Konstantin Makarychev, Yuri Makarychev, and Ilya Razenshteyn. Nonlinear dimension reduction via outer bi-lipschitz embeddings. In *Proceedings of the 50th ACM Symposium on the Theory of Computing (STOC '2018)*, pages 1088–1101, 2018.
- [39] Michael W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.
- [40] Konstantin Makarychev, Yuri Makarychev, and Ilya Razenshteyn. Performance of johnson-lindenstrauss transform for  $k$ -means and  $k$ -medians clustering. In *Proceedings of the 51th ACM Symposium on the Theory of Computing*

- (*STOC '2019*), 2019.
- [41] 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 '2021)*, 2021.
  - [42] Shyam Narayanan and Jelani Nelson. Optimal terminal dimensionality reduction in euclidean space. In *Proceedings of the 51th ACM Symposium on the Theory of Computing (STOC '2019)*, pages 1064–1069, 2019.
  - [43] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS '2006)*, 2006.
  - [44] Nariankadu D. Shyamalkumar and Kasturi Varadarajan. Efficient subspace approximation algorithms. *Discrete and Computational Geometry*, 47(1):44–63, 2012.
  - [45] Nariankadu D. Shyamalkumar and Kasturi Varadarajan. Efficient subspace approximation algorithms. In *Proceedings of the 18th ACM-SIAM Symposium on Discrete Algorithms (SODA '2007)*, 2007.
  - [46] Christian Sohler and David Woodruff. Strong coresets for  $k$ -median and subspace approximation. In *Proceedings of the 59th Annual IEEE Symposium on Foundations of Computer Science (FOCS '2018)*, pages 802–813, 2018.
  - [47] Murad Tukan, Xuan Wu, Samson Zhou, Vladimir Braverman, and Dan Feldman. New coresets for projective clustering and applications. *arXiv preprint arXiv:2203.04370*, 2022.
  - [48] Kasturi Varadarajan and Xin Xiao. A near-linear algorithm for projective clustering of integer points. In *Proceedings of the 33rd ACM-SIAM Symposium on Discrete Algorithms (SODA '2012)*, 2012.
  - [49] Kasturi Varadarajan and Xin Xiao. On the sensitivity of shape fitting problems. In *Proceedings of the 32nd International Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS '2012)*, 2012.
  - [50] David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.