A. The Combinatorial Complexity of the Sets Clustering

The following definition of a query space encapsulates all the ingredients required to formally define an optimization problem.

Definition A.1 (Query space; see Definition 4.2 in (Braverman et al., 2016))**.** *Let $\mathcal{P}$ be a set called input set. Let Q be be a (possibly infinite) set called query set. Let $f : P \times Q \to \mathbb{R}$ be a cost function. The tuple $(\mathcal{P}, Q, f)$ is called a* query space. *A* sets clustering query space *is a query space $(\mathcal{P}, Q, f)$ where $\mathcal{P}$ is an (n, m)-set, Q is the set $\mathcal{X}_k$, and $f = \tilde{D}$; see Section 2.*

In what follows we define some measure of combinatorial complexity for a query space.

Definition A.2 (Definition 4.5 in (Braverman et al., 2016))**.** *For a query space $(\mathcal{P}, Q, f)$, a query $C \in Q$ and $r \in [0, \infty)$ we define*

$$\text{range}(\mathcal{P}, C, r) = \left\{P \in \mathcal{P} \middle| f(P, c) \leq r\right\}.$$

Let $ranges(\mathcal{P}, Q, f) = \{\text{range}\,(\mathcal{P}, C, r)|C \in Q, r \geq 0\}$, the VC-dimension of $(P, ranges(\mathcal{P}, Q, f))$ is the smallest integer d' such that for every $\mathcal{H} \subseteq \mathcal{P}$ we have

$$\left|\left\{\text{range}(C, r)\middle|C \in Q, r \in [0, \infty)\right\}\right| \leq |\mathcal{H}|^{d'}.$$

The dimension of the query space $(\mathcal{P}, Q, f)$ is the VC-dimension of $(P, ranges(\mathcal{P}, Q, f))$.

Lemma A.3 (Variant of Theorem 8.4, (Anthony & Bartlett, 2009))**.** *Suppose h is a function from $\mathbb{R}^d \times \mathbb{R}^n$ to $\{0, 1\}$ and let*

$$H = \left\{h(a, x)\middle|a \in \mathbb{R}^d, x \in \mathbb{R}^n\right\}$$

be the class determined by h. Suppose that h can be computed by an algorithm that takes as an input a pair $(a, x) \in \mathbb{R}^d \times \mathbb{R}^n$ and returns $h(a, x)$ after no more than t operations of the following types:

- *the arithmetic operations $+, -, \times$, and $/$ on real numbers,*
- *jumps conditioned on $>, \geq, <, \leq, =$, and $\neq$ comparisons of real numbers, and*
- *outputs 0 or 1.*

Then the VC-dimension of H is $O\left(dt\right)$.

We now bound the dimension of a query space $(\mathcal{P}, \mathcal{X}_k, \tilde{D})$ as in Definition A.2.

Lemma A.4. *Let $(\mathcal{P}, \mathcal{X}_k, \tilde{D})$ be a sets clustering query space; see Definition A.1. Then the dimension d' of $(\mathcal{P}, \mathcal{X}_k, \tilde{D})$ is bounded by $\in O(md^2k^2)$.*

Proof. For $P \in \mathcal{P}$, $C \in \mathcal{X}_k$, and $r \in \mathbb{R}$, let $h_P(C, r) = 1$ if $\tilde{D}(P, C) \geq r$ and 0 otherwise. Then we observe that the VC-dimension of the class of functions $H = \left\{h_P : \mathcal{X}_k \times \mathbb{R} \to [0, \infty)\middle|P \in \mathcal{P}\right\}$ in Lemma A.3 is equivalent to the dimension d' of the given query space. Therefore, we now show that the VC-dimension of H is bounded by $O(md^2k^2)$.

Note that it takes $t = O(mdk)$ arithmetic operations to evaluate $h_P(C, r)$. Furthermore, any element in $\mathcal{X}_k \times \mathbb{R}$ can be represented as a vector in $(dk + 1)$-dimensional space. Hence by Lemma A.3, the VC-dimension of H is $O(dk \cdot mdk) = O(md^2k^2)$. □

B. Main theorems with full proof

B.1. Proof of Lemma 4.1

Lemma B.1. *Let $k \geq 1$ be an integer, $A, B \subseteq \mathcal{X}$ and $C \in \mathcal{X}_k$. If $\tilde{D}(A \cup B, C) \neq \tilde{D}(B, C)$ then $\tilde{D}(A \cup B, C) = \tilde{D}(A, C)$.*

Proof. By definition, $\tilde{D}(A \cup B, C) = \min\left\{\tilde{D}(A, C), \tilde{D}(B, C)\right\}$. By the assumption of the lemma, $\tilde{D}(A \cup B, C) \neq \tilde{D}(B, C)$. Therefore, $\tilde{D}(A \cup B, C) = \tilde{D}(A, C)$ □

Lemma B.2. *Let $A = \{a_1, \cdots, a_n\} \subseteq \mathcal{X}$ and put $b \in \mathcal{X}$. Let $B = (A \setminus \{a_1\}) \cup \{b\} = \{b, a_2, \cdots, a_n\} \subseteq \mathcal{X}$. Then for every $C \in \mathcal{X}_k$ we have that*

$$\tilde{D}(A, C) \leq \rho\left(\tilde{D}(B, C) + \tilde{D}(a_1, b)\right).$$

Proof. By definition, we have that

$$\begin{aligned}
&\tilde{D}(A, C)\\
&= \min\left\{\tilde{D}(a_1, C), \tilde{D}(A \setminus \{a_1\}, C)\right\}\\
&\leq \min\left\{\rho\left(\tilde{D}(a_1, b) + \tilde{D}(b, C)\right),\right.\\
&\qquad\qquad \left.\tilde{D}(A \setminus \{a_1\}, C)\right\}\\
&\leq \min\left\{\rho\left(\tilde{D}(a_1, b) + \tilde{D}(b, C)\right),\right.\\
&\qquad\qquad \left.\rho\left(\tilde{D}(A \setminus \{a_1\}, C) + \tilde{D}(a_1, b)\right)\right\}\\
&\leq \rho \min\left\{\tilde{D}(b, C), \tilde{D}(A \setminus \{a_1\}, C)\right\} + \rho\tilde{D}(a_1, b)\\
&= \rho\tilde{D}(B, C) + \rho\tilde{D}(a_1, b),
\end{aligned}$$

where the first inequality is by the weak triangle inequality by Lemma 2.2, and the last derivation is by the definition of B. □

Lemma 4.1. *Let $\mathcal{P}$ be an (n, m)-set, $k \geq 1$ be an integer and $(\mathcal{X}, \tilde{D})$ be as in Definition 2.1. Let $(\mathcal{P}^m, \mathcal{B}^m)$ be the* *output of a call to* RECURSIVE-ROBUST-MEDIAN$(\mathcal{P}, k)$*; see Algorithm 1. Then, for every $P \in \mathcal{P}^m$ we have that*

$$\sup_{C \in \mathcal{X}_k} \frac{\tilde{D}(P, C)}{\sum\limits_{Q \in \mathcal{P}} \tilde{D}(Q, C)} \in O(1) \cdot \left(\frac{1}{|\mathcal{P}^m|} \right).$$

Proof. In what follows, we use the variables and notations from Algorithm 1. Put $P \in \mathcal{P}^m$, $i \in [m]$, and consider the ith iteration of the "for" loop at Line 4 of Algorithm 1. Put $C \in \mathcal{X}_k$.

Let

$$\overline{\mathcal{P}}^{i-1} = \left\{ Q \in \mathcal{P}^{i-1} \middle| \tilde{D}(\mathrm{T}(Q, \mathcal{B}^{i-1}), C) = \tilde{D}(\mathcal{B}^{i-1}, C) \right\}$$

be the union of sets $Q \in \mathcal{P}^{i-1}$ whose closest point to the query C after the projection on $\mathcal{B}^{i-1}$ is one of the points of $\mathcal{B}^{i-1}$. First we prove that

$$\frac{\tilde{D}(\mathrm{T}(P, \mathcal{B}^{i-1}), C)}{\sum\limits_{Q \in \mathcal{P}^{i-1}} \tilde{D}(\mathrm{T}(Q, \mathcal{B}^{i-1}), C)} \leq 3\rho^2 \frac{\tilde{D}(\mathrm{T}(P, \mathcal{B}^{i}), C)}{\sum\limits_{Q \in \mathcal{P}^{i}} \tilde{D}(\mathrm{T}(Q, \mathcal{B}^{i}), C)} + \frac{4\rho}{|\mathcal{P}^i|} \tag{3}$$

by the following case analysis: **(i)** $\left|\overline{\mathcal{P}}^{i-1}\right| \geq \frac{|\mathcal{P}^{i-1}|}{2}$, i.e., more than half the sets satisfy that their closest point to C is amongst their projected points onto $\mathcal{B}^{i-1}$, and **(ii)** Otherwise, i.e., $\left|\overline{\mathcal{P}}^{i-1}\right| < \frac{|\mathcal{P}^{i-1}|}{2}$.

Case (i): $\left|\overline{\mathcal{P}}^{i-1}\right| \geq \frac{|\mathcal{P}^{i-1}|}{2}$. By Line 7 we have

$$\mathcal{P}^i \subseteq \mathcal{P}^{i-1} \subseteq \cdots \subseteq \mathcal{P}^0 = \mathcal{P}. \tag{4}$$

Therefore,

$$\sum_{Q \in \mathcal{P}^{i-1}} \tilde{D}(\mathrm{T}(Q, \mathcal{B}^{i-1}), C) \geq \sum_{Q \in \overline{\mathcal{P}}^{i-1}} \tilde{D}(\mathrm{T}(Q, \mathcal{B}^{i-1}), C) \tag{5}$$

$$= \sum_{Q \in \overline{\mathcal{P}}^{i-1}} \tilde{D}(\mathcal{B}^{i-1}, C) \geq \frac{\left|\mathcal{P}^{i-1}\right|}{2} \tilde{D}(\mathcal{B}^{i-1}, C), \tag{6}$$

where (5) holds since $\overline{\mathcal{P}}^{i-1} \subseteq \mathcal{P}^{i-1}$, the first derivation in (6) is by the definition of $\overline{\mathcal{P}}^{i-1}$, and the second derivation in (6) is by the assumption of Case (i). This proves (3) for Case (i) as

$$\frac{\tilde{D}(\mathrm{T}(P, \mathcal{B}^{i-1}), C)}{\sum\limits_{Q \in \mathcal{P}^{i-1}} \tilde{D}(\mathrm{T}(Q, \mathcal{B}^{i-1}), C)} \leq \frac{\tilde{D}(\mathcal{B}^{i-1}, C)}{\sum\limits_{Q \in \mathcal{P}^{i-1}} \tilde{D}(\mathrm{T}(Q, \mathcal{B}^{i-1}), C)}$$

$$\leq \frac{\tilde{D}(\mathcal{B}^{i-1}, C)}{\frac{|\mathcal{P}^{i-1}|}{2} \tilde{D}(\mathcal{B}^{i-1}, C)} = \frac{2}{|\mathcal{P}^{i-1}|} \leq \frac{2}{|\mathcal{P}^{i}|}, \tag{7}$$

where the first inequality holds since $\mathcal{B}^{i-1} \subseteq \mathrm{T}(P, \mathcal{B}^{i-1})$ by Definition 2.5, and the second inequality is by (6).

Case (ii): $\left|\overline{\mathcal{P}}^{i-1}\right| < \frac{|\mathcal{P}^{i-1}|}{2}$. Let $\gamma = 1/(2k)$. Let $\hat{\mathcal{P}}^{i-1}$, b^i and $\mathcal{P}^i$ be as defined in Lines 5, 6, and 7 respectively, and identify $\mathcal{B}^{i-1} = \left\{b^1, \cdots, b^{i-1}\right\}$ for $i \geq 2$ or $\mathcal{B}^{i-1} = \emptyset$ for $i = 1$. Let

$$OPT_i = \min_{C' \in \mathcal{X}_k} \sum_{\hat{P} \in \mathrm{closest}(\hat{\mathcal{P}}^{i-1}, C', 1/2)} \tilde{D}(\hat{P}, C'). \tag{8}$$

For every $Q \in \mathcal{P}^{i-1}$, substituting $A = \overline{\mathrm{proj}}(Q, \mathcal{B}^{i-1})$ and $B = \mathcal{B}^{i-1}$ in Lemma B.1 proves that

$$\begin{aligned} &\left\{ \begin{array}{c} Q \\ \in \mathcal{P}^{i-1} \end{array} \middle| \begin{array}{c} \tilde{D}\big(\overline{\mathrm{proj}}(Q, \mathcal{B}^{i-1}) \cup \mathcal{B}^{i-1}, C\big) \\ \neq \tilde{D}(\mathcal{B}^{i-1}, C) \end{array} \right\} \\ \subseteq &\left\{ \begin{array}{c} Q \\ \in \mathcal{P}^{i-1} \end{array} \middle| \begin{array}{c} \tilde{D}\big(\overline{\mathrm{proj}}(Q, \mathcal{B}^{i-1}) \cup \mathcal{B}^{i-1}, C\big) \\ = \tilde{D}\big(\overline{\mathrm{proj}}(Q, \mathcal{B}^{i-1}), C\big) \end{array} \right\} \end{aligned} \tag{9}$$

We now obtain that

$$\left| \left\{ \begin{array}{c} Q \\ \in \mathcal{P}^{i-1} \end{array} \middle| \begin{array}{c} \tilde{D}\left(\mathrm{T}(Q, \mathcal{B}^{i-1}), C\right) \\ = \tilde{D}\left(\overline{\mathrm{proj}}(Q, \mathcal{B}^{i-1}), C\right) \end{array} \right\} \right|$$

$$= \left| \left\{ \begin{array}{c} Q \\ \in \mathcal{P}^{i-1} \end{array} \middle| \begin{array}{c} \tilde{D}\big(\overline{\mathrm{proj}}(Q, \mathcal{B}^{i-1}) \cup \mathcal{B}^{i-1}, C\big) \\ = \tilde{D}\left(\overline{\mathrm{proj}}(Q, \mathcal{B}^{i-1}), C\right) \end{array} \right\} \right| \tag{10}$$

$$\geq \left| \left\{ \begin{array}{c} Q \\ \in \mathcal{P}^{i-1} \end{array} \middle| \begin{array}{c} \tilde{D}\big(\overline{\mathrm{proj}}(Q, \mathcal{B}^{i-1}) \cup \mathcal{B}^{i-1}, C\big) \\ \neq \tilde{D}\left(\mathcal{B}^{i-1}, C\right) \end{array} \right\} \right| \tag{11}$$

$$= \left| \left\{ Q \in \mathcal{P}^{i-1} \middle| \begin{array}{c} \tilde{D}\left(\mathrm{T}(Q, \mathcal{B}^{i-1}), C\right) \\ \neq \tilde{D}(\mathcal{B}^{i-1}, C) \end{array} \right\} \right| \tag{12}$$

$$= \left|\mathcal{P}^{i-1} \setminus \overline{\mathcal{P}}^{i-1}\right| \geq \frac{\left|\mathcal{P}^{i-1}\right|}{2}, \tag{13}$$

where (10) and (12) is by substituting $\mathcal{P} = Q$ and $\mathcal{B} = \mathcal{B}^{i-1}$ in Definition 2.5, (11) is by (9), the first derivation in (13) is by the definitions of $\mathcal{P}^{i-1}$ and $\overline{\mathcal{P}}^{i-1}$, and the last inequality is by the assumption of Case (ii).

Recall that by Line 5,

$$\hat{\mathcal{P}}^{i-1} = \left\{ \overline{\mathrm{proj}}(Q, \mathcal{B}^{i-1}) \middle| Q \in \mathcal{P}^{i-1} \right\},$$

and let

$$Z = \left\{ Q \in \mathcal{P}^{i-1} \middle| \overline{\mathrm{proj}}(Q, \mathcal{B}^{i-1}) \in \mathrm{closest}(\hat{\mathcal{P}}^{i-1}, C, 1/2) \right\}.$$

Since Z contains the $|Z| \leq \frac{|\mathcal{P}^{i-1}|}{2}$ sets $Q \in \mathcal{P}^{i-1}$ with the smallest $\tilde{D}(\overline{\mathrm{proj}}(Q, \mathcal{B}^{i-1}), C)$, for any set $Z' \subseteq \mathcal{P}^{i-1}$ such that $|Z'| \geq \frac{|\mathcal{P}^{i-1}|}{2}$, we have

$$\sum_{Q \in Z} \tilde{D}(\overline{\mathrm{proj}}(Q, \mathcal{B}^{i-1}), C) \leq \sum_{Q \in Z'} \tilde{D}(\overline{\mathrm{proj}}(Q, \mathcal{B}^{i-1}), C). \tag{14}$$

By the assumption of Case (ii),

$$\left|\mathcal{P}^{i-1} \setminus \overline{\mathcal{P}}^{i-1}\right| \geq \frac{\left|\mathcal{P}^{i-1}\right|}{2}, \tag{15}$$

and by the definition of Z, we have

$$\left\{\overline{\mathrm{proj}}(Q, \mathcal{B}^{i-1}) \middle| Q \in Z\right\} = \mathrm{closest}(\hat{\mathcal{P}}^{i-1}, C, 1/2). \tag{16}$$

Therefore,

$$\begin{aligned}
&\sum_{\hat{Q} \in \mathrm{closest}(\hat{\mathcal{P}}^{i-1}, C, 1/2)} \tilde{D}(\hat{Q}, C) \\
&= \sum_{Q \in Z} \tilde{D}(\overline{\mathrm{proj}}(Q, \mathcal{B}^{i-1}), C) \\
&\leq \sum_{Q \in \mathcal{P}^{i-1} \setminus \overline{\mathcal{P}}^{i-1}} \tilde{D}(\overline{\mathrm{proj}}(Q, \mathcal{B}^{i-1}), C),
\end{aligned} \tag{17}$$

where the first derivation is by (16) and the last derivation is by substituting $Z' = \mathcal{P}^{i-1} \setminus \overline{\mathcal{P}}^{i-1}$ in (14). By the definitions of $\mathcal{P}^{i-1}$ and $\overline{\mathcal{P}}^{i-1}$, for every $Q \in \mathcal{P}^{i-1} \setminus \overline{\mathcal{P}}^{i-1}$, we have

$$\tilde{D}(\mathrm{T}(Q, \mathcal{B}^{i-1}), C) = \tilde{D}(\overline{\mathrm{proj}}(Q, \mathcal{B}^{i-1}), C). \tag{18}$$

Hence,

$$\mathrm{OPT}_i \leq \sum_{\hat{Q} \in \mathrm{closest}(\hat{\mathcal{P}}^{i-1}, C, 1/2)} \tilde{D}(\hat{Q}, C) \tag{19}$$

$$\leq \sum_{Q \in \mathcal{P}^{i-1} \setminus \overline{\mathcal{P}}^{i-1}} \tilde{D}(\overline{\mathrm{proj}}(Q, \mathcal{B}^{i-1}), C) \tag{20}$$

$$= \sum_{Q \in \mathcal{P}^{i-1} \setminus \overline{\mathcal{P}}^{i-1}} \tilde{D}(\mathrm{T}(Q, \mathcal{B}^{i-1}), C) \tag{21}$$

$$\leq \sum_{Q \in \mathcal{P}^{i-1}} \tilde{D}(\mathrm{T}(Q, \mathcal{B}^{i-1}), C), \tag{22}$$

where (19) holds by the definition of OPT_i, (20) is by (17), and (21) is by (18).

Recall that $P \in \mathcal{P}^m$, identify

$$\mathrm{closepairs}(P, \mathcal{B}^m) = \left\{(\hat{p}_1, \hat{b}_1), \cdots, (\hat{p}_m, \hat{b}_m)\right\},$$

as in Definition 2.5 (i). Also by Definition 2.5, for every $i \in [m]$ we have

$$\begin{aligned}
&\tilde{D}(\overline{\mathrm{proj}}(P, \mathcal{B}^{i-1}), \hat{b}_i) \\
&= \tilde{D}(P \setminus \{\hat{p}_1, \cdots, \hat{p}_{i-1}\}, \hat{b}_i) \\
&= \tilde{D}(\hat{p}_i, \hat{b}_i).
\end{aligned} \tag{23}$$

Since $P \in \mathcal{P}^i$ and $\gamma = \frac{1}{2k}$, we have by Line 7 that

$$\overline{\mathrm{proj}}(P, \mathcal{B}^{i-1}) \in \mathrm{closest}(\hat{\mathcal{P}}^{i-1}, \left\{b^i\right\}, (1-\tau)\gamma/2). \tag{24}$$

Observe that in the definition of OPT_i in (8), the largest cluster in every set C' of k centers contains at least $\frac{|\hat{\mathcal{P}}^{i-1}|}{2k} = \gamma|\hat{\mathcal{P}}^{i-1}|$ points by the Pigeonhole Principle. Therefore, since the cost of the closest $(1-\tau)\gamma|\hat{\mathcal{P}}^{i-1}|$ sets for $\hat{b}^i$ is a 2-approximation for the optimal set of $\gamma|\hat{\mathcal{P}}^{i-1}|$ points, we have

$$\begin{aligned}
&\sum_{Q \in \mathrm{closest}(\hat{\mathcal{P}}^{i-1}, \left\{\hat{b}_i\right\}, (1-\tau)\gamma)} \tilde{D}(Q, \hat{b}_i) \\
&\leq 2 \min_{\{b\} \in \mathcal{X}_1} \sum_{Q \in \mathrm{closest}(\hat{\mathcal{P}}^{i-1}, \{b\}, \gamma)} \tilde{D}(Q, b) \\
&\leq 2 \cdot \mathrm{OPT}_i.
\end{aligned} \tag{25}$$

Therefore,

$$\tilde{D}(\hat{p}_i, \hat{b}_i) = \tilde{D}(\overline{\mathrm{proj}}(P, \mathcal{B}^{i-1}), \hat{b}_i) \tag{26}$$

$$\leq 2 \cdot \frac{\sum_{Q \in \mathrm{closest}(\hat{\mathcal{P}}^{i-1}, \left\{\hat{b}_i\right\}, (1-\tau)\gamma)} \tilde{D}(Q, \hat{b}_i)}{(1-\tau)\gamma\left|\hat{\mathcal{P}}^{i-1}\right|} \tag{27}$$

$$\leq 2 \cdot \frac{\sum_{Q \in \mathrm{closest}(\hat{\mathcal{P}}^{i-1}, \left\{\hat{b}_i\right\}, (1-\tau)\gamma)} \tilde{D}(Q, \hat{b}_i)}{|\mathcal{P}^i|} \tag{28}$$

$$\leq \frac{4\mathrm{OPT}_i}{|\mathcal{P}^i|}, \tag{29}$$

where (26) is by (23), (27) is by combining Markov's Inequality with (24), (28) follows since $|\mathcal{P}^i| = \frac{(1-\tau)\gamma}{2}|\mathcal{P}^{i-1}| \leq (1-\tau)\gamma|\mathcal{P}^{i-1}|$, and (29) is by (25).

Now, since the sets $\mathrm{T}(P, \mathcal{B}^{i-1})$ and $\mathrm{T}(P, \mathcal{B}^i)$ differ by at most one point, i.e.,

$$\mathrm{T}(P, \mathcal{B}^i) = \left(\mathrm{T}(P, \mathcal{B}^{i-1}) \setminus \{\hat{p}_i\}\right) \cup \left\{\hat{b}_i\right\},$$

by substituting $A = \mathrm{T}(P, \mathcal{B}^{i-1})$, and $B = \mathrm{T}(P, \mathcal{B}^i)$ in Lemma B.2, we obtain that

$$\tilde{D}(\mathrm{T}(P, \mathcal{B}^{i-1}), C) \leq \rho\tilde{D}(\mathrm{T}(P, \mathcal{B}^i), C) + \rho\tilde{D}(\hat{p}_i, \hat{b}_i). \tag{30}$$

By the previous inequality we obtain

$$\begin{aligned}
\frac{\tilde{D}(\mathrm{T}(P, \mathcal{B}^{i-1}), C)}{\sum_{Q \in \mathcal{P}^{i-1}} \tilde{D}(\mathrm{T}(Q, \mathcal{B}^{i-1}), C)} &\leq \rho\frac{\tilde{D}(\mathrm{T}(P, \mathcal{B}^i), C)}{\sum_{Q \in \mathcal{P}^{i-1}} \tilde{D}(\mathrm{T}(Q, \mathcal{B}^{i-1}), C)} \\
&\quad + \rho\frac{\tilde{D}(\hat{p}_i, \hat{b}_i)}{\sum_{Q \in \mathcal{P}^{i-1}} \tilde{D}(\mathrm{T}(Q, \mathcal{B}^{i-1}), C)}.
\end{aligned} \tag{31}$$

We now bound the rightmost term of (31) as

$$\rho \frac{\tilde{D}(\hat{p}_i,\hat{b}_i)}{\sum_{Q\in\mathcal{P}^{i-1}} \tilde{D}(\mathrm{T}(Q,\mathcal{B}^{i-1}),C)} \leq \rho\frac{\tilde{D}(\hat{p}_i,\hat{b}_i)}{\mathrm{OPT}_i} \tag{32}$$

$$\leq \rho\frac{4\mathrm{OPT}_i}{|\mathcal{P}^i|\,\mathrm{OPT}_i} = 4\rho\frac{1}{|\mathcal{P}^i|}, \tag{33}$$

where (32) is by (22), and the first derivation in (33) is by (29).

We now bound the middle term of (31). By identifying $\mathrm{closepairs}(Q,\mathcal{B}^m) = \left\{(\hat{q}_1,\hat{b}_1),\cdots,(\hat{q}_m,\hat{b}_m)\right\}$ for every $Q\in\mathcal{P}^i$, we have,

$$\sum_{Q\in\mathcal{P}^i} \tilde{D}(\mathrm{T}(Q,\mathcal{B}^i),C)$$

$$\leq \rho\sum_{Q\in\mathcal{P}^i} \tilde{D}(\mathrm{T}(Q,\mathcal{B}^{i-1}),C) + \rho\sum_{Q\in\mathcal{P}^i}\tilde{D}(\hat{q}_i,\hat{b}_i) \tag{34}$$

$$\leq \rho\sum_{Q\in\mathcal{P}^i} \tilde{D}(\mathrm{T}(Q,\mathcal{B}^{i-1}),C) + \rho\left|\mathcal{P}^i\right|\frac{2\mathrm{OPT}_i}{|\mathcal{P}^i|} \tag{35}$$

$$\leq \rho\sum_{Q\in\mathcal{P}^{i-1}} \tilde{D}(\mathrm{T}(Q,\mathcal{B}^{i-1}),C) + 2\rho\mathrm{OPT}_i \tag{36}$$

$$\leq (\rho+2\rho)\sum_{Q\in\mathcal{P}^{i-1}} \tilde{D}(\mathrm{T}(Q,\mathcal{B}^{i-1}),C), \tag{37}$$

where (34) follows similarly to (30), (35) holds similarly to (29) for the set Q instead of P, (36) holds since $\mathcal{P}^i\subseteq\mathcal{P}^{i-1}$ by (4) and (37) is by (22). Thus, by (37), the middle term of (31) is bounded by

$$\rho\frac{\tilde{D}(\mathrm{T}(P,\mathcal{B}^i),C)}{\sum_{Q\in\mathcal{P}^{i-1}}\tilde{D}(\mathrm{T}(Q,\mathcal{B}^{i-1}),C)} \leq 3\rho^2\frac{\tilde{D}(\mathrm{T}(P,\mathcal{B}^i),C)}{\sum_{Q\in\mathcal{P}^{i}}\tilde{D}(\mathrm{T}(Q,\mathcal{B}^{i}),C)}. \tag{38}$$

By combining (31), (33) and (38), we get that

$$\begin{aligned}&\frac{\tilde{D}(\mathrm{T}(P,\mathcal{B}^{i-1}),C)}{\sum_{Q\in\mathcal{P}^{i-1}}\tilde{D}(\mathrm{T}(Q,\mathcal{B}^{i-1}),C)}\\ &\leq 3\rho^2\frac{\tilde{D}(\mathrm{T}(P,\mathcal{B}^i),C)}{\sum_{Q\in\mathcal{P}^{i}}\tilde{D}(\mathrm{T}(Q,\mathcal{B}^{i}),C)} + 4\rho\frac{1}{|\mathcal{P}^i|}.\end{aligned} \tag{39}$$

Now (3) holds by taking the maximum between the bounds of Case (i) in (7), and the bound of Case (ii) in (39).

We can now apply (3) recursively over every $i\in[m]$ to obtain that

$$\frac{\tilde{D}(P,C)}{\sum_{Q\in\mathcal{P}}\tilde{D}(Q,C)} = \frac{\tilde{D}(\mathrm{T}(P,\mathcal{B}^0),C)}{\sum_{Q\in\mathcal{P}^0}\tilde{D}(\mathrm{T}(Q,\mathcal{B}^0),C)} \tag{40}$$

$$\leq (3\rho)^{2m}\frac{\tilde{D}(\mathrm{T}(P,\mathcal{B}^m),C)}{\sum_{Q\in\mathcal{P}^m}\tilde{D}(\mathrm{T}(Q,\mathcal{B}^m),C)} + 4\rho\sum_{i\in[m]}\frac{(3\rho^2)^{i-1}}{|\mathcal{P}^i|}. \tag{41}$$

Also, for every $Q\in\mathcal{P}^m$ observe that $|Q| = |\mathcal{B}^m| = m$, hence

$$\mathrm{T}(Q,\mathcal{B}^m) = \mathcal{B}^m = \left\{\hat{b}_1,\cdots,\hat{b}_m\right\}.$$

Thus, for every $Q\in\mathcal{P}^m$ and $C\in\mathcal{X}_k$

$$\tilde{D}(\mathrm{T}(Q,\mathcal{B}^m),C) = \tilde{D}\left(\left\{\hat{b}_1,\cdots,\hat{b}_m\right\},C\right) \tag{42}$$

Lemma 4.1 now holds as

$$\frac{\tilde{D}(P,C)}{\sum_{Q\in\mathcal{P}}\tilde{D}(Q,C)} \leq \frac{(3\rho^2)^m}{|\mathcal{P}^m|} + 4\rho\sum_{i\in[m]}\frac{(3\rho^2)^{i-1}}{|\mathcal{P}^i|} \tag{43}$$

$$\leq \frac{(3\rho^2)^m}{|\mathcal{P}^m|} + 4\rho\sum_{i\in[m]}\frac{(3\rho^2)^{i-1}}{|\mathcal{P}^m|} \tag{44}$$

$$\leq \frac{(3\rho^2)^m}{|\mathcal{P}^m|} + \frac{4\rho}{|\mathcal{P}^m|}\cdot\frac{(3\rho^2)^{m-1}-1}{(3\rho^2)-1} \tag{45}$$

$$\leq \frac{(3\rho^2)^m}{|\mathcal{P}^m|} + \frac{4\rho}{|\mathcal{P}^m|}\cdot(3\rho^2)^m \tag{46}$$

$$\leq \frac{5\rho(3\rho^2)^m}{|\mathcal{P}^m|}, \tag{47}$$

where (43) holds by plugging (42) in (40), (44) holds since $|\mathcal{P}^m|\leq\left|\mathcal{P}^i\right|$ for every $i\in[m]$, (45) holds by summing the geometric sequence, and inequalities (46) and (47) hold since $\rho\geq 1$. □

B.2. Proof of Theorem 4.2

Theorem 4.2. *Let $\mathcal{P}$ be an (n,m)-set, $k\geq 1$ be an integer, $(\mathcal{X},\tilde{D})$ be as in Definition 2.1, and $\varepsilon,\delta\in(0,1)$. Let $(\mathcal{S},v)$ be the output of a call to* CORESET$(\mathcal{P},k,\varepsilon,\delta)$. *Then*

(i) $|\mathcal{S}|\in O\left(\left(\frac{md\log n}{\varepsilon}\right)^2 k^{m+4}\right)$.

(ii) With probability at least $1-\delta$, $(\mathcal{S},v)$ is an ε-coreset for $(\mathcal{P},\mathcal{X}_k,\tilde{D})$; see Section 1.4.

(iii) (S,v) can be computed in $O(n\log(n)(k)^m)$ time.

Proof. **(i):** Let J denote the number of while iterations in Algorithm 2, and for every $j\in[J]$ let $\mathcal{P}^0_{(j)}$, $\mathcal{P}^m_{(j)}$ and

$\mathcal{B}^m_{(j)}$ denote respectively the sets $\mathcal{P}^0$, $\mathcal{P}^m$ and $\mathcal{B}^m$ at the jth while iteration of Algorithm 2.

By Line 7 of Algorithm 1, we observe that the output set $\mathcal{P}^m$ is of size $|\mathcal{P}^m| \geq \frac{|\mathcal{P}|}{(bk)^m}$ for some constant b, where $\mathcal{P}$ is the input set to the algorithm. Therefore, the size of $\mathcal{P}^m_j$ returned at Line 7 of algorithm 2 in the jth while iteration is

$$\left|\mathcal{P}^m_{(j)}\right| \geq \frac{\left|\mathcal{P}^0_{(j)}\right|}{(bk)^m}. \tag{48}$$

By (48) and Line 11 of Algorithm 2, we obtain that

$$\begin{aligned}
\left|\mathcal{P}^0_{(j+1)}\right| \leq \left|\mathcal{P}^0_{(j)}\right| - \left|\mathcal{P}^m_{(j)}\right| &\leq \left|\mathcal{P}^0_{(j)}\right| - \frac{\left|\mathcal{P}^0_{(j)}\right|}{(bk)^m} \\
&= \left|\mathcal{P}^0_{(j)}\right| \left(1 - \frac{1}{(bk)^m}\right) \\
&= \left|\mathcal{P}^0_{(1)}\right| \left(1 - \frac{1}{(bk)^m}\right)^j \\
&= n \left(1 - \frac{1}{(bk)^m}\right)^j,
\end{aligned} \tag{49}$$

where the second derivation is by (48). Combining that $\left|\mathcal{P}^0_{(J)}\right| \geq 1$ with (49) we conclude that

$$J \leq (bk)^m \log n. \tag{50}$$

Therefore, by Lines 9 and 14 of Algorithm 2, the total sensitivity computed at Line 16 of Algorithm 2 is equal to

$$\begin{aligned}
t = \sum_{P \in \mathcal{P}} s(P) &\leq \sum_{j \in [J]} \left(\sum_{P \in \mathcal{P}^m_{(j)}} \frac{b}{\left|\mathcal{P}^m_{(j)}\right|} \right) + O(1) \\
&= \sum_{j \in [J]} b + O(1) = Jb + O(1) \leq (bk)^{m+1} \log n.
\end{aligned}$$

By this and Line 17 of Algorithm 2,

$$|S| = \frac{(bk)^{m+1} \log n}{\varepsilon^2} \left(\log\left((bk)^{m+1} \log n\right) d' + \log\left(\frac{1}{\delta}\right) \right).$$

where $d' = O(md^2k^2)$ is the dimension of the sets clustering query space $(\mathcal{P}, \mathcal{X}_k, \tilde{D})$; see Section A. By simple derivations we obtain that:

$$|S| \in O\left(\left(\frac{md \log n}{\varepsilon} \right)^2 k^{m+4} \right).$$

(ii): The pair $(\mathcal{P}^m_{(j)}, \mathcal{B}^m_{(j)})$ satisfy Lemma 4.1 for every $j \in [J]$. Hence, with an appropriate b (determined from the proof of Lemma 4.1), for every $P \in \mathcal{P}^m_{(j)}$ the value $s(P)$ defined at Lines 9 and 14 satisfies for every $C \in \mathcal{X}_k$ that

$$s(P) = \frac{b}{\left|\mathcal{P}^m_{(j)}\right|} \geq \frac{\tilde{D}(P, C)}{\sum_{Q \in \mathcal{P}^0_{(j)}} \tilde{D}(Q, C)} \geq \frac{\tilde{D}(P, C)}{\sum_{Q \in \mathcal{P}} \tilde{D}(Q, C)}.$$

By Theorem 3.1, a sample S of $|S| \leq \frac{bt}{\varepsilon^2}\left(\log(t)d' + \log\left(\frac{1}{\delta}\right)\right)$ is an ε-coreset for (the sets clustering query space) $(\mathcal{P}, \mathcal{X}_k, \tilde{D})$. Therefore, by Theorem 3.1, the pair $(\mathcal{S}, v)$ computed at Lines 17–19 satisfies Property (ii) of Theorem 4.2.

Computational time. Consider a call $\textsc{Recursive-Robust-Median}(\mathcal{P}, k)$ to Algorithm 1 where $\mathcal{P}$ is an (n, m)-set. The ith iteration of the for loop at Line 4 takes $O\left(n\left(\frac{1}{(4k)}\right)^{i-1} + k^4\right)$ time. Summing over all the m iterations yields a total running time of $O(n + mk^4)$.

Consider the call $(\mathcal{P}^m, \mathcal{B}^m) := \textsc{Recursive-Robust-Median}(\mathcal{P}^0, k)$ at Line 7 of Algorithm 2, which dominates the running time of this algorithm. This call is made J times (in each of the J iterations of the while loop). The set $\mathcal{P}^0$ at the ith call is of size $s_i = O\left(n\left(1 - \frac{1}{(4k)}\right)^{i-1}\right)$. Therefore, the ith such call takes $O(s_i + mk^4)$ time. Summing this running time over every $i \in [J]$, where $J \leq (bk)^m \log n$ by (50), yields a total running time of

$$J \cdot mk^4 + n \sum_{i=1}^{J} \left(1 - \frac{1}{(4k)}\right)^{i-1} \in O\left(n \log(n) (bk)^m\right).$$

□

C. Polynomial Time Approximation Scheme

The following theorem states that given n polynomials in d (constant number of) variables of constant degree, then the space $\mathbb{R}^d$ can be decomposed into a polynomial (n^d) number of cells, such that for every d variables C from the cell Δ the sign sequence of all the polynomials is the same cell.

Theorem C.1 (Theorem 3.4 in (Chazelle et al., 1991))**.** *Let d be a constant and let $\mathcal{F} = \{\mathrm{pl}_1, \cdots, \mathrm{pl}_n\}$ be a set of n multivariate polynomials of constant degree with range $\mathbb{R}^d$ and image $\mathbb{R}$. It is possible to split $\mathbb{R}^d$ into $O\left(n^{2d-2}\right)$ cells $\Delta(\mathcal{F}) = \{\Delta_i\}$, with the property that for every polynomials pl_i and every cell Δ_j it holds that pl_i is either positive, negative, or equal to 0 on the entire cell Δ_j. This decomposition, including a set of points $A = \{a_i\}$ with $a_i \in \Delta_i$ can be found in time $O\left(n^{2d-1} \log n\right)$.*

C.1. Proof of Theorem 4.4

Theorem 4.4. *Let $\mathcal{P}$ be an (n, m)-set in $\mathbb{R}^d$, $w : \mathcal{P} \to [0, \infty)$ be a weights function, $k \geq 1$ be an integer, $\alpha \geq 1$ and $\delta \in [0, 1)$. Let $\tilde{D}$ be a loss function as in Definition 2.1 for $\mathcal{X} = \mathbb{R}^d$. Let* ALG *be an algorithm that solves the case where $k = m = 1$, i.e., it takes as input a set $Q \subseteq$*

$\mathcal{X}$, a weights function $u : Q \to [0, \infty)$ and the failure probability δ, and in time $T(n)$ outputs $\hat{c} \in \mathcal{X}$ that with probability at least $1 - \delta$ satisfies $\sum_{q \in Q} u(q) \cdot \tilde{D}(q, \hat{c}) \leq \alpha \cdot \min_{c \in \mathcal{X}} \sum_{q \in Q} u(q) \cdot \tilde{D}(q, c)$. Then in $T(n) \cdot (nmk)^{O(dk)}$ time we can compute $\hat{C} \in \mathcal{X}_k$ such that with probability at least $(1 - k \cdot \delta)$ we have

$$\sum_{P \in \mathcal{P}} w(P) \cdot \tilde{D}(P, \hat{C}) \leq \alpha \cdot \min_{C \in \mathcal{X}_k} \sum_{P \in \mathcal{P}} w(P) \cdot \tilde{D}(P, C).$$

Proof. What follows is a constructive proof for the theorem. Algorithm 4 gives a suggested implementation.

Identify $\mathcal{P} = \{P_1, \cdots, P_n\}$ where $P_i = \left\{p_1^i, \cdots, p_m^i\right\}$ for every $i \in [n]$.

First we define a set of $n^2 m^2 k^2$ polynomials as follows. For every $i, i' \in [n]$, $j, j' \in [m]$, $\ell, \ell' \in [k]$ and vector $x = (x_1^T | \cdots | x_k^T) \in \mathbb{R}^{dk}$ of dk unknowns ($x_1, \cdots, x_k$ are vectors in $\mathbb{R}^d$) , let

$$\mathrm{pl}_{i,j,\ell,i',j',\ell'}(x) = \left\|p_j^i - x_\ell\right\|^2 - \left\|p_{j'}^{i'} - x_{\ell'}\right\|^2$$

be a polynomial in those dk unknowns, of degree at most 2, and let $\mathcal{F}$ be a set that contains all those polynomials. Here, each polynomial in $\mathcal{F}$ contains up to $2d$ variables, and $|\mathcal{F}| = n^2 m^2 k^2$. A polynomial $\mathrm{pl}_{i,j,\ell,i',j',\ell'}(x)$ is positive iff $p_{j'}^{i'}$ is closer to $x_{\ell'}$ than the distance between p_j^i and x_ℓ. Therefore, given a possible assignment $x' = ({x'}_1^T | \cdots | {x'}_k^T) \in \mathbb{R}^{dk}$ for the dk unknowns, the vector of sign values of the polynomials in $\mathcal{F}$ when plugging x' corresponds to a clustering of $\mathcal{P}$ into k clusters centered at ${x'}_1^T, \cdots, {x'}_k^T$, and indicates which point in each input m-set is the closest to this cluster center, and vice versa, as follows. Given x', the first cluster $\mathcal{C}_1 \subseteq \mathbb{R}^d$ contains all the points p_j^i such that for every $j' \in [m]$ and $\ell' \in [k]$,

$$\left\|p_j^i - x_1\right\|^2 \leq \left\|p_{j'}^i - x_{\ell'}\right\|^2.$$

Which, by the definition of the polynomials in $\mathcal{F}$, means that for every $j' \in [m]$ and $\ell' \in [k]$,

$$\mathrm{sign}(\mathrm{pl}_{i,j,1,i,j',\ell'}(x')) = -1.$$

This enables us to compute the points $\mathcal{C}_1, \cdots, \mathcal{C}_k \subseteq \mathbb{R}^d$ of each cluster that are induced by the sign sequence of $\mathcal{F}$ when plugging x'. Given those clusters $\mathcal{C}_1, \cdots, \mathcal{C}_k \subseteq \mathbb{R}^d$, we can apply ALG to each such cluster $\mathcal{C}_i$ (since $m = k = 1$), to obtain, with probability at least $1 - \delta$, the optimal point $\hat{c}_i \in \mathbb{R}^d$ that minimizes $\sum_{p \in \mathcal{C}_i} \tilde{D}(p, z)$ over every $z \in \mathbb{R}^d$, and its cost $cost_i = \sum_{p \in \mathcal{C}_i} \tilde{D}(p, \hat{c}_i)$. The sum $\sum_{i=1}^k cost_i$ is the total cost of this clustering option of $\mathcal{P}$.

Since ALG is used to compute k centers of k clusters, the probability that $\hat{c}_1, \cdots, \hat{c}_k$ are the optimal centers is at least $1 - k\delta$.

By Theorem C.1, we can decompose $\mathbb{R}^{dk}$ into $|\Delta(\mathcal{F})| = (nmk)^{O(dk)}$ cells $\{\Delta_j\}$, such that the sign of each polynomial $\mathrm{pl}_i \in \mathcal{F}$ in an entire cell $\Delta_j \in \Delta(\mathcal{F})$ is the same, i.e., the sign sequence of all the polynomials in $\mathcal{F}$ is the same over the entire cell Δ'. Hence, the number of different such sign sequences is at most the number of different cells, which is $(nmk)^{O(dk)}$.

By iterating over every cell $\Delta' \in \Delta(\mathcal{F})$ and taking the sign sequence of the polynomials in $\mathcal{F}$ in this cell, we would have covered all the different sign sequences, which correspond to all the feasible clustering options of $\mathcal{P}$ into k clusters. For each option we can evaluate the total cost as described above, and pick the clustering with the smallest total cost.

The running time of such an algorithm is dominated by the computation of such an arrangement of $\mathbb{R}^{dk}$, and by calling ALG $|\Delta(\mathcal{F})|$ times; once for each region $\Delta' \in \Delta(\mathcal{F})$. Computing this arrangement takes $nmk^{O(dk)}$ time by Theorem C.1 and produces $|\Delta(\mathcal{F})| \in (nmk)^{O(dk)}$ cells. Now it takes $T(n) \cdot (nmk)^{O(dk)}$ total time for the calls to ALG. □

C.2. Proof of Corollary 4.5

Corollary 4.4 (PTAS for sets-k-means)**.** *Let $\mathcal{P}$ be an (n, m)-set, $k \geq 1$ be an integer, and put $\varepsilon \in \left(0, \frac{1}{2}\right]$ and $\delta \in (0, 1)$. Let* OPT *be the cost of the sets-k-means. Then in $n \log(n)(k)^m + \left(\frac{\log n}{\varepsilon} dmk^m\right)^{O(dk)}$ time we can compute $\hat{C} \in \mathcal{X}_k$ such that with probability at least $1 - k \cdot \delta$,*

$$\sum_{P \in \mathcal{P}} \min_{p \in P, c \in \hat{C}} \|p - c\|^2 \leq (1 + 4\varepsilon) \cdot \mathrm{OPT}.$$

Proof. We will first compute a coreset for the input $\mathcal{P}$ and the given cost function $\tilde{D}$ and query set $\mathcal{X}_k$, and then find the sets-k-means for the (weighted) coreset using Theorem 4.4.

Recall that in this sets-k-means problem, $\tilde{D}(P, C) = \min_{p \in P, c \in C} \|p - c\|^2$ for every $P, C \subseteq \mathbb{R}^d$.

Let $(\mathcal{S}, v)$ be an output of a call to CORESET$(\mathcal{P}, k, \varepsilon, \delta)$; see Algorithm 2. Then by Theorem 4.2, $(\mathcal{S}, v)$ is an ε-coreset for $(\mathcal{P}, \mathcal{X}_k, \tilde{D})$ of size $|\mathcal{S}| \in O\left(\left(\frac{mdk \log n}{\varepsilon}\right)^2 k^{O(m)}\right)$ with probability at least $1 - \delta$ which is computed in $O\left(n \log(n)(bk)^m\right)$; see Section 1.4.

Let $Q \subseteq \mathcal{X}$ be a set of size $|Q| = n$ and let $u : Q \to [0, \infty)$ be a weights function. Let ALG be an algorithm that takes Q and u as input and returns the point $c^* := \frac{\sum_{q \in Q} u(q) \cdot q}{\sum_{q \in Q} u(q)} \in \mathcal{X}$. Observe that c^* minimizes its sum of weighted squared distances to the points of Q, i.e.,

$$\begin{aligned}\sum_{q\in Q} u(q)\tilde{D}(q,c^*) &= \sum_{q\in Q} u(q)\left\|q-c^*\right\|^2\\ &= \min_{c\in\mathcal{X}}\sum_{q\in Q} u(q)\left\|q-c\right\|^2 = \min_{c\in\mathcal{X}}\sum_{q\in Q} u(q)\tilde{D}(q,c).\end{aligned}$$

Furthermore, observe that c^* can be computed in $T(n) = O(n)$ time.

Plugging $\mathcal{P} = \mathcal{S}$, $w = v$,Q, u, ALG, $\alpha = 1$ and $T(|\mathcal{S}|) = O(|\mathcal{S}|)$ in Theorem 4.4 yields that in $(|\mathcal{S}|\, mk)^{O(dk)} \in \left(\frac{\log n}{\varepsilon} dmk^m\right)^{O(dk)}$ time we can compute $\hat{C} \in \mathcal{X}_k$ such that with probability at least $1 - k \cdot \delta$,

$$\sum_{P\in\mathcal{S}} v(P)\cdot\tilde{D}(P,\hat{C}) = \min_{C\in\mathcal{X}_k}\sum_{P\in\mathcal{S}} v(P)\cdot\tilde{D}(P,C). \quad (51)$$

Hence, the total running time for obtaining $\hat{C}$ is $\left(\frac{\log n}{\varepsilon} dmk^m\right)^{O(dk)} + O\left(n\log(n)(bk)^m\right)$.

Corollary 4.5 now holds as

$$\begin{aligned}\sum_{P\in\mathcal{P}}\min_{p\in P, c\in\hat{C}}\left\|p-c\right\|^2 &= \sum_{P\in\mathcal{P}}\tilde{D}(P,\hat{C})\\ &\le \frac{1}{1-\varepsilon}\cdot\sum_{P\in\mathcal{S}} v(P)\cdot\tilde{D}(P,\hat{C}) && (52)\\ &\le (1+2\varepsilon)\cdot\sum_{P\in\mathcal{S}} v(P)\cdot\tilde{D}(P,\hat{C}) && (53)\\ &= (1+2\varepsilon)\cdot\min_{C\in\mathcal{X}_k}\sum_{P\in\mathcal{S}} v(P)\cdot\tilde{D}(P,C) && (54)\\ &\le (1+2\varepsilon)(1+\varepsilon)\cdot\min_{C\in\mathcal{X}_k}\sum_{P\in\mathcal{P}}\tilde{D}(P,C) && (55)\\ &\le (1+4\varepsilon)\cdot\min_{C\in\mathcal{X}_k}\sum_{P\in\mathcal{P}}\tilde{D}(P,C) && (56)\\ &= (1+4\varepsilon)\cdot\min_{C\in\mathcal{X}_k}\sum_{P\in\mathcal{P}}\min_{p\in P, c\in C}\left\|p-c\right\|^2,\end{aligned}$$

where (52) and (55) hold since $(\mathcal{S}, v)$ is an ε-coreset for $(\mathcal{P}, \mathcal{X}_k, \tilde{D})$, (53) and (56) hold since $\varepsilon \le \frac{1}{2}$ and (54) is by (51). □

C.3. Suggested implementation

In this section we give a suggested implementation for the constructive proof of Theorem 4.4; see Algorithm 4.

Overview of Algorithm 4. Algorithm 4 gets as input a set $\mathcal{P}$ of m-sets, an integer $k \ge 1$, an error parameter $\varepsilon \in (0,1)$ and the probability of failure $\delta \in (0,1)$. The algorithm returns as output a set $\hat{C} \in \mathcal{X}_k$ of k centers that approximate the optimal cost of the k-means for set

Algorithm 4 PTAS$(\mathcal{P}, w, k, \text{ALG})$

```
Input: An $(n,m)$-set $\mathcal{P}$, a weights function $w : \mathcal{P} \to [0,\infty)$, a positive integer $k$, and an algorithm ALG as in Theorem 4.4.
Output: A set $\hat{C} \in \arg\min_{C\in\mathcal{X}_k} \sum_{P\in\mathcal{P}} w(P)\tilde{D}(P,C)$;
        see Theorem 4.4.
Identify $\mathcal{P} = \{P_1, \cdots, P_n\}$ where $P_i = \{p_1^i, \cdots, p_m^i\}$ for every $i \in [n]$.
Define $w'(p) := w(P)$ for every $p \in P$ and $P \in \mathcal{P}$.
Let $x = (x_1^T|\cdots|x_k^T)^T \in \mathbb{R}^{dk}$ be a vector of $dk$ unknowns.
for every $i, i' \in [n], j, j' \in [m], \ell, \ell' \in [k]$ do
    $\text{pl}_{i,j,\ell,i',j',\ell'}(x) = \left\|p_j^i - x_\ell\right\|^2 - \left\|p_{j'}^{i'} - x_{\ell'}\right\|^2$
    {A polynomial of degree 2 containing up to $2d$ unknowns from $x$. If this polynomial is positive iff $p_{j'}^{i'}$ is closer to $x_{\ell'}$ than the distance between $p_j^i$ and $x_\ell$.}
    $\mathcal{F} := \mathcal{F} \cup \{\text{pl}_{i,j,\ell,i',j',\ell'}(x)\}$
end for
Compute a decomposition of $\mathbb{R}^{dk}$ into cells $\Delta(\mathcal{F}) = \{\Delta_j\}$ as described in Theorem C.1, and let $A$ contain a representative $a \in \Delta'$ from each cell $\Delta' \in \Delta(\mathcal{F})$.
$min = \infty$
for every $a \in A$ do
    $sum = 0$
    for every $\ell \in [k]$ do
        $\mathcal{C}_\ell := \left\{ p_j^i \;\middle|\; \begin{array}{c} i \in [n], j \in [m] \text{ s.t.} \\ \forall j' \in [m], \ell' \in [k] \\ \text{sign}\left(\text{pl}_{i,j,\ell,i,j',\ell'}(a)\right) = -1 \end{array} \right\}$
        {The points of cluster number $\ell$ defined by the sign sequence of the cell representative $a \in A$.}
        $(\hat{c}_\ell, cost_\ell) := \text{ALG}(\mathcal{C}_\ell, w')$.
        {Compute the optimal center $(k = 1)$ $\hat{c}_\ell$ of the set $\mathcal{C}_\ell \subseteq \mathbb{R}^d$ $(m = 1)$ and its cost $\text{cost}_\ell$, for a given cost function.}
        $sum = sum + cost_\ell$
    end for
    if $sum < min$ then
        $min = sum$
        $\hat{C} = \{\hat{c}_1, \cdots, \hat{c}_k\}$
    end if
end for
Return $\hat{C}$
```

D. Robust Median

D.1. Proof of Lemma 5.1

Algorithm 3 overview: The algorithm relies on the 2 following observations: (i) To compute a robust approximation of the entire data, it suffices to compute a robust approximation of a randomly sampled subset of this data of sufficient size; see Line 4 of Algorithm 3 and Lemma D.1, (ii) If b is a robust approximation of some input set of elements, then by the (weak) triangle inequality for singletons, one of those elements is a constant factor approximation for b; see Line 5 of Algorithm 3.

Lemma D.1. *Let $\mathcal{P}$ be an (n,m)-set, $k \geq 1$, $\delta, \gamma \in (0,1)$, and $\tau \in (0, 1/10)$. Pick uniformly, i.i.d, a (multi)-set $\mathcal{S}$ of*

$$|\mathcal{S}| = \frac{c}{\tau^4\gamma^2}\left(md^2 + \log\left(\frac{1}{\delta}\right)\right)$$

elements from $\mathcal{P}$, where c is a sufficiently large universal constant. Then with probability at least $1-\delta$, any $((1-\tau)\gamma, \tau, 2)$-median of $\mathcal{S}$ is also a $(\gamma, 4\tau, 2)$-median of $\mathcal{P}$.

Proof. For every $P \in \mathcal{P}$ and $b \in \mathcal{X}_1$ define $f_P(b) = \tilde{D}(P, b)$. Let $F = \left\{f_P \middle| P \in \mathcal{P}\right\}$ and $F_{\mathcal{S}} = \left\{f_P \middle| P \in \mathcal{S}\right\}$. Observe that by Definition 4.2 in (Feldman & Langberg, 2011), the dimension of the function space (F, X_1) is equivalent to the dimension $d' = md^2$ of the query space $(\mathcal{P}, X_1, \tilde{D})$. Since $F_{\mathcal{S}}$ is a random sample of $\frac{c}{\tau^4\gamma^2}\left(d' + \log\left(\frac{1}{\delta}\right)\right) = \frac{c}{\tau^4\gamma^2}\left(md^2 + \log\left(\frac{1}{\delta}\right)\right)$ functions, sampled i.i.d from F, Lemma D.1 now holds by Theorem 9.6 in (Feldman & Langberg, 2011) which states that a $((1-\tau)\gamma, \tau, 2)$-median of $F_{\mathcal{S}}$ (which in our case is a $((1-\tau)\gamma, \tau, 2)$-median of $\mathcal{S}$) is a $(\gamma, 4\tau, 2)$-median of F (which in our case is a $(\gamma, 4\tau, 2)$-median of $\mathcal{P}$). □

Lemma 5.1 (based on Lemma 9.6 in (Feldman & Langberg, 2011))**.** *Let $\mathcal{P}$ be an (n,m)-set, $k \geq 1$, $\delta \in (0,1)$ and $(\mathcal{X}, \tilde{D})$ be as in Definition 2.1. Let $q \in \mathcal{X}$ be the output of a call to* MEDIAN$(\mathcal{P}, k, \delta)$*; see Algorithm 3. Then with probability at least $1-\delta$, q is a $(1/(2k), 1/6, 2)$-median for P; see Definition 2.4. Furthermore, q can be computed in $O\left(tb^2k^4\log^2\left(\frac{1}{\delta}\right)\right)$ time, where t is the time it takes to compute $\tilde{D}(P,Q)$ for $P, Q \in \mathcal{P}$.*

Proof. Let $\gamma = 1/(2k)$ and $\tau = 1/24$. For a sufficient constant b, the random sample S in Line 4 satisfies Lemma D.1. Therefore,

$$\begin{aligned}&\text{a } (23/(48k), 1/24, 2)\text{-median of } \mathcal{S} \text{ is also a}\\ &(1/(2k), 1/6, 2)\text{-median of } \mathcal{P}.\end{aligned} \tag{57}$$

Let $q^*_{\mathcal{S}}$ be the $(23/(48k), 0, 0)$-median of $\mathcal{S}$, and let $q'_{\mathcal{S}}$ be the closest point in $\mathcal{S}$ to $q^*_{\mathcal{S}}$, i.e.,

$$q'_{\mathcal{S}} \in \operatorname*{arg\,min}_{q \in Q : Q \in \mathcal{S}} \tilde{D}(q^*_{\mathcal{S}}, q).$$

By the weak triangle inequality from Lemma 2.2, we have that $\tilde{D}(P, q'_{\mathcal{S}}) \leq 2\rho\tilde{D}(P, q^*_{\mathcal{S}})$ for every $P \in \mathcal{S}$, i.e., that $q'_{\mathcal{S}}$ is a 2-approximation for $q^*_{\mathcal{S}}$. This yields that $q'_{\mathcal{S}}$ is a $(23/(48k), 0, 2)$-median of $\mathcal{S}$, which is also a $(23/(48k), 1/6, 2)$-median of $\mathcal{S}$. Hence, one of the points of $\mathcal{S}$ is a $(23/(48k), 1/6, 2)$-median of $\mathcal{S}$. Therefore, the point q computed at Line 5 and returned in Line 6 is such a $(23/(48k), 1/6, 2)$-median of $\mathcal{S}$, which by (57) is also a $(1/(2k), 1/6, 2)$-median of $\mathcal{P}$.

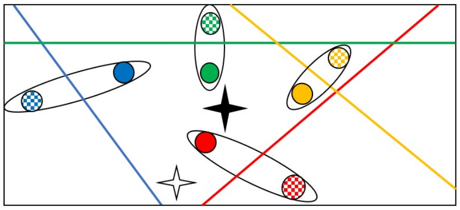

Figure 6. **`exact-mean` via sets Voronoi diagram.** A set of $n = 4$ pairs on the plane ($m = d = 2$) and its sets Voronoi diagram which is computed as follows: (i) A set voronoi diagram is computed for each pair (m-set) to obtain a set of hyperplanes, (ii) an arrangement of those hyperplanes is then computed, which results in a partition of $\mathbb{R}^2$ into the cells which are presented above. Each cell corresponds to a selection of representatives, one from each pair. The sets-mean c^* (solid star) is also the 1-mean of the representative points shown in solid circles, which correspond to this Voronoi cell. Any other point (empty star) inside the same Voronoi cell as c^* admits the same set of representatives. Therefore, to compute c^*, it suffices to exhaustive search over all the Voronoi.

The computation time of Algorithm 3 is dominated by Line 5, which can be implemented in $t|\mathcal{S}|^2 = tb^2k^4\log^2\left(\frac{1}{\delta}\right)$ time by simply computing the pairwise distances between every two sets in $\mathcal{S}$ and using order statistics. □

E. Implemented Algorithms

`exact-mean`$(\mathcal{P})$ is implemented by what we call *sets Voronoi diagram*; see Fig. 6.

k**`-means`**$(\mathcal{P}, k)$**.** We focused on the sets-k-means case (see Section 1 and Table 1), where the clustering algorithm we applied is a modified version of the the well know Lloyd algorithm (Lloyd, 1982) as follows. The algorithm starts by an initial k random centers $C \subseteq \left\{p \in P \middle| P \in \mathcal{P}\right\}$. It then assigns every $P \in \mathcal{P}$ to its closest center $c_P = \arg\min_{c \in C} \tilde{D}(P, c)$. Finally, it replaces every $c \in C$ with the sets-mean of the (possibly weighted) sets $\left\{P \in \mathcal{P} \middle| c_P = c\right\}$ in its cluster. It repeats this process till convergence, but no more than 12 iterations. The sets-mean is computed as follows.

`approx-mean`$(\mathcal{P}, t)$**.** As explained in Section 1, computing the sets-mean c^* is a non-trivial and time consuming task. However, at least $|\mathcal{P}|/2$ of the input sets $P \in \mathcal{P}$ satisfy that $\tilde{D}(P, c^*) \leq \frac{2\sum_{Q \in \mathcal{P}} \tilde{D}(Q, c^*)}{n}$. By the triangle inequality for singletons (Lemma 2.2), it follows immediately that the closest point $p \in P$ to c^* is a 3-approximation for c^*. Therefore, with probability at least $1/2$, one of the points of a randomly sampled input set is a good approximation. We can amplify this probability by sampling $t \geq 1$ such sets.

Handling sets of different sizes. For example in dataset (ii), each newspaper P_i consists of different number of paragraphs and hence is represented by a different number $|P_i|$ of vectors. Let z denote the maximal such set size. To compute a coreset for such dataset $\mathcal{P}$, we first partition $\mathcal{P}$ into z sets $\mathcal{P} = \mathcal{P}_1 \cup \cdots \cup \mathcal{P}_z$ where $\mathcal{P}_i$ contains all the sets $P \in \mathcal{P}$ of size $|P| = i$. Then, for every $i \in [z]$, we plug $\mathcal{P}^0 = \mathcal{P}_i$ at Lines 5– 14 of Algorithm 2 to compute $s(P)$ for every $P \in \mathcal{P}_i$. In other words, we compute the sensitivity bound for each set on its own. We compute the total sensitivity $t_i := \sum_{P \in \mathcal{P}_i} s_i(P)$ of each set $\mathcal{P}_i$ and $t := \sum_{i \in [m]} t_i$ to be their total. We then simply perform Lines 17– 21.