Oblivious Dimension Reduction for k-Means – Beyond Subspaces and the Johnson-Lindenstrauss Lemma [*]

Luca Becchetti[1], Marc Bury[2], Vincent Cohen-Addad[3], Fabrizio Grandoni[4], and Chris Schwiegelshohn[1]

[1]Sapienza, University of Rome, Italy
[2]Zalando SE, Germany
[3]CNRS, Paris, France
[4]IDSIA, USI-SUPSI, Switzerland

Abstract

We show that for n points in d-dimensional Euclidean space, a data oblivious random projection of the columns onto $m \in O\left(\frac{\log k + \log\log n}{\varepsilon^6} \log \frac{1}{\varepsilon}\right)$ dimensions is sufficient to approximate the cost of all k-means clusterings up to a multiplicative $(1 \pm \varepsilon)$ factor. The previous-best upper bounds on m are $O(\frac{\log n}{\varepsilon^2})$ given by a direct application of the Johnson-Lindenstrauss Lemma, and $O(\frac{k}{\varepsilon^2})$ given by [Cohen et al.-STOC'15].

We also prove the existence of a non-oblivious cost preserving sketch with target dimension $O\left(\frac{\log k}{\varepsilon^4} \log \frac{1}{\varepsilon}\right)$, improving on $\lceil \frac{k}{\varepsilon} \rceil$ [Cohen et al.-STOC'15]. Furthermore, we show how to construct strong coresets for the k-means problem of size $O(k \cdot \text{poly}(\log k, \varepsilon^{-1}))$. Previous constructions of strong coresets have size of order $k \cdot \min(d, k/\varepsilon)$.

Introduction

Random projections are a fundamental tool for dimensionality reduction, with numerous applications in streaming, compressed sensing, numerical linear algebra, graph sparsification, nearest neighbor search, privacy, and clustering. At a high level, we are given an input consisting of n vectors of dimension d, that induces a target set $\mathcal{X}$ of $n \times d$ matrices. We consider a proper distribution over random $d \times m$ matrices S (*sketching matrix*) such that, with good probability and for every $M \in \mathcal{X}$, the squared Frobenius norm[1] of M is approximately preserved after projection with S, i.e.

$$(1 - \varepsilon)\|M\|_F^2 \leq \|MS\|_F^2 \leq (1 + \varepsilon)\|M\|_F^2. \tag{1}$$

In this work, we consider the application of random projections to the problem of sketching the *(Euclidean) k-means objective*. Here, we are given n points $A_1, \ldots, A_n$ in d-dimensional Euclidean space (also represented as an $n \times d$ matrix A whose ith row is A_i). Our goal is to identify k points $c_1, \ldots, c_k$ (*centers*) so as to minimize the sum of squared distances of any point to the closest center, i.e. $\sum_{i=1}^{n} \min_{j=1}^{k} \|A_i - c_j\|^2$. In other words, the centers define a partition $C_1, \ldots, C_k$ of the points (*clustering*), where C_i contains the points whose closest center is c_i, and we wish to minimize $\sum_{j=1}^{k} \sum_{p \in C_j} \|p - c_j\|^2$. Every possible clustering is associated to exactly one $M \in \mathcal{X}$, where the rows of M correspond to the difference vector between each point and its associated center. Our goal is to determine the minimal target dimension m that achieves

---

[1]We recall that, for an $a \times b$ matrix M, the Frobenius norm of M is $\|M\|_F = \sqrt{\sum_{i=1}^{a} \sum_{j=1}^{b} M_{i,j}^2}$. For $b = 1$, this is equivalent to the Euclidean norm of a vector.

Eq. 1 for all $M \in \mathcal{X}$. Intuitively, this allows us to *reduce the dimension* of the problem from d to m, while approximately preserving its fundamental properties.

A particular interesting case is when the sketching matrix S is chosen in a data-oblivious way. In other words, the distribution over random matrices can be fixed a priori, without any knowledge of A. This makes random projections extremely useful in a number of areas such as streaming, distributed computing, and massively parallel frameworks like MapReduce.

There are two incomparable bounds for obliviously sketching the Euclidean k-means problem. The first bound of $O(\frac{\log n}{\varepsilon^2})$ is a direct consequence of the distributional Johnson-Lindenstrauss Lemma [38]. It states that when choosing $m \in O(\frac{\log n}{\varepsilon^2})$, all pairwise distances in A are approximately preserved. Hence the cost of every clustering is preserved via standard formulas, with probability arbitrarily close to 1. The other state of the art bound by Cohen et al. [24] yields $m \in O(\frac{k}{\varepsilon^2})$ and follows by sketching all orthogonal projections of rank k, of which k-means is a special case (see the related work for an overview).

A natural question is whether one can get a better dependence of m on k and $\log n$. For example, this was posed as an open question by Jelani Nelson during his plenary talk at HALG 2018. Cohen et al. [24] had given strong evidence that a better dependence might indeed be possible, showing that a projection onto $O\left(\frac{\log k}{\varepsilon^2}\right)$ dimensions provides (at most) a $(9 + \varepsilon)$ approximation. Our main result is as follows.

> For the Euclidean k-means problem, an oblivious random projection onto $O\left(\frac{\log k + \log \log n}{\varepsilon^6} \log \frac{1}{\varepsilon}\right)$ dimensions preserves the cost of any k-clustering up to a multiplicative $(1 \pm \varepsilon)$ factor.

Hence we improve the current state-of-the-art bounds for the range of $k \in \omega(\log \log n) \cap o(n)$. In the regime $k \in \Theta(\log n)$, where previous methods [24, 38] achieved the same bounds, the above bound is an exponential improvement (for constant ε).

While we consider the data-oblivious bound the more important result from both a theoretical and applied point of view, we are also able to show that using our techniques, there exists a data-dependent random projection onto $O\left(\frac{\log k}{\varepsilon^4} \log \frac{1}{\varepsilon}\right)$ dimensions with a multiplicative $(1 \pm \varepsilon)$ distortion. This improves on the $\lceil \frac{k}{\varepsilon} \rceil$ bound by Cohen et al. [24]. This latter bound has additional applications. For instance, combining our techniques with recent work on terminal embeddings [49] allows us to prove the existence of coresets for the k-means problem of size $O(k \cdot \text{poly}(\log k, \varepsilon^{-1}))$. Previously, only coresets of order $k \cdot \min(k, d)$ were known, see Table 1 in the appendix.

Related Work

There are three basic techniques used for linear dimension reduction: random projections, principal component analysis (PCA), or feature selection (sometimes called column selection).

Random Projections for Subspace Approximation and k-Means Following a tremendous amount of activity over the past decade, we now know that random projections achieve optimal dimension reduction in a number of regimes [6, 7, 37, 44, 45, 53]. They are also the only known method for oblivious dimension reduction. The fact that they do not depend on the data often makes them substantially quicker to apply especially when using sparse constructions. A folklore application of the Johnson-Lindenstrauss Lemma (see Lemma 2.6) states that a random projection onto $O(\varepsilon^{-2} \log n)$ dimensions also preserves the cost of any k-means clustering. Other than this result, most of the work has focused on low-rank approximation. Here, we are interested in computing a matrix A' of rank at most k such that $\|A - A'\|_F$ is minimized. The solution to this problem can be expressed analytically via the singular value decomposition $A := U \Sigma V^T$, where $U \in \mathbb{R}^{n \times d}$ has orthogonal columns[2], $V \in \mathbb{R}^{d \times d}$ has orthogonal rows, and Σ is a diagonal matrix where by convention $\Sigma_{1,1} \geq \Sigma_{2,2} \geq \ldots, \Sigma_{d,d} \geq 0$ the

---

[2] the inner product of any two columns of U is 0 and the Euclidean norm of any column of U is 1.

entries are non-increasingly ordered. It is well known that $A' = U_k \Sigma_k V_k^T$ is the optimal rank k approximation of A, where U_k is obtained by selecting only the first k columns of U, V_k is obtained by selection the first k rows of V and Σ_k is obtained by setting $\Sigma_{i,i} = 0$ for $i > k$. Using standard facts from linear algebra, we can alternatively express $A' = U_k U_k^T A$. In other words, computing the optimal solution $\min_{\text{rank } kA'} \|A - A'\|_F^2$ is equivalent to computing the optimal solution $\min_{\text{rank } k \text{ projection } X} \|A - XX^T A\|_F^2$. The problem is closely related to k-means, as the centroids of any k-means clustering are given via a specific rank k projection (see Section 2).

Sarlos [54] gave a data dependent random projection that achieves a $(1 \pm \varepsilon)$ approximation to the optimal rank k projection in $O(k \log k + k\varepsilon^{-1})$ dimensions and an oblivious projection onto $O(k \log k\varepsilon^{-2})$ dimensions[3]. Clarkson and Woodruff [21] improved this to $O(k\varepsilon^{-2})$ with a matching lower bound for oblivious methods later shown by Nelson and Nguyen [53]. The first paper to explicitly consider random projections for the k-means problem was authored Boutsides et al. [14], who showed that an oblivious random projection onto $O(k\varepsilon^{-2})$ dimensions preserved the k-means cost up to a $2 + \varepsilon$ factor. Cohen et al. [24] considerably improved this, showing that *all* rank k projections (i.e. in particular also k-means) are preserved up to $(1 \pm \varepsilon)$ factors by an oblivious projection onto $O(k/\varepsilon^2)$ dimensions. They also gave evidence that going below k and $\log n$ dimensions is possible for k-means, achieving a $9 + \varepsilon$ approximation in $O(\varepsilon^{-2} \log k)$ dimensions.

We note that there exists a vast amount of literature optimizing between running times, target dimension, and sparsity of random projection matrices. Some of these constructions could also be used for our result, at the cost of a slightly larger target dimension (up to $\text{polylog}(m)$ factors). The interested reader is referred to [3, 4, 5, 11, 20, 22, 23, 25, 28, 40, 48, 51, 52].

PCA Principal component analysis is arguably the most widely used form of dimension reduction. Not only does PCA reduce the intrinsic dimension of the point set, it also removes a substantial amount of noise. Indeed, this feature is the main reason why PCA is routinely used for learning, see [2, 8, 18, 27, 41, 42, 56]. Drineas et al. [29] were the first to apply PCA to the k-means problem, showing that a projection onto the first k components preserves the cost up to a factor of 2. This was substantially improved by Feldman et al. [33] and Cohen et al. [24], and we now know that a $(1 + \varepsilon)$ approximation is possible by projecting onto the first $\lceil k/\varepsilon \rceil$ components. Cohen et al. [24] also showed that this bound is tight, i.e. PCA *cannot* achieve a target dimension below k for k-means. Recently, Sohler and Woodruff [55] extended PCA-based methods for arbitrary powers of Euclidean distances such as k-median.

Column Selection The last technique we wish to survey are column selection methods. This form of dimension reduction has the advantage of being faster to compute than PCA and retaining the features and in particular the sparsity of the data set, at the cost of a slightly worse target dimension, see [10, 12, 13, 15, 26], with the current state of the art of $O(k/\varepsilon^2)$ columns being due to Cohen et al. [24].

Organization. The rest of this paper is organized as follows. In Section 2 we introduce some preliminary notions and tools. In Section 3, we present our high level ideas and outline the proof strategy. In Section 4 we describe the cluster decomposition at the heart of our analysis, and in Section 5 we show how to use it to bound the target dimension m. All the proofs that are omitted in the main body are given in Appendix A. Our data-dependent dimension reduction result is given in Appendix B. We briefly remark on k-means coresets in Appendix C.

Preliminaries

We use $\|M\|_F^2 = (1 \pm \varepsilon)\|N\|_F^2$ as an abbreviation for $(1 - \varepsilon)\|N\|_F^2 \leq \|M\|_F^2 \leq (1 + \varepsilon)\|N\|_F^2$. We use the shorthand $[i] = \{1, \ldots i\}$ to refer to the natural numbers up to a positive integer i.

---

[3]We note that if we only want to preserve the a $(1 + \varepsilon)$-approximation to the optimal rank k projection, but not the cost, a random projection onto $O(k/\varepsilon)$ dimensions is sufficient and necessary. See [21, 54] for details.

For an $n \times d$ matrix A, view the ith row A_i for $i \in [n]$ as a point in d-dimensional Euclidean space. Sometimes we will refer to a point set or the associated data matrix interchangeably by A. We let $|B|$ denote the number of rows of a matrix B. We say the cost of a point set P is the value of the optimal 1-means clustering of P. The cost of a clustering $C = \{C_1, \ldots, C_k\}$ is the sum of the costs of the clusters in C. Throughout this paper, we will use OPT to denote the cost of an optimal k-means clustering. The optimal center of any cluster C_i is the centroid $\mu(C_i) = \frac{1}{|C_i|} \sum_{p \in C_i} p$. This can be easily seen due to the following lemma.

Lemma 2.1. *[folklore] For any set of points A and any point p, the following equations hold:*

- $\sum_{i=1}^{|A|} \|A_i - p\|^2 = \sum_{i=1}^{|A|} \|A_i - \mu(A)\|^2 + |A| \cdot \|p - \mu(A)\|^2$

- $\sum_{i=1}^{|A|} \sum_{j=1}^{|A|} \|A_i - A_j\|^2 = 2|A| \cdot \sum_{i=1}^{|A|} \|A_i - \mu(A)\|^2$

This lemma enables us to express the k-means problem algebraically. We define the $n \times k$ clustering matrix X (using the shorthand rank k c.m. X) with entries

$$
X_{i,j} = \begin{cases} \frac{1}{\sqrt{|C_j|}} & \text{if } A_i \in C_j \\ 0 & \text{otherwise.} \end{cases}
$$

Three properties are of interest. First, every column has unit Euclidean norm. Second, the columns are pairwise orthogonal. Third, $XX^T A$ maps the ith row of A to the centroid of cluster C_j. Thus, we can express the k-means objective as

$$
\min_{\text{rank } k \text{ clustering matrix } X} \|A - XX^T A\|_F^2.
$$

We note that if we lift the constraint that X is a clustering matrix and instead require X only to be orthogonal, the problem is then known as the low rank subspace approximation problem. The best rank 1-clustering matrix that maps the rows of A to the centroid $\mu(A)$ is known as the center matrix. The resulting matrix of centroids can interchangeably be expressed as $\frac{1}{|A|} 11^T A$ and $1\mu(A)^T$, where 1 is the all 1 vector of appropriate dimension. We note that this operation is invariant when sketching the rows of A, i.e. $\mu(A)^T S = \mu(AS)^T$.

We will aim at proving the following guarantee similar to those proposed in earlier work by Feldman et al. [33] and Cohen et al. [24].

Definition 2.2 ((ε, k)-Means Cost Preserving Sketches)**.** *Let A be an $n \times d$ matrix corresponding to n points in d dimensional space. Let $c \geq 0$ be some fixed constant possibly depending on A. Then an $n \times m$ matrix B is an (ε, k)-means cost preserving sketch of A with offset κ if for any rank k clustering matrix X*

$$
(1 - \varepsilon) \cdot \|A - XX^T A\|_F^2 \leq \|B - XX^T B\|_F^2 + \kappa \leq (1 + \varepsilon)\|A - XX^T A\|_F^2.
$$

If $B := AS$ for some $d \times m$ matrix S sampled from a distribution $\mathcal{D}$ that does not depend on A, we say B is an oblivious *sketch.*

If $\kappa = 0$ (which is the case in Theorem 5.3), we will simply call B a (ε, k)-means cost preserving sketch. We next state a few useful and standard properties (proof in the appendix) of clustering matrices that we will extensively use throughout this paper.

Lemma 2.3. *Let A be a matrix, and let X and Y be clustering matrices of A. Then:*

1. $\|A - XX^T A\|_F^2 = \|A - 1\mu(A)^T\|_F^2 - \|XX^T A - 1\mu(A)^T\|_F^2 \leq \|A - 1\mu(A)^T\|_F^2.$

2. *If Y is a* refinement *of X, i.e., every cluster induced by Y is a subcluster of a cluster induced by X, then $XX^TYY^T = XX^T$ and*

$$\|A - XX^TA\|_F^2 = \|A - YY^TA\|_F^2 + \|YY^TA - XX^TYY^TA\|_F^2.$$

3. $\|XX^TA\|_F^2 \leq \|A\|_F^2.$

We will use the following approximate triangle inequality for squared Euclidean spaces. Similar statements can be found throughout k-means and coreset literature (see e.g. [19, 27, 33]).

Lemma 2.4 (Approximate Triangle Inequality). *For any $\varepsilon > 0$, and matrices A, B, C of equal dimension, we have $|\|A - C\|_F^2 - \|B - C\|_F^2| \leq \left(1 + \frac{1}{\varepsilon}\right)\|A - B\|_F^2 + \varepsilon\|A - C\|_F^2.$*

Random Projections. We require our sketching matrices to have the following bounds.

Theorem 2.5 (Distributional Johnson-Lindenstrauss Lemma, Theorem 2.4 [21]). *Let A be an $n \times d$ matrix. Then there exists a distribution $\mathcal{D}$ over linear mappings in $\mathbb{R}^{d \times m}$ such that for $S \sim \mathcal{D}$ and $m \in O\left(\frac{\log 1/\delta}{\varepsilon^2}\right)$, with probability $1 - \delta$,*

$$(1 - \varepsilon)\|A\|_F^2 \leq \|AS\|_F^2 \leq (1 + \varepsilon)\|A\|_F^2$$

Possible realizations of these matrices are random Gaussian matrices, dense Rademacher matrices [1], or sparse Rademacher matrices designed in [40]. Faster methods can sometimes also be used, at the cost of a slightly larger target dimension. The interested reader is referred to [3, 4, 5, 11, 23]. The super sparse embedding matrices of [22, 48, 51] *cannot* be used.

We note that this theorem already allows us to preserve the 1-means cost with target dimension $m \in O(\frac{\log 1/\delta}{\varepsilon^2})$ by applying it to the matrix $A - \mathbf{1}\mu(A)^T$. This theorem also gives rise to the following two simple, but very useful statements. The first is essentially the application of the Johnson-Lindenstrauss Lemma to the k-means problem.

Lemma 2.6. *Let A be an $n \times d$ matrix with ℓ distinct rows. Then there exists a distribution $\mathcal{D}$ over linear mappings in $\mathbb{R}^{d \times m}$ such that for $S \sim \mathcal{D}$ and $m \in O\left(\frac{\log(\ell/\delta)}{\varepsilon^2}\right)$, AS is an (ε, k)-means cost preserving sketch with probability $1 - \delta$.*

Generally, we cannot expect A to have any less than n distinct rows, therefore a naive application of the aforementioned lemma requires a target dimension of $\Omega(\log n)$. We will apply this lemma to YY^TA, where Y is a clustering matrix of rank $k' \ll n$. By definition of clustering matrices, YY^TA can have at most k' distinct rows.

The second immediate implication of Theorem 2.5 shows that even for an unbounded number of distinct rows most of the distances will be approximately preserved. We will use this lemma essentially to subsample the set of pairwise distances whenever a uniform subsample is sufficient to preserve the entire cost.

Lemma 2.7. *Let A be an $n \times d$ matrix. There exists a distribution $\mathcal{D}$ over linear mappings in $\mathbb{R}^{d \times m}$ such that for $S \sim \mathcal{D}$ and $m \in O\left(\frac{\log \frac{1}{\zeta\delta}}{\varepsilon^2}\right)$, with probability $1 - \delta$ at least a $(1 - \zeta)$-fraction of the pairs $A_i, A_j, i \neq j$ are such that $\|A_iS - A_jS\|^2 = (1 \pm \varepsilon)\|A_i - A_j\|^2.$*

Our Techniques

Most upper bounds for random projections apply the following basic proof scheme:

1. The Euclidean (resp. Frobenius) norm of any fixed vector (resp. matrix) is preserved up to a factor $(1 \pm \varepsilon)$ with probability $1 - \delta$ if the target dimension is $O\left(\frac{\log 1/\delta}{\varepsilon^2}\right)$.

2. Identify a vector set P such that preserving the norm of each vector in P preserves the cost of the entire problem. Then apply a union bound by setting $\delta = \frac{1}{|P|}$.

The first step is tight [37, 39], so all improvements on the target dimension are a result of sophisticated applications of the union bound. For k-means there are two known ways to proceed. First, it is well known that k-means is a (constrained) subspace approximation problem (see Section 2). As mentioned above, this approach can never achieve a bound better than $\Omega(k)$, meaning that specific properties of the k-means problem have to be exploited.

For the $\log n$ target dimension, we proceed in such a manner (Lemma 2.6). The famous general Johnson-Lindenstrauss Lemma [38] states that the pairwise distances between any set of n points are preserved by a random projection onto $O(\varepsilon^{-2} \log n)$ dimensions, which is also optimal [45]. It is well known that the centroid of a set of points is the optimal 1-means solution. A consequence of this fact is that the cost of clustering all points to the centroid is a non-negative linear combination of the pairwise distances of the points, see Lemma 2.1. Hence, the Johnson-Lindenstrauss Lemma also preserves the cost of k-means. A careful application of this lemma is the basis of both the previous result by Cohen et al. [24] and our work.

We first review the $(9 + \varepsilon)$ approximation by Cohen et al. [24]. Let A be the $n \times d$ matrix corresponding to n points in d dimensions. They condition on the fact that distances between centroids of the optimal k-means clustering are preserved, which, using Lemma 2.6, requires only $O(\frac{\log k}{\varepsilon^2})$ dimensions. The matrix of centroids is obtained by a suitable rank k clustering matrix X applied to A. For any given clustering matrix Y, they then consider the cost $\|A - YY^T A\|_F^2$ in terms of the image $\|XX^T A - YY^T XX^T A\|_F^2$ and the kernel $\|(I - XX^T)A - YY^T(I - XX^T)A\|_F^2$ of $XX^T A$, respectively. The cost of clustering in the former space is preserved, i.e.

$$\|(XX^T A - YY^T XX^T A)S\|_F^2 = (1 \pm \varepsilon) \cdot \|\|XX^T A - YY^T XX^T A\|_F^2.$$

The cost within the latter space can be upper bounded by

$$\|((I - XX^T)A - YY^T(I - XX^T)A)S\|_F^2 \leq \|(I - XX^T)AS\|_F^2 \leq (1 + \varepsilon) \cdot OPT.$$

Unfortunately, the cost cannot be decomposed in terms of image and kernel, i.e. $\|(A - YY^T A)S\|_F^2 = \|(XX^T A - YY^T XX^T A)S\|_F^2 + \|((I - XX^T)A - YY^T(I - XX^T)A)S\|$ does not hold in general. Here, Cohen et al. [24] apply a weaker form of the triangle inequality in squared Euclidean spaces, which with some calculation leads to a $9 + \varepsilon$ approximation. The analysis by Cohen et al. [24] is tight for the choice of distance vectors P they preserve, so we require a number of additional ideas.

The loss in approximation is mainly due to the Frobenius norm of the kernel $(I - XX^T)A$. Therefore, one might be tempted to decrease it by adding additional centers (see also Feldman et al. [33] for a similar idea for Bregman divergences). For every cluster C_i induced by X, we have two conditions.

a. Either the cost of C_i may be decreased. If this is the case, we decrease the cost until the kernel has norm less than $\varepsilon^2 \cdot \|(I - XX^T)A\|_F^2$, in which case the error incurred by applying the triangle inequality is relative to $O(\varepsilon) \cdot \|(I - XX^T)A\|_F^2$.

b. The cost of C_i cannot be decreased, even when adding a substantial amount of centers (say $\text{poly}(k)$).

One may hope that for a C_i of the second type, we could find a different proof strategy to preserve the cost. Unfortunately, this does not seem to be substantially easier than the general case. However, if we additionally require the points in C_i to have (roughly) equal distance from the center, we are able to provide such proof. To illustrate the idea, let us consider the n-simplex. For every clustering $C = \{C_1, \ldots C_k\}$ of the simplex, the cost of the clustering is dominated by the clusters of largest cardinality. If we preserve the cost of *any* cluster of size,

say $\frac{\varepsilon}{k} \cdot n$, then we preserve the cost of any clustering. Here our argument deviates from previous applications of the union bound. The bound on the target dimension above also implies that a $(1 - \delta)$-fraction of the pairwise distances between the points is preserved (Lemma 2.7). By setting $\delta \leq \left(\frac{\varepsilon}{k}\right)^2$, and noting that the distances in the simplex are all identical, we can ensure that most of the intra-cluster pairwise distances are always preserved for the large clusters, which implies that the cost of the large clusters is preserved. Our analysis simply extends this illustration for the simplex to arbitrary clusters obeying a equidistance condition. To ensure that the equidistance condition within C_i holds, we use a cluster decomposition that first grows $\log n$ balls of exponentially increasing radii centered around the centroid of C_i. For each ring induced by the difference of two subsequent balls, we apply steps a. and b. above.

A Cluster Decomposition

We will first outline a recursive procedure to subdivide the point set into clusters with carefully determined properties, see also Algorithm 1. A similar construction without the equidistance property was previously proposed by Feldman et al. [33] in the context of producing coresets for clustering problems including (but not limited to) k-means.

This procedure repeatedly computes optimal k-means clusterings but is only used for analyzing random projections in Section 5.

We will compute a $O(k \log n)$-ary tree T of constant (depending only on ε^{-1}) depth. The root of the tree represents the entire point set A. The nodes of the tree correspond to point aubsets. The children of a node $A' \subset A$ are a clustering of A', i.e. in particular a partition of A'. The leaves of T will also form a partition of A. For each leaf corresponding to the point set $L \subset A$, we use $|L|$ copies of the centroid $\mu(L)$ as a representative of L.

In more detail, let α, β, γ be sufficiently small constants depending on ε. We process each leaf A' of the current tree (starting with A) as follows. Let $r_i := \frac{\gamma}{|A'|} \|A' - 1\mu(A')^T\|_F^2 \cdot 2^i$. Then the *rings* of A' induced by the r_i are $R^0(A') := \{p \in A' \mid \|p - \mu(A')\|^2 \leq r_0\}$ and $R^i(A') := \{p \in A' \mid r_{i-1} < \|p - \mu(A')\|^2 \leq r_i\}$ for $i \in [\log \frac{n}{\gamma}]$. We will only write R^i instead of $R^i(A')$ when the parent node A' is clear from context. We compute an optimal k-clustering for every ring. If the cost of such a clustering of $R^i(A')$ is smaller than a $\frac{1}{1+\alpha}$-factor compared to the cost of clustering these to $\mu(A')$, we append the clustering and continue the recursion. The recursion stops in the following three cases. For points in $R^0(A')$, we always stop. If the clustering in $R^i(A')$, $i > 0$, does not become cheaper by at least a $1 + \alpha$ factor, we also stop. Finally, we always stop if the depth of A' in the tree is $1/\beta + 2$, see also Algorithm 1. The stopping criteria are summarized with the following property.

Algorithm 1 Sketching Tree

1: Initialize an $O(k \log \frac{n}{\gamma})$-ary tree T with root A.
2: Compute an optimal k-means clustering $C = \{C_1, \ldots C_k\}$
3: Append clusters $C_i \in C$ as children of A in T and initialize queue $Q = C$
4: **while** $Q \neq \emptyset$ **do**
5:     $A' \leftarrow pop(Q)$
6:     **if** depth of A' in T is less than $1/\beta + 2$ **then**
7:        Partition A' into rings $\{R^0, \ldots R^{\log \frac{n}{\gamma}}\}$
8:        Append R^0 as a child of A' in T
9:        **for** each ring $R^i \neq R^0$ **do**
10:           Compute an optimal k-means clustering of R^i with clusters $K(R^i) = \{K_1, \ldots K_k\}$ and clustering matrix Z^i.
11:           Append clusters in $K(R^i)$ as children of A'
12:           **if** $\|R^i - Z^i(Z^i)^T R^i\|_F^2 \cdot (1 + \alpha) < \|R^i - 1\mu(A')^T\|_F^2$ **then**
13:              Add clusters of $K(R^i)$ to Q

Definition 4.1. *We define:*

1. *Let $\mathcal{L}_{low}$ be the leaves at depth $1/\beta + 2$.*

2. *Let $\mathcal{L}_{inner}$ be the points sets corresponding to rings R^0.*

3. *Let $\mathcal{L}_{exp}$ be the point sets corresponding to rings R^i, $i > 0$, for which we did not continue the recursion.*

The parent of a node L in T is denoted by $p(L)$.

Property 4.2. *For each ring R^0, we have $\|R^0 - 1\mu(R^0)^T\|_F^2 \leq \gamma\|p(R^0) - 1\mu(p(R^0))^T\|_F^2$. Furthermore, For each such R^i, $i > 0$, in $\mathcal{L}_{exp}$, we have $\|R^i - Z^i(Z^i)^T R^i\|_F^2 \cdot (1+\alpha) \geq \|R^i - 1\mu(p(R^i))^T\|_F^2$.*

The following bound on the size of the tree T is an immediate consequence of the stopping criterion of the algorithm, see also line 6 of Algorithm 1.

Observation 4.3. *T has at most $\sum_{i=1}^{1/\beta+2}(k(1+\log\frac{n}{\gamma}))^i \in (k\log\frac{n}{\gamma})^{O(1/\beta)}$ many nodes.*

We first show that the point sets obeying the first two properties have small cost.

Lemma 4.4. $$\sum_{L \in \mathcal{L}_{low} \cup \mathcal{L}_{inner}} \|L - 1\mu(L)^T\|_F^2 \leq \left(\gamma \cdot \frac{1+\alpha}{\alpha} + \left(\frac{1}{1+\alpha}\right)^{1/\beta}\right) \cdot OPT.$$

We define Y to be the clustering matrix induced by the leaves of T, note that the rank k' of Y is larger than k (of order $(k\log n/\gamma)^{O(1/\beta)}$). Let X be an arbitrary rank k clustering matrix. The next lemma will be used to show that under the conditions of Property 4.2 and using Lemma 4.4, the Frobenius norm of $XX^T(A - YY^TA)$ is bounded.

Lemma 4.5. *Let A be an $n \times d$ matrix and let $\alpha > 0$ be a constant. Let Y be a rank k' clustering matrix of A with clusters $C = \{C_1, \ldots C_{k'}\}$. Moreover:*

1. *Let C_{exp} be the subset of C containing all clusters satisfying $\|C_i - 1\mu(C_i)^T\|_F^2 \leq (1+\alpha) \cdot \|C_i - ZZ^TC_i\|_F^2$ for any rank k clustering matrix Z. Further, define $\Delta_{exp} := \alpha \cdot \sum_{C_i \in C_{exp}} \|C_i - 1\mu(C_i)^T\|_F^2$.*

2. *Let $C_{cheap} = C \setminus C_{exp}$. Define $\Delta_{cheap} := \sum_{C_i \in C_{cheap}} \|C_i - 1\mu(C_i)^T\|_F^2$.*

Then for any rank k clustering matrix X of A, we have

$$\|XX^T(I - YY^T)A\|_F^2 \leq \Delta_{exp} + \Delta_{cheap}.$$

Finally, we use the following bound on $\|YY^TA - XX^TYY^TA\|_F^2$. The proof and statement is very related to the $9+\varepsilon$ approximation used by Cohen et al. [24]. While we will later have tighter estimates, this will nevertheless be useful. Note that the assumption $\|A - YY^TA\|_F^2 \leq OPT$ is always satisfied, as our tree is initialized with the optimal k-means clustering.

Lemma 4.6. *Suppose $\|A - YY^TA\|_F^2 \leq OPT \leq \|A - XX^TA\|_F^2$.*
Then $\|YY^TA - XX^TYY^TA\|_F^2 \leq 9 \cdot \|A - XX^TA\|_F^2$.

Analysis of Oblivious Random Projections

Our main technical lemma now applies Lemma 2.7 to points sets with the properties satisfied by any leaf in $\mathcal{L}_{exp}$, see Property 4.2. In these cases, we do not require the full power of a union bound to show that the k-means cost is well approximated.

Lemma 5.1. *Let A be an $n \times d$ matrix, let p be a point and let α, ε be sufficiently small constants. Suppose that the following two conditions hold:*

1. *$\|A_i - p\|^2 \leq 2 \cdot \frac{1}{|A|}\|A - 1p^T\|_F^2$ for all $i \in [n]$.*

2. *$\|A - XX^TA\|_F^2(1 + \alpha) \geq \|A - 1p^T\|_F^2$ for all rank k cluster matrices X.*

Then there exists a distribution $\mathcal{D}$ over linear mappings in $\mathbb{R}^{d \times m}$ with $m \in O\left(\frac{\log k/(\varepsilon\delta)}{\varepsilon^2}\right)$ such that for $S \sim \mathcal{D}$ and for all rank k clustering matrices X of A, with probability at least $1 - \delta$,

$$\|AS - XX^TAS\|_F^2 \cdot (1 + \alpha + \varepsilon) \geq \|AS - 1p^TS\|_F^2 \geq \|AS - 1\mu(A)^TS\|_F^2.$$

The high level argument is as follows. Suppose that the rank k clustering matrix Z is the minimizer of $\|AS - ZZ^TAS\|_F^2$, and let $C = \{C_1, \ldots C_k\}$ be the set of clusters induced by Z. We consider C_i to be *cheap* if its cost for the original points in A was at most an $O\left(\frac{\varepsilon}{k}\right)$-fraction of $\|A - 1p^T\|_F^2$. C_i is considered to be *expensive* otherwise. The total contribution of the cheap clusters can be at most an $O(\varepsilon)$-fraction of $\|A - ZZ^TA\|_F^2$, so what remains to be shown is that the cost of the remaining clusters is lower bounded.

This is shown in the following lemma, where we prove that the cost of expensive clusters are always preserved. The crucial observation is that these clusters always contain many points.

Lemma 5.2. *Assume the conditions of Lemma 5.1 hold and suppose the target dimension m of a random projection S is in $O\left(\frac{\log 1/(\varepsilon\eta\delta)}{\varepsilon^2}\right)$, where $\eta, \varepsilon, \delta > 0$. Then with probability $1 - \delta$, for all set of points $P \subset A$, i.e. a subset of rows of A, satisfying*

$$\|P - 1\mu(P)^T\|_F^2 \geq \eta \cdot \|A - 1p^T\|_F^2, \tag{2}$$

we have

$$\|PS - 1\mu(P)^TS\|_F^2 \geq (1 - \varepsilon) \cdot \|P - 1\mu(P)^T\|_F^2.$$

Proof. We first upper bound the cost of P. Observe that, for $x, y \in P$, we deterministically have:

$$\|x - y\|^2 \leq 2\|x - p\|^2 + 2\|y - p\|^2 \overset{\substack{Ass.\ 1 \\ of\ Lemma\ 5.1}}{\leq} \frac{8}{|A|}\|A - 1p^T\|_F^2, \tag{3}$$

hence

$$\|P - 1\mu(P)^T\|_F^2 \overset{Lem.\ 2.1}{=} \frac{1}{2|P|}\sum_{x \in P}\sum_{y \in P}\|x - y\|^2 \overset{(3)}{\leq} |P|\frac{4}{|A|}\|A - 1p^T\|_F^2 \overset{(2)}{\Rightarrow} |P| \geq \frac{1}{4}\eta \cdot |A|. \tag{4}$$

Next, we set $\zeta < \frac{\varepsilon \cdot \eta^3}{8}$ and apply Lemma 2.7. It follows that all but a ζ-fraction of the pairwise distances in A are approximately preserved up to $(1 \pm \varepsilon)$ factors with probability $1 - \delta$, if $m \in O\left(\frac{\log\frac{1}{\zeta\delta}}{\varepsilon^2}\right) = O\left(\frac{\log\frac{1}{\varepsilon\eta\delta}}{\varepsilon^2}\right)$. For our choice of ζ, we then *deterministically* have

$$\zeta \cdot \binom{|A|}{2} \leq \zeta \cdot \frac{|A|^2}{2} \overset{(4)}{\leq} \zeta\frac{|P|^2 2}{\eta^2} \leq \frac{\varepsilon\eta}{4}|P|^2 \tag{5}$$

which implies that all but an $\varepsilon\eta/8$-fraction of the pairwise distances in P are preserved. Let $D_{good}(PS)$ and $D_{bad}(PS)$ be the set of pairs of points of P whose distances are preserved and

are not preserved, resp., up to a $(1 \pm \varepsilon)$ factor after projection. We lower bound the distances of $D_{bad}(PS)$ by 0. We have

$$\|PS - 1\mu(P)^T S\|_F^2 \overset{Lem.\ 2.1}{=} \frac{1}{2|P|} \sum_{x \in P} \sum_{y \in P} \|xS - yS\|^2$$

$$\geq \quad \frac{1}{2|P|} \sum_{(x,y) \in D_{good}(PS)} \|xS - yS\|^2 \overset{Lem.\ 2.7}{\geq} \frac{1-\varepsilon}{2|P|} \sum_{(x,y) \in D_{good}(PS)} \|x - y\|^2$$

$$\geq \quad \frac{1-\varepsilon}{2|P|} \left(\sum_{x \in P} \sum_{y \in P} \|x - y\|^2 - \sum_{(x,y) \in D_{bad}(PS)} \|x - y\|^2 \right)$$

$$\overset{(5)}{\geq} \quad \frac{1-\varepsilon}{2|P|} \left(\sum_{x \in P} \sum_{y \in P} \|x - y\|^2 - \frac{\varepsilon\eta}{4} |P|^2 \max_{x,y} \|x - y\|^2 \right)$$

$$\overset{(3)}{\geq} \quad \frac{1-\varepsilon}{2|P|} \left(\sum_{x \in P} \sum_{y \in P} \|x - y\|^2 - \frac{\varepsilon\eta}{4} \cdot |P|^2 \cdot \frac{8}{|A|} \|A - 1p^T\|_F^2 \right)$$

$$\overset{Lem.\ 2.1}{=} \quad \frac{1-\varepsilon}{2|P|} \left(2|P| \|P - 1\mu(P)^T\|_F^2 - 2\varepsilon\eta \cdot |P|^2 \cdot \frac{1}{|A|} \|A - 1p^T\|_F^2 \right)$$

$$\overset{(2)}{\geq} \quad (1-\varepsilon) \cdot \left(\|P - 1\mu(P)^T\|_F^2 - \varepsilon\|P - 1\mu(P)^T\|_F^2 \right) \geq (1 - 2\varepsilon) \cdot \|P - 1\mu(P)^T\|_F^2$$

Rescaling ε completes the proof. $\qquad \square$

Theorem 5.3. *Let A be an $n \times d$ matrix corresponding to n points in d-dimensional Euclidean space. Then there exists an oblivious (ε, k)-means cost preserving sketch $AS \in \mathbb{R}^{n \times m}$ with $m \in O\left((\log k + \log \log n) \frac{\log \varepsilon^{-1}}{\varepsilon^6} \right)$.*

Proof. Let α, β, γ, and ε' be constants depending on ε to be determined later. Let $\mathcal{L} = \{L_1, \ldots L_{k'}\}$ be the clustering induced by the leaves when running Algorithm 1 on A with parameters α, β, and γ. We further use Y to denote the clustering matrix induced by $\mathcal{L}$. Finally, define $c := \sum_{L \in \mathcal{L}_{exp}} \|L - 1\mu(L)^T\|_F^2$. We will condition on the following events. The number of events depends on α, β, γ.

1. Let $\mathcal{E}_1$ be the event that the 1-means cost of all the leaves L of T is preserved, i.e.:

$$\forall L \in \mathcal{L}, \quad \|LS - 1\mu(L)^T S\|_F^2 = (1 \pm \varepsilon') \|L - 1\mu(L)^T\|_F^2$$

2. Let $\mathcal{E}_2$ be the event that the pairwise distances between all rows of $YY^T A$ are preserved and for any clustering matrix X, we have

$$\|(YY^T A - XX^T YY^T A)S\|_F^2 = (1 \pm \varepsilon') \|YY^T A - XX^T YY^T A\|_F^2.$$

3. Let $\mathcal{E}_3$ be the event that for all leaves $L \in \mathcal{L}_{exp}$, for all rank k clustering matrices Z of appropriate dimension, we have

$$\|LS - ZZ^T LS\|_F^2 \cdot (1 + \alpha + \varepsilon') \geq \|LS - 1\mu(L)^T S\|_F^2.$$

For any rank k clustering matrix X, we bound the distortion incurred by S as follows:

$$\left| \|A - XX^T A\|_F^2 - \|AS - XX^T AS\|_F^2 \right|$$

$$\leq \quad \left| \|A - XX^T A\|_F^2 - (\|YY^T A - XX^T YY^T A\|_F^2 + c) \right| \tag{6}$$

$$+ \quad \left| \|YY^T A - XX^T YY^T A\|_F^2 + c - (\|YY^T AS - XX^T YY^T AS\|_F^2 + c) \right| \tag{7}$$

$$+ \quad \left| \|AS - XX^T AS\|_F^2 - (\|YY^T AS - XX^T YY^T AS\|_F^2 + c) \right|. \tag{8}$$

We bound the terms (6), (7) and (8) separately. To apply Lemma 4.5, let us first consider Δ_{cheap} and Δ_{exp} as given by the trees. For the former, we define $\Delta_{cheap} = \sum_{L \in \mathcal{L}_{low} \cup \mathcal{L}_{inner}} \|L - \mathbb{1}\mu(L)^T\|_F^2$. For the latter, we define $\Delta_{exp} = \sum_{L \in \mathcal{L}_{exp}} \max_{\substack{\text{rank } k \\ \text{c.m. } Z}} \|L - \mathbb{1}\mu(L)^T\|_F^2 - \|L - ZZ^T L\|_F^2$.

Using Property 4.2 and Lemma 2.3, we then have $\Delta_{exp} \leq \alpha \cdot \sum_{L \in \mathcal{L}_{exp}} \|L - \mathbb{1}\mu(L)^T\|_F^2 = \alpha \cdot c$.

After projecting, we derive upper bounds for these values denoted by Δ_{exp}^S and Δ_{cheap}^S. Conditioned on event $\mathcal{E}_1$, we have $\|LS - \mathbb{1}\mu(L)^T S\|_F^2 \leq (1 + \varepsilon') \cdot \|LS - \mathbb{1}\mu(L)^T S\|_F^2$ and hence $\Delta_{cheap}^S := (1 + \varepsilon') \cdot \Delta_{cheap}$. For Δ_{exp}, we observe, conditioned on events $\mathcal{E}_1$ and $\mathcal{E}_3$ that $\sum_{L \in \mathcal{L}_{exp}} \max_{\substack{\text{rank } k \\ \text{c.m. } Z}} \|LS - \mathbb{1}\mu(L)^T S\|_F^2 - \|LS - ZZ^T LS\|_F^2 \leq (\alpha + \varepsilon') \cdot \sum_{L \in \mathcal{L}_{exp}} \|LS - \mathbb{1}\mu(L)^T S\|_F^2 \leq (\alpha + \varepsilon') \cdot (1 + \varepsilon') \sum_{L \in \mathcal{L}_{exp}} \|L - \mathbb{1}\mu(L)^T\|_F^2 = (\alpha + \varepsilon')(1 + \varepsilon') \cdot c$ and hence we set $\Delta_{exp}^S := (\alpha + \varepsilon')(1 + \varepsilon') \cdot c$.

Further, for every cluster L_i induced by Y, let Z_i be the rank k clustering matrix induced by X in L_i. Define the rank $k \cdot k'$ clustering matrix Z as the concatenation of all Z_i and let $A' = ZZ^T A$. Due to the second statement of Lemma 2.3, we then have $XX^T ZZ^T = XX^T$ and $YY^T ZZ^T = YY^T$.

We will use the following lemma (proof in the appendix).

Lemma 5.4. *Let B be either the realization of a random sketching matrix S or the $d \times d$ identity matrix. Then for matrices A', X, Y, and Z and the constant c defined as above, and conditioned on events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ happening for S, we have:*

1. $\|AB - XX^T AB\|_F^2 = \|AB - ZZ^T AB\|_F^2 + \|A'B - XX^T A'B\|_F^2$

2. $\|A'B - YY^T A'B\|_F^2 \leq \Delta_{cheap}^S + \Delta_{exp}^S$

3. $\left| \|AB - ZZ^T AB\|_F^2 - c \right| \leq 3\Delta_{exp}^S + \Delta_{cheap}^S$.

For Term (8), we have

$$\left| \|AS - XX^T AS\|_F^2 - (\|YY^T AS - XX^T YY^T AS\|_F^2 + c) \right|$$

$$\overset{\text{1. of } Lem.\ 5.4}{=} \left| \|A'S - XX^T A'S\|_F^2 + \|AS - ZZ^T AS\|_F^2 - (\|YY^T A'S - XX^T YY^T A'S\|_F^2 + c) \right|$$

$$\overset{\text{3. of } Lem.\ 5.4}{\leq} \left| \|A'S - XX^T A'S\|_F^2 - \|YY^T A'S - XX^T YY^T A'S\|_F^2 \right| + 2\Delta_{exp}^S + \Delta_{cheap}^S$$

$$\overset{Lem.\ 2.1}{=} \left| \|A'S - XX^T YY^T A'S\|_F^2 - \|XX^T YY^T A'S - XX^T A'S\|_F^2 \right.$$
$$\left. - \|YY^T A'S - XX^T YY^T A'S\|_F^2 \right| + 3\Delta_{exp}^S + \Delta_{cheap}^S$$

$$\overset{Lem.\ 4.5}{\leq} \left| \|A'S - XX^T YY^T A'S\|_F^2 - \|YY^T A'S - XX^T YY^T A'S\|_F^2 \right| + 4\Delta_{exp}^S + 2\Delta_{cheap}^S$$

$$\overset{Lem.\ 2.4}{\leq} \left(1 + \frac{1}{\varepsilon}\right) \|A'S - YY^T A'S\|_F^2 + \varepsilon\|YY^T A'S - XX^T YY^T A'S\|_F^2 + 4\Delta_{exp}^S + 2\Delta_{cheap}^S$$

$$\overset{\text{2. of } Lem.\ 5.4}{\leq} \left(1 + \frac{1}{\varepsilon}\right) (\Delta_{cheap}^S + \Delta_{exp}^S) + \varepsilon\|YY^T AS - XX^T YY^T AS\|_F^2 + 4\Delta_{exp}^S + 2\Delta_{cheap}^S$$

$$\overset{\text{Event } \mathcal{E}_2}{\leq} \left(5 + \frac{1}{\varepsilon}\right) (\Delta_{cheap}^S + \Delta_{exp}^S) + \varepsilon(1 + \varepsilon')\|YY^T A - XX^T YY^T A\|_F^2$$

$$\overset{Lem.\ 4.6}{\leq} \left(5 + \frac{1}{\varepsilon}\right) (\Delta_{cheap}^S + \Delta_{exp}^S) + 9(1 + \varepsilon') \cdot \varepsilon\|A - XX^T A\|_F^2 \tag{9}$$

The same bound for Term (6) can be derived in a completely analogous way (a slight modification in the chain of the inequalities can get rid of the leading factor 9 in front of $\|A - XX^T A\|_F^2$, which we omit for conciseness). We therefore have

$$\left| \|A - XX^T A\|_F^2 - (\|YY^T A - XX^T YY^T A\|_F^2 + c) \right|$$

$$\leq \left(5 + \frac{1}{\varepsilon}\right) (\Delta_{cheap}^S + \Delta_{exp}^S) + 9\varepsilon\|A - XX^T A\|_F^2. \tag{10}$$

For Term (7), we use Lemma 2.6 and condition on Event $\mathcal{E}_2$, i.e. the pairwise distances between all the rows of YY^TA are preserved. This yields

$$|\|YY^TA - XX^TYY^TA\|_F^2 + c - (\|YY^TAS - XX^TYY^TAS\|_F^2 + c)|$$
$$\overset{\text{Event } \mathcal{E}_2}{\leq} \varepsilon' \cdot \|YY^TA - XX^TYY^TA\|_F^2 \overset{Lem.\ 4.6}{\leq} 9\varepsilon' \cdot \|A - XX^TA\|_F^2. \tag{11}$$

Recall that $\Delta_{cheap}^S = (1 + \varepsilon') \cdot \Delta_{cheap} \leq 2 \cdot \left(\gamma \cdot \frac{1+\alpha}{\alpha} + \left(\frac{1}{1+\alpha} \right)^{1/\beta} \right)$ OPT due to Lemma 4.4. Furthermore due to Property 4.2, $\Delta_{exp}^S \leq (\alpha + \varepsilon')(1 + \varepsilon') \cdot c \leq (\alpha + \varepsilon')(1 + \varepsilon')(1 + \alpha) \cdot \text{OPT} \leq 4(\alpha + \varepsilon') \cdot \text{OPT}$. Combining this with (10), (9), and (11), we obtain

$$\|A - XX^TA\|_F^2 - \|AS - XX^TAS\|_F^2| \leq \left(10 + \frac{2}{\varepsilon} \right) \cdot (\Delta_{cheap}^S + \Delta_{exp}^S) + 18(\varepsilon' + \varepsilon) \cdot \|A - XX^TA\|_F^2$$

$$\leq \frac{12}{\varepsilon} \cdot \left(2 \cdot \left(\gamma \cdot \frac{1+\alpha}{\alpha} + \left(\frac{1}{1+\alpha} \right)^{1/\beta} \right) + 4(\alpha + \varepsilon') \right) \text{OPT} + 18(\varepsilon' + \varepsilon) \cdot \|A - XX^TA\|_F^2$$

We set $\alpha = \varepsilon' = \varepsilon^2$, $\beta = \frac{\varepsilon^2}{2\log 1/\varepsilon^2}$, and $\gamma = \varepsilon^4$. Using the fact that $\ln(1 + \alpha) \geq \alpha/2$, we have $\left(\frac{1}{1+\alpha} \right)^{1/\beta} \leq \varepsilon^2$. Then the factor in front of OPT is bounded from above by 96ε. Rescaling ε, (and consequently α, ε', β, and γ) proves the desired approximation guarantee.

To conclude the proof, we show that our success probability is $1 - \delta$. We take a union bound over the probability of events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ not happening. Note that the target dimension is at least $c' \cdot \left(\frac{\log\left(\frac{1}{\varepsilon\delta}(k \log \frac{n}{\varepsilon^4})^{O(\varepsilon^{-2}\log\varepsilon^{-2})} \right)}{\varepsilon^4} \right)$ for some large enough constant c', and so at least $c^* \cdot \left(\frac{\log \frac{k \cdot |T|}{\varepsilon\delta}}{\varepsilon^4} \right)$ for a large enough constant c^*.

To obtain a bound on the probability of event $\mathcal{E}_1$ happening, we apply Theorem 2.5. By observation 4.3, the tree has size $|T| = (k \log \frac{n}{\gamma})^{(O(1/\beta))} = (k \log \frac{n}{\varepsilon})^{O(\varepsilon^{-2}\log\varepsilon^{-1})}$. Therefore, since the target dimension is at least $c^* \cdot \left(\frac{\log \frac{k \cdot |T|}{\varepsilon\delta}}{\varepsilon^4} \right)$, we have that event $\mathcal{E}_1$ happens with probability at least $1 - \delta/3$. We claim that the success probability of event $\mathcal{E}_2$ is at least $1 - \delta/3$ as well. Indeed, the number of rows in YY^TA is bounded by the size of the tree and so applying Lemma 2.6 for YY^TA with target dimension at least $c^* \cdot \left(\frac{\log \frac{|T|}{\delta}}{\varepsilon^4} \right)$ yields the claim. Finally, event $\mathcal{E}_3$ also happens with probability at least $1 - \delta/3$. Indeed, following Property 4.2, we invoke Lemma 5.1 with target dimension $c^* \cdot \left(\log k + \log \log n + \log \frac{1}{\delta} \right) \frac{\log \frac{1}{\varepsilon}}{\varepsilon^6}$. This is enough to get success probability at least $1 - \delta/3$ since by Observation 4.3 we have $c^* \cdot \left(\frac{\log \frac{k \cdot |T|}{\varepsilon\delta}}{\varepsilon^4} \right) \geq c' \cdot \left(\frac{\log \frac{k \cdot |\mathcal{L}_{exp}|}{\varepsilon\delta}}{\varepsilon^4} \right)$. $\qquad \square$

References

[1] Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.

[2] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *Learning Theory, 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005, Proceedings*, pages 458–469, 2005.

[3] Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing, Seattle, WA, USA, May 21-23, 2006*, pages 557–563, 2006.

[4] Nir Ailon and Edo Liberty. Fast dimension reduction using Rademacher series on dual BCH codes. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2008, San Francisco, California, USA, January 20-22, 2008*, pages 1–9, 2008.

[5] Nir Ailon and Edo Liberty. An almost optimal unrestricted fast Johnson-Lindenstrauss transform. *ACM Trans. Algorithms*, 9(3):21:1–21:12, 2013.

[6] Noga Alon and Bo'az Klartag. Optimal compression of approximate inner products and dimension reduction. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 639–650, 2017.

[7] Alexandr Andoni, Piotr Indyk, and Mihai Patrascu. On the optimality of the dimensionality reduction method. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21-24 October 2006, Berkeley, California, USA, Proceedings*, pages 449–458, 2006.

[8] Pranjal Awasthi and Or Sheffet. Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 15th International Workshop, APPROX 2012, and 16th International Workshop, RANDOM 2012, Cambridge, MA, USA, August 15-17, 2012. Proceedings*, pages 37–49, 2012.

[9] Maria-Florina Balcan, Steven Ehrlich, and Yingyu Liang. Distributed k-means and k-median clustering on general communication topologies. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 1995–2003, 2013.

[10] Joshua D. Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-ramanujan sparsifiers. *SIAM J. Comput.*, 41(6):1704–1721, 2012.

[11] Jean Bourgain, Sjoerd Dirksen, and Jelani Nelson. Toward a unified theory of sparse dimensionality reduction in Euclidean space. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 499–508, 2015.

[12] Christos Boutsidis and Malik Magdon-Ismail. Deterministic feature selection for k-means clustering. *IEEE Trans. Information Theory*, 59(9):6099–6110, 2013.

[13] Christos Boutsidis, Michael W. Mahoney, and Petros Drineas. Unsupervised feature selection for the k-means clustering problem. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada.*, pages 153–161, 2009.

[14] Christos Boutsidis, Anastasios Zouzias, and Petros Drineas. Random projections for k-means clustering. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada.*, pages 298–306, 2010.

[15] Christos Boutsidis, Anastasios Zouzias, Michael W. Mahoney, and Petros Drineas. Randomized dimensionality reduction for k-means clustering. *IEEE Trans. Information Theory*, 61(2):1045–1062, 2015.

[16] Vladimir Braverman, Dan Feldman, and Harry Lang. New frameworks for offline and streaming coreset constructions. *CoRR*, abs/1612.00889, 2016.

[17] Vladimir Braverman, Harry Lang, Keith Levin, and Morteza Monemizadeh. Clustering problems on sliding windows. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 1374–1390, 2016.

[18] S. Charles Brubaker and Santosh Vempala. Isotropic PCA and affine-invariant clustering. In *49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008, October 25-28, 2008, Philadelphia, PA, USA*, pages 551–560, 2008.

[19] Ke Chen. On coresets for k-median and k-means clustering in metric and Euclidean spaces and their applications. *SIAM J. Comput.*, 39(3):923–947, 2009.

[20] Kenneth L. Clarkson, Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, Xiangrui Meng, and David P. Woodruff. The fast cauchy transform and faster robust linear regression. *SIAM J. Comput.*, 45(3):763–810, 2016.

[21] Kenneth L. Clarkson and David P. Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)*, pages 205–214, 2009.

[22] Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA*, pages 81–90, 2013.

[23] Michael B. Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 278–287, 2016.

[24] Michael B. Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 163–172, 2015.

[25] Michael B. Cohen, T. S. Jayram, and Jelani Nelson. Simple analyses of the sparse Johnson-Lindenstrauss transform. In *1st Symposium on Simplicity in Algorithms, SOSA 2018, January 7-10, 2018, New Orleans, LA, USA*, pages 15:1–15:9, 2018.

[26] Michael B. Cohen, Cameron Musco, and Christopher Musco. Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, pages 1758–1777, 2017.

[27] Vincent Cohen-Addad and Chris Schwiegelshohn. On the local structure of stable clustering instances. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 49–60, 2017.

[28] Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse Johnson-Lindenstrauss transform. In *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010*, pages 341–350, 2010.

[29] Petros Drineas, Alan M. Frieze, Ravi Kannan, Santosh Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56(1-3):9–33, 2004.

[30] Michael Elkin, Arnold Filtser, and Ofer Neiman. Terminal embeddings. *Theor. Comput. Sci.*, 697:1–36, 2017.

[31] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 569–578, 2011.

[32] Dan Feldman, Morteza Monemizadeh, and Christian Sohler. A PTAS for k-means clustering based on weak coresets. In *Proceedings of the 23rd ACM Symposium on Computational Geometry, Gyeongju, South Korea, June 6-8, 2007*, pages 11–18, 2007.

[33] 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 Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, New Orleans, Louisiana, USA, January 6-8, 2013*, pages 1434–1453, 2013.

[34] Gereon Frahling and Christian Sohler. Coresets in dynamic geometric data streams. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing (STOC)*, pages 209–217, 2005.

[35] Sariel Har-Peled and Akash Kushal. Smaller coresets for k-median and k-means clustering. *Discrete & Computational Geometry*, 37(1):3–19, 2007.

[36] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing, Chicago, IL, USA, June 13-16, 2004*, pages 291–300, 2004.

[37] T. S. Jayram and David P. Woodruff. Optimal bounds for Johnson-Lindenstrauss transforms and streaming problems with subconstant error. *ACM Trans. Algorithms*, 9(3):26:1–26:17, 2013.

[38] William Johnson and Joram Lindenstrauss. Extensions of lipschitz maps into a hilbert space. 26:189–206, 01 1984.

[39] Daniel M. Kane, Raghu Meka, and Jelani Nelson. Almost optimal explicit Johnson-Lindenstrauss families. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 14th International Workshop, APPROX 2011, and 15th International Workshop, RANDOM 2011, Princeton, NJ, USA, August 17-19, 2011. Proceedings*, pages 628–639, 2011.

[40] Daniel M. Kane and Jelani Nelson. Sparser Johnson-Lindenstrauss transforms. *J. ACM*, 61(1):4:1–4:23, 2014.

[41] Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. *SIAM J. Comput.*, 38(3):1141–1156, 2008.

[42] Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pages 299–308, 2010.

[43] Michael Langberg and Leonard J. Schulman. Universal ε-approximators for integrals. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*, pages 598–607, 2010.

[44] Kasper Green Larsen and Jelani Nelson. The Johnson-Lindenstrauss lemma is optimal for linear dimensionality reduction. In *43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy*, pages 82:1–82:11, 2016.

[45] Kasper Green Larsen and Jelani Nelson. Optimality of the Johnson-Lindenstrauss Lemma. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 633–638, 2017.

[46] Sepideh Mahabadi, Konstantin Makarychev, Yury Makarychev, and Ilya P. Razenshteyn. Nonlinear dimension reduction via outer bi-lipschitz extensions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 1088–1101, 2018.

[47] Konstantin Makarychev, Yury Makarychev, Maxim Sviridenko, and Justin Ward. A bi-criteria approximation algorithm for k-means. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2016, September 7-9, 2016, Paris, France*, pages 14:1–14:20, 2016.

[48] Xiangrui Meng and Michael W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 91–100, 2013.

[49] Alan Mislove, Massimiliano Marcon, P. Krishna Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Internet Measurement Conference, IMC 2007, San Diego, California, USA, October 24-26, 2007*, pages 29–42, 2007.

[50] Shyam Narayanan and Jelani Nelson. Optimal terminal dimensionality reduction in euclidean space. *CoRR*, abs/1810.09250, 2018.

[51] Jelani Nelson and Huy L. Nguyen. OSNAP: faster numerical linear algebra algorithms via sparser subspace embeddings. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS, Berkeley, CA, USA*, pages 117–126, 2013.

[52] Jelani Nelson and Huy L. Nguyen. Sparsity lower bounds for dimensionality reducing maps. In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 101–110, 2013.

[53] Jelani Nelson and Huy L. Nguyên. Lower bounds for oblivious subspace embeddings. In *Automata, Languages, and Programming - 41st International Colloquium, ICALP 2014, Copenhagen, Denmark, July 8-11, 2014, Proceedings, Part I*, pages 883–894, 2014.

[54] Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 143–152, 2006.

[55] Christian Sohler and David P. Woodruff. Strong coresets for k-median and subspace approximation: Goodbye dimension. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 802–813, 2018.

[56] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *J. Comput. Syst. Sci.*, 68(4):841–860, 2004.

Omitted Proofs from Main Body

Proof of Lemma 2.3. For the first statement, let $\{C_1, \ldots C_k\}$ be the clusters of A induced by X. We have

$$
\begin{aligned}
\|A - 1\mu(A)^T\|_F^2 \quad &= \quad \sum_{i=1}^{k} \sum_{A_j \in C_i} \|A_j - \mu(A)\|^2 \\
&\overset{Lem.\ 2.1}{=} \quad \sum_{i=1}^{k} \left(\sum_{A_j \in C_i} \|A_j - \mu(C_i)\|^2 \right) + |C_i| \cdot \|\mu(C_i) - \mu(A)\|^2 \\
&= \quad \sum_{i=1}^{k} \|C_i - 1\mu(C_i)^T\|_F^2 + \|1\mu(C_i)^T - 1\mu(A)^T\|_F^2 \\
&= \quad \|A - XX^T A\|_F^2 + \|XX^T A - 1\mu(A)^T\|_F^2.
\end{aligned}
$$

The claim follows by rearranging and noting that $\|XX^T A - 1\mu(A)^T\|_F^2$ can never be negative[4].

For the second statement, consider a cluster C_i induced by X and $K_{i,1} \ldots K_{i,\ell}$ to be the subclusters of C_i induced by Y. To see $XX^T YY^T = XX^T$, we have for all C_i

$$
\mu(C_i) = \frac{1}{|C_i|} \sum_{x \in C_i} x = \frac{1}{|C_i|} \sum_{j=1}^{\ell} \sum_{x \in K_{i,j}} x = \frac{1}{|C_i|} \sum_{j=1}^{\ell} |K_{i,j}| \mu(K_{i,j}) = \frac{1}{\sum_{j=1}^{\ell} |K_{i,j}|} \sum |K_{i,j}| \mu(K_{i,j}).
$$

Summing up over all C_i and applying the first claim completes the proof of the second claim.

For the third statement, we use the fact that XX^T is an orthogonal projection, which can only decrease the norm. $\qquad\square$

Proof of Lemma 2.4. We prove the case where $A, B, C \in \mathbb{R}$ (i.e. are scalars). The general claim follows as the absolute value of the sum is upper bounded by the sum of absolute values.

The claim holds if $A = B$ (LHS zero, RHS non-negative), $A = C$, or $B = C$ (RHS is always at least $\|A - B\|^2$).

For $C < A < B$ ($C < B < A$ analogously), we have

$$
\begin{aligned}
& (B - C)^2 - (A - C) = (B - A + A - C)^2 - (A - C)^2 \leq x(b - a)^2 + \varepsilon(A - C)^2 \\
\Leftrightarrow \quad & 2(B - A) \cdot (A - C) - \varepsilon(A - C)^2 \leq (x - 1)(B - A)^2 \\
\Leftrightarrow \quad & 2(A - C) - \varepsilon \frac{(A - C)^2}{B - A} \leq (x - 1)(B - A).
\end{aligned}
$$

The LHS is maximal for $(A - C) = (B - A)/\varepsilon$. Then $(B - A)^2/\varepsilon \leq (x - 1)(B - A)^2$ which holds for $x \geq 1 + 1/\varepsilon$.

Now, let $A < C < B$ and $B - C > C - A$ (analogously $B < C < A$ and $A - C > C - B$). Then

$$
\begin{aligned}
& (B - C)^2 - (C - A)^2 = (B - A - (C - A))^2 - (C - A)^2 \leq x(B - A)^2 + \varepsilon(C - A)^2 \\
\Leftrightarrow \quad & (B - A)^2 - 2(B - A)(C - A) \leq x(B - A)^2 + \varepsilon(C - A)^2 \\
\Leftrightarrow \quad & -2(B - A)(C - A) - \varepsilon(C - A)^2 \leq (x - 1)(B - A)^2,
\end{aligned}
$$

which holds for $x \geq 1$.

---

[4]We note that this equation is a special case of the Pythagorean theorem, as $\frac{1}{|A|} 11^T$ is always a subspace of XX^T.

Lastly, let $A < C < B$ and $C - A > B - C$ (analogously $B < C < A$ and $C - B > B - C$). Then

$$
\begin{aligned}
& (C-A)^2 - (B-C)^2 = (C-A)^2 - (B - A - (C-A))^2 \leq x(B-A)^2 + \varepsilon(C-A)^2 \\
\Leftrightarrow \quad & 2(B-A)(C-A) - (B-A)^2 \leq x(B-A)^2 + \varepsilon(C-A)^2 \\
\Leftrightarrow \quad & 2(B-A)(C-A) - \varepsilon(C-A)^2 \leq (x+1)(B-A)^2 \\
\Leftrightarrow \quad & 2(C-A) - \varepsilon\frac{(C-A)^2}{B-A} \leq (x+1)(B-A)
\end{aligned}
$$

As in the first case, the LHS is maximal if $(C-A) = (B-A)/\varepsilon$, which holds if $x \geq 1 + 1/\varepsilon$. $\quad\square$

Proof of Lemma 2.6. We first note that a linear mapping always preserves the distances between two identical rows (i.e. the zero vector is always mapped to the zero vector). We therefore only have to consider the non-zero difference vectors between two rows.

The second part of Lemma 2.1 states that the k-means cost can be expressed as a non-negative linear combination of the pairwise distances. Since there are only ℓ distinct rows, there are only $\binom{\ell}{2} \leq \ell^2$ pairwise distances. The claim then follows by applying Theorem 2.5 with $\delta' = \delta/\ell^2$. $\quad\square$

Proof of Lemma 2.7. Due to Theorem 2.5, for every vector $x \in \mathbb{R}^d$ the Euclidean norm is preserved up to $(1 \pm \varepsilon)$ factors with probability $1 - \zeta\delta$. This also implies that on expectation at most a $\zeta\delta$ fraction of the distances are not preserved. Using Markov's inequality, at most a ζ-fraction of the distances are not preserved with probability $1 - \delta$. $\quad\square$

Proof of Lemma 4.4. We say that a node $P \in T$ is a parent node if it has at least one child. Define I^ℓ to be the parent nodes (i.e. the nodes with children) of T at level ℓ. We will first show that the sum of all 1-means costs for the parent nodes in I^ℓ satisfies

$$
\sum_{P \in I^\ell} \|P - \mathbf{1}\mu(P)^T\|_F^2 \leq \left(\frac{1}{1+\alpha}\right)^{i-1} \cdot \text{OPT}. \tag{12}
$$

We prove this claim by induction. For level 1, we only have the optimal clusters $C = \{C_1, \ldots C_k\}$ of A. Suppose all of them have children. Then $\sum_{j=1}^k \|C_j - \mathbf{1}\mu(C_j)^T\|_F^2 = \text{OPT} = \left(\frac{1}{1+\alpha}\right)^0 \text{OPT}$.

For the inductive step, we assume that this claim holds up to level ℓ and we consider level $\ell + 1$. Let $A' \in I^\ell$ and let R^i be the ring associated with a child P of A. For P to be an parent node, the check in line 12 was true, i.e. for the clustering matrix Z^i we have $\|R^i - Z^i(Z^i)^T R^i\|_F^2 \cdot (1+\alpha) \leq \|R^i - \mathbf{1}\mu(A')^T\|_F^2$. Therefore

$$
\begin{aligned}
& \sum_{A' \in I^\ell} \sum_{P \in children(A') \cap I^{\ell+1}} \|P - \mathbf{1}\mu(P)^T\|_F^2 \\
\overset{Line\ 12}{\leq} \quad & \frac{1}{1+\alpha} \sum_{A' \in I^\ell} \sum_{P \in children(A') \cap I^{\ell+1}} \|P - \mathbf{1}\mu(A')^T\|_F^2 \leq \frac{1}{1+\alpha} \sum_{A' \in I^\ell} \|A' - \mathbf{1}\mu(A')^T\|_F^2 \\
\leq \quad & \frac{1}{1+\alpha} \cdot \left(\frac{1}{1+\alpha}\right)^{\ell-1} \text{OPT} = \left(\frac{1}{1+\alpha}\right)^\ell \text{OPT}.
\end{aligned}
$$

Let us now consider the cost of all leaves in $\mathcal{L}_{low}$, i.e. the leaves at the lowest level $1/\beta + 2$. Their cost is upper bounded by the cost of their parent nodes, so using (12)

$$
\sum_{L \in \mathcal{L}_{low}} \|L - \mathbf{1}\mu(L)^T\|_F^2 \leq \left(\frac{1}{1+\alpha}\right)^{1/\beta} \cdot \sum_{j=1}^k \|C_j - \mathbf{1}\mu(C_j)^T\|_F^2. \tag{13}
$$

Let us define $p(L)$ to be the parent node of any $L \in \mathcal{L}_{inner}$. We have

$$\sum_{L \in \mathcal{L}_{inner}} \|L - 1\mu(L)^T\|_F^2 = \sum_{i=1}^{1/\beta} \sum_{L \in \mathcal{L}_{inner} \cap I^i} \|L - 1\mu(L)^T\|_F^2$$

$$\leq \sum_{i=1}^{1/\beta} \sum_{L \in \mathcal{L}_{inner} \cap I^i} \gamma \cdot \|p(L) - 1\mu(p(L))^T\|_F^2$$

$$\overset{(12)}{\leq} \gamma \cdot \sum_{i=1}^{1/\beta} \left(\frac{1}{1+\alpha}\right)^i \sum_{j=1}^{k} \|C_j - 1\mu(C_j)^T\|_F^2 \leq \gamma \cdot \frac{1+\alpha}{\alpha} \sum_{j=1}^{k} \|C_j - 1\mu(C_j)^T\|_F^2, \quad (14)$$

where the first inequality follows from the item of Definition 4.1. Summing (13) and (14) completes the proof. $\square$

Proof of Lemma 4.5. Let $K = \{K_1, \ldots K_k\}$ be the clustering induced by X on A. Define the row indexes of C_i as $Ind(C_i) := \{\ell \in [n] \mid A_\ell \in C_i\}$. We first modify the rows of A to obtain a new matrix A'. For every $A_\ell \in C_i \cap K_j$, we set $A'_\ell = \mu(C_i \cap K_j)$. We use Z_i to refer to the clustering matrix mapping C_i to the centroids $\mu(C_i \cap K_j)$, and Z to denote the overall clustering matrix obtained by appending the columns of the Z_i. We define $A' = ZZ^T A$. Since Z is a refinement of X and Y, by the second item of Lemma 2.3 we have $XX^T A = XX^T A'$ and $YY^T A = YY^T A'$. Thus

$$\|XX^T(I - YY^T)A\|_F^2 = \|XX^T A - XX^T YY^T A\|_F^2 = \|XX^T A' - XX^T YY^T A'\|_F^2$$

$$= \|XX^T(I - YY^T)A'\|_F^2 \overset{Lem. \ 2.3}{\leq} \|(I - YY^T)A'\|_F^2.$$

Let us now consider the rows of $(I - YY^T)A'$ induced by the clusters in C_{exp}. We have

$$\sum_{C_i \in C_{exp}} \sum_{\ell \in Ind(C_i)} \|A'_\ell - (YY^T A')_\ell\|^2 = \sum_{C_i \in C_{exp}} \|Z_i Z_i^T C_i - 1\mu(C_i)^T\|_F^2$$

$$\overset{Lem. \ 2.3}{=} \sum_{C_i \in C_{exp}} \|C_i - 1\mu(C_i)^T\|_F^2 - \|C_i - Z_i Z_i^T C_i\|_F^2$$

$$\leq \alpha \cdot \sum_{C_i \in C_{exp}} \|C_i - Z_i Z_i^T C_i\|_F^2 \overset{Lem. \ 2.3}{\leq} \alpha \cdot \sum_{C_i \in C_{exp}} \|C_i - 1\mu(C_i)^T\|_F^2 = \Delta_{exp} \quad (15)$$

where the first equation holds by definition of A' and Z_i and the first inequality due to the definition of C_{exp}. For C_{cheap}, we have

$$\sum_{C_i \in C_{cheap}} \sum_{\ell \in Ind(C_i)} \|A'_\ell - (YY^T A')_\ell\|^2$$

$$= \sum_{C_i \in C_{cheap}} \|Z_i Z_i^T C_i - 1\mu(C_i)^T\|_F^2$$

$$\overset{Lem. \ 2.3}{=} \sum_{C_i \in C_{cheap}} \|C_i - 1\mu(C_i)^T\|_F^2 - \|C_i - Z_i Z_i^T C_i\|_F^2$$

$$\leq \sum_{C_i \in C_{cheap}} \|C_i - 1\mu(C_i)^T\|_F^2 = \Delta_{cheap}, \quad (16)$$

The claim follows from (15) and (16). $\square$

Proof of Lemma 4.6.

$$\|YY^TA - XX^TYY^TA\|_F^2$$
$$= \|YY^TA - A + A - XX^TA + XX^TA - XX^TYY^TA\|_F^2$$
$$\leq (\|YY^TA - A\|_F + \|A - XX^TA\|_F + \|XX^TA - XX^TYY^TA\|_F)^2$$
$$= (\|A - YY^TA\|_F + \|A - XX^TA\|_F + \|XX^T(A - YY^TA)\|_F)^2$$
$$\overset{3.\ of\ Lem.\ 2.3}{\leq} (2\|A - YY^TA\|_F + \|A - XX^TA\|_F)^2 \leq 9\|A - XX^TA\|_F^2$$

$\square$

Proof of Lemma 5.1. Let us condition on the fact that the 1-means clustering and the cost of clustering to the center p is approximately preserved, i.e. $\|AS - 1p^TS\|_F^2 = (1 \pm \varepsilon)\|A - 1p^T\|_F^2$. Clearly, since this is a fixed matrix, this will happen with probability $1 - \delta$ due to Theorem 2.5.

We can now turn our attention to $\|AS - ZZ^TAS\|_F^2$, where Z is the optimal rank k clustering matrix of AS. We partition the clusters induced by Z into cheap clusters C_{cheap} with cost at most $\frac{\varepsilon}{k(1+\alpha)} \cdot \|A - 1p^T\|_F^2$ and the remaining expensive clusters C_{exp}. We will apply Lemma 5.2 with $\eta \geq \frac{\varepsilon}{k(1+\alpha)}$, i.e. for a target dimension $m \in O\left(\frac{\log \frac{1}{\varepsilon\eta\delta}}{\varepsilon^2}\right) = O\left(\frac{\log \frac{k}{\varepsilon\delta}}{\varepsilon^2}\right)$, the projection decreases the cost of any expensive cluster by no more than an $(1 - \varepsilon)$ factor with probability $1 - \delta$. The total contribution of the cheap clusters is

$$\sum_{C_i \in C_{cheap}} \|C_i - 1\mu(C_i)^T\|_F^2 \leq \frac{\varepsilon}{1+\alpha} \cdot \|A - 1p^T\|_F^2 \overset{Ass.\ 2}{\leq} \varepsilon \cdot \|A - ZZ^TA\|_F^2, \qquad (17)$$

hence the expensive clusters incur all but an ε-fraction of the cost of $\|A - ZZ^TA\|_F^2$. Putting everything together, we have

$$\|AS - ZZ^TAS\|_F^2 = \sum_{C_i \in C_{exp}} \|C_iS - 1\mu(C_i)^TS\|_F^2 + \sum_{C_i \in C_{cheap}} \|C_iS - 1\mu(C_i)^TS\|_F^2$$

$$\geq \sum_{C_i \in C_{exp}} \|C_iS - 1\mu(C_i)^TS\|_F^2 \overset{Lem.\ 5.2}{\geq} (1 - \varepsilon) \sum_{C_i \in C_{exp}} \|C_i - 1\mu(C_i)^T\|_F^2$$

$$\overset{(17)}{\geq} (1 - \varepsilon)^2 \|A - ZZ^TA\|_F^2 \overset{Ass.\ 2}{\geq} (1 - \varepsilon)^2 \cdot \frac{1}{1+\alpha}\|A - 1p^T\|_F^2$$

$$\overset{Thm\ 2.5}{\geq} (1 - \varepsilon)^3 \cdot \frac{1}{1+\alpha}\|AS - 1p^TS\|_F^2$$

$$\overset{\varepsilon \leq 1/7, \alpha \leq 1}{\Rightarrow} \|AS - 1p^TS\|_F^2 \leq (1 + \alpha + 9\varepsilon) \cdot \|AS - ZZ^TAS\|_F^2.$$

By the union bound, we have a success probability of at least $1 - 2\delta$. Rescaling ε and δ concludes the proof. $\square$

Proof of Lemma 5.4. For (1), we applying the second statement of Lemma 2.3 and the Pythagorean theorem:

$$\|AB - XX^TAB\|_F^2 = \|AB\|_F^2 - \|XX^TAB\|_F^2$$
$$= \|AB\|_F^2 - \|ZZ^TAB\|_F^2 + \|ZZ^TAB\|_F^2 - \|XX^TAB\|_F^2$$
$$= \|AB - ZZ^TAB\|_F^2 + \|ZZ^TAB\|_F^2 - \|XX^TZZ^TAB\|_F^2$$
$$= \|AB - ZZ^TAB\|_F^2 + \|ZZ^TAB - XX^TZZ^TAB\|_F^2$$
$$= \|AB - ZZ^TAB\|_F^2 + \|A'B - XX^TA'B\|_F^2.$$

For (2), we condition on $\|L_iS - Z_iZ_i^TLS\|_F^2 \cdot (1 + \alpha + \varepsilon') \geq \|L_iS - \mu(L_i)S\|_F^2$ (Condition 3) for all $L_i \in \mathcal{L}_{exp}$.

$$\|A'B - YY^TA'B\|_F^2$$

$$= \sum_{L_i \in \mathcal{L}_{low} \cup \mathcal{L}_{inner}} \|Z_iZ_i^TL_iB - 1\mu(L_i)^TB\|_F^2 + \sum_{L_i \in \mathcal{L}_{exp}} \|Z_iZ_i^TL_iB - 1\mu(L_i)^TB\|_F^2$$

$$\overset{Lem.\ 2.3}{\leq} \sum_{L_i \in \mathcal{L}_{low} \cup \mathcal{L}_{inner}} \|L_iB - 1\mu(L_i)^TB\|_F^2 + \sum_{L \in \mathcal{L}_{exp}} \|L_iB - 1\mu(L_i)^TB\|_F^2 - \|LB - Z_iZ_i^TL_iB\|_F^2$$

$$\overset{Event\ \mathcal{E}_3}{\leq} \sum_{L_i \in \mathcal{L}_{low} \cup \mathcal{L}_{inner}} \|L_iB - 1\mu(L_i)^TB\|_F^2 + (\alpha + \varepsilon') \sum_{L_i \in \mathcal{L}_{exp}} \|L_iB - Z_iZ_i^TL_iB\|_F^2$$

$$\leq \Delta_{exp}^S + \Delta_{cheap}^S.$$

For (3) and again assuming Condition 3, we have

$$\big|\|AB - ZZ^TAB\|_F^2 - c\big| = \left| \sum_{L_i \in \mathcal{L}_{low} \cup \mathcal{L}_{inner}} \|L_iB - Z_iZ_i^TL_iB\|_F^2 + \sum_{L_i \in \mathcal{L}_{exp}} \|L_iB - Z_iZ_i^TL_iB\|_F^2 - c \right|$$

$$\overset{Event\ \mathcal{E}_1}{\leq} \left| \sum_{L_i \in \mathcal{L}_{exp}} \|L_iB - Z_iZ_i^TL_iB\|_F^2 - \|L_iB - 1\mu(L_i)^TB\|_F^2 \right| + \Delta_{cheap}^S.$$

We will bound $\left| \sum_{L_i \in \mathcal{L}_{exp}} \|L_iB - Z_iZ_i^TL_iB\|_F^2 - \|L_iB - 1\mu(L_i)^TB\|_F^2 \right|$ assuming that the arguments are always positive or always negative. The entire sum may then be bounded by the sum of both derived values. If $\|LB - Z_iZ_i^TLB\|_F^2 - \|LN - 1\mu(L)^TB\|_F^2$ is positive, then conditioning on event $\mathcal{E}_1$ the difference is at most $\varepsilon' \cdot c \leq \Delta_{exp}^S$. If the sign is negative, we have

$$\big|\|AB - ZZ^TAB\|_F^2 - c\big| \overset{Events\ \mathcal{E}_1, \mathcal{E}_3}{\leq} \sum_{L \in \mathcal{L}_{exp}} \left| \left(\frac{1 - \varepsilon'}{1 + \alpha + \varepsilon'} - 1 \right) \right| \|L - 1\mu(L)^T\|_F^2 + \Delta_{cheap}^S$$

$$\leq (\alpha + 2\varepsilon') \sum_{L \in \mathcal{L}_{exp}} \|L - 1\mu(L)^T\|_F^2 + \Delta_{cheap}^S \leq 2\Delta_{exp}^S + \Delta_{cheap}^S.$$

$\square$

Data Dependent Dimension Reduction

In this section we show that an explicit construction of the sketching tree from Section 4 combined with a random projection achieves allows to reduce the target dimension to $O_\varepsilon(\log k)$ dimensions. The main difference is that we do not require to preserve the cost of the expensive leaves in an oblivious manner. This allows us to store the cost of these points in the offset c, and we apply the random projection onto YY^TA, instead of A.

In the following we use $\mathrm{OPT}_k(A')$ to denote the cost of an optimal k means clustering on a point set A' and write (a, b)-approximation if a clustering has cost less than $a \cdot \mathrm{OPT}_k$ and uses at most $b \cdot k$ centers. Since we aim at making Algorithm 1 constructive, we use the following result by Makarychev et al. [47].

Theorem B.1 ([47]). *There exists a polynomial time algorithm for k-means that computes a $(1 + \varepsilon, O(1/\varepsilon))$-approximation.*

The algorithm now proceeds very similar to the one proposed in Section 4. For every node A' of the tree, we use the bi-criteria approximation to obtain a clustering with cost at most $(1 + \alpha/3) \cdot \mathrm{OPT}_k(A')$. If the cost decreases, we continue to do so. If the cost does not decrease, the procedure stops, i.e. we skip the partitioning into rings. The stopping criteria are summarized thus:

Algorithm 2 Constructive Sketching Tree

1: Initialize an $O(k/\alpha)$-ary tree T with root A.
2: Compute a $(1 + \alpha/3, O(1/\alpha))$-approximate clustering $C = \{C_1, \ldots C_{O(k/\alpha)}\}$ and initialize queue $Q := C$
3: **while** $Q \neq \emptyset$ **do**
4:     $A' \leftarrow pop(Q)$
5:     **if** depth of A' in T is less than $1/\beta + 2$ **then**
6:         Compute a $(1 + \alpha/3, O(1/\alpha))$-approximate clustering of A' with clustering matrix X and clusters $\{C_1, \ldots, C_{O(k/\alpha)}\}$
7:         **if** $\|A' - XX^T A'\|_F^2 (1 + \alpha/3) \leq \|A' - 1\mu(A')^T\|_F^2$ **then**
8:            Append $\{C_1, \ldots, C_{O(k/\alpha)}\}$ as children of A' and add them to Q

Property B.2. *The leaves computed by Algorithm 1 will have one of the following properties.*

1. *Leaf L_i is at depth $1/\beta + 2$. Denote the entire set of leaves at depth $1/\beta + 2$ by $\mathcal{L}_{low}$.*

2. *Let $\mathcal{L}_{exp}$ be the point sets for which we did not continue the recursion. For each such leaf L_i, we have $\|L_i - Z_i Z_i^T L_i\|_F^2 \cdot (1 + \alpha) \geq \|L_i - 1\mu(L_i)^T\|_F^2$ for any rank k clustering matrix Z_i.*

We can apply Observation 4.3 and Lemmas 4.4 and 4.5 to the resulting clustering just like in Section 4. The only difference is a slight adjusting of parameters.

Observation B.3. *T has at most $O\left((k/\alpha)^{O(1/\beta)}\right)$ many nodes.*

Lemma B.4.
$$\sum_{L \in \mathcal{L}_{low}} \|L - 1\mu(L)^T\|_F^2 \leq \left(\frac{1}{1 + \alpha/3}\right)^{1/\beta} \cdot OPT_k.$$

For any $\alpha^{-1}, \beta^{-1} \in \text{poly}(\varepsilon^{-1})$, Algorithm 2 can be run in polynomial time. Using a very similar line of reasoning as in Theorem 5.3, we then obtain the following theorem.

Theorem B.5. *Let A be an n by d matrix corresponding to n points in d-dimensional Euclidean space. There exists a linear dimension reduction with offset c given by a matrix $B \in \mathbb{R}^{n \times m}$ where $m \in O\left(\frac{\log k + \log 1/\delta}{\varepsilon^4} \log \frac{1}{\varepsilon}\right)$ such that for all rank k clustering matrices X,*

$$(1 - \varepsilon)\|A - XX^T A\|_F^2 \leq \|B - XX^T B\|_F^2 + c \leq (1 + \varepsilon)\|A - XX^T A\|_F^2$$

with probability $1 - \delta$. For every fixed constant ε, the dimension reduction can be computed in polynomial time.

Proof. Let Y be the clustering matrix induced by the set of leaves. Set $c = \sum_{\mathcal{L}_\ell \in \mathcal{L}_{exp}} \|\mathcal{L}_\ell - 1\mu(\mathcal{L}_\ell)^T\|_F^2$. We sample $S \in \mathbb{R}^{d \times m}$ from a distribution satisfying the bounds given by Theorem 2.5 and Lemma 2.6 and set $B := YY^T AS$. We invoke Lemma 2.6 for $YY^T A$, i.e. the pairwise distances between all rows of $YY^T A$ are preserved and for any clustering matrix X, we have with probability $1 - \delta$

$$\|(YY^T A - XX^T YY^T A)S\|_F^2 = (1 \pm \varepsilon)\|YY^T A - XX^T YY^T A\|_F^2. \tag{18}$$

To use Lemma 4.5, we again consider Δ_{exp} and Δ_{cheap}. The former is equal to $\alpha \cdot c$ by Property B.2. The latter is at most $\left(\frac{1}{1 + \alpha/3}\right)^{1/\beta} \cdot OPT_k$ due to Lemma B.4. Further, for every cluster (leaf) L_i induced by Y, let Z_i be the rank k clustering matrix induced by X in $\mathbb{L}_i$. Define the rank $k \cdot k'$ clustering matrix Z as the concatenation of all Z_i and let $A' = ZZ^T A$.

Again, applying the second statement of Lemma 2.3, we then have $XX^TZZ^T = XX^T$ and $YY^TZZ^T = YY^T$. Further following Lemma 5.4, we have the identities

$$\|A - XX^TA\|_F^2 = \|A - ZZ^TA\|_F^2 + \|A' - XX^TA'\|_F^2 \tag{19}$$

$$\|A' - YY^TA'\|_F^2 \le \Delta_{cheap} + \Delta_{exp} \tag{20}$$

$$\left|\|A - ZZ^TA\|_F^2 - c\right| \le 3\Delta_{exp} + \Delta_{cheap}. \tag{21}$$

We then have

$$
\begin{aligned}
&\left|\|A - XX^TA\|_F^2 - (\|B - XX^TB\|_F^2 + c)\right| \\
=\ &\left|\|A - XX^TA\|_F^2 - (\|YY^TAS - XX^TYY^TAS\|_F^2 + c)\right| \\
\overset{(18)}{\le}\ &\left|\|A - XX^TA\|_F^2 - (\|YY^TA - XX^TYY^TA\|_F^2 + c)\right| + \varepsilon\|YY^TA - XX^TYY^TA\|_F^2 \\
\overset{(19)}{=}\ &\left|\|A' - XX^TA'\|_F^2 + \|A - ZZ^TA\|_F^2 - (\|YY^TA' - XX^TYY^TA'\|_F^2 + c)\right| \\
&+ \varepsilon\|YY^TA - XX^TYY^TA\|_F^2 \\
\overset{(21)}{=}\ &\left|\|A' - XX^TA'\|_F^2 - \|YY^TA' - XX^TYY^TA'\|_F^2\right| + 3\Delta_{exp} + \Delta_{cheap} \\
&+ \varepsilon\|YY^TA - XX^TYY^TA\|_F^2 \\
\overset{Lem.\ 2.1}{\le}\ &\left|\|A' - XX^TYY^TA'\|_F^2 - \|YY^TA' - XX^TYY^TA'\|_F^2\right| \\
&+ \|XX^TA' - XX^TYY^TA'\|_F^2 + 3\Delta_{exp} + \Delta_{cheap} + \varepsilon\|YY^TA - XX^TYY^TA\|_F^2 \\
\overset{Lem.\ 2.4}{\le}\ &\left(1 + \frac{1}{\varepsilon}\right)\left|\|A' - YY^TA'\|_F^2 + \varepsilon \cdot \|YY^TA' - XX^TYY^TA'\|_F^2\right| \\
&+ \|XX^T(I - YY^T)A'\|_F^2 + 3\Delta_{exp} + \Delta_{cheap} + \varepsilon\|YY^TA - XX^TYY^TA\|_F^2 \\
\overset{Lem.\ 4.5}{\le}\ &\left(1 + \frac{1}{\varepsilon}\right)\left|\|A' - YY^TA'\|_F^2 + \varepsilon \cdot \|YY^TA' - XX^TYY^TA'\|_F^2\right| \\
&+ 4\Delta_{exp} + 2\Delta_{cheap} + \varepsilon\|YY^TA - XX^TYY^TA\|_F^2 \\
\overset{(20)}{\le}\ &\left(1 + \frac{1}{\varepsilon}\right)(\Delta_{cheap} + \Delta_{exp}) + 4\Delta_{exp} + 2\Delta_{cheap} + 2\varepsilon\|YY^TA - XX^TYY^TA\|_F^2 \\
\overset{Lem.\ 4.6}{\le}\ &\frac{5}{\varepsilon}(\Delta_{exp} + \Delta_{cheap}) + 2\varepsilon \cdot 9(1 + \alpha/3)^3 \cdot \|A - XX^TA\|_F^2 \\
\overset{Lem.\ B.4}{\le}\ &\frac{5}{\varepsilon}\left(\left(\frac{1}{1 + \alpha/3}\right)^{1/\beta} \cdot \text{OPT} + \alpha \cdot \text{OPT}\right) + 162\varepsilon \cdot \|A - XX^TA\|_F^2 \\
\le\ &\left(\frac{5}{\varepsilon}\left(\left(\frac{1}{1 + \alpha/3}\right)^{1/\beta} + \alpha\right) + 162\varepsilon\right) \cdot \|A - XX^TA\|_F^2
\end{aligned}
$$

We set $\alpha = \varepsilon^2$ and $\beta = \frac{\varepsilon^2}{6\log 1/\varepsilon^2}$. Then the above bounds yields $172\varepsilon \cdot \|A - XX^TA\|_F^2$. Rescaling ε completes the proof of the approximation. By Observation B.3, YY^TA has $O\left((k/\alpha)^{O(1/\beta)}\right)$ distinct rows. Hence a target dimension of $\frac{\log O\left((k/\alpha)^{O(1/\beta)}\right)}{\varepsilon^2} \in O\left(\frac{\log k/\varepsilon}{\varepsilon^4}\log\frac{1}{\varepsilon}\right)$ is sufficient. $\qquad\square$

Theorem B.5 can be used in a distributed environment to solve k-means in a dimension efficient way given that the number of servers are less than $O(\log n)$, in which case Theorem 5.3 gives better bounds. Assume that the number of servers is at most t. Each server runs Algorithm 2 locally. The servers then sample a suitable sketching matrix of target dimension $O\left(\frac{\log t \cdot k/\varepsilon}{\varepsilon^4}\log\frac{1}{\varepsilon}\right)$. By the union bound, this sketch preserves the pairwise rows between all rows. After this, any distributed clustering algorithm may be run on the (now low-dimensional) point set. In the following, we describe a more substantial application of Algorithm 2 to coresets.

A Brief Remark on Coresets for k-Means

Coresets are a loosely defined concept for aggregating and compressing data that has found applications for numerous problems beyond clustering. Generally speaking coresets aim to summarize the data set such that we can answer any given query up to a small (typically $(1 \pm \varepsilon)$ factor) distortion. Most (but not all) coreset definitions satisfy composability. In other words, coresets are closed under union, i.e. given coresets P_1 of A_1 and P_2 of A_2, then $P_1 \cup P_2$ is a coreset of $A_1 \cup A_2$. This feature makes coresets extremely flexible and applicable in a variety of settings such as distributed computing and streaming. The most general and powerful definition of coresets for k-means is due to Feldman et al. [33]:

Definition C.1. *Let A be a set of n points in d dimensional Euclidean space, let k be a non-negative integer, let c_A be a constant possibly depending on A, and let $\varepsilon > 0$. Then a set P is an (ε, k)-coreset if there exists a weight function $w : P \to \mathbb{R}^+$ such that for any candidate set of centers C*

$$\left| \sum_{p \in A} \min_{\mu \in C} \|p - \mu\|^2 - \left(\sum_{q \in P} \min_{\mu \in C} w(q) \cdot \|q - \mu\|^2 + c_A \right) \right| \leq \varepsilon \cdot \sum_{p \in A} \min_{\mu \in C} \|p - \mu\|^2$$

Reference	Size (Number of Points)
Low Dimensions	
Har-Peled, Mazumdar (STOC'04) [36] Frahling, Sohler (STOC'05) [34]	$O(k\varepsilon^{-d} \log n)$
Har-Peled, Kushal (DCG'07) [35]	$O(k^3 \varepsilon^{-(d+1)})$
High Dimensions	
Chen (Sicomp'09) [19]	$O(d^2 k^2 \varepsilon^{-2} \log^5 n)$
Langberg, Schulman (SODA'10) [43]	$O(d^2 k^3 \varepsilon^{-2})$
Feldman, Langberg (STOC'11) [31]	$O(dk\varepsilon^{-4})$
Feldman et al. (SODA'13) [33]	$O(k^2 \varepsilon^{-6})$
Cohen et al. (STOC'15) [24]	$O(k^2 \varepsilon^{-5})$
here	$O(k\varepsilon^{-8})$

Table 1: Comparison of memory demands, where lower order factors are suppressed and the memory to store a d-dimensional point is not specified. The constructions for high dimensions do not treat d as a constant and succeed with constant probability.

Intuitively, we consider a set of points P to be a k-means coreset of A, if for any set of candidate centers C of size at most k the (possibly weighted) cost of clustering P to C is approximately equal to the cost of clustering A to to C. The definition is similar to Definition 2.2, but there is a crucial difference. The coreset guarantee applies to all locally optimal assignments to k centers in d-dimensions. The cost-preserving sketch guarantee applies to the means of all possible clustering. Neither definition is trivially stronger than the other.

Work on coresets for k-means includes [9, 16, 17, 19, 32, 33, 34, 35, 36, 43], with the current state of the art being the sensitivity framework due to Feldman and Langberg [31]. For an overview of the current bounds, we refer to Table 1. The sensitivity framework yields coresets of size $O(kd\varepsilon^{-4})$ points. Using dimension reduction techniques by Feldman et al. [33] and Cohen et al. [24], the dependency on d may be replaced by k/ε. We note that Theorem B.5 cannot be applied as a black box, as the original space is not preserved after a random projection onto $O_\varepsilon(\log k)$ dimensions, whereas the dimension-reduction techniques by [24, 33] reduce the rank of A in the original space. We will instead use terminal embeddings that preserve the entire space.

Definition C.2 (Terminal Embeddings). *Let $\varepsilon \in (0,1)$ and let $A \subset \mathbb{R}^d$ be arbitrary with $|A|$ having size $n > 1$. Than a mapping $f : \mathbb{R}^d \to \mathbb{R}^m$ is a* terminal embedding *of*

$$\forall x \in P \forall y \in \mathbb{R}^d, \ \|x - y\| \leq \|f(x) - f(y)\| \leq (1 + \varepsilon)\|x - y\|^2.$$

Terminal embeddings for Euclidean spaces were studied in the work [30, 46, 50]. In particular, Mahabadi et al. [46] showed that a target dimension of $m \in O(\varepsilon^{-4} \log n)$ was sufficient, which was very recently by Narayanan and Nelson [50] to $m \in O(\varepsilon^{-2} \log n)$, which is optimal.

Terminal embeddings preserve coresets up to small distortion; given a point set A and a coreset $P \subset A$ of A, then $f(P)$ remains a coreset of $f(A)$ if f is a terminal embedding of A. Vice versa, given a coreset $P' \subset f(A)$ of $f(A)$, we also know that $f^{-1}(P')$ [5] is a coreset of A. Combining the results from section B with terminal embeddings yields the following theorem.

Theorem C.3. *Let A be a set of n points in d dimensional Euclidean space and let k be a non-negative integer. Then there exists an (ε, k)-coreset for the k-means problem consisting of at most $O(k \log(k/\varepsilon) \cdot \varepsilon^{-8} \log \varepsilon^{-1})$ points.*

Proof. We apply a terminal embedding onto the leaves of the sketching tree from Section B. From the proof of Theorem B.5, $YY^T A$ is both a cost preserving sketch (albiet with no reduction in dimension) and a coreset of A with offset c. Since $YY^T A$ has at most $O(k^{O(\varepsilon^{-2} \log \varepsilon^{-1})})$ rows, this results in a target dimension of order $m \in O(\varepsilon^{-4} \log \varepsilon^{-1} \log k)$, due to Theorem 1.1. of Narayanan and Nelson [50]. We then compute a coreset P in the embedded space. Applying the algorithm by Feldman and Langberg [31] (see Table 1) results in a coreset of size $O(k \log k \varepsilon^{-8} \log \varepsilon^{-1})$. For any set of centers $C \in \mathbb{R}^d$, we therefore have $\min_{\mu \in C} \|x - \mu\|^2 = (1 \pm \varepsilon) \min_{\mu \in C} \|f(x) - f(\mu)\|^2$ for all $x \in P$. Hence $f^{-1}(P)$ is a 3ε coreset of $YY^T A$, and therefore a 7ε coreset of A with offset. Rescaling ε completes the proof. $\square$

---

[5]The mapping f is not invertable. We use the notation to refer to the points of A in the original space.