An Empirical Evaluation of k-Means Coresets

Chris Schwiegelshohn
Department of Computer Science, Aarhus University, Denmark

Omar Ali Sheikh-Omar
Department of Computer Science, Aarhus University, Denmark

Abstract

Coresets are among the most popular paradigms for summarizing data. In particular, there exist many high performance coresets for clustering problems such as k-means in both theory and practice. Curiously, there exists no work on comparing the quality of available k-means coresets.

In this paper we perform such an evaluation. There currently is no algorithm known to measure the distortion of a candidate coreset. We provide some evidence as to why this might be computationally difficult. To complement this, we propose a benchmark for which we argue that computing coresets is challenging and which also allows us an easy (heuristic) evaluation of coresets. Using this benchmark and real-world data sets, we conduct an exhaustive evaluation of the most commonly used coreset algorithms from theory and practice.

2012 ACM Subject Classification Theory of computation → Data compression; Information systems → Clustering

Keywords and phrases coresets, k-means coresets, evaluation, benchmark

Digital Object Identifier 10.4230/LIPIcs.ESA.2022.84

Related Version *Full Version*: `https://arxiv.org/pdf/2207.00966`

Supplementary Material *Software (Source Code)*: `https://github.com/sheikhomar/eval-k-means-coresets`, archived at `swh:1:dir:53066aa034ea87cdf2fd2f5cb2077400aaf341c3`

Funding *Chris Schwiegelshohn*: Independent Research Fund Denmark (DFF) Sapere Aude Research Leader grant No 1051-00106B.
Omar Ali Sheikh-Omar: Innovation Fund Denmark under grant agreement No 0153-00233A.

1 Introduction

The design and analysis of scalable algorithms has become an important research area over the past two decades. This is particularly important in data analysis, where even polynomial running time might not be enough to handle proverbial *big data* sets. One of the main approaches to deal with the scalability issue is to compress or sketch large data sets into smaller, more manageable ones. The aim of such compression methods is to preserve the properties of the original data, up to some small error, while significantly reducing the number of data points.

Among the most popular and successful paradigms in this line of research are *coresets* [40]. Informally, given a data set A, a coreset $\Omega \subset A$ with respect to a given set of queries Q and query function $f : A \times Q \to \mathbb{R}_{\geq 0}$ approximates the behaviour of A for all queries up to some multiplicative distortion D via $\sup_{q \in Q} \max\left(\frac{f(\Omega,q)}{f(A,q)}, \frac{f(A,q)}{f(\Omega,q)}\right) \leq D$. Coresets have been applied to a number of problems such as computational geometry [2, 9], linear algebra [30, 34], and machine learning [36, 41]. But the by far most intensively studied and arguably most successful applications of the coreset framework is the k-clustering problem.

Here we are given n points A with (potential unit) weights $w : A \to \mathbb{R}_{\geq 0}$ in some metric space with distance function dist and aim to find a set of k centers C such that

$$\text{cost}_A(C) := \frac{1}{n} \sum_{p \in A} \min_{c \in C} w(p) \cdot \text{dist}^z(p, c)$$

is minimized. The most popular variant of this problem is probably the k-means problem in d-dimensional Euclidean space where $z = 2$ and $\text{dist}(x, y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$.

A (k, ε)-coreset is now a subset $\Omega \subset A$ with weights $w : \Omega \to \mathbb{R}_{\geq 0}$ such that for any set of k centers C

$$\sup_C \max \left(\frac{\text{cost}_A(C)}{\text{cost}_\Omega(C)}, \frac{\text{cost}_\Omega(C)}{\text{cost}_A(C)} \right) \leq 1 + \varepsilon. \tag{1}$$

The coreset definition in Equation (1) provides an upper bound for the distortion of all candidate solutions i.e., all possible sets of k centers. A *weak coreset* is a relaxed guarantee that holds for optimal or nearly optimal clusterings of A instead of all clusterings.

In a long line of work spanning the last 20 years [4, 8, 10, 14, 15, 19, 20, 26, 25, 27, 29, 8, 33, 44], the size of coresets has been steadily improved with the current state of the art yielding a coreset with $\tilde{O}(k\varepsilon^{-2} \cdot \min(d, k, \varepsilon^{-2}))$ points for a distortion $D \leq (1 + \varepsilon)$ due to [13][1].

While we have a good grasp of the theoretical guarantees of these algorithms, our understanding of the empirical performance is somewhat lacking. There exist a number of coreset implementations, but it is usually difficult to assess which implementation summarizes the data best. To accurately evaluate a given coreset, we would need to come up with a k clustering C which results in a maximal distortion. Solving this problem is likely difficult: related questions such as deciding whether a 3-dimensional point set A is an ε-net of a set B with respect to convex ranges is co-NP hard [24].

Due to this difficulty, a common heuristic for evaluating coresets is as follows [1, 22]. First, compute a coreset Ω with the available algorithm(s) using some input data A. Then, run an optimization algorithm on Ω to compute a k clustering. The *best* coreset algorithm is considered to be the one which yields a clustering with the smallest cost.

This practice has substantial drawbacks. The first is that this evaluation method conflates the two separate tasks of coreset construction and optimization. It is important to note that the first step of virtually all coreset algorithms is a low-cost (bicriteria) constant factor approximation, i.e. a solution with $\beta \cdot k$ clusters that costs at most $\alpha \cdot \text{OPT}$, where OPT is the cost of an optimal k clustering. Given that this initial solution has an α approximation to the cost, a routine calculation shows that the additive error of the coreset, i.e. the maximum difference $|\text{cost}_A(C) - \text{cost}_B(C)|$ over all solutions C is at most $O(\alpha) \cdot \text{cost}_A(C)$. In particular, in the case that the initial bicriteria approximation has $\alpha \ll 2$, which is not too difficult to achieve with more than k centers, any γ approximation algorithm will find solutions with approximation factor $O(\gamma + \alpha) \cdot \text{OPT}$. In particular, the distortion may be unbounded, for example if B only consists of the k centers, while simply returning B itself yields a low cost clustering. Thus, it is difficult to measure coreset quality in this way.

The second drawback is that this practice will mainly measure the performance of the optimization algorithm, rather than the performance of the coreset algorithms. During its execution it might simply not consider any solution with high distortion. For example, if the

[1] We use $\tilde{O}(x)$ to hide $\log^c x$ terms for any constant c.

approximation factor γ of the solution returned by the algorithm is large then this solution (as well as any even higher cost solution considered during the algorithm's execution) will have a low distortion.

The third drawback of this evaluation method is that it does not consider the main use cases of coresets, nor the full power of their guarantee. Indeed, if speeding up the computation of an optimization algorithm, one would hardly need a strong coreset; approximating the cost of every candidate solution, as weaker coreset definitions (or indeed a bicriteria approximation) would be suitable as well. A coreset's main and most powerful feature is *composability*, i.e. given two disjoint point sets X and Y, the union of a coreset of X and a coreset of Y is a coreset. Composability is what enables coresets to scale to massively parallel computation models and enables simple streaming algorithms via the merge and reduce technique. To which degree a coreset is composable is generally not a property of an optimal clustering of the point set, as optimal solutions C_X of X or C_Y of Y may have little in common with an optimal solution of $X \cup Y$.

The purpose of this study is to systematically evaluate the quality of various coreset algorithms for k-means. As such, we develop a new evaluation procedure which estimates the distortion of coreset algorithms. On real-world data sets, we observe that while the evaluated coreset algorithms are generally able to find solutions with comparable costs, there is a stark difference in their distortions. This shows that differences between optimization and compression are readily observable in practice.

As a complement to our evaluation procedure on real-world data sets, we propose a benchmark framework for generating synthetic data sets. We argue why this benchmark has properties that results in hard instances for all known coreset constructions. We also show how to efficiently estimate the distortion of a candidate coreset on the benchmark.

2 Coreset Algorithms

Though the algorithms vary in details, coreset constructions come in one of the following two flavours:

1. **Movement-based constructions:** Such algorithms compute a coreset Ω with T points given some input point set A such that $\text{cost}_\Omega(C) \ll \text{OPT}$, where OPT is the cost of an optimal k-means clustering of A. The coreset guarantee then follows as a consequence of the triangle inequality. These algorithms all have an exponential dependency on the dimension d, and therefore have been overtaken by sampling-based methods. Nevertheless, these constructions are more robust to various constrained clustering formulations [28, 43] and continue to be popular. Examples from theory include [23, 26].
2. **Importance sampling:** Points are sampled proportionate to their impact on the cost of any given candidate solution. The idealized distribution samples proportionate to the sensitivity which for a point p is defined as $sens(p) := \sup_C \frac{\min_{c \in C} \text{dist}^2(p,c)}{\text{cost}_A(C)}$ and weighted by their inverse sampling probability. The sensitivities are hard to compute exactly but much work exists on how to find other distributions with very similar properties. In terms of theoretical performance, sensitivity sampling has largely replaced movement-based constructions, see for example [19, 33].

Of course, there exist algorithms that draw on techniques from both, see for example [15]. In what follows, we will survey implementations of various coreset constructions that we will evaluate later.

StreamKM++ [1]. The popular k-means++ algorithm [3] computes a set of centers K by iteratively sampling a point p in A proportionate to $\min_{q \in K} \text{dist}^2(p,q)$ and adding it to K. The procedure terminates once the desired number of centers has been reached. The

first center is typically picked uniformly at random. The StreamKM++ paper runs the k-means++ algorithms for T iterations, where T is the desired coreset size. At the end, every point q in K is weighted by the number of points in A closest to it. While the construction has elements of importance sampling, the analysis is largely movement-based. The provable bound required for the algorithm to compute a coreset is $O\left(\frac{k\log n}{\delta^{d/2}\varepsilon^d}\cdot\log^{d/2}\frac{k\log n}{\delta^{d/2}\varepsilon^d}\right)$. Despite its simplicity, its running time compares unfavourably to all other constructions.

BICO [22]. BICO combines the very fast, but poor quality clustering algorithm BIRCH [47] with the movement-based analysis from [23, 26]. The clustering is organized by way of a hierarchical decomposition: When adding a point p to one of the coreset points Ω at level i, it first finds the closest point q in Ω. If p is too far away from q, a new cluster is opened with center at p. Otherwise p is either added to the same cluster as q, or, if adding p to q's cluster increases the clustering cost beyond a certain threshold, the algorithm attempts to add p to the child-clusters of q. The procedure then continues recursively. The provable bound required for the algorithm to compute a coreset is $O\left(k\varepsilon^{-d-2}\log n\right)$.

Ray Maker [25]. The algorithm computes an initial solution with k centers which is a constant factor approximation of the optimal clustering. Around each center, $O(1/\epsilon^{d-1})$ random rays are created which span the hyperplane. Next, each point $p\in A$ is snapped to its closest ray resulting in a set of one-dimensional points associated with each ray. Afterwards, a coreset is created for each ray by computing an optimal 1D clustering with k^2/ϵ^2 centers and weighing each center by the number of points in each cluster. The final coreset is composed of the coresets computed for all the rays. The provable bound required for the algorithm to compute a coreset is $O(k^3\cdot\varepsilon^{-d-1})$. The algorithm has recently received some attention due to its applicability to the fair clustering problem [28].

Sensitivity Sampling [19]. The simplest implementation of sensitivity sampling first computes an $(O(1),O(1))$ bicriteria approximation[2], for example by running k-means++ for $2k$ iterations [46]. Let K be the $2k$ clustering thus computed and let K_i be an arbitrary cluster of K with center q_i. Subsequently, the algorithm picks points proportionate to $\frac{\mathrm{dist}^2(p,q)}{\mathrm{cost}_{K_i}(\{q_i\})}+\frac{1}{|K_i|}$ and weighs any point by its inverse sampling probability. Let $|\hat{K}_i|$ be the estimated number of points in the sample. Finally, the algorithm weighs each q_i by $(1+\varepsilon)\cdot|K_i|-|\hat{K}_i|$. The provable bound required for the algorithm to compute a coreset is $\tilde{O}\left(kd\varepsilon^{-4}\right)$ ([19]), $\tilde{O}\left(k\varepsilon^{-6}\right)$ ([29]), or $\tilde{O}\left(k^2\varepsilon^{-4}\right)$ ([8]).

Group Sampling [15]. First, the algorithm computes an $O(1)$ approximation (or a bicriteria approximation) K. Subsequently, the algorithm preprocesses the input into groups such that (1) for any two points $p,p'\in K_i$, their cost is identical up to constant factors and (2) for any two clusters K_i,K_j, their cost is identical up to constant factors. In every group, Group Sampling now samples points proportionate to their cost. The authors of [15] show that there always exist a partitioning into $\log^2 1/\varepsilon$ groups. Points not contained in a group are snapped to their closest center q in K. q is weighted by the number of points snapped to it. The provable bound required for the algorithm to compute a coreset is $\tilde{O}\left(k\varepsilon^{-2}\min(d,k,\varepsilon^{-2})\right)$ ([13]). While this improves over sensitivity sampling, it is generally slower and not as easy to implement.

[2] An (α,β) bicriteria approximation computes an α approximation using $\beta\cdot k$ many centers.

Finally, we note that some of the more popular algorithms in theory have not been mentioned here. For example, Chen's [10] construction is particularly popular among theoreticians. The Group Sampling algorithm by [15] is an extension and improvement of Chen's method. Thus, the performance of Group Sampling is also indicative of Chen's algorithm.

Dimension Reduction

Finally, we also combine coreset constructions with a variety of dimension reduction techniques. Starting with [17], a series of results [4, 5, 6, 7, 12, 16, 20, 21, 32, 37, 44] explored the possibility of using dimension reduction methods for k-clustering, with a particular focus on principal component analysis (PCA) and random projections. The seminal paper by Feldman, Schmidt, and Sohler [20] was the first to use dimension reduction to obtain smaller coresets for k-means. Movement-based coresets in particular often have an exponential dependency on the dimension, which can be alleviated with some form of dimension reduction, both in theory [43] and in practice [31]. There are essentially two main dimension reduction techniques for coresets.

Principal Component Analysis. Feldman, Schmidt, and Sohler [20] showed that projecting an input A onto the first $O(k/\varepsilon^2)$ principal components is a coreset. This coreset still consists of n points, but they now lie in low dimension. The analysis was subsequently tightened by [12] and extended to other center-based cost functions by [44]. Although its target dimension is generally worse than those based on random projections and terminal embeddings, there is nevertheless reasons for using PCA regardless: It removes noise and thus may make it easier to compute a high quality coreset. For more applications of PCA to k-means clustering, we refer to

Terminal Embeddings. Given a set of points A in $\mathbb{R}^D$, a terminal embedding $f : \mathbb{R}^D \to \mathbb{R}^d$ preserves the pairwise distance between any point $p \in A$ and any point $q \in \mathbb{R}^D$ up to a $(1 \pm \varepsilon)$ factor. The statement is related to the famous Johnson-Lindenstrauss lemma but it is stronger as it does not apply to only the pairwise distances of A. Nevertheless, the same target dimension is sufficient. Terminal embeddings were studied by [11, 18, 35, 42], with Narayanan and Nelson [42] achieving an optimal target dimension of $O(\varepsilon^{-2} \log n)$, where n is the number of points. We note that terminal embeddings, combined with an iterative application of the coreset construction from [8], can reduce the target dimension to a factor $\tilde{O}(\varepsilon^{-2} \log k)$. This is mainly of theoretical interest, as in practice the deciding factor wrt the target dimension is the precision, rather than dependencies on $\log n$ and $\log k$. For applications to coresets, we refer to [4, 15, 29]. For an empirical evaluation of random projections, which form the basis of all known terminal embeddings, we refer to Venkatsubramanian and Wang [45].

3 Benchmark Construction

In this section, we describe our benchmark. We start by describing the aims of the benchmark, followed by giving the construction. Our aim is to generate a data set containing many clusterings with the following properties.

1. The benchmark has many clusterings that, in a well defined sense, are highly dissimilar. Specifically, we want the overlap between any two clusters of different clusterings to be small.

2. The different clusterings have very similar and low cost. This ensures that despite the solutions being different in terms of composition and center placement, a good coreset has to consider them equally regarding distortion.
3. The clusterings are induced by a minimal cost assignment of input points to a set of centers in $\mathbb{R}^d$. This final property ensures that the coreset guarantee has to apply to these clusterings.

To generate the benchmark, we now use the following construction. The benchmark has a parameter α which controls the number of points and dimensions of the generated data instance. For a given value of k, the benchmark instance consists of $n = k^\alpha$ points and $d = \alpha \cdot k$ dimensions, i.e. we will construct and $n \times d$ matrix A where every row corresponds to an input point and every column corresponds to one of the dimensions.

Let $\mathbb{1}_k$ be the k-dimensional all-one vector and v_i^1 be the k-dimensional vector with entries $(v_i^1)_j = \begin{cases} -\frac{1}{k} & \text{if } i \neq j \\ \frac{k-1}{k} & \text{if } i = j \end{cases}$. For $\ell \leq \alpha$, recursively define the k^ℓ dimensional vector $v_i^\ell = v_i^{\ell-1} \otimes \mathbb{1}_k$, where $\otimes$ denotes the Kronecker product, i.e. $v_i^{\ell-1} \otimes \mathbb{1}_k = \begin{bmatrix} (v_i^{\ell-1})_1 \cdot \mathbb{1}_k \\ (v_i^{\ell-1})_2 \cdot \mathbb{1}_k \\ \vdots \\ (v_i^{\ell-1})_1 \cdot \mathbb{1}_k \end{bmatrix}$. Finally, set the t-th column of A, for $t = a \cdot k + b$, $a \in \{0, \dots \alpha - 1\}$ and $b \in \{1, \dots k\}$, to be $\mathbb{1}_{k^{\alpha-a+1}} \otimes v_b^{a+1}$.

To get a better feel for the construction, we have given two small example instances for $k = 2$ and $k = 3$ in Figue Figure 1.

$$\begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \qquad \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

Figure 1 Benchmark construction for $k = 2$ and $\alpha = 3$ (left) and $k = 3$ and $\alpha = 2$ (right).

Properties of the Benchmark

We now summarize the key properties of the benchmark. To this end, we require a few notions. Let A be the input matrix. We slightly abuse notation and refer to A_i as both the ith point as well as the ith row of the matrix A. For a clustering $\mathcal{C} = \{C_1, \dots, C_k\}$, we define that the $n \times k$ indicator matrix $\tilde{X}$ induced by $\mathcal{C}$ via $\tilde{X}_{i,j} = \begin{cases} 1 & \text{if } A_i \in C_j \\ 0 & \text{else.} \end{cases}$ Furthermore, we will also use the $n \times k$ normalized clustering matrix X defined as $X_{i,j} = \begin{cases} \frac{1}{\sqrt{|C_i|}} & \text{if } A_i \in C_j \\ 0 & \text{else.} \end{cases}$

We also recall the following lemma which will allow us to express the k-means cost of a clustering $\mathcal{C}$ with optimally chosen centers in terms of the cost of X and A.

▶ **Lemma 1** (Folklore). *Let A be an arbitrary set of points and let $\mu(A) = \frac{1}{|A|}\sum_{p\in A} p$ be the mean. Then $\sum_{p\in A}\|p-c\|^2 = |A|\cdot\|\mu(A)-c\|^2 + \sum_{p\in A}\|p-\mu(A)\|^2$ for any point c.*

This lemma proves that for any given cluster C_j, the mean is the optimal choice of center. We also note that any two distinct columns of X are orthogonal. Furthermore $\frac{1}{n}\mathbf{1}\mathbf{1}^T A$ copies the mean into every entry of A. Combining these two observations, we see that the matrix XX^TA maps the ith row of A onto the mean of the cluster it is assigned to. Finally, define the Frobenius norm of an $n\times d$ A by $\|A\|_F = \sqrt{\sum_{i=1}^n\sum_{j=1}^d A_{i,j}^2}$. Then the k-means cost of the clustering $\mathcal{C}$ is precisely $\|A-XX^TA\|_F^2$.

We also require the following distance measure on clusterings as proposed by Meila [38, 39]. Given two clusterings $\mathcal{C}$ and $\mathcal{C}'$, the $k\times k$ confusion matrix M is defined as $M_{i,j} = |C_i\cap C'_j|$. Furthermore for the indicator matrices $\tilde{X}$ and $\tilde{X}'$ induced by $\mathcal{C}$ and $\mathcal{C}'$ we have the identity $M = \tilde{X}^T\tilde{X}'$. Denote by Π_k the set of all permutations over k elements. Then the distance between $\mathcal{C}$ and $\mathcal{C}'$ is defined as $d(\mathcal{C},\mathcal{C}') = 1-\frac{1}{n}\max_{\pi\in\Pi_k}\sum_{i=1}^k M_{i,\pi(i)}$. Observe that for clusters that are identical, their distance is 0. The maximum distance between any two k clusterings is always $\frac{k-1}{k}$.

The solutions we consider are given as follows. For the columns $a\cdot k+1,\dots(a+1)\cdot k$, we define the clustering $\mathcal{C}^a = \{C_1^a,\dots C_k^a\}$ with $A_i\in C_j^a$ if and only if $A_{i,j}>0$. Let $\tilde{X}^a$ and X^a denote the indicator matrix and clustering matrix, respectively, as induced by $\mathcal{C}^a$. These clusterings satisfy the properties we stated at the beginning of this section, that is:

1. The distance between these clustering is $1-\frac{1}{k}$, i.e. it is maximized.
2. The clusterings have equal cost and the centers in each clustering have equal cost.
3. The clusterings are induced by a set of centers in $\mathbb{R}^d$.

Benchmark Evaluation

We now describe how we use the benchmark to measure the distortion of a coreset. Assume for now that the coresets are subsets of the original input points. The extension to coresets that do not consist of input points is described at the end of this section.

Consider the clustering $\mathcal{C}^a = \{C_1^a,\dots C_k^a\}$ for some a and let Ω with weights $w:\Omega\to\mathbb{R}_{\geq 0}$ be the coreset and let $\delta>0$ be a parameter. Note that there are α many such clusterings, for each value of a. We use $w(C_i^a\cap\Omega) := \sum_{p\in C_i^a\cap\Omega} w(p)$ to denote the mass of points of C_i^a in Ω. For every cluster C_i^a with $w(C_i^a\cap\Omega)\geq|C_i|(1-\delta)$, we place a center at $\mu(C_i^a)$. Conversely, if $w(C_i^a\cap\Omega)<|C_i^a|(1-\delta)$, we do not place a center at $\mu(C_i^a)$. We call such clusters *deficient*. Let $\mathcal{S}$ be the centers of these deficient clusters.

We now compare the cost as computed on the coreset and the true cost of $\mathcal{S}$. Due to Lemma 1 and the fact that all clusters have equal cost, we may write for any deficient cluster C_i^a $\operatorname{cost}_{C_i^a}(\mathcal{S}) = \operatorname{cost}_{C_j^a}(\{\mu(C_j^a)\}) + k^{\alpha-1}\|\mu(C_j^a)-\mu(C_h^a)\|_2^2$, where C_h^a is a non-deficient cluster. Thus, the cost is $\operatorname{cost}_{C_i^a}(\mathcal{S})\approx\left(1+\frac{2}{\alpha}\right)\cdot\operatorname{cost}_{C_j^a}(\{\mu(C_j^a)\})$.

Conversely, the cost on the coreset is

$$\operatorname{cost}_{\Omega\cap C_i^a}(\mathcal{S}) \approx \frac{w(C_i^a\cap\Omega)}{\operatorname{cost}_{C_j^a}(\{\mu(C_j^a)\})}\left(1+\frac{2}{\alpha}\right)\cdot\operatorname{cost}_{C_j^a}(\{\mu(C_j^a)\}).$$

Thus for each deficient clustering individually, the distortion will be close to $\frac{k^{\alpha-1}}{w(C_i^a\cap\Omega)}\geq\frac{1}{1-\varepsilon}$. If there are many deficient clusters, then this will also be the overall distortion. For all possible (suitably discretized) thresholds for deficiency, i.e. all values of δ, we can now identify the clustering $\mathcal{C}^a$ with a maximum number of deficient clusters and use the aforementioned construction to get a lower bound on the distortion.

To extend this evaluation to coresets where the points are not part of the input, we consider a point $p \in \Omega$ to be in C_i^a if it is closer to $\mu(C_i^a)$ than to to $\mu(C_j^a)$.

4 Experiments

In this section, we present how we evaluated different algorithms. First, we propose our evaluation procedure which gauges the quality of coresets. Then, we describe the data sets used for the empirical evaluation and our experimental setup. Finally, we detail the outcome of the experiments and our interpretation of the results.

Evaluation Procedure

Accurately evaluating a k-means coreset of a real-world data set requires constructing a solution (a set of k centers) which results in a maximal distortion. Finding such a solution, however, is difficult. Instead, we can estimate the quality of a given coreset by finding meaningful candidate solutions.

A first attempt can be to randomly generate candidate solutions. It is not readily apparent how to define a distribution of meaningful solutions from which to sample. One could, for instance, generate k random points inside the convex hull or the minimum enclosing ball (MEB) of a coreset Ω. Convex hulls in high dimensions are infeasible to compute, so we sample a center by choosing random convex combination of the centers of the initial bicriteria approximation computed for every coreset. A better way to generate candidate solutions turns out to be k-means++, where we sample k points with respect to the k-means++ distribution and use the resulting centers as a solution. The main advantage of this approach is that k-means++ can uncover natural cluster structures in the data, which uniform sampling generally does not. For all variants, we generated 5 candidate solutions, where the candidate solution with the largest distortion being a lower bound for the true distortion of the coreset.

Given the usefulness of evaluating coresets on real-world data sets, it can be tricky to gauge the general performance of coreset algorithms using only a small selection of data sets. For this reason, we used our benchmark to complement the evaluation on real-world data sets. The benchmark accomplishes two important tasks. First, the benchmark allows us to quickly find a bad solution because both good and bad clusterings are known a priori. It is unclear how to find bad clusterings for real-world data sets. Second, it is easier to make a fair comparison of different coreset constructions because the benchmark is known to generate hard instances for all known coreset algorithms. This cannot be said for real-world data sets. For the benchmark, we computed the distortion following the evaluation procedure described in Section 3.

Every randomized coreset construction was repeated 10 times. We aggregated the reported maximum distortions for every run by taking the average over all 10 evaluations. It is important to not aggregate the distortions here by taking the maximum over all runs: If one run of the coreset algorithm fails but the others succeed, then such an aggregation predicts far worse distortion than what we could typically expect.

Data sets

We conducted experiments on five real-world data sets *Census*, *Covertype*, *Tower*, *Caltech*, *NYTimes*, and four instances of our benchmark. Benchmark instances were generated to match approximately the sizes of the real-world data sets. The sizes of the considered data sets are given in Table 1.

Table 1 The sizes of the real-world datasets used for the experimental evaluation.

	Data points	Dimensions
Caltech	3,680,458	128
Census	2,458,285	68
Covertype	581,012	54
NYTimes	500,000	102,660
Tower	4,915,200	3

Table 2 The parameter values and the sizes of the benchmark instances used for the experimental evaluation.

k	α	Data points	Dimensions
10	6	1,000,000	60
20	5	3,200,000	100
30	4	810,000	120
40	4	2,560,000	160

The *Census*[3] dataset is a small subset of the Public Use Microdata Samples from 1990 US census. It consists of demographic information encoded as 68 categorical attributes of 2,458,285 individuals.

Covertype[4] is comprised of cartographic descriptions and forest cover type of four wilderness areas in the Roosevelt National Forest of Northern Colorado in the US. It consists of 581,012 records, 54 cartographic variables and one class variable. Although *Covertype* was originally made for classification tasks, it is often used for clustering tasks by removing the class variable [1].

The data set with the fewest number of dimensions is *Tower*[5]. This data set consists of 4,915,200 rows and 3 features as it is a 2,560 by 1,920 picture of a tower on a hill where each pixel is represented by a RGB color value.

Inspired by [22], *Caltech* was created by computing SIFT features from the images in the Caltech101[6] image database. This database contains pictures of objects partitioned into 101 categories. Disregarding the categories, we concatenated the 128-dimensional SIFT vectors from each image into one large data matrix with 3,680,458 rows and 128 columns.

NYTimes[7] is a dataset composed of the bag-of-words (BOW) representations of 300,000 news articles from The New York Times. The vocabulary size of the text collection is 102,660. Due to the BOW encoding, *NYTimes* has a very large number of dimensions and is highly sparse. To make processing feasible, we reduced the number of dimensions to 100 using terminal embeddings.

[3] `https://archive.ics.uci.edu/ml/datasets/US+Census+Data+(1990)`
[4] `https://archive.ics.uci.edu/ml/datasets/covertype`
[5] `http://homepages.uni-paderborn.de/frahling/coremeans.html`
[6] `http://www.vision.caltech.edu/Image_Datasets/Caltech101/`
[7] `https://archive.ics.uci.edu/ml/datasets/bag+of+words`

Preprocessing & Experimental Setup

To understand how denoising effects the quality of the outputted coresets, we applied Principal Component Analysis (PCA) on *Caltech*, *Census*, *Covertype*, and *NYTimes* by using the k singular vectors corresponding to the largest singular values. We did not perform any preprocessing on *Tower* due to its low dimensionality.

We followed the same experimental procedure with respect to the choice of parameter values for the algorithms as prior works [1, 22]. For the target coreset size T, we experimented with $T = mk$ for $m = \{50, 100, 200, 500\}$. On *Caltech*, *Census*, *Covertype*, and *NYTimes*, we used values k in $\{10, 20, 30, 40, 50\}$, while for *Tower* we used larger cluster sizes $k \in \{20, 40, 60, 80, 100\}$. On the benchmark, we used $k \in \{10, 20, 30, 40\}$.

We implemented Sensitivity Sampling, Group Sampling, Ray Maker, and StreamKM++ in C++. The source code can be found on GitHub[8]. For BICO, we used the authors' reference implementation[9]. The source code was compiled with gcc 9.3.0. The experiments were performed on a machine with Intel Core i9 10940X 3.3GHz 14-Core and 2x DDR4 PC3200 128GB RAM.

Outcome of Experiments

We observed that in the majority of our experiments, varying the coreset sizes does not significantly change the performance profiles of individual algorithms when comparing them against each other. Therefore, in the following sections, we focus on a cross-section of the experiments where $m = 200$ i.e., coreset sizes $T = 200k$. For numerical results including variances of all the experiments and tables containing distortions, costs and running times, we refer to the full version of this paper.

In Figure 2, we summarized the distortions of the experiments with coreset sizes $T = 200k$. All five algorithms are matched on the *Tower* dataset. The worst distortions across the algorithms are close to 1, and performance between the algorithms is negligible. The performance difference between sampling-based and movement-based methods become more pronounced as the number of dimensions increase. On *Covertype* with its 54 features, Ray Maker performs the worst followed by BICO and Group Sampling while Sensitivity Sampling and StreamKM++ perform the best. Differences in performance are more noticeable on *Census*, *Caltech*, and *NYTimes* where methods based on importance sampling perform much better. Sensitivity Sampling and Group Sampling perform the best, StreamKM++ come in second while BICO and Ray Maker perform the worst across these data sets. On the *Benchmark*, Ray Maker is the worst while Sensitivity Sampling and Group Sampling are the best. StreamKM++ performs also very well compared to BICO.

Interpretation of Experimental Results

Optimization versus Compression. While all five algorithms are equally matched when optimizing on the candidate coresets, coreset quality performance differ significantly (see Figure 2). For all data sets, the obtained costs differed insignificantly for all values of k, irrespective of the coreset algorithm used, while distortions varied strongly, depending on the coreset algorithm.

[8] `https://github.com/sheikhomar/eval-k-means-coresets`
[9] `https://ls2-www.cs.tu-dortmund.de/grav/en/bico`

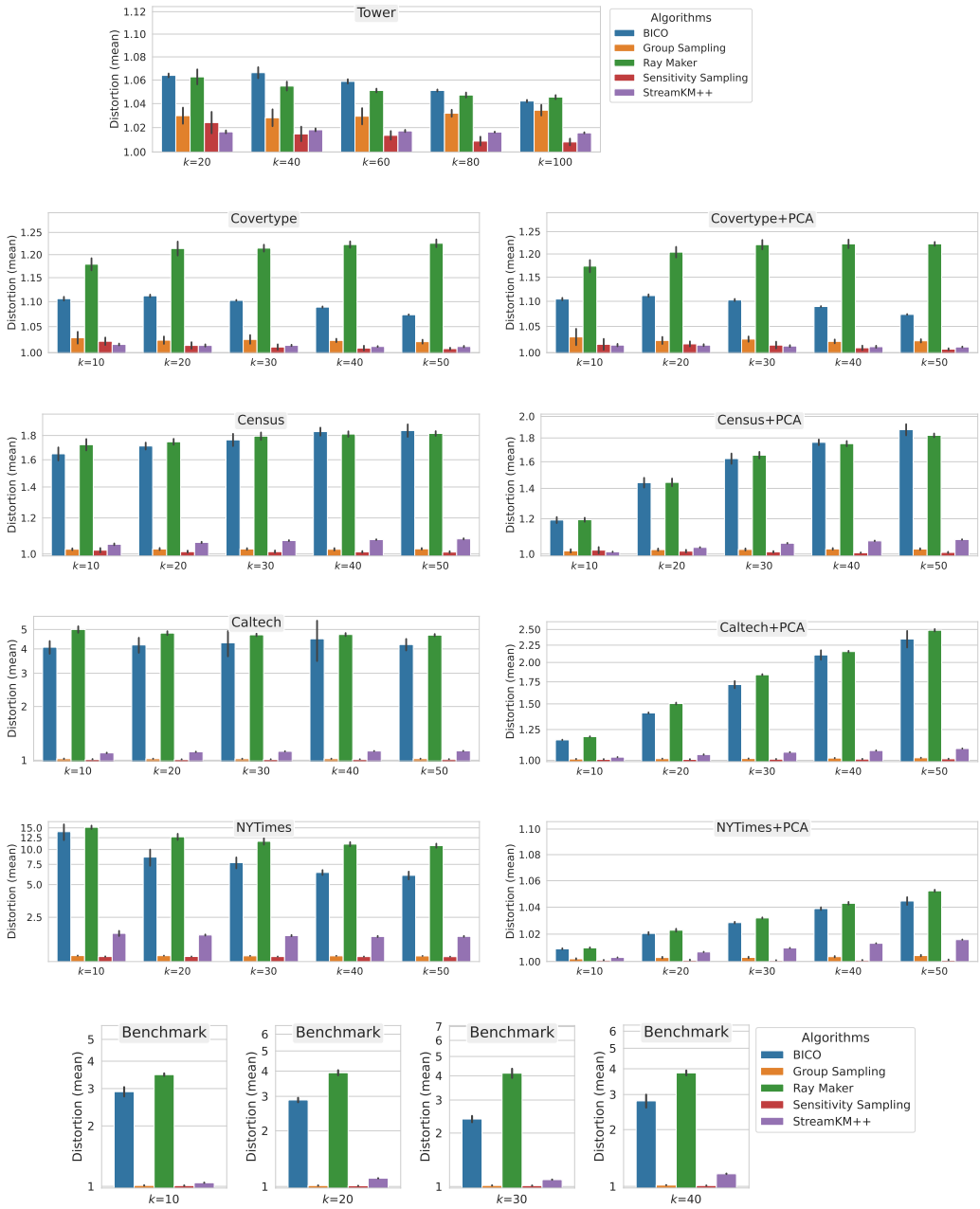

Figure 2 The average distortions of the evaluated coreset algorithms with coreset size $T = 200k$ on five real-world data sets and on four benchmark instances. Black bars indicate standard deviations. Notice that the axis is non-linear as otherwise the bars for Sensitivity Sampling and Group Sampling would disappear on the plots as their distortions are close to 1.

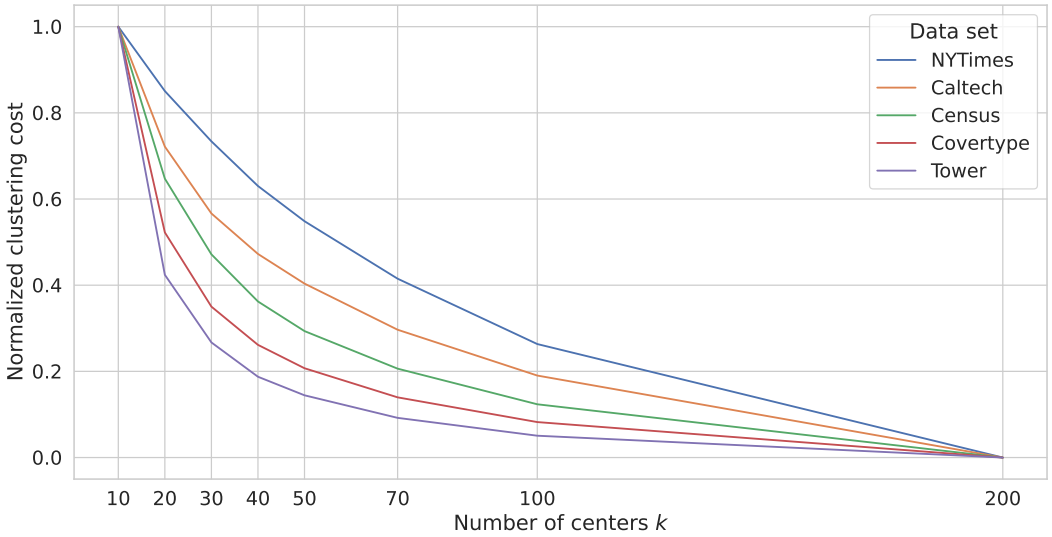

Figure 3 Depicts how clustering costs of five real-world data sets decrease as the number of centers increase. Plotting the cost curve allows us to study whether we can observe a difference between coreset construction and optimization in a data set when evaluating a coreset based on cost.

Nevertheless, the cost drop with increasing values of k is a predictor for the quality of certain coresets. It is not uncommon for the k-means cost of real-world data sets to drop significantly for larger values of k. Figure 3 illustrates this behavior for several real-world data sets. The more the curve bends, the less of a difference there is between computing a coreset and a clustering with low cost. For data sets with an L-shaped cost curve, a coreset algorithm adding more centers to the coreset will seem to be performing well when evaluating it based on the outcome of the optimization. *Tower* is a good example of a data set where optimization is very close to compression. Its cost curve bends the most which indicates that adding more centers help reduce the cost. One of the strengths of the benchmark is that there is no way of reducing the cost without capturing the right subclusters within a benchmark instance. This means that the cost does not decrease markedly beyond a certain value of k even if more centers are added.

For BICO, Ray Maker, and StreamKM++, there is a correlation between the steepness of the cost curve for a data set and the distortion of the generated coreset. On data sets where the curve is less steep, we observed higher distortions. The effect is more pronounced for BICO and Ray Maker than for StreamKM++. Importance sampling approaches (Group Sampling and Sensitivity Sampling) seem to be free from this behavior as they consistently generate high quality coresets irrespective of the shape of cost curve.

Movement-based versus Sampling-based Approaches. In general, movement-based constructions perform the worst in terms of coreset quality. We observed that BICO and Ray Maker have the highest distortions across all data sets including on the benchmark instances. Among the sampling-based algorithms, Sensitive Sampling performs well with Group Sampling generally being competitive. This runs contrary to theory where Group Sampling has the better (currently known) theoretical bounds. StreamKM++ is an interesting case. Like the movement-based methods, its distortion increases with the dimension. Nevertheless, it generally performs significantly better than BICO and Ray Maker. This can be attributed to the fact that the coreset produced by StreamKM++ consists entirely of

k-means++ centers weighted by the number of points of a minimal cost assignment. This is similar to movement-based algorithms such as BICO. Nevertheless, it also retains some of the performance from pure importance schemes.

In practice as well as in theory, the distortion of movement-based algorithms are affected by the dimension. By comparison, sampling-based algorithms are affected very little. Theoretically, there should not exist a difference, as the sampling bounds are independent of the dimension. What little effect can be observed is likely due to PCA making it easier to find low cost solutions that form the backbone of all coreset constructions. StreamKM++ is an interesting case, as it is still affected by the dimension, though less than the other movement based methods.

A notable exception is the benchmark. Here, sensitivity sampling generally found the lowest cost clustering, with BICO finding the second lowest cost clustering. This happens *despite* BICO generally having a worse distortion than for example Group Sampling or StreamKM++.

Impact of PCA. On almost all our data sets, the performance improves when input data is preprocessed with PCA, especially for the movement-based algorithms. Empirically, the more noise is removed (i.e., small k value), the lower the distortion. Notice that k is the number of principal components that the input data is projected on to. The rest of the low variance components are treated as noise and removed. Method utilizing sampling (Group Sampling, Sensitivity Sampling and StreamKM++) are less effected by the preprocessing step. On *Covertype*, PCA does not change the distortions by much because almost all the variance in the data is explained by the first five principal components. On *Caltech* and *NYTimes*, the quality of the coresets by BICO and Ray Maker improves greatly because the noise removal is more aggressive. Even if the quality is much better for movement-based coreset constructions due to PCA, importance sampling methods are still superior when it comes to the quality of the compression. Summarizing, all methods benefit from PCA, and in case of movement-based constructions, we consider PCA a necessary preprocessing step. For the sampling-based methods, the computational expense of using PCA in preprocessing does not seem justify the comparatively meager gains in coreset distortion.

5 Conclusion

In this work, we studied how to assess the quality of k-means coresets computed by state-of-the-art algorithms. Previous work generally measured the quality of optimization algorithms run on the coreset, which we empirically observed to be a poor indicator of coreset quality. For real-world data sets, we sampled candidate clusterings and evaluated the worst case distortion on them. Complementing this, we also proposed a benchmark framework which generates hard instances for known k-means coreset algorithms. Our experiments indicate a general advantage for algorithms based on importance sampling over movement-based methods. Despite movement-based methods running on very efficient code, it is necessary to complement them with rather expensive dimension reduction methods, rendering what efficiency they might have over importance sampling somewhat moot.

Two results bear further investigation. First, the currently known provable coreset sizes for Sensitivity Sampling are worse than those provable via Group Sampling. Empirically, we observed the opposite: While Group Sampling is competitive, Sensitivity Sampling always outperforms it. Since Group Sampling requires somewhat cumbersome computational overhead, practical applications should prefer Sensitivity Sampling. In light of these results, a theoretical analysis for Sensitivity Sampling matching the performance of Group Sampling would be welcome.

The second point of interest focuses on the performance of StreamKM++. The distortion of this algorithm is significantly better than what one would expect from its theoretical analysis. Empirically, StreamKM++ is notably better than the other movement-based constructions across all data sets, and especially on high dimensional data. While it is not competitive to the pure importance sampling algorithms, there are several reasons for investigating it further. It essentially only requires running k-means++ for additional iterations, which is already a nearly ubiquitous algorithm for the k-means problem. Although the other sampling-based coreset algorithms can also be readily implemented, doing so might be cumbersome. In particular, the theoretically (but not empirically) best algorithm Group Sampling requires extensive preprocessing steps. This begs the question whether there exist a better theoretical analysis for StreamKM++.

In addition, StreamKM++ currently weighs each point by the number of points assigned to it. It may also be possible to improve the performance of the algorithm in both theory and practice by using a different weighting scheme. We leave this as an open problem for future research.

References

1 Marcel R. Ackermann, Marcus Märtens, Christoph Raupach, Kamil Swierkot, Christiane Lammersen, and Christian Sohler. Streamkm++: A clustering algorithm for data streams. *ACM Journal of Experimental Algorithmics*, 17(1), 2012. doi:10.1145/2133803.2184450.

2 Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. Geometric approximation via coresets. In *Combinatorial and computational geometry, MSRI*, pages 1–30. University Press, 2005.

3 David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007*, pages 1027–1035, 2007. URL: http://dl.acm.org/citation.cfm?id=1283383.1283494.

4 Luca Becchetti, Marc Bury, Vincent Cohen-Addad, Fabrizio Grandoni, and Chris Schwiegelshohn. Oblivious dimension reduction for k-means: beyond subspaces and the johnson-lindenstrauss lemma. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, pages 1039–1050, 2019. doi:10.1145/3313276.3316318.

5 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. URL: http://papers.nips.cc/paper/3724-unsupervised-feature-selection-for-the-k-means-clustering-problem.

6 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. URL: http://papers.nips.cc/paper/3901-random-projections-for-k-means-clustering.

7 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. doi:10.1109/TIT.2014.2375327.

8 Vladimir Braverman, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering in excluded-minor graphs and beyond. In Dániel Marx, editor, *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual Conference, January 10 - 13, 2021*, pages 2679–2696. SIAM, 2021. doi:10.1137/1.9781611976465.159.

9 Timothy M. Chan. Dynamic coresets. *Discret. Comput. Geom.*, 42(3):469–488, 2009. doi:10.1007/s00454-009-9165-3.

10 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.

11 Yeshwanth Cherapanamjeri and Jelani Nelson. Terminal embeddings in sublinear time. In *62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021, Denver, CO, USA, February 7-10, 2022*, pages 1209–1216. IEEE, 2021. doi:10.1109/FOCS52979.2021.00118.

12 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.

13 Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, and Chris Schwiegelshohn. Towards optimal lower bounds for k-median and k-means coresets. In Stefano Leonardi and Anupam Gupta, editors, *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022*, pages 1038–1051. ACM, 2022. doi:10.1145/3519935.3519946.

14 Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. Improved coresets and sublinear algorithms for power means in euclidean spaces. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 21085–21098, 2021. URL: https://proceedings.neurips.cc/paper/2021/hash/b035d6563a2adac9f822940c145263ce-Abstract.html.

15 Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. A new coreset framework for clustering. In Samir Khuller and Virginia Vassilevska Williams, editors, *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 169–182. ACM, 2021.

16 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. doi:10.1109/FOCS.2017.14.

17 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. doi:10.1023/B:MACH.0000033113.59016.96.

18 Michael Elkin, Arnold Filtser, and Ofer Neiman. Terminal embeddings. *Theor. Comput. Sci.*, 697:1–36, 2017. doi:10.1016/j.tcs.2017.06.021.

19 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.

20 Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca, and projective clustering. *SIAM J. Comput.*, 49(3):601–657, 2020. doi:10.1137/18M1209854.

21 Zhili Feng, Praneeth Kacham, and David P. Woodruff. Dimensionality reduction for the sum-of-distances metric. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 3220–3229. PMLR, 2021. URL: http://proceedings.mlr.press/v139/feng21a.html.

22 Hendrik Fichtenberger, Marc Gillé, Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. BICO: BIRCH meets coresets for k-means clustering. In *Algorithms - ESA 2013 - 21st Annual European Symposium, Sophia Antipolis, France, September 2-4, 2013. Proceedings*, pages 481–492, 2013.

23 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.

24 Panos Giannopoulos, Christian Knauer, Magnus Wahlström, and Daniel Werner. Hardness of discrepancy computation and ϵ-net verification in high dimension. *J. Complex.*, 28(2):162–176, 2012. doi:10.1016/j.jco.2011.09.001.

25 Sariel Har-Peled and Akash Kushal. Smaller coresets for k-median and k-means clustering. *Discret. Comput. Geom.*, 37(1):3–19, 2007. doi:10.1007/s00454-006-1271-x.

26 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.

27 Lingxiao Huang, Shaofeng H.-C. Jiang, Jian Li, and Xuan Wu. Epsilon-coresets for clustering (with outliers) in doubling metrics. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 814–825, 2018. doi:10.1109/FOCS.2018.00082.

28 Lingxiao Huang, Shaofeng H.-C. Jiang, and Nisheeth K. Vishnoi. Coresets for clustering with fairness constraints. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7587–7598, 2019. URL: https://proceedings.neurips.cc/paper/2019/hash/810dfbbebb17302018ae903e9cb7a483-Abstract.html.

29 Lingxiao Huang and Nisheeth K. Vishnoi. Coresets for clustering in euclidean spaces: importance sampling is nearly optimal. In Konstantin Makarychev, Yury Makarychev, Madhur Tulsiani, Gautam Kamath, and Julia Chuzhoy, editors, *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pages 1416–1429. ACM, 2020. doi:10.1145/3357713.3384296.

30 Piotr Indyk, Sepideh Mahabadi, Shayan Oveis Gharan, and Alireza Rezaei. Composable core-sets for determinant maximization problems via spectral spanners. In Shuchi Chawla, editor, *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020*, pages 1675–1694. SIAM, 2020. doi:10.1137/1.9781611975994.103.

31 Jan-Philipp W. Kappmeier, Daniel R. Schmidt, and Melanie Schmidt. Solving k-means on high-dimensional big data. In *Experimental Algorithms - 14th International Symposium, SEA 2015, Paris, France, June 29 - July 1, 2015, Proceedings*, pages 259–270, 2015. doi:10.1007/978-3-319-20086-6_20.

32 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. doi:10.1109/FOCS.2010.35.

33 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. doi:10.1137/1.9781611973075.50.

34 Alaa Maalouf, Ibrahim Jubran, and Dan Feldman. Fast and accurate least-mean-squares solvers. In *Advances in Neural Information Processing Systems*, pages 8307–8318, 2019.

35 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. doi:10.1145/3188745.3188828.

36 Tung Mai, Cameron Musco, and Anup Rao. Coresets for classification – Simplified and strengthened. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 11643–11654, 2021. URL: https://proceedings.neurips.cc/paper/2021/hash/6098ed616e715171f0dabad60a8e5197-Abstract.html.

37 Konstantin Makarychev, Yury Makarychev, and Ilya P. Razenshteyn. Performance of johnson-lindenstrauss transform for k-means and k-medians clustering. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, pages 1027–1038, 2019. doi:10.1145/3313276.3316350.

38 Marina Meila. Comparing clusterings: an axiomatic view. In Luc De Raedt and Stefan Wrobel, editors, *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, volume 119 of *ACM International Conference Proceeding Series*, pages 577–584. ACM, 2005. doi:10.1145/1102351.1102424.

39 Marina Meila. The uniqueness of a good optimum for k-means. In William W. Cohen and Andrew W. Moore, editors, *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, volume 148 of *ACM International Conference Proceeding Series*, pages 625–632. ACM, 2006. doi:10.1145/1143844.1143923.

40 Alexander Munteanu and Chris Schwiegelshohn. Coresets-methods and history: A theoreticians design pattern for approximation and streaming algorithms. *Künstliche Intell.*, 32(1):37–53, 2018. doi:10.1007/s13218-017-0519-3.

41 Alexander Munteanu, Chris Schwiegelshohn, Christian Sohler, and David P. Woodruff. On coresets for logistic regression. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 6562–6571, 2018. URL: https://proceedings.neurips.cc/paper/2018/hash/63bfd6e8f26d1d3537f4c5038264ef36-Abstract.html.

42 Shyam Narayanan and Jelani Nelson. Optimal terminal dimensionality reduction in euclidean space. In Moses Charikar and Edith Cohen, editors, *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, pages 1064–1069. ACM, 2019. doi:10.1145/3313276.3316307.

43 Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. Fair coresets and streaming algorithms for fair k-means. In *Approximation and Online Algorithms - 17th International Workshop, WAOA 2019, Munich, Germany, September 12-13, 2019, Revised Selected Papers*, pages 232–251, 2019. doi:10.1007/978-3-030-39479-0_16.

44 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. doi:10.1109/FOCS.2018.00081.

45 Suresh Venkatasubramanian and Qiushi Wang. The johnson-lindenstrauss transform: An empirical study. In Matthias Müller-Hannemann and Renato Fonseca F. Werneck, editors, *Proceedings of the Thirteenth Workshop on Algorithm Engineering and Experiments, ALENEX 2011, Holiday Inn San Francisco Golden Gateway, San Francisco, California, USA, January 22, 2011*, pages 164–173. SIAM, 2011.

46 Dennis Wei. A constant-factor bi-criteria approximation guarantee for k-means++. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 604–612, 2016.

47 Tian Zhang, Raghu Ramakrishnan, and Miron Livny. BIRCH: A new data clustering algorithm and its applications. *Data Min. Knowl. Discov.*, 1(2):141–182, 1997. doi:10.1023/A:1009783824328.