KI - Künstliche Intelligenz (2018) 32:37–53
https://doi.org/10.1007/s13218-017-0519-3

TECHNICAL CONTRIBUTION

Coresets-Methods and History: A Theoreticians Design Pattern for Approximation and Streaming Algorithms

Alexander Munteanu[1] · Chris Schwiegelshohn[1]

Received: 17 February 2017 / Accepted: 4 December 2017 / Published online: 19 December 2017

Abstract
We present a technical survey on the state of the art approaches in data reduction and the coreset framework. These include geometric decompositions, gradient methods, random sampling, sketching and random projections. We further outline their importance for the design of streaming algorithms and give a brief overview on lower bounding techniques.

Keywords Coresets · Sketching · Sampling · Gradient methods · Streaming · Clustering · Regression · Subspace approximation

1 Introduction

More is more is one of the central tenets associated with Big Data. More data means more information from which we hope to gain a better understanding of an underlying truth. The artificial intelligence and machine learning communities have focused on modeling and learning increasingly elaborate statistical models for obtaining new knowledge from the data. However, in the era of Big Data, scalability has become essential for any learning algorithm. In light of this, research has begun to focus on aggregation tools that provide trade-offs between information and space requirement. Note, that these are also extremely valuable for practical applications. The design of a data aggregation can be decoupled from the design of the actual learning algorithm. This is commonly referred to as the sketch-and-solve paradigm. The main idea is to reduce the size of the data first, to have a sketch whose size has only little or even no dependency on the initial size[1]: "Big Data turns into tiny data!" Then a learning algorithm is applied to the little sketch only. Its complexity measures like running time, memory, and communication are thus significantly less dependent on the size of the initial data. In most cases, the learning algorithm remains unchanged. For a given data aggregation, the main challenge is to bound the trade-off between its size and its loss of information.

What constitutes relevant information is, of course, application dependent. Coresets are an algorithmic framework towards quantifying such trade-offs for various objective functions of interest. Generally speaking, we are given a data set A, a set of candidate solutions C, and some optimization function f. Our aim is to find a significantly smaller data set S such that for all candidate solutions $c \in C$, $f(S, c)$ approximates $f(A, c)$.

Coresets turn out be extremely useful in the context of both, approximation and streaming algorithms. Given that a problem admits a coreset whose size is independent of the input, a polynomial time approximation scheme (PTAS) can often be derived via brute force enumeration. While such a PTAS will not necessarily be of practical relevance, this example illustrates how coresets can make computationally expensive algorithms more viable. In a streaming setting, we are given the further restriction that we do not have a random access to the data, but must process the data points one by one. It turns out that if a problem admits a coreset, there exists a black box reduction to streaming algorithms commonly referred to as the merge-and-reduce principle, cf. [49]. It is thus not necessary to develop a special streaming algorithm for that problem.

Coresets have been studied for several base problems that arise in many applications of artificial intelligence and machine learning. Some examples include clustering [4, 17,

✉ Chris Schwiegelshohn
chris.schwiegelshohn@tu-dortmund.de

Alexander Munteanu
alexander.munteanu@tu-dortmund.de

[1] Department of Computer Science, TU Dortmund, 44227 Dortmund, Germany

[1] using the words of [49]

42, 50, 75], classification [56, 59, 89], regression [34, 40, 41, 54], and the smallest enclosing ball problem [6, 15, 16, 48]; we refer to [84] for a recent extensive literature overview. If one of these base problems arises in some application, a coreset construction can often be used in a black box manner to improve the performance of the learning algorithm. This survey, in contrast, is intended to give the reader an overview over existing methods to design and analyze new coreset constructions for individual learning tasks of interest. We believe such an methodological introduction to coresets, and their benefits in the design of fast approximation- and streaming algorithms can be very valuable to the artificial intelligence community.

This survey is now organized into the following parts. In Sect. 2, we describe some basic facts and notations and give a formal definition of the strong coreset guarantee. In Sect. 3, we provide an overview on the four most important techniques used in their construction:

- Geometric decompositions.
- Gradient descent.
- Random sampling.
- Sketching and projections.

For each technique we will consider an example problem and a coreset construction. We further give an overview on results in this line of research. In Sect. 4, we show how coresets may be used in the context of streaming algorithms. In Sect. 5, we discuss lower bounds on coreset sizes, and provide example problems for which coresets do not exist. We conclude with some open problems.

2 Preliminaries

We assume $0 < \varepsilon, \delta \leq \frac{1}{2}$ for all approximation resp. failure parameters in this paper. We first give the definition of the *strong coreset* guarantee.

Definition 1 (*Strong Coresets*) Let A be a subset of points of some universe U and let C be a set of candidate solutions. Let $f : U \times C \to \mathbb{R}^{\geq 0}$ be a non-negative measurable function. Then a set $S \subset U$ is an ε-coreset, if for all candidate solutions $c \in C$ we have

$$|f(A,c) - f(S,c)| \leq \varepsilon \cdot f(A,c).$$

This first paper to formalize the coreset framework is that of Agarwal et al. [4]. In many cases, the construction of a coreset is a randomized process. Here, we account for the probability of failure via the parameter δ, which in turn influences the size of S. We are interested in coresets S whose size is significantly smaller than $|A|$. Specifically, we aim to find an S such that $|S| \in \text{polylog}(|A|)$[2] or even such that $|S|$ has no dependency on $|A|$. The size of $|S|$ will almost always depend on ε, and often on properties of the universe U (e.g. a dependency on d if the points lie d-dimensional Euclidean space) or on properties of f (e.g. a dependency on k, if when we are dealing with the k-means objective function). If there exists no constant $c > 0$ such that $|S| \in o(|A|^{1-c})$, we say that no coreset exists.

There exists another notion of coreset whose use is not as clear cut as the strong coreset. Roughly speaking, we only require that a solution computed on S is a $(1+\varepsilon)$ approximation to the optimum solution of A. In this case, we say that S is a *weak coreset*, though we emphasize that there exist no unifying definition on the guarantees of a weak coreset and there are many variations encountered in literature.

We note that while S often is a subset of A, it does not necessarily have to be and we will see examples for this. In most cases the universe U is a Euclidean space, where the ℓ_2 norm for $\mathbb{R}^d$ is defined as $\|x\| := \sqrt{\sum_{i=1}^{d} |x_i|^2}$.

We will frequently use a standard result on ball-covers for Euclidean spaces. An ε-ball-cover of the unit sphere is a set of points B, such that for any point p in the unit sphere the distance to any point of B is at most ε.

Lemma 1 ([85]) *Let U be the unit sphere in d-dimensional Euclidean space. Then for every* $0 < \varepsilon < 1$ *there exists an ε-ball-cover B of size* $\left(1 + \frac{2}{\varepsilon}\right)^d$, *i.e. for every point* $p \in U$

$$\min_{b \in B} \|p - b\| \leq \varepsilon.$$

We note that while this bound is sharp, there exists no known efficient method of constructing such a ball-cover. In most cases, we require only the existence of a small ball-cover for the analysis. If an algorithm uses an actual ball-cover, there are multiple options to construct one using $\varepsilon^{-O(d)}$ points, see Chazelle for an extensive overview [26].

3 Construction Techniques

We have identified four main approaches used to obtain coresets. This is not meant to be an absolute classification of algorithms, but rather an overview of what techniques are currently available and an illustration of how we may apply them.

[2] $\text{polylog}(|A|) = \bigcup_{c>1} O(\log^c |A|)$

3.1 Geometric Decompositions

Assume that we are given a point set A in Euclidean space. A simple way to reduce the size of A is to compute a discretization S of the space, snap each point of A to its nearest neighbor in S, and use the (weighted) set S to approximate the target function. Many of these decompositions are based on packing arguments and range spaces. We will illustrate one approach via the k-means problem, based on the papers by Har-Peled and Mazumdar [58] and Fichtenberger et al. [51].

Definition 2 (*k-Means Clustering*) Let A be set of n points in Euclidean space and let C be a set of points with $|C| = k$. The k-means objective asks for a set C minimizing

$$f(A,C) := \sum_{x\in A} \min_{c\in C} \|x-c\|^2.$$

We first take a closer look at the objective function for Euclidean k-means. Let OPT be the value of the optimum solution. We first compute a 10-approximation[3] C' to the k-means problem, which can be done in polynomial time [68].

Consider now a sequence of balls with exponentially increasing radius centered around each point of C', starting at radius $\frac{1}{n}\cdot\mathrm{OPT}$ and ending at $10\cdot\mathrm{OPT}$. Our discretization will consist of a suitably scaled ε-ball-cover of each ball. To prove correctness, we require the following generalization of the triangle inequality.

Lemma 2 (Generalized Triangle Inequality [71]) *Let a, b, c be points in Euclidean space. Then for any $\varepsilon \in (0,1)$ we have*

$$\left|\|a-c\|^2 - \|b-c\|^2\right| \le \frac{12}{\varepsilon}\cdot\|a-b\|^2 + 2\varepsilon\cdot\|a-c\|^2.$$

We now show that the cost of using the discretization is bounded.

Lemma 3 *Let A be a set of points, let B^i be the ball with radius $r_i := \frac{2^i}{n}\cdot\sum_{x\in A}\|x\|^2$ centered at the origin and let S^i be the $\frac{\varepsilon}{3}$-ball-cover of B^i. Denote by $S = \bigcup_{i=0}^{\log 10n} S^i$. Then $\sum_{x\in A} \min_{s\in S} \|x-s\|^2 \le \varepsilon^2\cdot\sum_{x\in A}\|x\|^2$.*

Proof Denote by A_{close} the points with squared Euclidean norm at most $\frac{1}{n}\cdot\sum_{x\in A}\|x\|^2$ and by A_{far} the remaining points. Since $|A_{close}| \le n$, we have $\sum_{x\in A_{close}} \min_{s\in S^0} \|x-s\|^2 \le |A_{close}|\cdot\frac{1}{n}\cdot\sum_{x\in A}\|x\|^2\cdot\frac{\varepsilon^2}{9} \le \frac{\varepsilon^2}{9}\cdot\sum_{x\in A}\|x\|^2$. For the points in A_{far}, consider any point x in $B^i\setminus B^{i-1}$ for $i\in\{1,\ldots,\log 10n\}$. We have $\min_{s\in S^i} \|x-s\|^2 \le \frac{\varepsilon^2}{9}r_i^2 \le \frac{4\varepsilon^2}{9}r_{i-1}^2 \le \frac{4\varepsilon^2}{9}\cdot\|x\|^2$. Summing up over all points, we have $\sum_{x\in A} \min_{s\in S} \|x-s\|^2 \le \frac{\varepsilon^2}{9}\cdot\sum_{x\in A}\|x\|^2 + \frac{4\varepsilon^2}{9}\cdot\sum_{x\in A_{far}}\|x\|^2 < \varepsilon^2\cdot\sum_{x\in A}\|x\|^2$.

□

We can reapply this analysis for each center of the 10-approximation C'. Observe that the cost of points A_c assigned to a center $c\in C'$ is $\sum_{x\in A_c}\|x-c\|^2$, which is equal to $\sum_{x\in A}\|x\|^2$ if we treat c as the origin. Combining this lemma with Lemma 2 and rescaling ε shows that S is a coreset.

Theorem 1 *For any set of n points A Euclidean space, there exists a coreset for k-means consisting of $O(k\varepsilon^{-d}\log n)$ points, where d is the (constant) dimension.*

Proof For each of the k centers from the original 10-factor approximation we have a total of $\log 10n$ balls with varying radii. For each such ball of radius r, we compute an $\frac{\varepsilon}{16}\cdot r$ ball-cover. For any point $x\in A$, let $B(x)$ be the nearest point in the union of all ball-covers. By Lemma 3, we have $\sum_{x\in A}\|x-B(x)\|^2 \le \left(\frac{\varepsilon}{16}\right)^2\cdot 10\cdot\mathrm{OPT}$. Now consider an arbitrary set of centers C. We have

$$\begin{aligned}
&\left|\sum_{x\in A}\min_{c\in C}\|x-c\|^2 - \sum_{x\in A}\min_{c\in C}\|B(x)-c\|^2\right| \\
&\le \frac{12}{\varepsilon}\sum_{x\in A}\|x-B(x)\|^2 + 2\varepsilon\cdot\sum_{x\in A}\min_{c\in C}\|x-c\|^2 \\
&\le \frac{12}{\varepsilon}(\varepsilon/16)^2\cdot 10\cdot\mathrm{OPT} + 2\varepsilon\cdot\sum_{x\in A}\min_{c\in C}\|x-c\|^2 \\
&< 2\varepsilon\cdot\mathrm{OPT} + 2\varepsilon\cdot\sum_{x\in A}\min_{c\in C}\|x-c\|^2 \\
&\le 4\varepsilon\cdot\sum_{x\in A}\min_{c\in C}\|x-c\|^2,
\end{aligned}$$

where the first inequality follows from Lemma 2, the second inequality follows follows from Lemma 1, and the last inequality follows from $\mathrm{OPT}\le\sum_{x\in A}\min_{c\in C}\|x-c\|^2$ for any set of centers C.

Rescaling ε by a factor of 1 / 4 gives the proof. The space bound now follows from Lemma 1 and the fact that we compute a $(\varepsilon/64)$-ball-cover $k\cdot\log(10n)$ times. □

Bibliographic Remarks Algorithms based on a discretization and then snapping points to the closest point of the discretization are particular common for extent approximation problems such as maintaining ε-kernels, the directional width, the diameter, and the minimum enclosing ball

[3] Any constant factor would do. We only fix the constant for sake of exposition.

problem. In fact, the notion behind coresets was introduced in the context of these problems in the seminal paper by Agarwal et al. [4]. Since then, a number of papers have progressively reduced the space bound required to maintain coresets [8, 23, 24, 99] with Arya and Chan giving an algorithm that stores $O(\varepsilon^{-(d-1)/2})$ points [14] in d-dimensional Euclidean space. We will briefly outline in Sect. 5 that this space bound is indeed optimal.

The geometric approach was also popular when coresets were introduced to k-median and k-means clustering and generalizations, see for instance [44, 51, 52, 57, 58]. Due to the exponential dependency inherent in all known constructions, the focus later shifted to sampling. Nevertheless, geometric arguments for correctness proofs are necessary for almost all algorithms, and are present in all the other examples presented in this survey. For an extensive and more complete overview of purely geometric algorithms for coreset construction, we recommend the survey by Agarwal et al. [5]

3.2 Gradient Descent

A quite popular and often only implicitly used technique to build coresets is derived from convex optimization. We begin with an example where a simple application of the well-known sub-gradient method yields the desired coreset. Consider the problem of finding the smallest enclosing ball of a set of points, which is equivalent to the 1-center problem in Euclidean space.

Definition 3 Given $P \subset \mathbb{R}^d$, the smallest enclosing ball problem (SEB) consist in finding a center $c^* \in \mathbb{R}^d$ that minimizes the cost function

$$f(P,c) = \max_{p \in P} \|c-p\|.$$

A standard approach for minimizing the convex function f is to start at an arbitrary point c_0 and to perform iterative updates of the form $c_{i+1} = c_i - s \cdot g(c_i)$ where $s \in \mathbb{R}^{>0}$ is an appropriately chosen step size and $g(c_i) \in \partial f(c_i)$ is a sub-gradient of f at the current point c_i. We will refer to this as the *sub-gradient method*. This method has been formally developed by several mathematicians in the sixties. See [83, 90, 92] for details and historical remarks. Now, consider the following theorem regarding the convergence behavior of the sub-gradient method.

Theorem 2 ([83]) *Let f be convex and Lipschitz continuous with constant L and $\|c_0 - c^*\| \le R$. Let $f^* = f(c^*)$ be the optimal solution. Let $f_l^* = \min_{i=1..l}$ be the best solution among l iterations of the sub-gradient algorithm with step size $s = \frac{R}{\sqrt{l+1}}$. Then*

$$f_l^* - f^* \le \frac{LR}{\sqrt{l+1}}.$$

From Theorem 2 we learn, that if our function has a small Lipschitz constant and our initial starting point is not too far from the optimal solution, then the best center among a small number of iterations of the sub-gradient algorithm will have a radius that is close to the optimal radius of the smallest enclosing ball. To assess the parameters more closely we begin with finding a sub-gradient at each step. It is easy to see that for any $p_{\max} \in \operatorname{argmax}_{p \in P} \|c-p\|$ attaining the maximum distance $g(c) = \frac{c-p_{\max}}{\|c-p_{\max}\|}$ is a sub-gradient of f at c, i.e. $g(c) \in \partial f(c)$. Also note that $c - p_{\max} \neq 0$ unless P is a singleton, in which case the problem is trivial since the only input point is the optimal center. Thus, in all interesting cases, $g(c)$ is well-defined. Also note that $g(c)$ is by definition a normalized vector, i.e., $\|g(c)\| = 1$ which implies that f is L-Lipschitz with $L = 1$, since by definition of a sub-gradient and applying the Cauchy-Schwarz inequality (CSI) we have

$$f(c) - f(x) \le g(c)^T(c-x) \overset{\text{CSI}}{\le} \|g(c)\|\|c-x\| \le 1 \cdot \|c-x\|.$$

This leads to the following corollary.

Corollary 1 *We can choose a starting point and an appropriate step size, such that f_l^*, the best solution among $l = O(\frac{1}{\varepsilon^2})$ iterations of the sub-gradient method for the smallest enclosing ball problem satisfies*

$$f_l^* - f^* \le \varepsilon f^*.$$

Proof Note that if we pick an arbitrary input point as our starting point $c_0 = p_0 \in P$ and another point $p_1 \in P$ that maximizes the distance to p_0 then we can deduce that $\|p_0 - c^*\| \le f^* \le \|p_0 - p_1\| := R \le 2f^*$. So, in the light of Theorem 2, R is a good approximation for determining an appropriate step size, and at the same time acts as a good parameter for the upper bound on the error. Now plugging $l = \frac{4}{\varepsilon^2} - 1 = O(\frac{1}{\varepsilon^2})$ and $s = \frac{R}{\sqrt{l+1}} = \frac{\varepsilon}{2}R$ into Theorem 2, we get

$$f_l^* - f^* \le \frac{LR}{\sqrt{l+1}} \le \frac{2f^*}{\sqrt{l+1}} \le \varepsilon f^*.$$

□

Now we would like to argue that the set of points S, that the sub-gradient algorithm selects in each iteration as furthest point from the current center, forms a (weak) coreset for the smallest enclosing ball of P. We will need the following well-known fact for our further analysis.

Lemma 4 ([17]) *Let $B(c, r)$ be the smallest enclosing ball of a point set $P \subset \mathbb{R}^d$, then any closed halfspace that contains c must also contain at least one point from P that is at distance r from c.*

Using this lemma we can argue that any $(1+\varepsilon)$-approximate center must be within $\sqrt{\varepsilon f^*}$ distance to the optimal center.

Lemma 5 *Let $B(c, r)$ be the smallest enclosing ball of a point set $P \subset \mathbb{R}^d$. Let $\tilde{c}$ be a center such that for all $p \in P$ we have $\|c - p\| \leq (1+\varepsilon)r$. Then $\|\tilde{c} - c\| \leq \sqrt{3\varepsilon} r$.*

Proof Let $\tilde{c} - c$ be normal to the hyperplane H passing through c and let H^- be the halfspace whose border is H containing c but not containing $\tilde{c}$. By Lemma 4 we know that there must be some $p \in P \cap H^-$ such that $\|c - p\| = r$. By the law of cosines we have $\|\tilde{c} - c\|^2 = \|\tilde{c} - p\|^2 - \|p - c\|^2 + 2\|\tilde{c} - c\|\|c - p\| \cos\alpha$, where α is the angle between $\tilde{c} - c$ and $p - c$. We further know that $\cos\alpha \leq 0$ since $p \in H^-$. Thus, $\|\tilde{c} - c\|^2 \leq (1+\varepsilon)^2 r^2 - r^2 = (2\varepsilon + \varepsilon^2) r^2 \leq 3\varepsilon r^2$. Taking the square root yields the claim. □

Now we are ready to prove the main result, namely that running $O(\frac{1}{\varepsilon^4})$ iterations, S is indeed a weak ε-coreset for the smallest enclosing ball of P.

Theorem 3 *Let S be the set of points that the sub-gradient method chooses as furthest points from the current center in each of $O(\frac{1}{\varepsilon^4})$ iterations. Then S is an ε-coreset for the smallest enclosing ball of P.*

Proof Note, that starting with the same point c_0, the decisions made by the algorithm on input S is the same as on input P. Therefore Corollary 1 applies to both of the sets S and P. (Ties in furthest point queries are assumed to be resolved lexicographically.) Now consider the optimal center c^* for P and the optimal center c_S for S as well as the best solution $\tilde{c}$ found by the sub-gradient method. We can run the gradient algorithm for a number of $l = O(\frac{1}{\varepsilon^4})$ iterations to get a center $\tilde{c}$ which is a $(1+\varepsilon^2)$-approximation for both sets. By Lemma 5 we have that $\|c^* - \tilde{c}\| \leq \sqrt{3\varepsilon f^*}$ and $\|c_S - \tilde{c}\| \leq \sqrt{3\varepsilon f^*}$. So, by the triangle inequality $\|c^* - c_S\| \leq \|c_S - \tilde{c}\| + \|c^* - \tilde{c}\| \leq 2\sqrt{3\varepsilon f^*}$. Rescaling ε by a factor of $\frac{1}{2\sqrt{3}}$ and again leveraging the triangle inequality yields $\forall p \in P : \|c_S - p\| \leq \|c_S - c^*\| + \|c^* - p\| \leq (1+\varepsilon) f^*$ □

Bibliographic Remarks The presented result is rather weak compared to the optimal coreset size of $\lceil \frac{1}{\varepsilon} \rceil$, cf. [16]. However, we chose to present the method in this way since it shows that the gradient method in its basic form can already yield constant size coresets that do neither depend on the number of input points nor on their dimension. Therefore we see this as a generic method for dimensionality reduction and the design of coresets for computational problems. Putting more effort into geometrically analyzing every single step of the gradient algorithm, i.e., leveraging more problem specific structure in the proof of Nesterov's theorem (Thm. 2, cf. [83]), can lead to even better results which we want to survey here. Badoiu and Clarkson present in their short and elegant seminal work [15] that their gradient algorithm converges to within εf^* error in $O(\frac{1}{\varepsilon^2})$ iterations similar to Corollary 1. Their analysis is stronger in the sense that it implicitly shows that the points selected by the algorithm form an ε-coreset of size $O(\frac{1}{\varepsilon^2})$. Such a result was obtained before in [17] by picking a point in each iteration which is *far enough* away from the center of the smallest enclosing ball of the current coreset. Taking the *furthest point* instead, yields the near-optimal bound of $\frac{2}{\varepsilon}$ from [15]. The complexity was settled with matching upper and lower bounds of $\lceil \frac{1}{\varepsilon} \rceil$ in [16]. The smallest enclosing ball problem is by far not the only one for which looking at gradient methods can help. For instance Har-Peled et al. [60] have used similar construction methods to derive coresets for support vector machine training. Clarkson [29] has generalized the method by unifying and strengthening the aforementioned results (among others) into one framework for coresets, sparse approximation and convex optimization. A probabilistic sampling argument was already used in [17] as a dimensionality reduction technique for the 1-median problem which led to the first linear time approximation algorithms for the k-median problem in Euclidean and more general metric spaces [3, 69]. The method itself is closely related to stochastic sub-gradient methods where uniform sub-sampling is used for obtaining an unbiased estimator for the actual sub-gradient in each step [92]. Similar approaches are currently under study to reduce the exponential dependency on the dimension for the probabilistic smallest enclosing ball problem [48].

3.3 Random Sampling

The arguably most straightforward way to reduce the size of dataset is to simply pick as many points as permissible uniformly at random. Though it is rare for an algorithm to produce a coreset in this straightforward fashion, we will see that for outlier-resistant problems this approach can already provide reasonable results. In the following we will examine the geometric median, which is one of the few problems for which uniform sampling gives any form of guarantee. Thereafter, we will show how to obtain strong coresets using more involved sampling distributions.

Definition 4 (*Geometric Median*) Let A be a set of points in $\mathbb{R}^d$. The geometric median $m(A)$ minimizes

$$f(A,m) := \sum_{x\in A} \|x - m(A)\|.$$

We will show that uniform sampling is sufficient to obtain a weak-coreset guarantee, The exposition for uniform sampling follows Thorup [94] and Ackerman et al. [3].

Lemma 6 *Let A be a set of points in $\mathbb{R}^d$ and let S be a uniform sample of A. Then for any point b with $\sum_{x\in A}\|x-b\| \geq (1+\frac{4\varepsilon}{5})\sum_{x\in A}\|x-m(A)\|$, we have*

$$\mathbb{P}\left[\sum_{x\in S}\|x-b\| < \sum_{x\in S}\|x-m(A)\| + \frac{\varepsilon|S|}{5|A|}\sum_{x\in A}\|x-m(A)\|\right] \leq \exp\left(-\frac{\varepsilon^2|S|}{144}\right).$$

Proof We have $\sum_{x\in A}(\|x-b\| - \|x-m(A)\|) > \frac{4\varepsilon}{5}\sum_{x\in A}\|x-m(A)\|$ and $\sum_{x\in A}(\|x-b\| - \|x-m(A)\|) > \frac{4\varepsilon}{5}/(1+\frac{4\varepsilon}{5})\sum_{x\in A}\|x-b\|$. Then

$$\begin{aligned}
&\sum_{x\in A}(\|x-b\| - \|x-m(A)\|)\\
&\quad > \left(3-\frac{4\varepsilon}{5}\right)\frac{\varepsilon}{5}\sum_{x\in A}\|x-m(A)\| + \frac{\varepsilon}{5}\sum_{x\in A}\|x-b\|\\
&\quad \geq \left(2-\frac{4\varepsilon}{5}\right)\frac{\varepsilon}{5}\sum_{x\in A}\|x-m(A)\| + \frac{\varepsilon}{5}\sum_{x\in A}\|p-m(A)\|\\
&\quad \geq \frac{\varepsilon}{5}\sum_{x\in A}\|x-m(A)\| + \frac{\varepsilon}{5}\left(\sum_{x\in A}\left(\|p-m(A)\| + \frac{\varepsilon}{5}\|x-m(A)\|\right)\right)
\end{aligned} \tag{1}$$

Now consider the random variable $X = \sum_{x\in S}\frac{\|x-b\|-\|x-m(A)\|+\|m(A)-b\|}{2\cdot\left(\|m(A)-b\|+\frac{\varepsilon}{5|A|}\sum_{x\in A}\|x-m(A)\|\right)}$. Note that $\sum_{x\in S}\|x-b\| \leq \sum_{x\in S}\|x-m(A)\| + \frac{\varepsilon|S|}{5|A|}\sum_{x\in A}\|x-b\| \Leftrightarrow |X| \leq |S|/2$. Due to the triangle inequality, each summand is in between 0 and 1. Furthermore, $\mathbb{E}[X] = \frac{|S|}{|A|}\sum_{x\in A}\frac{\|x-b\|-\|x-m(A)\|+\|m(A)-b\|}{2\cdot\left(\|m(A)-b\|+\frac{\varepsilon}{5|A|}\sum_{x\in A}\|x-m(A)\|\right)}$.

Using Eq. 1, we have

$$\begin{aligned}
\mathbb{E}[X] =& \frac{|S|}{2}\cdot\frac{1}{|A|}\sum_{x\in A}\frac{\|x-b\|-\|x-m(A)\|+\|m(A)-b\|}{\|m(A)-b\|+\frac{\varepsilon}{5|A|}\sum_{x\in A}\|x-m(A)\|}\\
\geq& \frac{|S|}{2}\cdot\frac{1}{|A|}\sum_{x\in A}\frac{\left(1+\frac{\varepsilon}{5}\right)\cdot\left(\|m(A)-b\|+\frac{\varepsilon}{5|A|}\sum_{x\in A}\|x-m(A)\|\right)}{\|m(A)-b\|+\frac{\varepsilon}{5|A|}\sum_{x\in A}\|x-m(A)\|}\\
=& \frac{|S|}{2}\cdot\left(1+\frac{\varepsilon}{5}\right)
\end{aligned}$$

Applying the Chernoff bound, we have

$$\begin{aligned}
\mathbb{P}[X \leq |S|/2] \leq& \mathbb{P}\left[X \leq \left(1-\frac{\varepsilon}{6}\right)\cdot E[X]\right]\\
\leq& \exp\left(-\frac{\varepsilon^2\cdot\mathbb{E}[X]}{72}\right) \leq \exp\left(-\frac{\varepsilon^2\cdot|S|}{144}\right).
\end{aligned}$$

□

We wish to use Lemma 6 to apply a union bound on all candidate points. In a finite metric consisting of n points, we can set $|S| \geq 144\varepsilon^{-2}\log n$ and are done. In the continuous setting, this is not as straightforward. We will use a suitably scaled ball-cover and argue that this already suffices to prove the claim for all points.

Theorem 4 *Let A be a set of n points in $\mathbb{R}^d$ and let S be a uniform sample of A with $|S| \in \Omega(d\varepsilon^{-2}\log\frac{d}{\varepsilon})$. Let $m(A)$ and $m(S)$ be the geometric medians of A and S, respectively. Then with constant probability, we have*

$$\sum_{x\in A}\|x-m(S)\| \leq (1+\varepsilon)\cdot\sum_{x\in A}\|x-m(A)\|.$$

Proof Denote by $\text{OPT} := \sum_{x\in A}\|x-m(A)\|$. By Markov's inequality and a union bound over S, all points in S will be contained in the ball B of radius $4|S|\frac{\text{OPT}}{n}$ centered around $m(A)$ with probability at least 1 / 4. Obviously, this means that the optimal median $m(S)$ of |S| will be contained in B. Let C be a $\frac{\varepsilon}{5}\frac{\text{OPT}}{n}$-ball-cover of B. Setting $|S| \geq k\cdot d\varepsilon^{-2}\log\frac{d}{\varepsilon}$ for a sufficiently large absolute constant k ensures that Lemma 6 holds for all points in C with probability at least 1 / 4.

Now let $c \in C$ be the closest point in the ball-cover to $m(S)$. We have

$$\begin{aligned}
\sum_{x\in S}\|x-c\| \leq& \sum_{x\in S}\|x-m(S)\| + \frac{\varepsilon|S|}{5|A|}\sum_{x\in A}\|x-m(A)\|\\
\leq& \sum_{x\in S}\|x-m(A)\| + \frac{\varepsilon|S|}{5|A|}\sum_{x\in A}\|x-m(A)\|
\end{aligned}$$

By Lemma 6 we have $\sum_{x\in A}\|x-c\| \leq (1+\frac{4\varepsilon}{5})\cdot\sum_{x\in A}\|x-m(A)\|$. Applying the triangle inequality, then yields $\sum_{x\in A}\|x-m(S)\| \leq \sum_{x\in A}\|x-c\| + n\cdot\|c-m(S)\| \leq (1+\frac{4\varepsilon}{5}+\frac{\varepsilon}{5})\sum_{x\in A}\|x-m(A)\|$. □

Obviously, the uniform sampling approach can not yield a strong coreset even for the geometric median problem. The reason is that the cost can rely on a small constant number of input points even in one dimension. Given any input P of size $n-2$, we can place two additional points $p_{\min} < \min P$ and $p_{\max} > \max P$ arbitrarily far from the points of P without affecting the median. Clearly, any strong coreset must contain these points to preserve the cost for all possible centers, but the probability of sampling one of them is only $\frac{2}{n}$. This implies that a strong coreset based on uniform sampling must have linear size. To overcome this limitation, Langberg and Schulman [71] introduced the notion of sensitivity for families of functions. For simplicity of presentation, we consider only the special case of the geometric median problem. For a much more general view on sensitivity sampling for numerous other functions the reader is referred to the original literature [45, 71]

Definition 5 (*Sensitivity*) We define the sensitivity of a point $x \in P$ as

$$s(x) = \sup_{c \in \mathbb{R}^d} \frac{\|x - c\|}{\sum\limits_{p \in P} \|p - c\|}.$$

We define the total sensitivity as $S(P) = \sum_{x \in P} s(x)$.

Informally, the sensitivity of a point measures how important it is for preserving the cost over all possible solutions. Note, that from the definition

$$\|x - c\| \leq s(x) \sum_{p \in P} \|p - c\| \tag{2}$$

follows for all $x \in P$ and all centers $c \in \mathbb{R}^d$. The total sensitivity will turn out to be a crucial measure for the complexity of sampling from the distribution given by the sensitivities of the input points. We begin with describing the sampling scheme. The following importance or weighted sampling approach yields an unbiased estimator for the cost of any fixed center $f(P, c) = \sum_{x \in P} \|x - c\|$. We sample a point $x \in P$ from the distribution $q(x) = \frac{s(x)}{S(P)}$ and set $T = \frac{\|x-c\|}{q(x)}$. A straightforward calculation of its expected value $\mathbb{E}[T] = \sum \frac{\|x-c\|}{q(x)} \cdot q(x) = \sum\limits_{x \in P} \|x - c\|$ shows that T is unbiased.

Next we bound its variance which is a crucial parameter in deriving concentration results for our sampling procedure.

Lemma 7 $\mathbb{V}[T] \leq (S(P) - 1)\mathbb{E}[T]^2$

Proof Let $S = S(P)$. Now $\frac{1}{\mathbb{E}[T]^2}\mathbb{V}[T]$ equals

$$\begin{aligned}
&\frac{1}{\mathbb{E}[T]^2} \sum_{x \in P} \left(\frac{\|x - c\|}{q(x)} - \mathbb{E}[T] \right)^2 \cdot q(x) \\
&= \frac{1}{\mathbb{E}[T]^2} \sum_{x \in P} \left(\frac{\|x - c\| \cdot S}{s(x)} - \mathbb{E}[T] \right)^2 \cdot \frac{s(x)}{S} \\
&= \frac{1}{\mathbb{E}[T]^2} \sum_{x \in P} \left(\frac{\|x - c\|^2 \cdot S}{s(x)} - 2\mathbb{E}[T]\|x - c\| + \frac{s(x)\mathbb{E}[T]^2}{S} \right) \\
&= \left(\frac{1}{\mathbb{E}[T]^2} \sum_{x \in P} \frac{\|x - c\|^2 \cdot S}{s(x)} \right) - \left(\frac{2}{\mathbb{E}[T]} \sum_{x \in P} \|x - c\| \right) + \left(\sum_{x \in P} \frac{s(x)}{S} \right) \\
&= \left(\frac{1}{\mathbb{E}[T]^2} \sum_{x \in P} \frac{\|x - c\|^2 \cdot S}{s(x)} \right) - 2 + 1 \leq \frac{S}{\mathbb{E}[T]} \sum_{x \in P} \|x - c\| - 1 = S - 1\,.
\end{aligned}$$

The inequality follows from Eq. (2). □

In our next lemma we will see why the total sensitivity plays such an important role for the random sampling proportional to the sensitivities.

Lemma 8 *Let $\varepsilon > 0$. Let R be a random sample of size $m \geq \frac{3S}{\varepsilon^2} \ln(\frac{2\kappa}{\delta})$ drawn i.i.d. from P proportional to the distribution q. Then for any fixed set of centers $C \subset \mathbb{R}^d$ of size $|C| \leq \kappa$ we have*

$$\mathbb{P}\left[\exists c \in C : \left| \frac{1}{m} \sum_{x \in R} \frac{\|x - c\|}{q(x)} - \sum_{x \in P} \|x - c\| \right| \geq \varepsilon \sum_{x \in P} \|x - c\| \right] \leq \delta.$$

Proof Fix any center $c \in \mathbb{R}^d$. Let $T_i = \frac{\|x-c\|}{q(x)} = \frac{\|x-c\|S}{s(x)}$ with probability $q(x)$ and let $T = \sum_{i=1}^m T_i$ be their sum. By linearity we have $\mathbb{E}[T] = m\mathbb{E}[T_i] = m \sum_{x \in P} \|x - c\|$. By independence of the individual samples and Lemma 7 we have $\mathbb{V}[T] = m\mathbb{V}[T_i] \leq m(S - 1)\mathbb{E}[T_i]^2$. Note that by Eq. (2) it holds that $0 \leq T_i = \frac{\|x-c\|S}{s(x)} \leq S \sum_{x \in P} \|x - c\| = S\mathbb{E}[T_i]$ and therefore $|T_i - \mathbb{E}[T_i]| \leq M := S\mathbb{E}[T_i]$ is a bound that holds almost surely for each i. Now, by an application of Bernstein's inequality [20] we have $\mathbb{P}[|T - \mathbb{E}[T]| > \varepsilon\mathbb{E}[T]] \leq 2 \exp\left(-\frac{\varepsilon^2\mathbb{E}[T]^2}{2\mathbb{V}[T] + \frac{2M}{3}\varepsilon\mathbb{E}[T]} \right)$.We bound the exponent by

$$\begin{aligned}
\frac{\varepsilon^2\mathbb{E}[T]^2}{2\mathbb{V}[T] + \frac{2M}{3}\varepsilon\mathbb{E}[T]} &\geq \frac{\varepsilon^2 m^2 \mathbb{E}[T_i]^2}{2m(S - 1)\mathbb{E}[T_i]^2 + \frac{2}{3}\varepsilon m S\mathbb{E}[T_i]^2} \\
&\geq \frac{\varepsilon^2 m}{2(S - 1) + \frac{2}{3}\varepsilon S} \geq \frac{\varepsilon^2 m}{3S} \geq \ln\left(\frac{2\kappa}{\delta} \right).
\end{aligned}$$

It follows that $\mathbb{P}[|T - \mathbb{E}[T]| > \varepsilon\mathbb{E}[T]] \leq \frac{\delta}{\kappa}$ which implies our claim by taking a union bound over the set C of size $|C| \leq \kappa$. □

Lemma 8 shows that the number of samples that we need, depends linearly on the total sensitivity as well as logarithmically on the number of centers for which we need a guarantee. First we want to bound the total sensitivity for the geometric median problem. If for every input point, there is some center such that the points' contribution to the cost is large, say bounded below by a constant, then we are lost since in that case the total sensitivity sums up to $\Omega(n)$. The next lemma shows that this cannot happen. Actually it turns out that the total sensitivity can be bounded by a constant.

Lemma 9 $S(P) = \sum_{x\in P} s(x) \le 6$.

Proof Let c^* be the optimal center and let $\Delta = \sum_{x\in P} \|x - c^*\| > 0$ denote the optimal cost. Let $c \in \mathbb{R}^d$ be any center, let $\Gamma = \sum_{x\in P} \|x - c\| \ge \Delta$ denote its cost and let $\rho = \|c - c^*\|$ be its distance to the optimal center. Now consider a ball B centered at c^* with radius $\frac{2\Delta}{n}$. By the Markov inequality we have that the number of points inside the ball is at least $|P \cap B| \ge \frac{n}{2}$. Therefore we have $\Gamma \ge \frac{n}{2}\max(0, \rho - \frac{2\Delta}{n})$. Combining both lower bounds we have $\Gamma \ge \frac{1}{2}[\frac{n}{2}\max(0, \rho - \frac{2\Delta}{n}) + \Delta]$. Now,

$$\begin{aligned} s(x) =& \sup_{c\in\mathbb{R}^d} \frac{\|x-c\|}{\sum_{x\in P}\|x-c\|} \le \sup_{c\in\mathbb{R}^d} \frac{\rho + \|x - c^*\|}{\frac{1}{2}[\frac{n}{2}\max(0, \rho - \frac{2\Delta}{n}) + \Delta]} \\ =& \sup_{\rho\ge 0} \frac{\rho + \|x-c^*\|}{\frac{1}{2}[\frac{n}{2}\max(0,\rho - \frac{2\Delta}{n}) + \Delta]} = \sup_{\rho \ge \frac{2\Delta}{n}} \frac{\rho + \|x - c^*\|}{\frac{1}{2}[\frac{n}{2}(\rho - \frac{2\Delta}{n}) + \Delta]} \\ =& \sup_{\rho\ge\frac{2\Delta}{n}} \frac{4\rho + 4\|x-c^*\|}{n\rho} = \sup_{\rho\ge\frac{2\Delta}{n}} \frac{4}{n} + \frac{4\|x-c^*\|}{n\rho} := \sup_{\rho\ge\frac{2\Delta}{n}} G(\rho). \end{aligned}$$

G, restricted to $\rho \ge \frac{2\Delta}{n}$ is clearly monotonously decreasing in ρ and is thus maximized at $G(\frac{2\Delta}{n}) = \frac{4}{n} + \frac{2\|x-c^*\|}{\Delta}$. We can conclude that $S(P)$ is bounded by $\sum_{x\in P} s(x) \le \sum_{x\in P} \left(\frac{4}{n} + \frac{2\|x-c^*\|}{\Delta}\right) = 4 + 2\,\frac{\sum_{x\in P}\|x-c^*\|}{\Delta} = 6$. □

Now we know that the total sensitivity contributes only a constant factor to our sampling complexity. Before we move to the main result of this section, we will need another technical lemma. It is not immediately clear that we can use the triangle inequality for our samples, due to reweighting. We will therefore establish a relaxation of the triangle inequality for the entire weighted subsample.

Lemma 10 *Let $\varepsilon > 0$. Let R be a random sample of size $m \ge \frac{18}{\varepsilon^2}\ln(\frac{2}{\delta})$ drawn i.i.d. from P proportional to the distribution q and reweighted by $w(x) = \frac{1}{mq(x)}$. Then with probability at least $1 - \delta$ for any choice of two centers $c, c' \in \mathbb{R}^d$ we have*

1. $\mathrm{cost}(R, c) \le \mathrm{cost}(R, c') + (1+\varepsilon)n\|c - c'\|$
2. $\mathrm{cost}(R, c) \ge (1-\varepsilon)n\|c - c'\| - \mathrm{cost}(R, c')$.

Proof The tricky part is to bound the total weight of the sample $\frac{1}{m}\sum_{x\in R}\frac{1}{q(x)}$. In the light of Lemmas 7 and 8, we define the random variables $T_i = \frac{1}{q(x)}$. Their expectation is $\mathbb{E}[T_i] = \sum_{x\in P}\frac{1}{q(x)}\cdot q(x) = n$. To bound their variance, we first need an upper bound on $\frac{1}{s(x)}$. To this end, consider a ball $B \supset P$ containing all input points, centered at the optimal center c^*. Let Δ be the diameter of B and let $b \in B$ be the closest point to c located on the surface of B. Note that $\forall p \in P : \|p - b\| \le \Delta$. Then

$$\begin{aligned} \frac{1}{s(x)} =& \inf_{c\in\mathbb{R}^d} \frac{\sum_{p\in P}\|p-c\|}{\|x-c\|} \le \inf_{c\notin B} \frac{\sum_{p\in P}\|p-b\| + \|b-c\|}{\|b-c\|} \\ \le& \inf_{c\notin B}\left(1 + \frac{\Delta}{\|b-c\|}\right)n = n. \end{aligned} \tag{3}$$

Now, $\mathbb{V}[T_i] \le (S-1)\mathbb{E}[T_i]$ follows via very similar calculations as in Lemma 7. Inequality (2) is simply replaced by Inequality (3) in the derivation.

Following the arguments from Lemma 8 closely, we can conclude that with high probability $|\frac{1}{m}\sum_{x\in R}\frac{1}{q(x)} - n| \le \varepsilon n$ holds.

Back to the main claims, the standard triangle inequality yields

$$\begin{aligned} \frac{1}{m}\sum_{x\in R}\frac{\|x-c\|}{q(x)} \le& \frac{1}{m}\sum_{x\in R}\frac{\|x-c'\| + \|c-c'\|}{q(x)} \\ \le& \frac{1}{m}\sum_{x\in R}\frac{\|x-c'\|}{q(x)} + (1+\varepsilon)n\|c-c'\|. \end{aligned}$$

The second claim follows similarly using $\|x - c\| \ge \|c - c'\| - \|x - c'\|$ in the numerator. □

It remains to bound the number of centers for which we will need a guarantee. This seems impossible at first glance, since the strong coreset property asks to hold for every center $c \in \mathbb{R}^d$ which has infinite cardinality. In the following we will see, that again, it is sufficient to consider a ball of appropriate radius centered at the optimum and decompose it by an ε-ball-cover. Lemma 8 is applied to these points only. Now, for every center inside the ball, there is a point of the cover close to it for which the coreset guarantee holds. For the other centers we can argue that they are so far from the optimum center that their distance dominates the cost for both, the original point set as well as for the coreset. This will establish the coreset property for every possible center.

Theorem 5 *Let $\varepsilon > 0$, $c \in \mathbb{R}^d$. Let R be a random sample of P of size $m \ge \Theta(\frac{d}{\varepsilon^2}\ln\frac{1}{\varepsilon\delta})$ drawn i.i.d. proportional to the* *distribution q and reweighted by $w(x) = \frac{1}{mq(x)}$. Then R is a strong coreset for the geometric median of P with probability at least $1 - \delta$.*

Proof Let c^* be the optimal center for P and let $\mathrm{OPT} = \mathrm{cost}(P, c^*)$ denote the optimal cost. Consider the closed ball B of radius $r = \frac{\mathrm{OPT}}{\varepsilon n}$ centered at c^*. Let C be the set of center points of a $\varepsilon\frac{\mathrm{OPT}}{n}$-ball-cover of B. For technical reasons we add c^* to C. Recall from Lemma 1 that $|C| \le \kappa \le \left(1 + \frac{2}{\varepsilon^2}\right)^d + 1 = \left(\frac{1}{\varepsilon}\right)^{O(d)}$. We apply Lemma 8 to C with $m = \frac{18}{\varepsilon^2}\ln(\frac{2\kappa}{\delta}) = \Theta(\frac{d}{\varepsilon^2}\ln\frac{1}{\varepsilon\delta})$. Thus we can assume that with probability $1 - \delta$ for all $\tilde{c} \in C$ simultaneously and in particular for c^* we have $|\mathrm{cost}(R, \tilde{c}) - \mathrm{cost}(P, \tilde{c})| \le \varepsilon\, \mathrm{cost}(P, \tilde{c})$. Also we have that the two triangle inequalities from Lemma 10 hold with probability $1 - \delta$.

Now let $c \in \mathbb{R}^d$ be any center. We distinguish between the cases $c \in B$ and $c \notin B$. In the former case let $\tilde{c} \in C$ be the closest center to c in our cover, i.e., $\|c - \tilde{c}\| \le \varepsilon\frac{\mathrm{OPT}}{n}$. Using Lemma 10 we have

$$\begin{aligned}\mathrm{cost}(R, c) \le& \mathrm{cost}(R, \tilde{c}) + (1+\varepsilon)n \cdot \|c - \tilde{c}\|\\ \le& (1+\varepsilon)\,\mathrm{cost}(P, \tilde{c}) + (1+\varepsilon)n \cdot \|c - \tilde{c}\|\\ \le& (1+\varepsilon)\,(\mathrm{cost}(P, c) + n \cdot \|c - \tilde{c}\|) + (1+\varepsilon)n \cdot \|c - \tilde{c}\|\\ \le& (1+\varepsilon)\,\mathrm{cost}(P, c) + 2(1+\varepsilon)n \cdot \|c - \tilde{c}\|\\ \le& (1+\varepsilon)\,\mathrm{cost}(P, c) + 3\varepsilon\,\mathrm{OPT}\\ \le& (1+4\varepsilon)\,\mathrm{cost}(P, c).\end{aligned}$$

The lower bound of $(1 - 4\varepsilon)\,\mathrm{cost}(P, c)$ can be derived similarly and is omitted for brevity of presentation. We are left to deal with the case $c \notin B$. The inequality to keep in mind is $\|c - c^*\| > \frac{\mathrm{OPT}}{\varepsilon n}$ or, equivalently $\mathrm{cost}(P, c^*) = \mathrm{OPT} < \varepsilon n\|c - c^*\|$. From

$$\begin{aligned}\mathrm{cost}(R, c) \le& \mathrm{cost}(R, c^*) + (1+\varepsilon)n\|c - c^*\|\\ \le& (1+\varepsilon)\,(\mathrm{cost}(P, c^*) + n\|c - c^*\|)\\ \le& (1+\varepsilon)(\varepsilon n\|c - c^*\| + n\|c - c^*\|)\\ \le& (1+3\varepsilon)n\|c - c^*\|\end{aligned}$$

and

$$\begin{aligned}\mathrm{cost}(P, c) \ge& n\|c - c^*\| - \mathrm{cost}(P, c^*)\\ \ge& n\|c - c^*\| - \varepsilon n\|c - c^*\|\\ \ge& (1-\varepsilon)n\|c - c^*\|\end{aligned}$$

it follows that $\frac{\mathrm{cost}(R,c)}{\mathrm{cost}(P,c)} \le \frac{1+3\varepsilon}{1-\varepsilon} \le 1 + 8\varepsilon$.

The lower bound of $1 - 3\varepsilon$ can be verified very similarly and is again omitted from the presentation. Rescaling ε and δ by absolute constants concludes the proof. □

Bibliographic Remarks The arguably first paper to use random sampling for coreset construction is due to Chen [27]. Like the result on k-means described in Sect. 3.1, he considered balls of exponential increasing radii. Instead of using a ball-cover, he showed that a uniform random sample suffices to approximate the cost and thereby gave the first coresets for k-means and k-median with polynomial dependency on k, d, ε, and $\log n$. The sensitivity sampling approach was introduced by Langberg and Schulman [71] and further improved by Feldman and Langberg [45] and Braverman, Feldman and Lang [22].

The technique is now one of the crown jewels of coreset construction, giving the best currently available parameters for a wide range of problems including ℓ_p regression, k-means and k-median clustering, and low rank subspace approximation. It has defined a unified framework for several importance sampling approaches from the early literature of sampling based sublinear approximation algorithms and coresets. One of the first such works is due to Clarkson [28] on ℓ_1 regression. After a series of studies on sampling based randomized linear algebra [36–38], the approach could also be adapted for ℓ_2 regression [40, 41] and was later generalized in [34] to ℓ_p regression for $p \in [1, \infty)$. In accordance with the statistical meaning of the sampling weights, they were later called (statistical) leverage scores. It was an open problem proposed in [40] to even approximate the ℓ_2 leverage scores in less time than needed to solve the regression problem exactly. This problem was resolved in [39] via the oblivious random projection approach we are going to detail in the next section. Low rank subspace approximation was often treated implicitly using the same approaches leading to weak coresets, cf. [91] for a technical review. The first strong coresets of polynomial size for ℓ_p subspace approximation are due to Feldman et al. [47] again achieved via weighted sampling techniques related to the notion of sensitivity. More recently, the sampling based methods from randomized linear algebra [41] were leveraged to develop coresets in [79] for dependency networks [61] and Poisson dependency networks [55].

3.4 Sketches and Projections

It is often useful to view a dataset as a matrix, where the ith row corresponds to the ith point. In the following, let us assume that we have n points in d-dimensional Euclidean space, i.e. we are given a matrix $A \in \mathbb{R}^{n \times d}$.

A sketch of A is a linear projection obtained by multiplying A with a sketching matrix $S \in \mathbb{R}^{m \times n}$ for some $m \ll n$. Our goal is to design a sketching matrix such that SA retains the key properties of A. Many of the previous coreset constructions can also be viewed in terms of sketching matrices. For instance, a sampling algorithm can be viewed as choosing a diagonal matrix $S \in \mathbb{R}^{n\times n}$ where the diagonal entries are 1 (or some weight) if the point was picked and 0 otherwise in which case the row can as well be deleted.

We are more interested in *oblivious* sketching matrices, that is the sketching matrix S can be constructed ahead of time without viewing the data. Though it might seem surprising that this yields any results, sketching is now regarded as one of the central tools for streaming. We will illustrate the power of oblivious sketching in the context of subspace embedding and then show how it may be applied to linear regression.

Definition 6 Let $U \in \mathbb{R}^{n\times d}$ be a matrix with orthogonal unit columns and let $\varepsilon > 0$. An ε-subspace embedding of U is a matrix $SU \in \mathbb{R}^{m\times d}$ for $m \ll n$ such that for any vector $x \in \mathbb{R}^d$, we have

$$|\|SUx\|^2 - \|Ux\|^2| \leq \varepsilon \cdot \|x\|^2.$$

The central ingredient is based around the seminal Johnson-Lindenstrauss lemma [64].

Lemma 11 (Distributional Johnson-Lindenstrauss Lemma) *There exists a distribution D over $m \times n$ matrices with $m \in O(\varepsilon^{-2}\log(1/\delta))$ such that for a matrix Π drawn from D and any fixed vector $x \in \mathbb{R}^n$ we have*

$$\mathbb{P}[(1-\varepsilon)\|x\| \leq \frac{1}{\sqrt{m}}\|\Pi x\| \leq (1+\varepsilon)\|x\|] \geq 1-\delta.$$

The classic variant of the distributional Johnson-Lindenstrauss Lemma further states that an adequate distribution independently draws each entry of the matrix as a Gaussian with mean 0 and variance 1. Any distribution with mean 0 and variance 1 may be used, and nowadays, it is more common to draw entries from the Rademacher distribution, where each entry is with probability 1 / 2 either 1 or −1. We will briefly discuss advantages of various sketching matrices at the end of this section.

Lemma 11 now gives us a powerful dimension reduction tool for Euclidean spaces. Consider, for instance, the case where we are given η points in an arbitrary number of dimensions. Each distance between two points can be represented by a vector and there are at most η^2 such vectors. Using Lemma 11, we can achieve an $(1 \pm \varepsilon)$-approximate embedding into $O(\varepsilon^{-2}\log\eta)$ dimensions with probability at least 2 / 3 by setting $m \geq c \cdot \varepsilon^{-2}\log(\eta^2)$, where c is some absolute constant.

For subspaces, the application of the union bound is not as straightforward as there are infinitely many vectors, even in a 1-dimensional subspace. We first observe that any vector x may be rewritten as $\|x\| \cdot \frac{x}{\|x\|}$ and hence it is sufficient to consider vectors with unit norm. This solves the subspace approximation problem in a single dimension, but even in 2 dimensions we have infinitely many unitary vectors. Instead, we show that applying the union bound on ball-covers is sufficient. Recall that there exists an ε-ball-cover of size $(1+2/\varepsilon)^d$ (c.f. Lemma 1). Applying the union bound now gives us $m \geq \frac{c\log(1+2/\varepsilon)^d}{\varepsilon^2} = O(d\varepsilon^{-2}\log\varepsilon^{-1})$. A slightly more detailed analysis will allow us to remove the $\log\varepsilon^{-1}$ factor. The main argument is that we can write every vector as linear combination of the vectors of a 1 / 2-cover. To see this, consider a 1 / 2-cover $\{x_0, \ldots x_{5^d}\}$ and an arbitrary unit vector y. There exists some x_i such that $\|y - x_i\| \leq 1/2$. We then consider the vector $\|y - x_i\|x_j$ closest to $(y - x_i)$. Since $\|\frac{1}{\|y-x_i\|}(y - x_i) - x_j\| \leq 1/2$, we then have $\|(y - x_i) - \|y - x_i\|x_j\| \leq 1/4$. In each subsequent step, we halve the length of the next vector added and by iterating this sequence into infinity, the length of the remaining vector $\lim y - \sum \alpha_i x_i$ is 0.

Hence, there exists a subspace approximation of size $O(d/\varepsilon^2)$. Now let us consider linear regression, where we aim to minimize $\|Ax - b\|$ with $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$. Since all possible vectors $Ax - b$ lie in a $(d+1)$-dimensional subspace spanned by the columns of A and b, an oblivious subspace embedding is a coreset for linear regression.

Theorem 6 *Let $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$. Choose $S \in \mathbb{R}^{m\times n}$ to be a Rademacher matrix, where $m \in O(d\varepsilon^{-2})$. Then with constant probability, for all $x \in \mathbb{R}^d$, we have*

$$(1-\varepsilon)\|Ax - b\| \leq \|S(Ax - b)\| \leq (1+\varepsilon)\|Ax - b\|.$$

Bibliographic Remarks Oblivious subspace embeddings were first introduced by Sarlos [91] both as a means for solving regression and as means for computing low rank approximations. The upper bound of $O(d/\varepsilon^2)$ on the target dimension for oblivious sketches was (implicitly) given in the work by Clarkson and Woodruff [31], and this bound was later proved to be optimal by Nelson and Nguyen [82], though even smaller bounds are possible if the sketch is not oblivious, see for instance [33, 49]. It is worth noting that if we are only interested in approximating the optimum using the sketch for linear regression, somewhat akin to a weak coreset guarantee, a target dimension of $O(d/\varepsilon)$ is sufficient and necessary [31].

Projections and sketches also play an important role for coresets in Euclidean k-means [33, 49]. Cohen et al. [33] showed that an oblivious sketch of target dimension $O(k/\varepsilon^2)$ is cost-preserving. As a corollary, this implies that for any coreset construction, the dependency on d may be replaced by a dependency on k/ε^2.

The computation time for all sketching approaches can be generally regarded as acceptable. The smallest target dimension is achievable by computing the SVD or a sufficiently good low rank approximation [33, 49]. While this can be done in polynomial time, it is more expensive than the oblivious sketching methods we describe in the following. Conceptually, there are two basic approaches to multiply sketching matrices faster. The first one is to improve the sketching dimension. This, however, is known to be impossible for various regimes, see [12, 63, 65, 72] and very recently for any embedding method by Larsen and Nelson [73]. The other direction is to make the sketching matrix sparse.

Sparse matrices tend to distort sparse vectors. Clearly, the Johnson-Lindenstrauss guarantee cannot hold in such cases. To remedy this problem, Ailon and Chazelle [9] proposed to first increase the density of the input vectors by preforming a randomized Fourier Transform before multiplying with a very sparse embedding matrix. This approach, called the *Fast Johnson Lindenstrauss Transform*, was later improved and refined, see for instance [10, 11, 21, 95] and references therein.

The second approach for sparse Johnson Lindenstrauss transforms is based around sophisticated ways of combining Count Sketch estimators. The Count Sketch estimator [25] originally proposed for finding heavy hitters is essentially a single row of a Rademacher matrix. Instead of naively repeating the Count Sketch and averaging the results as in a standard Rademacher matrix, we aim to partition the entries of the vectors that are to be embedded, apply a Count Sketch for each partition and thereafter aggregate the results. The degree of sparsity that may be achieved using various partitioning schemes have been studied in [2, 32, 35, 66, 81]. We want to highlight the work by Clarkson and Woodruff [32] who can achieve an embedding in essentially input sparsity time, however at the cost of slightly larger target dimension $\frac{d^2}{\varepsilon^2}$. The squared dependency on d was also shown to be necessary by Nelson and Nguyen [80].

Another related direction is generalizing to ℓ_p subspace embeddings. A first step was done by Woodruff and Sohler [93] who designed the first subspace embedding for ℓ_1 via Cauchy random variables. The method is in principle generalizable to using p-stable distributions and was improved in [30, 77]. The idea is that the sum of such random variables forms again a random variable from the same type of distribution leading to concentration results for the ℓ_p norm under study. At the same time it is inherently limited to $1 \leq p \leq 2$ as in [77], since no such distributions exist for $p > 2$, cf. [96]. The first attempts to generalize to $p > 2$ [30] had nearly linear size, namely $n/\text{poly}(d)$, which clearly was not satisfying. A remedy came with a manuscript of Andoni [13], who discovered the *max stability* of inverse exponential random variables as a means to embed $\ell_p, p > 2$ with little distortion into ℓ_∞. Combining this with the work on ℓ_2 embeddings of Clarkson and Woodruff [32] culminated in oblivious subspace embeddings for all p (into ℓ_2 resp. ℓ_∞) and only poly(d) distortion [96]. Note, that the embedding dimension for $p > 2$ is $n^{1-2/p}\text{poly}(d)$ which improved upon the previous $n/\text{poly}(d)$ and is close to optimal given the lower bound of $\Omega(n^{1-2/p})$ [86]. The desirable $(1 \pm \varepsilon)$ distortion can be achieved using the embeddings for preconditioning and sampling proportional to the ℓ_p leverage scores [30, 34, 96].

Current research has moved beyond strict algorithmic and optimization related characteristics to the study of statistical properties [76, 88]. In particular, not only maximum likelihood estimators are approximated under random projections. Geppert et al. [54] showed that in important classes of Bayesian regression models, the whole structure of the posterior distribution is preserved. This yields much faster algorithms for the widely applicable and flexible, but at the same time computationally demanding Bayesian machinery.

Recently a series of optimization software based on random projections and sketches have appeared. A parallel least squares regression solver LSRN was developed in [78, 97]. An implementation of some of the presented sketching techniques named RaProR was made available for the statistics programming language R [53, 54, 87].

4 Streaming

Streaming algorithms process the data point by point (or entry by entry) and aim to (approximately) answer queries to the data using as little space as possible. Though we describe the coreset constructions in the last chapters with no streaming implementation in mind, it turns out that if we can find a strong coreset for a problem, there also exists a streaming algorithm with little overhead in terms of space requirement. Once a coreset construction is known for a problem, it is often not necessary to develop a specific streaming algorithm, since one can rely on the following black box reduction:

For most functions[4], coresets have the very useful property of being closed under union, that is, for two point sets, the union of coresets for both point sets is a coreset for the entire point set. To get an idea of how the reduction works, assume that we partition the input sequence into $n/\log n$ batches of size $\log n$. These batches form the leaves of a binary tree of height $h \leq \log n$. Whenever we process an entire batch, we compute a coreset. Whenever we have computed a coreset for two children of a node, we aggregate them by recomputing a coreset of the union of the children

[4] Any function such that for any two disjoint point sets A and B and any candidate solution c we have $f(A, c) + f(B, c) \leq f(A \cup B, c)$. All problems mentioned in this survey have this property.

and storing it in the parent node. The children can be deleted at this point. Thus, we only have to store at most two coresets at each level bounding the number of coresets in memory to $O(\log n)$.

So, this framework comes not for free, but its blow up remains bounded. If we apply the merging step as a black box, the coreset contained at the root node, i.e. the final output of the algorithm will have an approximation guarantee of $(1+\varepsilon)^{\log n}$. Rescaling ε by $2\log n$, we have the desired $\left(1+\frac{\varepsilon}{2\log n}\right)^{\log n} \leq \exp\left(\frac{\varepsilon}{2}\right) \leq (1+\varepsilon)$ approximation ratio. Finally, many known constructions require randomization and have some adjustable failure probability δ. To limit the overall failure probability when processing a stream, δ is rescaled by the number of coreset constructions. Since the space dependency on δ is typically $\log\frac{1}{\delta}$, we incur another factor of $O(\log n)$. In total, the space dependency on $\log n$ is increased by a factor of $\log^{c+p} n$, where c is the exponent of ε in the offline coreset construction and $p = 1$ if the construction is deterministic and $p = 2$ if it is randomized.

Heinrich et al. have extended this streaming construction of coresets to an asymptotic error guarantee of $\varepsilon \to 0$ as $n \to \infty$ while the memory remains bounded by polylog(n). This has led to the notion of *asymptotically exact streaming algorithms* [62].

The framework, called *merge and reduce*, was originally introduced by Bentley and Saxe [19] and first applied to streaming by Agarwal, Har-Peled and Varadarajan in their seminal paper [4]. Nowadays, many papers have given more efficient streaming implementations for their coresets. For extent approximation algorithms and k-center clustering, we now have constructions with no dependency on n, see [8, 23, 24, 99], with the currently best algorithms storing $O(\varepsilon^{-(d-1)/2})$ points [14] for ε-kernels and $O(k\varepsilon^{-d})$ points for k-center [98]. The dependency on the dimension d is exponential for all these algorithms and Agarwal and Sharathkumar [7] showed that no algorithm with polynomial dependency on d can exist, see also Sect. 5, unless one is willing to drop the $(1+\varepsilon)$ guarantee for a weaker fixed constant.

k-median and k-means clustering thus far are more reliant on the merge and reduce technique. Certain geometric decompositions avoid this, see [51, 52], but have a much larger offline space dependency compared to other constructions. See Table 1 for an overview.

The oblivious sketching algorithms avoid the merge and reduce framework entirely and immediately translate to the streaming algorithm. In fact, they can operate even when processing deletions and entry-wise modifications to an input matrix. Li, Nelson and Woodruff [74] showed that that essentially any such streaming algorithm may be reformulated in terms of linear sketches.

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

Algorithm	Offline memory	Streaming memory
Low dimensions		
[58]	$O(k\varepsilon^{-d}\log n)$	$O(k\varepsilon^{-(d+1)}\log^{2d+2} n)$
[57]	$O(k^3\varepsilon^{-(d+1)})$	$O(k^3\varepsilon^{-(d+1)}\log^{d+2} n)$
[52]	$O(k\varepsilon^{-d}\log n)$	$O(k\varepsilon^{-(d+2)}\log^4 n)$
[51]	$O(k\varepsilon^{-(d+2)}\log n)$	$O(k\varepsilon^{-(d+2)}\log n)$
High dimensions		
[27]	$O(d^2k^2\varepsilon^{-2}\log^5 n)$	$O(d^2k^2\varepsilon^{-2}\log^9 n)$
[46]	$O(k^2\varepsilon^{-5})$	$O(k^2\varepsilon^{-5}\log^7 n)$
[71]	$O(d^2k^3\varepsilon^{-2})$	$O(d^2k^3\varepsilon^{-2}\log^4 n)$
[45]	$O(dk\varepsilon^{-4})$	$O(dk\varepsilon^{-4}\log^6 n)$

[46] produces a weak coreset from which an $(1+\varepsilon)$-approximation can be recovered. Any dependency on d may be replaced by $k\varepsilon^{-2}$ via Theorem 12 of Cohen et al. [33]

5 Lower Bounds

Lower bounds for coresets come in two flavors: (1) space complexity in terms of points or bits and (2) impossibility results. We will sketch examples for both. First, let us consider the extent approximation problem.

Definition 7 Let A be a set of points in $\mathbb{R}^d$ and let $\varepsilon > 0$. An extent approximation is a subset S of A such that for any unit vector w, we have

$$\left|\min_{x\in A} w^T x - \max_{x\in A} w^T x\right| \leq (1+\varepsilon)\cdot\left|\min_{x\in S} w^T x - \max_{x\in S} w^T x\right|$$

Arya and Chan gave an algorithm storing $O(\varepsilon^{-(d-1)/2})$ points. We will briefly outline why this is indeed optimal (see Agarwal et al. [5] for details). Consider the unit sphere and the spherical cap with angular radius $\sqrt{\varepsilon}$. The height of this cap is $1-\cos(\sqrt{\varepsilon}) \leq \frac{\varepsilon}{2}$. Thus, the extent approximation must contain at least one point from every cap. The radius of each such cap is $\sin(\sqrt{\varepsilon}) = \Theta(\sqrt{\varepsilon})$. We know that the bound of Lemma 1 is asymptotically tight, i.e., we require $\Omega(\sqrt{\varepsilon}^{-d})$ space for a $\sqrt{\varepsilon}$-ball-cover in d-dimensional space. Combining this with the fact that the unit sphere is a $d-1$-dimensional space, we know that $\Omega(\sqrt{\varepsilon}^{-(d-1)}) = \Omega(\varepsilon^{-(d-1)/2})$ points are necessary.

We now present an impossibility result for logistic regression.

Definition 8 (*Logistic Regression*) Let A be a set of points in $\mathbb{R}^d$. Then logistic regression objective aims to find a vector w minimizing

$$\sum_{x_i \in A} \ln\left(1+\exp(-w^T x_i)\right).$$

Note, that we assume that the label $y_i \in \{-1, 1\}$ of point x_i is already folded into x_i.

Let us first recall the reduction of streaming algorithms to the construction of strong coresets from the last section. Taking it the opposite direction, a lower bound for streaming algorithms gives us a lower bound for strong coresets. Thus, a wide variety of tools from communication complexity becomes available to us. A complete review of all techniques is well out of scope, for further reading we refer to the book by Kushilevitz and Nisan [70]. We will focus on the following communication problem.

The *indexing problem* is a two party communication game, where the first player Alice has a binary bit string $x \in \{0, 1\}^n$ and the second player Bob has an index $k \in \{1, \ldots, n\}$. Alice is allowed to send one message to Bob, whereupon Bob has to output the kth bit. The number of bits of the transmitted message required by any randomized protocol succeeding with probability at least 2 / 3 over the random choices of the players is in $\Omega(n)$, see [1].

We will use the indexing problem to show that no strong coresets for logistic regression exist.

We first show a reduction to the convex hull membership problem. Here, the stream consist of sequence of points A and at any given time we want to be able to answer whether a given point x lies in the convex hull $C(A)$ or not.

Lemma 12 *Let P be a set of n points in $\mathbb{R}^2$. Let A be a subset of P arriving one after the other in a stream. Then any single pass randomized algorithm deciding with probability 2 / 3 whether some point $b \in P$ lies in $C(A)$ requires at least $\Omega(n)$ space.*

Proof Let $x \in \{0, 1\}^n$ be Alice's bit string and let k be Bob's index. For each $i \in \{1, \ldots, n\}$, define the point $p_i = (\sin(i/n), \cos(i/n))$. By construction, all points lie on the unit sphere and therefore p_k is in the convex hull of any point set $\bigcup_{i \in I} p_i$ with $I \subseteq \{1, \ldots, n\}$ if and only if $k \in I$. For each entry $x_i = 1$, Alice constructs p_i. She then runs a streaming algorithm on all generated points and sends the memory of the streaming algorithm to Bob. Bob then checks whether p_k is in the convex hull generated by Alice. Since this solves the indexing problem, the communication complexity of indexing is a lower bound to the space complexity of convex hull membership. □

Corollary 2 *Let A and B be two sets of a total n points in 2-dimensional space arriving in a data stream. Then any single pass randomized algorithm deciding with probability 2 / 3 whether A and B are linearly separable requires at least $\Omega(n)$ space.*

We now return to logistic regression. We have the following theorem.

Theorem 7 *For any $\delta > 0$ and any integer n there exists a set of n points $C = A \cup B$, such that any strong ε-coreset of C for logistic regression must consist of $\Omega(n^{1-\delta})$ points.*

Proof Let A and B be linearly separable, we have at least one misclassified point. The cost of this point is lower bounded by $\ln(1 + \exp(0)) = \ln(2)$. Otherwise, let w be a separating hyperplane. Then $\lim_{\|w\| \to \infty} \sum_{x_i \in A} \ln(1 + \exp(-w^T x_i)) = 0$. Given a single pass randomized algorithm for logistic regression we can distinguish between these two cases. Corollary 2 implies that the space complexity must therefore be $\Omega(n)$. Since the merge and reduce framework incurs a polylog(n) blowup, this implies a lower bound of $\Omega(n/\text{polylog}(n)) \subset \Omega(n^{1-\delta})$ for the space complexity of strong coresets for logistic regression. □

A very similar impossibility result was derived recently for Poisson regression by reduction from communication complexity problems [79]. The paper discusses and demonstrates how coresets can be useful in practice anyway, going beyond the worst-case perspective.

6 Conclusion and Open Problems

We have outlined techniques and limitations of coreset construction methods. To summarize their benefits in a short statement: Coresets are arguably the state of the art technique to turn[5] "Big Data into tiny data!" Their design and analysis should be considered whenever a new statistical model or learning task is designed. They make the algorithmic assessment more efficient saving time, space, communication, and energy, to tackle the most common resource restrictions associated with Big Data.

There are further lines of research employing coresets, concerning topics like privacy issues [43] or distributed computation [18, 67, 97], that we have not covered here, but encourage the interested reader to investigate. In the following, we would like to conclude with three important open problems.

Problem 1 Let A be a set of n points in d-dimensional Euclidean space. Is it possible to deterministically compute

[5] using the words of [49]

a coreset S for the k-median or k-means objective in polynomial time, such that $|S| \in \text{poly}(k, \varepsilon^{-1}, d, \log n)$?

All known coreset constructions for k-median and k-means clustering with polynomial dependency on k, ε^{-1}, d, and $\log n$ are randomized. The best known deterministic constructions are either exponential in d, or exponential in k and ε^{-1}. Since we know that small coresets exist, it is possible to compute them by brute force enumeration, which is deterministic but also clearly infeasible.

Problem 2 Let A be a set of n points in d-dimensional Euclidean space. Is it possible to compute a coreset S for the geometric median objective such that $|S|$ is independent of d and n?

It is a simple exercise to show that coresets with no dependency on n, d, or for that matter ε exist for the centroid or mean of a point set. Indeed, there exist coresets for the more general Euclidean k-means problem of size $\text{poly}(k/\varepsilon)$ [33, 49]. The algorithms and proofs are heavily reliant on connections to linear algebra and in particular the fact that Euclidean k-means is a constrained form of low-rank approximation with respect to the Frobenius norm. The unconstrained low-rank approximation of a matrix can be determined via singular value decomposition (SVD). Computational and mathematical aspects of the SVD are well understood. In contrast, far less is known by the k-median analogue of a low-rank approximation and its computation is invariably harder than the SVD. Currently, we neither know of a lower bound stating that a dependency of d is required for a geometric median coreset, nor of a coreset construction consisting of even $\exp(\varepsilon^{-1})$ many points. It is known that for the minimum enclosing ball (i.e. 1-center clustering), an exponential dependency on d is necessary.

Problem 3 Let A be a subset of points of some universe U and let C be a set of candidate solutions. Let $f : U \times C \to \mathbb{R}^{\geq 0}$ be a non-negative measurable function. Let S be the smallest possible ε-coreset with respect to A and f. Is it possible to always compute a ε-coreset S', such that $|S'| \leq \alpha \cdot |S|$ for some approximation factor α?

Though we have algorithms that are optimal in the worst case for some problems, the worst case complexity may nevertheless be unappealing, such as is the case for extent problems. It would be nice to have algorithms with instance optimal running times and output sizes. To be more specific, consider the minimum enclosing ball problem. Is it possible to compute a coreset S' in time $O(|A| + \alpha \cdot |S|)$ such that $|S'| \leq \alpha \cdot |S|$, where S is the optimal coreset and α a small (ideally constant) factor?

Acknowledgements This work has been supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 “Providing Information by Resource-Constrained Analysis”, projects A2 and C4.

References

1. Ablayev F (1996) Lower bounds for one-way probabilistic communication complexity and their application to space complexity. Theor Comput Sci 157(2):139–159
2. Achlioptas D (2003) Database-friendly random projections: Johnson-lindenstrauss with binary coins. J Comput Syst Sci 66(4):671–687
3. Ackermann MR, Blömer J, Sohler C (2010) Clustering for metric and nonmetric distance measures. ACM Trans Algorithm 6(4):59:1–59:26
4. Agarwal PK, Har-Peled S, Varadarajan KR (2004) Approximating extent measures of points. J ACM 51(4):606–635
5. Agarwal PK, Har-Peled S, Varadarajan KR (2005) Geometric approximation via coresets. Combinatorial and computational geometry, MSRI. Cambridge University Press, Cambridge, pp 1–30
6. Agarwal PK, Sharathkumar R (2015) Streaming algorithms for extent problems in high dimensions. Algorithmica 72(1):83–98. https://doi.org/10.1007/s00453-013-9846-4
7. Agarwal PK, Sharathkumar R (2015) Streaming algorithms for extent problems in high dimensions. Algorithmica 72(1):83–98
8. Agarwal PK, Yu H (2007) A space-optimal data-stream algorithm for coresets in the plane. In: Proceedings of the 23rd ACM Symposium on Computational Geometry, Gyeongju, South Korea, June 6-8, 2007, pp 1–10 (2007)
9. Ailon N, Chazelle B (2006) 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, pp 557–563
10. Ailon N, Liberty E (2008) 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, pp 1–9
11. Ailon N, Liberty E (2013) An almost optimal unrestricted fast johnson-lindenstrauss transform. ACM Trans Algorithms 9(3):21:1–21:12
12. Alon N (2003) Problems and results in extremal combinatorics-i. Discrete Math 273(1–3):31–53
13. Andoni A (2013) High frequency moments via max-stability. Retreived online on 02/16/2017 from http://web.mit.edu/andoni/www/papers/fkStable.pdf
14. Arya S, Chan TM (2014) Better ε-dependencies for offline approximate nearest neighbor search, euclidean minimum spanning trees, and ε-kernels. In: 30th Annual Symposium on Computational Geometry, SOCG’14, Kyoto, Japan, June 08 - 11, 2014, pp 416
15. Badoiu M, Clarkson KL (2003) Smaller core-sets for balls. In: Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2003, pp 801–802

16. Badoiu M, Clarkson KL (2008) Optimal core-sets for balls. Comput Geom 40(1):14–22. https://doi.org/10.1016/j.comgeo.2007.04.002
17. Badoiu M, Har-Peled S, Indyk P (2002) Approximate clustering via core-sets. In: Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC), pp 250–257
18. Balcan M, Ehrlich S, Liang Y (2013) 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., pp 1995–2003
19. Bentley JL, Saxe JB (1980) Decomposable searching problems i: static-to-dynamic transformation. J Algorithms 1(4):301–358
20. Bernstein S (1946) Theory of probabilities. Gostechizdat, Moscow-Leningrad
21. Boutsidis C, Gittens A (2013) Improved matrix algorithms via the subsampled randomized hadamard transform. SIAM J Matrix Anal Appl 34(3):1301–1340
22. Braverman V, Feldman D, Lang H (2016) New frameworks for offline and streaming coreset constructions. CoRR **abs/1612.00889**
23. Chan TM (2002) Approximating the diameter, width, smallest enclosing cylinder, and minimum-width annulus. Int J Comput Geometry Appl 12(1–2):67–85
24. Chan TM (2006) Faster core-set constructions and data-stream algorithms in fixed dimensions. Comput Geom 35(1–2):20–35
25. Charikar M, Chen KC, Farach-Colton M (2004) Finding frequent items in data streams. Theor Comput Sci 312(1):3–15
26. Chazelle B (2001) The discrepancy method—randomness and complexity. Cambridge University Press, Cambridge
27. Chen K (2009) On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. SIAM J Comput 39(3):923–947
28. Clarkson KL (2005) Subgradient and sampling algorithms for ℓ_1 regression. In: Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2005, Vancouver, British Columbia, Canada, January 23-25, 2005, pp 257–266. Society for Industrial and Applied Mathematics, SIAM
29. Clarkson KL (2010) Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. ACM Trans Algorithms 6(4):63:1–63:30
30. Clarkson KL, Drineas P, Magdon-Ismail M, Mahoney MW, Meng X, Woodruff DP (2016) The Fast Cauchy Transform and faster robust linear regression. SIAM J Comput 45(3):763–810
31. Clarkson KL, Woodruff DP (2009) Numerical linear algebra in the streaming model. In: Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC), pp 205–214
32. Clarkson KL, Woodruff DP (2013) Low rank approximation and regression in input sparsity time. In: Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, pp 81–90
33. Cohen MB, Elder S, Musco C, Musco C, Persu M (2015) 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, pp 163–172
34. Dasgupta A, Drineas P, Harb B, Kumar R, Mahoney MW (2009) Sampling algorithms and coresets for ℓ_p regression. SIAM J Comput 38(5):2060–2078
35. Dasgupta A, Kumar R, Sarlós T (2010) 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, pp 341–350
36. Drineas P, Kannan R, Mahoney MW (2006) Fast monte carlo algorithms for matrices I: approximating matrix multiplication. SIAM J Comput 36(1):132–157
37. Drineas P, Kannan R, Mahoney MW (2006) Fast monte carlo algorithms for matrices II: computing a low-rank approximation to a matrix. SIAM J Comput 36(1):158–183
38. Drineas P, Kannan R, Mahoney MW (2006) Fast monte carlo algorithms for matrices III: computing a compressed approximate matrix decomposition. SIAM J Comput 36(1):184–206
39. Drineas P, Magdon-Ismail M, Mahoney MW, Woodruff DP (2012) Fast approximation of matrix coherence and statistical leverage. J Mach Learn Res 13:3475–3506
40. Drineas P, Mahoney MW, Muthukrishnan S (2006) Sampling algorithms for ℓ_2 regression and applications. In: Proc. of SODA, pp 1127–1136. URL http://dl.acm.org/citation.cfm?id=1109557.1109682
41. Drineas P, Mahoney MW, Muthukrishnan S (2008) Relative-error CUR matrix decompositions. SIAM J Matrix Anal Appl 30(2):844–881. https://doi.org/10.1137/07070471X
42. Feldman D, Faulkner M, Krause A (2011) Scalable training of mixture models via coresets. In: Proc. of NIPS
43. Feldman D, Fiat A, Kaplan H, Nissim K (2009) Private coresets. In: Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, May 31 - June 2, 2009, pp 361–370
44. Feldman D, Fiat A, Sharir M (2006) Coresets forweighted facilities and their applications. In: 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21-24 October 2006, Berkeley, California, USA, Proceedings, pp 315–324
45. Feldman D, Langberg M (2011) 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, pp 569–578
46. Feldman D, Monemizadeh M, Sohler C (2007) 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, pp 11–18
47. Feldman D, Monemizadeh M, Sohler C, Woodruff DP (2010) Coresets and sketches for high dimensional subspace approximation problems. In: Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010, pp 630–649
48. Feldman D, Munteanu A, Sohler C (2014) Smallest enclosing ball for probabilistic data. In: 30th Annual Symposium on Computational Geometry, SOCG'14, pp. 214–223
49. Feldman D, Schmidt M, Sohler C (2013) 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, pp 1434–1453
50. Feldman D, Schmidt M, Sohler C (2013) Turning big data into tiny data: Constant-size coresets for k-means, PCA and projective clustering. In: Proc. of SODA, pp 1434–1453
51. Fichtenberger H, Gillé M, Schmidt M, Schwiegelshohn C, Sohler C (2013) BICO: BIRCH meets coresets for k-means clustering. In: Algorithms - ESA 2013 - 21st Annual European Symposium, Sophia Antipolis, France, September 2-4, 2013. Proceedings, pp 481–492
52. Frahling G, Sohler C (2005) Coresets in dynamic geometric data streams. In: Proceedings of the 37th Annual ACM Symposium on Theory of Computing (STOC), pp 209–217
53. Geppert LN, Ickstadt K, Munteanu A, Quedenfeld J, Sohler C (2015) RaProR: Random Projections for Bayesian linear Regression, R-package, Version 1.0 (2015). URL http://ls2-www.cs.uni-dortmund.de/projekte/RaProR/

54. Geppert LN, Ickstadt K, Munteanu A, Quedenfeld J, Sohler C (2017) Random projections for Bayesian regression. Stat Comput 27(1):79–101
55. Hadiji F, Molina A, Natarajan S, Kersting K (2015) Poisson dependency networks: gradient boosted models for multivariate count data. MLJ 100(2–3):477–507
56. Har-Peled S (2015) A simple algorithm for maximum margin classification, revisited. arXiv **1507.01563**. URL http://arxiv.org/abs/1507.01563
57. Har-Peled S, Kushal A (2007) Smaller coresets for k-median and k-means clustering. Discret Comput Geom 37(1):3–19
58. Har-Peled S, Mazumdar S (2004) 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, pp 291–300
59. Har-Peled S, Roth D, Zimak D (2007) Maximum margin coresets for active and noise tolerant learning. In: Proc. of IJCAI, pp 836–841
60. Har-Peled S, Roth D, Zimak D (2007) Maximum margin coresets for active and noise tolerant learning. In: IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007, pp 836–841
61. Heckerman D, Chickering D, Meek C, Rounthwaite R, Kadie C (2000) Dependency networks for density estimation, collaborative filtering, and data visualization. J Mach Learn Res 1:49–76
62. Heinrich M, Munteanu A, Sohler C (2014) Asymptotically exact streaming algorithms. CoRR **abs/1408.1847**
63. Jayram TS, Woodruff DP (2013) Optimal bounds for johnson-lindenstrauss transforms and streaming problems with subconstant error. ACM Trans. Algorithms 9(3):26:1–26:17
64. Johnson WB, Lindenstrauss J (1984) Extensions of Lipschitz mappings into a Hilbert space. Contemp Math 26(189–206):1–1
65. Kane DM, Meka R, Nelson J (2011) 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, pp 628–639
66. Kane DM, Nelson J (2014) Sparser johnson-lindenstrauss transforms. J ACM 61(1):4:1–4:23
67. Kannan R, Vempala S, Woodruff DP (2014) Principal component analysis and higher correlations for distributed data. In: Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014, pp 1040–1057
68. Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY (2004) A local search approximation algorithm for k-means clustering. Comput Geom 28(2–3):89–112
69. Kumar A, Sabharwal Y, Sen S (2010) Linear-time approximation schemes for clustering problems in any dimensions. J ACM 57(2):5:1–5:32. http://dx.doi.org/10.1145/1667053.1667054
70. Kushilevitz E, Nisan N (1997) Communication complexity. Cambridge University Press, New York
71. Langberg M, Schulman LJ (2010) 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, pp 598–607
72. Larsen KG, Nelson J (2016) 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, pp 82:1–82:11
73. Larsen KG, Nelson J (2017) Optimality of the johnson-lindenstrauss lemma. In: 58th IEEE Annual Symposium on Foundations of Computer Science (FOCS), pp 633–638
74. Li Y, Nguyen HL, Woodruff DP (2014) Turnstile streaming algorithms might as well be linear sketches. In: Proceedings of the 46th Annual ACM Symposium on Theory of Computing (STOC), pp 174–183
75. Lucic M, Bachem O, Krause A (2016) Strong coresets for hard and soft bregman clustering with applications to exponential family mixtures. In: Proc. of AISTATS, pp 1–9
76. Ma P, Mahoney MW, Yu B (2015) A statistical perspective on algorithmic leveraging. J Mach Learn Res 16:861–911
77. Meng X, Mahoney MW (2013) 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, pp 91–100
78. Meng X, Saunders MA, Mahoney MW (2014) LSRN: A parallel iterative solver for strongly over—or underdetermined systems. SIAM J Sci Comput 36(2):C95–C118
79. Molina A, Munteanu A, Kersting K (2018) Core dependency networks. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI). AAAI Press
80. Nelson J, Nguyen HL (2013) OSNAP: faster numerical linear algebra algorithms via sparser subspace embeddings. In: 54th Annual IEEE Symposium on Foundations of Computer Science, FOCS, Berkeley, CA, USA, pp 117–126
81. Nelson J, Nguyen HL (2013) Sparsity lower bounds for dimensionality reducing maps. In: Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013, pp 101–110
82. Nelson J, Nguyên HL (2014) 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, pp 883–894
83. Nesterov Y (2004) Introductory lectures on convex optimization: a basic course. Applied optimization. Springer, New York
84. Phillips JM (2017) Coresets and sketches. In: Goodman JE, O'Rourke J, Tóth CD (eds) Handbook of Discrete and Computational Geometry, 3rd edn. Chapman and Hall/CRC, London, pp 1265–1284
85. Pisier G (1999) The volume of convex bodies and Banach space geometry. Cambridge Tracts in Mathematics. 94, Cambridge University Press, Cambridge
86. Price E, Woodruff DP (2012) Applications of the shannon-hartley theorem to data streams and sparse recovery. In: Proceedings of the 2012 IEEE International Symposium on Information Theory, ISIT 2012, Cambridge, MA, USA, July 1-6, 2012, pp 2446–2450
87. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2014). URL http://www.R-project.org
88. Raskutti G, Mahoney MW (2015) Statistical and algorithmic perspectives on randomized sketching for ordinary least-squares. In: Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, pp 617–625
89. Reddi SJ, Póczos B, Smola AJ (2015) Communication efficient coresets for empirical loss minimization. In: Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI 2015, July 12-16, 2015, Amsterdam, The Netherlands, pp 752–761
90. Rockafellar RT (1970) Convex analysis. Princeton mathematical series. Princeton University Press, Princeton
91. Sarlós T (2006) Improved approximation algorithms for large matrices via random projections. In: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS), pp 143–152
92. Shor NZ (1985) Minimization methods for non-differentiable functions. Springer series in computational mathematics. Springer, Berlin. Transl. from the Russian, Kiev, Naukova Dumka, 1979

93. Sohler C, Woodruff DP (2011) Subspace embeddings for the ℓ_1 -norm with applications. In: Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011, pp 755–764
94. Thorup M (2004) Quick k-median, k-center, and facility location for sparse graphs. SIAM J Comput 34(2):405–432
95. Tropp JA (2011) Improved analysis of the subsampled randomized hadamard transform. Adv Adapt Data Anal 3(1–2):115–126
96. Woodruff DP, Zhang Q (2013) Subspace embeddings and ℓ_p -regression using exponential random variables. In: COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA, pp 546–567
97. Yang J, Meng X, Mahoney MW (2016) Implementing randomized matrix algorithms in parallel and distributed environments. Proc IEEE 104(1):58–92
98. Zarrabi-Zadeh H (2008) Core-preserving algorithms. In: Proceedings of the 20th Annual Canadian Conference on Computational Geometry, Montréal, Canada, August 13-15
99. Zarrabi-Zadeh H (2011) An almost space-optimal streaming algorithm for coresets in fixed dimensions. Algorithmica 60(1):46–59