On the Sensitivity of Shape Fitting Problems*

Kasturi Varadarajan[1] and Xin Xiao[2]

1 **Department of Computer Science**
University of Iowa, Iowa City, IA 52242, USA
kasturi-varadarajan@uiowa.edu

2 **Department of Computer Science**
University of Iowa, Iowa City, IA 52242, USA
xin-xiao@uiowa.edu

Abstract

In this article, we study shape fitting problems, ϵ-coresets, and total sensitivity. We focus on the (j,k)-projective clustering problems, including k-median/k-means, k-line clustering, j-subspace approximation, and the integer (j,k)-projective clustering problem. We derive upper bounds of total sensitivities for these problems, and obtain ϵ-coresets using these upper bounds. Using a dimension-reduction type argument, we are able to greatly simplify earlier results on total sensitivity for the k-median/k-means clustering problems, and obtain positively-weighted ϵ-coresets for several variants of the (j,k)-projective clustering problem. We also extend an earlier result on ϵ-coresets for the integer (j,k)-projective clustering problem in fixed dimension to the case of high dimension.

1998 ACM Subject Classification F.2.2 Analysis of Algorithms and Problem Complexity

Keywords and phrases Coresets, shape fitting, k-means, subspace approximation

Digital Object Identifier 10.4230/LIPIcs.FSTTCS.2012.486

1 Introduction

In this article, we study shape fitting problem, coresets, and in particular, total sensitivity. A shape fitting problem is specified by a triple $(\mathbb{R}^d, \mathcal{F}, \mathrm{dist})$, where $\mathbb{R}^d$ is the d-dimensional Euclidean space, $\mathcal{F}$ is a family of subsets of $\mathbb{R}^d$, and $\mathrm{dist} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ is a continuous function that we will refer to as a *distance* function. We also assume that (a) $\mathrm{dist}(p,q) = 0$ if and only if $p = q$, and (b) $\mathrm{dist}(p,q) = \mathrm{dist}(q,p)$. We refer to each $F \in \mathcal{F}$ as a *shape*, and we require each shape F to be a non-empty, closed, subset of $\mathbb{R}^d$. We define the *distance* of a point $p \in \mathbb{R}^d$ to a shape $F \in \mathcal{F}$ to be $\mathrm{dist}(p,F) = \min_{q \in F} \mathrm{dist}(p,q)$. An instance of a shape fitting problem is specified by a finite point set $P \subset \mathbb{R}^d$. We slightly abuse notation and use $\mathrm{dist}(P,F)$ to denote $\sum_{p \in P} \mathrm{dist}(p,F)$ when P is a set of points in $\mathbb{R}^d$. The goal is to find a shape which best fits P, that is, a shape minimizing $\sum_{p \in P} \mathrm{dist}(p,F)$ over all shapes $F \in \mathcal{F}$. This is referred to as the L_1 fitting problem, which is the main focus of this paper. In the L_∞ fitting problem, we seek to find a shape $F \in \mathcal{F}$ minimizing $\max_{p \in P} \mathrm{dist}(p,F)$.

In this paper, we focus on the (j,k)-projective clustering problem. Given non-negative integers j and k, the family of shapes is the set of k-tuples of affine j-subspaces (that is, j-flats) in $\mathbb{R}^d$. More precisely, each shape is the union of some k j-flats. The underlying distance function is usually the z^{th} power of the Euclidean distance, for a positive real number

* This material is based upon work supported by the National Science Foundation under Grant No. 0915543.

32nd Int'l Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2012).
Editors: D. D'Souza, J. Radhakrishnan, and K. Telikepalli; pp. 486–497

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

z. When $j = 0$, $\mathcal{F}$ is the set of all k-point sets of $\mathbb{R}^d$, so the $(0, k)$-projective clustering problem is the k-median clustering problem when the distance function is the Euclidean distance, and it is the k-means clustering problem when the distance function is the square of the Euclidean distance; when $j = 1$, the family of shapes is the set of k-tuples of lines in $\mathbb{R}^d$; when $k = 1$, $(j, 1)$-projective clustering is the subspace approximation problem, where the family of shapes is the set of j-flats. Other than these projective clustering problems where j or k is set to specific values, another variant of the (j, k)-projective clustering problem is the integer (j, k)-projective clustering problem, where we assume that the input points have integer coordinates (but there is no restriction on j and k), and the magnitude of these coordinates is at most n^c, where n is the number of input points and $c > 0$ is some constant. That is, the points are in a polynomially large integer grid.

An ϵ-coreset for an instance P of a shape fitting problem is a weighted set S, such that for any shape $F \in \mathcal{F}$, the summation of distances from points in P approximates the weighted summation of the distances from points in S up to a multiplicative factor of $(1 \pm \epsilon)$. A more precise definition (Definition 1) follows later. Coresets can be considered as a succinct representation of the point set; in particular, in order to obtain a $(1 + \epsilon)$-approximation solution fitting P, it is sufficient to find a $(1 + \epsilon)$-approximation solution for the coreset S. One usually seeks a small coreset, whose *size* $|S|$ is independent of the cardinality of P. Coresets of size $o(n)$ for the (j, k)-projective clustering problem for general j and k are not known to exist. However, the k-median/k-means clustering, k-line clustering, j-subspace approximation, and integer (j, k)-projective clustering problems admit small coresets.

Langberg and Schulman [10] introduced a general approach to coresets via the notion of *sensitivity* of points in a point set, which provides a natural way to set up a probability distribution $\Pr\cdot$ on P. Roughly speaking, the sensitivity of a point with respect to a point set measures the importance of the point, in terms of fitting shapes in the given family of shapes $\mathcal{F}$. Formally, the sensitivity of point p in a point set P is defined by $\sigma_P(p) := \sup_{F\in\mathcal{F}} \mathrm{dist}(p, F)/\mathrm{dist}(P, F)$. (In the degenerate case where the denominator in the ratio is 0, the numerator is also 0, and we take the ratio to be 0; the reader should feel free to ignore this technicality.) The total sensitivity of a point set P is defined by $\mathfrak{S}_P := \sum_{p\in P} \sigma_P(p)$. The nice property of quantifying the "importance" of a point in a point set is that for any $F \in \mathcal{F}$, $\mathrm{dist}(p, F)/\mathrm{dist}(P, F) \le \sigma_P(p)$. Setting the probability of selecting p to be $\sigma_P(p)/\mathfrak{S}_P$, and the weight of p to be $\mathfrak{S}_P/\sigma_P(p)$, $\forall p \in P$, one can show that the variance of the sampling scheme is $O((\mathfrak{S}_P)^2)$. When $\mathfrak{S}_P$ is $o(n)$, (for example, a constant or logarithmic in terms of $n = |P|$), one can obtain an ϵ-coreset by sampling a small number of points. Langberg and Schulman [10] show that the total sensitivity of any (arbitrarily large) point set $P \subset \mathbb{R}^d$ for k-median/k-means clustering problem is a constant, depending only on k, independent of the cardinality of P and the dimension of the Euclidean space where P and $\mathcal{F}$ are from. Using this, they derived a coreset for these problems with size depending polynomially on d and k and independent of n. Their work can be seen as evolving from earlier work on coresets for the k-median/k-means and related problems via other low variance sampling schemes [3, 4, 7, 5].

Feldman and Langberg [6] relate the notion of an ϵ-coreset with the well-studied notion of an ϵ-approximation of range spaces. They use a "functional representation" of points: consider a family of functions $\mathcal{P} = \{f_p(\cdot)|p \in P\}$, where each point p is associated with a function $f_p : X \to \mathbb{R}$. The target here is to pick a small subset $S \subseteq P$ of points, and assign weights appropriately, so that $\sum_{p\in S} w_p f_p(x)$ approximates $\sum_{p\in P} f_p(x)$ at every $x \in X$. When X is $\mathcal{F}$ and $f_p(F) = \mathrm{dist}(p, F)$, this is just the original ϵ-coreset for P. However, $f_p(\cdot)$ can be any other function defined over $\mathcal{F}$, for example, $f_p(\cdot)$ can be the "residue distance"

of p, *i.e.,* $f_p(F) = |\text{dist}(p, F) - \text{dist}(p', F)|$, where p' is the projection of p on the optimum shape F^* fitting P. The definitions of sensitivities and total sensitivity easily carry over in this setting: $\sigma_{\mathcal{P}}(f_p) = \sup_{x \in X} f_p(x) / \sum_{f_q \in \mathcal{P}} f_q(x)$ (which coincides with $\sigma_P(p)$ when $f_p(\cdot)$ is $\text{dist}(p, \cdot)$), and $\mathfrak{S}_{\mathcal{P}} = \sum_{f_p \in \mathcal{P}} \sigma_{\mathcal{P}}(f_p)$ (which coincides with $\mathfrak{S}_P$ similarly). One of the results in [6] is that an approximating subset $S \subseteq P$ can be computed with the size $|S|$ upper bounded by the product of two quantities: $(\mathfrak{S}_{\mathcal{P}})^2$, and another parameter, the "dimension" (see Definition 3) of a certain range space induced by $\mathcal{P}$, denoted $\dim(\mathcal{P})$. We remark that $\dim(\mathcal{P})$ depends on d, which is the dimension of Euclidean space where P is from, and some other parameters related to X; when X is the family of shapes for the (j, k)-projective clustering problem, $\dim(\mathcal{P})$ also depends on j and k. This connection allows them to use many results from the well-studied area of ϵ-approximation of range spaces (such as deterministic construction of small ϵ-approximation of range spaces), thus constructing smaller coreset deterministically, and removes some routine analysis in the traditional way of obtaining coresets via random sampling.

1.1 Our Results

In this article, we prove upper bounds of total sensitivities for the (j, k)-projective clustering problems. In particular, we show a careful analysis of computing total sensitivities for shape fitting problems in high dimension. Total sensitivity $\mathfrak{S}_P$ for a point set $P \subset \mathbb{R}^d$ may depend on d: consider the shape fitting problem where the family of shapes is the set of hyperplanes, and P is a point set of size d in general position. Then clearly $\sigma_P(p) = 1$ (since there always exists a hyperplane containing all $d - 1$ points other than p), so $\mathfrak{S}_P = d$.

One question that arises naturally is that whether the dependence of the total sensitivity on the dimension d is essential. To answer this question, we show that if the distance function is Euclidean distance, or the z^{th} power of Euclidean distance for $z \in [1, \infty)$, then the total sensitivity function of a shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \text{dist})$ in the high dimensional space $\mathbb{R}^d$ is roughly the same as that of the low-dimensional variant $(\mathbb{R}^{d'}, \mathcal{F}', \text{dist})$, where d' is the "intrinsic" dimension of the shapes in $\mathcal{F}$, and $\mathcal{F}'$ consists of shapes contained in the low dimensional space $\mathbb{R}^{d'}$. A reification of this statement is that the total sensitivity function of the (j, k)-projective clustering is independent of d. For the (j, k)-projective clustering problems, the shapes are intrinsically low dimensional: each k-tuple of j-flats is contained in a subspace of dimension at most $k(j + 1)$. As we will see, the total sensitivity function for $(\mathbb{R}^d, \mathcal{F}, \text{dist})$, where $\mathcal{F}$ is the family of k-tuples of j-flats in $\mathbb{R}^d$, is of the same magnitude as the total sensitivity function of $(\mathbb{R}^{f(j,k)}, \mathcal{F}', \text{dist})$, where $f(j, k)$ is a function of j and k (which is independent of d), and $\mathcal{F}'$ is the family of k-tuples of j-flats in $\mathbb{R}^{f(j,k)}$.

We sketch our approach to upper bound the total sensitivity of the (j, k)-projective clustering. We first make the observation (Theorem 7 below) that the total sensitivity of a point set P is upper bounded by a constant multiple of the total sensitivity of $P' = \text{proj}(P, F^*)$, which is the projection of P on the optimum shape F^* fitting P in $\mathcal{F}$. The computation of total sensitivity of P' is very simple in certain cases; for example, for k-median clustering, P' is a multi-set which contains k distinct points, whose total sensitivity can be directly bounded by k. Therefore, we are able to greatly simplify the proofs in [10]. Another more important use of this observation is that it allows us to get a dimension-reduction type result for the (j, k)-projective clustering problems: note that although the point set and the shapes might be in a high dimension space $\mathbb{R}^d$, the projected point set P' lies in a

subspace of dimension $(j+1)k$ (since each k-tuple of j-flats is contained in a subspace of dimension at most $(j+1)k$), which is small under the assumption that both j and k are constant. Therefore, $\mathfrak{S}_P$, which usually depends on d if one directly computes it in a high dimensional space, depends only on j and k, since $\mathfrak{S}_P$ is $O(\mathfrak{S}_{P'})$.

Our method for bounding the total sensitivity directly translates into a template for computing ϵ-coresets:

1. Compute F^*, the optimal shape fitting P. (It suffices to use an approximately optimal shape.) Compute P', the projection of P onto F^*.
2. Compute a bound on the sensitivity of each point in P' with respect to P'. Since the ambient dimension is $O(jk)$, we may use a method that yields bounds on $\mathfrak{S}_{P'}$ with dependence on the ambient dimension. Use Theorem 7 to translate this into a bound for $\sigma_P(p)$ for each $p \in P$.
3. Sample points from P with probabilities proportional to $\sigma_P(p)$ to obtain a coreset, as described in [10, 6].

We now point out the difference between our usage of total sensitivity in the construction of coresets and the method in [6]. The construction of coresets in [6] may also be considered as based on total sensitivity, however in a very different way:

1. First obtain a small weighted point set $S \subseteq P$, such that $\mathrm{dist}(P,F) - \mathrm{dist}(P',F)$ is approximately the same as $\mathrm{dist}(S,F) - \mathrm{dist}(S',F)$ (S' is $\mathrm{proj}\,(S,F^*)$) for every $F \in \mathcal{F}$.
2. Then compute an ϵ-coreset $Q' \subseteq P'$ for the projected point set P', that is, $\mathrm{dist}(Q',F)$ approximates $\mathrm{dist}(P',F)$ for every $F \in \mathcal{F}$. (Since P' is from a low-dimensional subspace, the ambient dimension is small, and the computation can exploit this.)

Therefore, for each $F \in \mathcal{F}$, $\mathrm{dist}(P,F) = (\mathrm{dist}(P,F) - \mathrm{dist}(P',F)) + \mathrm{dist}(P',F) \approx (\mathrm{dist}(S,F) - \mathrm{dist}(S',F)) + \mathrm{dist}(Q',F)$.

Thus the weighted set $Q' \cup S \cup S'$ is a coreset for P, but notice that the points in S' have negative weights. In contrast, the weights of points in the coreset in our construction are positive. The advantage of getting coresets with positive weights is that in order to get an approximate solution to the shape fitting problem, we may run algorithms or heuristics developed for the shape fitting problem on the coreset, such as [1]. When points have negative weights, on the other hand, some of these heuristics do not work or need to be modified appropriately.

Another useful feature of the coresets obtained via our results is that the coreset is a subset of the original point set. When each point stands for a data item, the coreset inherits a natural interpretation. See [11] for a discussion of this issue in a broader context.

The sizes of the coresets in this paper are somewhat larger than the size of coresets in [6]. Roughly speaking, the size of the coreset in [6] is $f_1(d) + f_2(j,k)$, where $f_1(d)$ (respectively $f_2(j,k)$) is a function depending only on d (respectively j and k) for the (j,k)-projective clustering problem, while the coreset size in our paper is $f_1(d) \cdot f_2(j,k)$.

Organization of this paper: In this article, we focus on the construction that establishes small total sensitivity for various shape fitting problems, and the size of the resulting coreset. For clarity, we omit the description of algorithms for computing such bounds on sensitivity. Efficient algorithms result from the construction using a methodology that is now well-understood. Also because the weights for points in the coreset are nonnegative, the coreset lend itself to streaming settings, where points arrive one by one as $p_1, p_2, \cdots$ [9][6]. In Section 2, we present necessary definitions used through this article, and summarize related results from [6] and [12]. In Section 3, we prove the upper bound of total sensitivity of an instance of

a shape fitting problem in high dimension by its low dimensional projection. In Sections 4, 5, 6, and 7, we apply the upper bound from Section 3 to k-median/k-means, clustering, k-line clustering, j-subspace approximation, and the integer (j,k)-projective clustering problem, respectively, to obtain upper bounds for their total sensitivities, and the size of the resulting ϵ-coresets.

2 Preliminaries

In this section, we formally define some of the concepts studied in this article, and state crucial results from previous work. We begin by defining an ϵ-coreset.

▶ **Definition 1** (ϵ-coreset of a shape fitting problem)**.** Given an instance $P \subset \mathbb{R}^d$ of a shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \mathrm{dist})$, and $\epsilon \in [0,1]$, an ϵ-coreset of P is a (weighted) set $S \subseteq P$, together with a weight function $w : S \to \mathbb{R}^+$, such that for any shape F in $\mathcal{F}$, it holds that $|\mathrm{dist}(P,F) - \mathrm{dist}(S,F)| \leq \epsilon \cdot \mathrm{dist}(P,F)$, where by definition, $\mathrm{dist}(P,F) = \sum_{p\in P} \mathrm{dist}(p,F)$, and $\mathrm{dist}(S,F) = \sum_{p\in S} w(p)\mathrm{dist}(p,F)$. The size of the weighted coreset S is defined to be $|S|$.

We note that in the literature, the requirement that the weights be non-negative, as well as the requirement that the coreset S be a subset of the original instance P, are sometimes relaxed. We include these requirements in the definition to emphasize that the coresets constructed here do satisfy them. We now define the sensitivities of points in a shape fitting instance, and the total sensitivity of the instance.

▶ **Definition 2** (Sensitivity of a shape fitting instance [10])**.** Given an instance $P \subset \mathbb{R}^d$ of a shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \mathrm{dist})$, the sensitivity of a point p in P is $\sigma_P(p) := \inf\{\beta \geq 0 | \mathrm{dist}(p,F) \leq \beta \mathrm{dist}(P,F), \forall F \in \mathcal{F}\}$.

Note that an equivalent definition is to let $\sigma_P(p) = \sup_{F\in\mathcal{F}} \mathrm{dist}(p,F)/\mathrm{dist}(P,F)$, with the understanding that when the denominator in the ratio is 0, the ratio itself is 0.

The total sensitivity of the instance P, is defined by $\mathfrak{S}_P := \sum_{p\in P} \sigma_P(p)$. The total sensitivity function of the shape fitting problem is $\mathfrak{S}_n := \sup_{|P|=n} \mathfrak{S}_P$.

We now need a somewhat technical definition in order to be able to state an important earlier result from [6]. On a first reading, the reader is welcome to skip the detailed definition.

▶ **Definition 3** (The dimension of a shape fitting instance [6])**.** Let $P \subset \mathbb{R}^d$ be an instance of a shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \mathrm{dist})$. For a weight function $w : P \to \mathbb{R}^+$, consider the set system $(P, \mathcal{R})$, where $\mathcal{R}$ is a family of subsets of P defined as follows: each element in $\mathcal{R}$ is a set of the form $R_{F,r}$ for some $F \in \mathcal{F}$ and $r \geq 0$, and $R_{F,r} = \{p \in P \mid w_p \cdot \mathrm{dist}(p,F) \leq r\}$. That is, $R_{F,r}$ is the set of those points in P whose weighted distance to the shape F is at most r. The dimension of the instance P of the shape fitting problem, denoted by $\dim(P)$, is the smallest integer m, such that for any weight function w and $A \subseteq P$ of size $|A| = a \geq 2$, we have: $|\{A \cap R_{F,r} | F \in \mathcal{F}, r \geq 0\}| \leq a^m$.

For instance, in the (j,k)-projective clustering problem with the underlying distance function dist being the z^{th} power of the Euclidean distance, the dimension $\dim(P)$ of any instance P is $O(jdk)$, independent of $|P|$ [6]. This is shown by methods similar to the ones used to bound the VC-dimension of geometric set systems. In fact, this bound is the only fact that we will need about the dimension of a shape fitting instance.

The following theorem recalls the connection established in [6] between coresets and sensitivity via the above notion of dimension.

▶ **Theorem 4** (Connection between total sensitivity and ϵ-coreset [6]). *Given any n-point instance $P \subset \mathbb{R}^d$ of a shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \text{dist})$, and any $\epsilon \in (0,1]$, there exists an ϵ-coreset for P of size $O\left(\left(\frac{\mathfrak{S}_n}{\epsilon}\right)^2 \dim(P)\right)$.*

Finally, we will need known bounds on the total sensitivity of (j,k)-projective clustering problem. These earlier bounds involve the dimension d corresponding to shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \text{dist})$.

▶ **Theorem 5** (Total sensitivity of (j,k)-projective clustering problem in fixed dimension [12]). *We have the following upper bounds of total sensitivities for the (j,k)-projective clustering problem $(\mathbb{R}^d, \mathcal{F}, \text{dist})$, where* dist *is the z-th power of the Eucldiean distance for $z \in (0,\infty)$.*

- *$j = 1$ (k-line center): $\mathfrak{S}_n$ is $O(k^{f(k,d)} \log n)$, where $f(d,k)$ is a function depending only on d and k.*
- *integer (j,k)-projective clustering problem: For any n-point instance P, with each coordinate being an integer of magnitude at most n^c for any constant $c > 0$, $\mathfrak{S}_P$ is $O((\log n)^{f(d,j,k)})$, where $f(d,j,k)$ is a function depending only on d, j, and k.*

3 Bounding the Total Sensitivity via Dimension Reduction

In this section, we show that the total sensitivity of a point set P is of the same order as that of $\text{proj}(P, F^*)$, which is the projection of P onto an optimum shape F^* from $\mathcal{F}$ fitting P. This result captures the fact that total sensitivity of a shape fitting problem quantifies the complexity of shapes, in the sense that total sensitivity depends on the dimension of smallest subspace containing each shape, regardless of the dimension of the ambient space where P is from.

▶ **Definition 6** (projection of points on a shape). For a shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \text{dist})$, define $\text{proj} : \mathbb{R}^d \times \mathcal{F} \to \mathbb{R}^d$, where $\text{proj}(p, F)$ is the projection of p on a shape F, that is, $\text{proj}(p, F)$ is a point in F which is nearest to p, with ties broken arbitrarily. That is, $\text{dist}(p, \text{proj}(p, F)) = \min_{q \in F} \text{dist}(p, q)$. We abuse the notation to denote the multi-set $\{\text{proj}(p, F) \,|\, p \in P\}$ by $\text{proj}(P, F)$ for $P \subset \mathbb{R}^d$.

We first show that $\mathfrak{S}_P$ is $O(\mathfrak{S}_{\text{proj}(P,F^*)})$, where F^* is an optimum shape fitting P from $\mathcal{F}$. In particular, this implies that when F^* is a low-dimensional object, the total sensitivity of $P \subset \mathbb{R}^d$ can be upper bounded by the total sensitivity of a point set contained in a low dimension subspace.

▶ **Theorem 7** (Dimension reduction, computing the total sensitivity of a point set in high dimensional space with the projected lower dimensional point set). *Given an instance P of a shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \text{dist})$, let F^* denote a shape that minimizes $dist(P, F)$ over all $F \in \mathcal{F}$. Let p' denote* proj (p, F^*) *and let P' denote* proj (P, F^*)*. Assume that the distance function satisfies the relaxed triangle inequality: $dist(p,q) \le \alpha(dist(p,r) + dist(r,q))$ for any $p, q, r \in \mathbb{R}^d$ for some constant $\alpha \ge 1$. Then*

1. *the following inequality holds: $\mathfrak{S}_P \le 2\alpha^2 \mathfrak{S}_{P'} + \alpha$.*
2. *if $dist(P, F^*) = 0$, then $\sigma_P(p) = \sigma_{P'}(p')$ for each $p \in P$. If $dist(P, F^*) > 0$, then $\sigma_P(p) \le \left(\alpha \frac{dist(p,p')}{dist(P,F^*)} + 2\alpha^2 \sigma_{P'}(p')\right)$.*

Proof. If $\text{dist}(P, F^*) = 0$, then $P = P'$, and clearly both parts of the theorem hold.

Let us consider the case where $\text{dist}(P, F^*) > 0$. By definition,

$$\sigma_P(p) = \inf\{\beta \ge 0 \mid \text{dist}(p, F) \le \beta \text{dist}(P, F), \forall F \in \mathcal{F}\},$$
$$\sigma_{P'}(p') = \inf\{\beta' \ge 0 \mid \text{dist}(p', F) \le \beta' \text{dist}(P', F), \forall F \in \mathcal{F}\}.$$

Let F be an arbitrary shape in $\mathcal{F}$. Then we have

$$\begin{aligned}
\mathrm{dist}(p,F) &\le \alpha\mathrm{dist}(p,p') + \alpha\mathrm{dist}(p',F) \\
&\le \alpha\mathrm{dist}(p,p') + \alpha\sigma_{P'}(p')\mathrm{dist}(P',F) \\
&\le \alpha\mathrm{dist}(p,p') + 2\alpha^2\sigma_{P'}(p')\mathrm{dist}(P,F) \\
&= \alpha\frac{\mathrm{dist}(p,p')}{\mathrm{dist}(P,F)}\cdot\mathrm{dist}(P,F) + 2\alpha^2\sigma_{P'}(p')\mathrm{dist}(P,F) \\
&\le \alpha\frac{\mathrm{dist}(p,p')}{\mathrm{dist}(P,F^*)}\cdot\mathrm{dist}(P,F) + 2\alpha^2\sigma_{P'}(p')\mathrm{dist}(P,F) \\
&= \left(\alpha\frac{\mathrm{dist}(p,p')}{\mathrm{dist}(P,F^*)} + 2\alpha^2\sigma_{P'}(p')\right)\mathrm{dist}(P,F).
\end{aligned}$$

The first inequality follows from the relaxed triangle inequality, the second inequality follows from the definition of sensitivity of p' in P', and third inequality follows from the fact that $\mathrm{dist}(P',F) = \sum_{p'\in P'}\mathrm{dist}(p',F) \le \sum_{p\in P}\alpha\left(\mathrm{dist}(p,F)+\mathrm{dist}(p,p')\right) = \alpha(\mathrm{dist}(P,F) + \mathrm{dist}(P,F^*)) \le 2\alpha\mathrm{dist}(P,F)$, since $\mathrm{dist}(P,F^*) \le \mathrm{dist}(P,F)$.

Thus the second part of the theorem holds. Now,

$$\begin{aligned}
\mathfrak{S}_P &= \sum_{p\in P}\sigma_P(p) \\
&\le \sum_{p\in P}\left(\alpha\frac{\mathrm{dist}(p,p')}{\mathrm{dist}(P,F^*)} + 2\alpha^2\sigma_{P'}(p')\right) \\
&= \alpha + 2\alpha^2\mathfrak{S}_{P'}.
\end{aligned}$$

◀

We make a remark regarding the value of α in Theorem 7 when the distance function is z^{th} power of Euclidean distance. It is used in Sections 4, 5, 6, and 7 when we derive upper bounds of total sensitivities for various shape fitting problems.

▶ Remark (Value of α when $\mathsf{dist}(\cdot,\cdot) = (\|\cdot\|_2)^z$). Let $z \in (0,\infty)$. Suppose $\mathrm{dist}(p,q) = (\|p-q\|_2)^z$. When $z \in (0,1)$, the weak triangle inequality holds with $\alpha = 1$; when $z \ge 1$, the weak triangle inequality holds with $\alpha = 2^{z-1}$. For a proof, see, for example, [8].

Theorem 7 bounds the total sensitivity of an instance P of a shape fitting problem $(\mathbb{R}^d,\mathcal{F},\mathrm{dist})$ in terms of the total sensitivity of P'. Suppose that there is an $m_2 \ll d$ so that each shape $F \in \mathcal{F}$ is in some subspace of dimension m_2. In the (j,k)-projective clustering problem, for example, $m_2 = k(j+1)$. Then note that P' is contained in a subspace of dimension m_2. Furthermore, when dist is the z^{th} power of the Euclidean distance, it turns out that for many shape fitting problems the sensitivity of P' can be bounded as if the shape fitting problem was housed in $\mathbb{R}^{2m_2}$ instead of $\mathbb{R}^d$. To see why this is the case for the (j,k)-projective clustering problem, fix an arbitrary subspace G of dimension $\min\{d,2m_2\}$ that contains P'. Then for for any $F \in \mathcal{F}$, there is an $F' \in \mathcal{F}$ such that (a) F' is contained in G, and (b) $\mathrm{dist}(p',F') = \mathrm{dist}(p',F)$ for all $p' \in P'$.

The following theorem summarizes this phenomenon. For simplicity, it is stated for the (j,k)-projective clustering problem, even though the phenomenon itself is somewhat more general.

▶ **Theorem 8** (Sensitivity of a lower dimensional point set in a high dimensional space)**.** *Let P' be an n-point instance of the (j,k)-projective clustering problem $(\mathbb{R}^d,\mathcal{F},\mathrm{dist})$, where* dist *is the z^{th} power of the Euclidean distance, for some $z \in (0,\infty)$. Assume that P' is*

contained in a subspace of dimension m_1. (Note that for each shape $F \in \mathcal{F}$, there is a subspace of dimension $m_2 = k(j+1)$ containing it.) Let G be any subspace of dimension $m = \min\{m_1 + m_2, d\}$ containing P'; fix an orthonormal basis for G, and for each $p' \in P'$, let $p'' \in \mathbb{R}^m$ be the coordinates of p' in terms of this basis. Let $P'' = \{p'' \mid p' \in P'\}$, and view P'' as an instance of the (j,k)-projective clustering problem $(\mathbb{R}^m, \mathcal{F}', \text{dist})$, where $\mathcal{F}'$ is the set of all k-tuples of j-subspaces in $\mathbb{R}^m$, and dist *is the z^{th} power of the Eucldiean distance. Then, $\sigma_{P'}(p') = \sigma_{P''}(p'')$ for each $p' \in P'$, and $\mathfrak{S}_{P'} = \mathfrak{S}_{P''}$.*

4 k-median/k-means Clustering Problem

In this section, we derive upper bounds for the total sensitivity function for the k-median/k-means problems, and its generalizations, where the distance function is z^{th} power of Euclidean distance, using the approach in Section 4. These bounds are similar to the ones derived by Langberg and Schulman [10], but the proof is much simplified. For the rest of the article, dist is assumed to be the z^{th} power of the Euclidean distance.

▶ **Theorem 9** (Total sensitivity of $(0,k)$-projective clustering)**.** *Consider the shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \text{dist})$, where $\mathcal{F}$ is the set of all k-point subsets of $\mathbb{R}^d$. We have the following upper bound on the total sensitivity:*

$$\mathfrak{S}_n \leq 2^{2z-1}k + 2^{z-1}, \qquad z \geq 1,$$
$$\mathfrak{S}_n \leq 2k + 1, \qquad z \in (0,1).$$

In particular, the total sensitivity of the k-median problem (which corresponds to the case when $z = 1$) is at most $2k+1$, and the total sensitivity of the k-means problem (which corresponds to the case when $z = 2$) is $8k + 2$.

Proof. Let P be an arbitrary n-point set. Apply Theorem 7, and note that $\text{proj}(P, C^*)$, where C^* is an optimum set of k centers, contains at most k distinct points. Assume that $C^* = \{c_1^*, c_2^*, \cdots, c_k^*\}$. Let P_i be the set of points in P whose projection is c_i^*, that is, $P_i = \{p \in P | \text{proj}(p, C^*) = c_i^*\}$. It is easy to see that the summation of sensitivities of the $|P_i|$ copies of c_i^* is at most 1: for any k-point set C in $\mathbb{R}^d$, $|P_i| \cdot \frac{\text{dist}(c_i^*, C)}{\text{dist}(C^*, C)} = \frac{|P_i|\text{dist}(c_i^*, C)}{\sum_{j=1}^{k} |P_j|\text{dist}(c_j^*, C)} \leq 1$.

Therefore, the total sensitivity of $\text{proj}(P, C^*)$ is at most k. Substituting α from the remark after Theorem 7, we get the above result. ◀

▶ **Theorem 10** (ϵ-coreset for $(0,k)$-projective clustering)**.** *Consider the shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \text{dist})$, where $\mathcal{F}$ is the set of all k-point subsets of $\mathbb{R}^d$. For any n-point instance P, there is an ϵ-coreset of size $O(k^3 d\epsilon^{-2})$.*

Proof. Observe that the $\dim(P)$ is $O(kd)$. Using Theorem 4, and Theorem 9, we obtain the above result. ◀

5 k-line Clustering Problem

In this section, we derive upper bounds on the total sensitivity function for the k-line clustering problem, that is, the $(1,k)$-projective clustering problem.

▶ **Theorem 11** (Total sensitivity for k-line clustering problem)**.** *Consider the shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \text{dist})$, where $\mathcal{F}$ is the set of k-tuple of lines. The total sensitivity function, $\mathfrak{S}_n$, is $O(k^{f(k)} \log n)$, where $f(k)$ is a function the depends only on k.*

Proof. Let P be an arbitrary n-point set. Let K^* denote an optimum set of k lines fitting P. Using Theorems 7 and 8, it suffices to bound the sensitivity of an n-point instance of a k-line clustering problem housed in $\mathbb{R}^{4k}$. By Theorem 5, the total sensitivity of this latter shape fitting problem is $O(k^{f(k)} \log n)$, where $f(k)$ is a function depending only on k. Therefore, $\mathfrak{S}_n$ is $O(k^{f(k)} \log n)$.

(Alternatively, one could use a recent result in [8]. Let P' denote the projection of P into K^*. Since K^* is a union of k lines, we can upper bound the sensitivity of P' by k times the sensitivity of an n-point set that lies on a *single line*. The sensitivity of an n-point set that lies on a single line can be upper bounded by the sensitivity of an n-point set for the *weighted* $(0, k)$-projective clustering problem, for which the sensitivity bound is $O(k^{f(k)} \log n)$ as shown in [8].) ◀

Notice that for k-line clustering problem, the bound on the total sensitivity depends logarithmically on n. We give below a construction of a point set that shows that this is necessary, even for $d = 2$.

▶ **Theorem 12** (The upper bound of total sensitivity for k-line clustering problem is tight)**.** *For every $n \geq 2$, there exists an n-point instance of the k-line clustering problem $(\mathbb{R}^2, \mathcal{F}, \mathrm{dist})$, where* dist *is the Euclidean distance, such that the total sensitivity of P is $\Omega(\log n)$.*

Proof. We construct a point set P of size n, together with n shapes $F_i \in \mathcal{F}$, $i = 1, \cdots, n$, such that $\sum_{i=1}^{n} \mathrm{dist}(p_i, F_i)/\mathrm{dist}(P, F_i)$ is $\Omega(\log n)$. Note that this implies that $\mathfrak{S}_P$ is at least $\Omega(\log n)$. Let P be the following point set in $\mathbb{R}^2$: $p_i = (1/2^{i-1}, 0)$, for $i = 1, \cdots, n$. Let F_i be a pair of lines: one vertical line and one horizontal line, where the vertical line is the y-axis, and the horizontal line is $\{(x, 1/2^i) | x \in \mathbb{R}\}$.

Consider the point p_i, where $i = 1, \cdots, n$. We show that $\mathrm{dist}(p_i, F_i)/\mathrm{dist}(P, F_i)$ is at least $1/(2 + i)$, for $i = 1, \cdots, n$. For $j \leq i$, note that $\mathrm{dist}(p_j, F_i) = 1/2^i$: since the distance from p_j to the horizontal line in F_i is $1/2^i$ and the distance to the vertical line is $1/2^{j-1}$, $\mathrm{dist}(p_j, F_i) = \min\{1/2^{j-1}, 1/2^i\} = 1/2^i$. For $i + 1 \leq j \leq n$, on the other hand, $\mathrm{dist}(p_j, F_i) = 1/2^{j-1}$. Therefore, $\sum_{j=i+1}^{n} \mathrm{dist}(p_j, F_i) = \sum_{j=i+1}^{n} 1/2^{j-1} = (1/2^{i-1}) \cdot (1 - (1/2)^{n-i})$. Thus, we have

$$\sigma_P(p_i) = \sup_{F \in \mathcal{F}} \frac{\mathrm{dist}(p_i, F)}{\mathrm{dist}(P, F)} \geq \frac{\mathrm{dist}(p_i, F_i)}{\mathrm{dist}(P, F_i)} = \frac{1/2^i}{(1/2^{i-1} - 1/2^{n-1}) + i \cdot (1/2^i)} > \frac{1}{2 + i}$$

Therefore, $\mathfrak{S}_P \geq \sum_{i=1}^{n} \sigma_P(p_i) > \sum_{i=1}^{n} \frac{1}{2+i}$, which is $\Omega(\log n)$. ◀

▶ **Theorem 13** (ϵ-coreset for k-line clustering problem)**.** *Consider the shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \mathrm{dist})$, where $\mathcal{F}$ is the set of all k-tuples of lines in $\mathbb{R}^d$. For any n-point instance P, there is an ϵ-coreset with size $O(k^{f(k)} d (\log n)^2 / \epsilon^2)$.*

Proof. This result follows from Theorem 11, Theorem 4, and the fact that $\dim(P)$ in this case is $O(kd)$. ◀

6 Subspace approximation

In this section, we derive upper bounds on the sensitivity of the subspace approximation problem, that is, the $(j, 1)$-projective clustering problem. For the applications of Theorems 7 and 8 in the other sections, we use existing bounds on the sensitivity that have a dependence on the dimension d. For the subspace approximation problem, however, we derive here the dimension-dependent bounds on sensitivity by generalizing an argument from [10] for the

case $j = d-1$ and $z = 2$. This derivation is somewhat technical. With these bounds in hand, the derivation of the dimension-independent bounds is readily accomplished in a manner similar to the other sections.

6.1 Dimension-dependent bounds on Sensitivity

We first recall the notion of an (α, β, z)-*conditioned basis* from [5], and state one of its properties (Lemma 15). We will use standard matrix terminilogy: m_{ij} denotes the entry in the i-th row and j-th column of M, and $M_{i\cdot}$ is the i-th row of M.

▶ **Definition 14.** Let M be an $n \times m$ matrix of rank ρ. Let $z \in [1, \infty)$, and $\alpha, \beta \geq 1$. An $n \times \rho$ matrix A is an (α, β, z)-conditioned basis for M if the column vectors of A span the column space of M, and additionally A satisfies that: (1) $\sum_{i,j} |a_{ij}|^z \leq \alpha^z$, (2) for all $u \in \mathbb{R}^\rho$, $\| u \|_{z'} \leq \beta \| Au \|_z$, where $\| \cdot \|_{z'}$ is the dual norm for $\| \cdot \|_z$ (*i.e.* $1/z + 1/z' = 1$).

▶ **Lemma 15.** *Let M be an $n \times m$ matrix of rank ρ. Let $z \in [1, \infty)$. Let A be an (α, β, z)-conditioned basis for M. For every vector $u \in \mathbb{R}^m$, the following inequality holds: $|M_{i\cdot} u|^z \leq (\| A_{i\cdot} \|_z^z \cdot \beta^z) \| Mu \|_z^z$.*

Proof. We have $M = A\tau$ for some $\rho \times m$ matrix τ. Then,

$$|M_{i\cdot} u|^z = |A_{i\cdot} \tau u|^z \leq \| A_{i\cdot} \|_z^z \cdot \| \tau u \|_{z'}^z \leq \| A_{i\cdot} \|_z^z \cdot \beta^z \| A\tau u \|_z^z = \| A_{i\cdot} \|_z^z \cdot \beta^z \| Mu \|_z^z.$$

The second step is Holder's inequality, and the third uses the fact that A is (α, β, z)-conditioned. ◀

Using Lemma 15, we derive an upper bound on the total sensitivity when each shape is a hyperplane.

▶ **Lemma 16** (total sensitivity for fitting a hyperplane)**.** *Consider the shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \text{dist})$ where $\mathcal{F}$ is the set of all $(d-1)$-flats, that is, hyperplanes. The total sensitivity of any n-point set is $O(d^{1+z/2})$ for $1 \leq z < 2$, $O(d)$ for $z = 2$, and $O(d^z)$ for $z > 2$.*

Proof. We can parameterize a hyperplane with a vector in $\mathbb{R}^{d+1}$, $u = \begin{bmatrix} u_1 & \cdots & u_{d+1} \end{bmatrix}^T$: the hyperplane determined by u is $h_u = \{x \in \mathbb{R}^d | \sum_{i=1}^d u_i x_i + u_{d+1} = 0\}$, where x_i denotes the i^{th} entry of the vector x. Without loss of generality, we may assume that $\sum_{i=1}^d u_i^2 = 1$. The Euclidean distance to h_u from a point $q \in \mathbb{R}^d$ is $\text{dist}(q, h_u) = |\sum_{i=1}^d u_i q_i + u_{d+1}| / \sqrt{\sum_{i=1}^d u_i^2} = |\sum_{i=1}^d u_i q_i + u_{d+1}|$. (the second equality follows from the assumption that $\sum_{i=1}^d u_i^2 = 1$.)

Let $P = \{p_1, p_2, \ldots, p_n\} \subseteq \mathbb{R}^d$ be any set of n points. Let $\tilde{p_i}$ denote the row vector $\begin{bmatrix} p_i^T & 1 \end{bmatrix}$, and let M be the $n \times (d+1)$ matrix whose i^{th} row is $\tilde{p_i}$. Then, $\text{dist}(p_i, h_u) = |M_{i\cdot} u|^z$, and $\text{dist}(P, h_u) = \sum_{i=1}^n |M_{i\cdot} u|^z = \| Mu \|_z^z$. Then using Lemma 15, we have $\sigma_P(p_i) = \sup_u \frac{|M_{i\cdot} u|^z}{\| Mu \|_z^z} \leq \| A_{i\cdot} \|_z^z \cdot \beta^z$, where A is an (α, β, z)-conditioned basis for M. Thus,

$$\mathfrak{S}_P = \sum_{i=1}^n \sigma_P(p_i) \leq \beta^z \sum_{i=1}^n \| A_{i\cdot} \|_z^z = \beta^z \sum_{i,j} |a_{ij}|^z = (\alpha\beta)^z.$$

For $1 \leq z < 2$, M has $((d+1)^{1/z+1/2}, 1, z)$-conditioned basis; for $z = 2$, M has $((d+1)^{1/2}, 1, z)$-conditioned basis; for $z > 2$, M has $((d+1)^{1/z+1/2}, (d+1)^{1/z'-1/2}, z)$-conditioned basis [5]. Thus the total sensitivity for the three cases are $(d+1)^{1+z/2}$, $d+1$, and $(d+1)^z$, respectively. ◀

It is now easy to derive dimension-dependent bounds on the sensitivity when each shape is a j-subspace.

▶ **Corollary 17** (Total sensitivity for fitting a j-subspace). *Consider the shape fitting problem* $(\mathbb{R}^d, \mathcal{F}, \mathrm{dist})$ *where* $\mathcal{F}$ *is the set of all* j*-flats. The total sensitivity of any* n*-point set is* $O(d^{1+z/2})$ *for* $1 \leq z < 2$*,* $O(d)$ *for* $z = 2$*, and* $O(d^z)$ *for* $z > 2$*.*

Proof. Denote by $\mathcal{F}'$ the set of hyperplanes in $\mathbb{R}^d$. Let $P \subseteq \mathbb{R}^d$ be an arbitrary n-point set. We first show that $\sigma_{P,\mathcal{F}}(p) \leq \sigma_{P,\mathcal{F}'}(p)$, where the additional subscript is being used to indicate which shape fitting problem we are talking about (hyperplanes or j-flats). Let p be an arbitrary point in P. Let $F_p \in \mathcal{F}$ denote the j-subspace such that $\sigma_{P,\mathcal{F}}(p) = \mathrm{dist}(p, F_p)/\mathrm{dist}(P, F_p)$. Let $\mathrm{proj}\,(p, F_p)$ denote the projection of p on F_p. Consider the hyperplane F' containing F_p and orthogonal to the vector $p - \mathrm{proj}\,(p, F_p)$. We have $\mathrm{dist}(p, F') = \mathrm{dist}(p, F_p)$, whereas $\mathrm{dist}(q, F') \leq \mathrm{dist}(q, F_p)$ for each $q \in P$. Therefore, $\sigma_{P,\mathcal{F}'}(p) \geq \mathrm{dist}(p, F')/\mathrm{dist}(P, F') \geq \mathrm{dist}(p, F_p)/\mathrm{dist}(P, F_p) = \sigma_{P,\mathcal{F}}(p)$. It follows that $\mathfrak{S}_{P,\mathcal{F}} \leq \mathfrak{S}_{P,\mathcal{F}'}$. The statement in the corollary now follows from Lemma 16. ◀

6.2 Dimension-independent Bounds on the Sensitivity

We now derive dimension-independent upper bounds for the total sensitivity for the j-subspace fitting problem.

▶ **Theorem 18** (Total sensitivity for j-subspace fitting problem). *Consider the shape fitting problem* $(\mathbb{R}^d, \mathcal{F}, \mathrm{dist})$ *where* $\mathcal{F}$ *is the set of all* j*-flats. The total sensitivity of any* n*-point set is* $O(j^{1+z/2})$ *for* $1 \leq z < 2$*,* $O(j)$ *for* $z = 2$*, and* $O(j^z)$ *for* $z > 2$*.*

Proof. Use Theorem 7, note that the projected point set P' is contained in a j-subspace. Further, each shape is a j-subspace. So, applying Theorem 8 and Corollary 17, the total sensitivity is $O(j^{2+z/2})$ or $z \in [1, 2)$, $O(j)$ for $z = 2$ and $O(j^z)$ for $z > 2$. ◀

Using Theorem 18 and the fact that $\dim(P)$ for the j-subspace fitting problem is $O(jd)$, we obtain small ϵ-coresets:

▶ **Theorem 19** (ϵ-coreset for j-subspace fitting problem). *Consider the shape fitting problem* $(\mathbb{R}^d, \mathcal{F}, \mathrm{dist})$ *where* $\mathcal{F}$ *is the set of all* j*-flats. For any* n*-point set, there exists an* ϵ*-coreset whose size is* $O(j^{3+z}d\epsilon^{-2})$ *for* $z \in [1, 2)$*,* $O(j^3 d\epsilon^{-2})$ *for* $z = 2$ *and* $O(j^{2z+1}d\epsilon^{-2})$ *for* $z \geq 2$*.*

Proof. The result follows from Theorem 18, and Theorem 4. ◀

We note that for the case $j = d - 1$ and $z = 2$, a linear algebraic result from [2] yields a coreset whose size is an improved $O(d\epsilon^{-2})$.

7 The (j, k) integer projective clustering

▶ **Theorem 20.** *Consider the shape fitting problem* $(\mathbb{R}^d, \mathcal{F}, \mathrm{dist})$*, where* $\mathcal{F}$ *is the set of* k*-tuples of* j*-flats. Let* $P \subset \mathbb{R}^d$ *be any* n*-point instance with integer coordinates, the magnitude of each coordinate being at most* n^c*, for some constant* c*. The total sensitivity* $\mathfrak{S}_P$ *of* P *is* $O((\log n)^{f(k,j)})$*, where* $f(k, j)$ *is a function of only* k *and* j*. There exists an* ϵ*-coreset for* P *of size* $O((\log n)^{2f(k,j)}kjd\epsilon^{-2})$*.*

Proof. Observe that the projected point set $P' = \mathrm{proj}\,(P, \{J_1^*, \cdots, J_k^*\})$, where $\{J_1^*, \cdots, J_k^*\}$ is an optimum k-tuple of j-flats fitting P, is contained in a subspace of dimension $O(jk)$. Using Theorem 5, Theorem 8, and Theorem 7, the total sensitivity $\mathfrak{S}_P$ is upper bouned by

$O((\log n)^{f(k,j)})$, where $f(k,j)$ is a function of k and j. (A technical complication is that the coordinates of P', in the appropriate orthonormal basis, may not be integers. This can be addressed by rounding them to integers, at the expense of increasing the constant c. A similar procedure is adopted in [12], and we omit the details here.)

Using Theorem 4 and the fact that $\dim(P)$ is $O(djk)$, we obtain the bound on the coreset. ◀

8 Acknowledgements.

We thank the anonymous reviewers and Dan Feldman for their insightful feedback.

References

1 Pankaj K. Agarwal and Nabil H. Mustafa. k-Means projective clustering. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '04, pages 155–165, New York, NY, USA, 2004. ACM.

2 Joshua D. Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-Ramanujan sparsifiers. In *STOC*, pages 255–262, 2009.

3 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.

4 Kenneth L. Clarkson. Subgradient and sampling algorithms for ℓ_1 regression. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 257–266, 2005.

5 Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W. Mahoney. Sampling algorithms and coresets for ℓ_p regression. *SIAM J. Comput.*, 38(5):2060–2078, 2009.

6 Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *STOC*, pages 569–578. For an updated version, see `http://arxiv.org/abs/1106.1379v1`, 2011.

7 Dan Feldman, Morteza Monemizadeh, and Christian Sohler. A PTAS for k-means clustering based on weak coresets. In *SCG '07: Proceedings of the twenty-third annual symposium on Computational geometry*, pages 11–18, New York, NY, USA, 2007. ACM.

8 Dan Feldman and Leonard J. Schulman. Data reduction for weighted and outlier-resistant clustering. In *SODA*, pages 1343–1354, 2012.

9 Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, STOC '04, pages 291–300, New York, NY, USA, 2004. ACM.

10 Michael Langberg and Leonard J. Schulman. Universal ϵ-approximators for integrals. In *SODA*, pages 598–607, 2010.

11 Michael W. Mahoney and Petros Drineas. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

12 Kasturi Varadarajan and Xin Xiao. A near-linear algorithm for projective clustering integer points. In *SODA*, pages 1329–1342, 2012.