CHAPTER 15

Epsilon-Approximations: Functional Case

Given a set P of n points in $\mathbb{R}^d$ and a parameter $\epsilon > 0$, let A be an ϵ-approximation of the primal set system induced on P by balls in $\mathbb{R}^d$. Then clearly A can be used to approximate P 'combinatorially' with respect to balls:

> Let $P_{q,r} = \text{Ball}\,(q,r) \cap P$ be the set of points of P contained in the ball of radius r centered at q; similarly set $A_{q,r} = \text{Ball}\,(q,r) \cap A$.
>
> Then for any $q \in \mathbb{R}^d$ and $r > 0$, $|P_{q,r}|$ can be approximated by $|A_{q,r}| \cdot \frac{|P|}{|A|}$, since by the definition of ϵ-approximations,

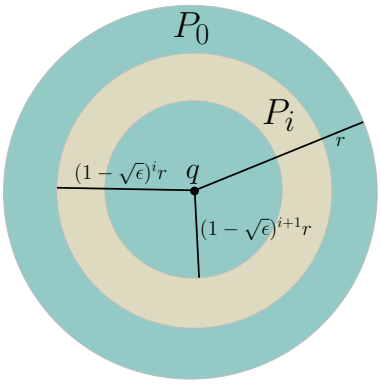
$$|A_{q,r}| = \frac{|P_{q,r}|\,|A|}{|P|} \pm \epsilon|A| \quad \text{or equivalently,}$$

$$|P_{q,r}| = |A_{q,r}| \cdot \frac{|P|}{|A|} \pm \epsilon|P|. \tag{15.1}$$

The new idea in this chapter is the observation that since for any $q \in \mathbb{R}^d$, Equation (15.1) holds for *every* radius r, the set A can also be used to approximate the sum of distances from q to the points of P. In particular, here is another property that holds for A: for any $q \in \mathbb{R}^d$ and $r > 0$,

$$\left| \frac{\sum_{p \in P_{q,r}} \text{dist}\,(p,q)}{n} - \frac{\sum_{p \in A_{q,r}} \text{dist}\,(p,q)}{|A|} \right| \leq 3\epsilon r. \tag{15.2}$$

To see the intuition for Equation (15.2), we sketch the proof for a weaker bound of $3\sqrt{\epsilon}r$.



Recall that we say A is an ϵ-approximation of a set $P' \subseteq P$ if $|P' \cap A| = \frac{|P'|\,|A|}{|P|} \pm \epsilon|A|$.

Partition $P_{q,r}$ into disjoint sets $P_0, P_1, \ldots$, where $p \in P_i$ if and only if

$$\operatorname{dist}(p,q) \in \left(\left(1-\sqrt{\epsilon}\right)^{i+1} r, \left(1-\sqrt{\epsilon}\right)^{i} r \right].$$

That is, P_i is the set of points of P lying in the region

$$\operatorname{Ball}\left(q, \left(1-\sqrt{\epsilon}\right)^{i} r\right) \setminus \operatorname{Ball}\left(q, \left(1-\sqrt{\epsilon}\right)^{i+1} r\right).$$

Now the sum of distances of the points of A to q can be approximated by summing up over the P_i's:

$$\left(\sum_{i=0}^{\infty} \left(1-\sqrt{\epsilon}\right)^{i+1} r \cdot |P_i \cap A|\right) < \sum_{p \in A_{q,r}} \operatorname{dist}(p,q) \leq \left(\sum_{i=0}^{\infty} \left(1-\sqrt{\epsilon}\right)^{i} r \cdot |P_i \cap A|\right).$$

We remark that we only need to do the above sum till index $i = \frac{1}{\sqrt{\epsilon}} \ln \frac{1}{\epsilon}$, as after that the average sum of distances is most ϵr in any case. Using the fact that A is a 2ϵ-approximation of each P_i (by Claim 13.6),

$$\begin{aligned}
\left(\sum_{i=0}^{\infty} \left(1-\sqrt{\epsilon}\right)^{i+1} r \left(\frac{|P_i|\,|A|}{n} - 2\epsilon|A|\right)\right) &< \sum_{p \in A_{q,r}} \operatorname{dist}(p,q) \leq \\
&\left(\sum_{i=0}^{\infty} \left(1-\sqrt{\epsilon}\right)^{i} r \left(\frac{|P_i|\,|A|}{n} + 2\epsilon|A|\right)\right) \\
\left(1-\sqrt{\epsilon}\right) \left(\sum_{i=0}^{\infty} \left(1-\sqrt{\epsilon}\right)^{i} \left(\frac{|P_i|}{n} - 2\epsilon\right)\right) &< \frac{\sum_{p \in A_{q,r}} \operatorname{dist}(p,q)}{r\,|A|} \leq \\
&\left(\sum_{i=0}^{\infty} \left(1-\sqrt{\epsilon}\right)^{i} \left(\frac{|P_i|}{n} + 2\epsilon\right)\right).
\end{aligned}$$

Using the fact that $2\epsilon \sum_{i=0}^{\infty} \left(1-\sqrt{\epsilon}\right)^{i} = 2\epsilon \cdot \frac{1}{1-(1-\sqrt{\epsilon})} = 2\sqrt{\epsilon}$,

$$\begin{aligned}
\left(1-\sqrt{\epsilon}\right) \left(\left(\frac{1}{n} \sum_{i=0}^{\infty} \left(1-\sqrt{\epsilon}\right)^{i} |P_i|\right) - 2\sqrt{\epsilon}\right) &< \frac{\sum_{p \in A_{q,r}} \operatorname{dist}(p,q)}{r\,|A|} \leq \\
&\left(\frac{1}{n} \sum_{i=0}^{\infty} \left(1-\sqrt{\epsilon}\right)^{i} |P_i|\right) + 2\sqrt{\epsilon}.
\end{aligned}$$

Similarly approximating $\sum_{p \in P_{q,r}} \operatorname{dist}(p,q)$ over the P_i's gives

$$\begin{aligned}
\left(1-\sqrt{\epsilon}\right) \left(\sum_{i=0}^{\infty} \left(1-\sqrt{\epsilon}\right)^{i} r\,|P_i|\right) &< \sum_{p \in P_{q,r}} \operatorname{dist}(p,q) \leq \\
&\left(\sum_{i=0}^{\infty} \left(1-\sqrt{\epsilon}\right)^{i} r\,|P_i|\right).
\end{aligned}$$

Dividing by rn,

$$\begin{aligned}
\left(1-\sqrt{\epsilon}\right) \left(\frac{1}{n} \sum_{i=0}^{\infty} \left(1-\sqrt{\epsilon}\right)^{i} |P_i|\right) &< \frac{\sum_{p \in P_{q,r}} \operatorname{dist}(p,q)}{rn} \leq \\
&\left(\frac{1}{n} \sum_{i=0}^{\infty} \left(1-\sqrt{\epsilon}\right)^{i} |P_i|\right).
\end{aligned}$$

These together imply the desired bound:

$$\begin{aligned}
\left| \frac{\sum_{p \in P_{q,r}} \operatorname{dist}(p,q)}{r\,n} - \frac{\sum_{p \in A_{q,r}} \operatorname{dist}(p,q)}{r\,|A|} \right| &\leq 2\sqrt{\epsilon} + \sqrt{\epsilon} \left(\frac{1}{n} \sum_{i=0}^{\infty} \left(1-\sqrt{\epsilon}\right)^{i} |P_i|\right) \\
&< 2\sqrt{\epsilon} + \sqrt{\epsilon} \left(\frac{1}{n} \sum_{i=0}^{\infty} |P_i|\right) \\
&\leq 3\sqrt{\epsilon}.
\end{aligned}$$

As we will see later, the improvement to $3\epsilon r$ presented later in this chapter follows with more precise calculations.

Consider now another application of the same idea on a slightly more complicated distance function where the point q is replaced by a set of k points. For any set X of k points in $\mathbb{R}^d$ and $p \in P$, define

$$\operatorname{dist}(p, X) = \min_{q \in X} \operatorname{dist}(p, q). \tag{15.3}$$

Further let

$$P_{X,r} = \{p \in P : \operatorname{dist}(p, X) \leq r\}.$$

Observe that $P_{X,r}$ is the set of points of P that lie in the union of the k balls of radius r centered at the points of X. Denote this union by $\operatorname{Ball}(X, r)$.

As earlier, our goal is to estimate, for any given $X \in \left(\mathbb{R}^d\right)^k$ and $r \geq 0$, the expression

$$\sum_{p \in P_{X,r}} \operatorname{dist}(p, X).$$

Not surprisingly, if A is an ϵ-approximation of the set system induced on P by the union of k balls, then one can show that

$$\left| \frac{\sum_{p \in P_{X,r}} \operatorname{dist}(p, X)}{n} - \frac{\sum_{p \in A_{X,r}} \operatorname{dist}(p, X)}{|A|} \right| \leq 3\epsilon r.$$

The first result of this chapter is a more general statement which implies both the above two instances.

The second result is its application to an algorithmic problem central to several domains: the k-median clustering problem, where given a set P of points in $\mathbb{R}^d$ and an integer parameter $k > 0$, the goal is to partition the points of P into k clusters based on certain geometric criteria.

1. A Functional View of Approximations

A common error of judgment among mathematicians is the confusion between telling the truth and giving a logically correct presentation. The two objectives are antithetical and hard to reconcile. Most presentations obeying the current Diktats of linear rigor are a long way from telling the truth; any reader of such a presentation is forced to start writing on the margin, or deciphering on a separate sheet of paper.

The truth of any piece of mathematical writing consists of realizing what the author is "up to"; it is the tradition of mathematics to do whatever it takes to avoid giving away this secret.

Gian-Carlo Rota

Recall the following statement.

Let P be a set of n points in $\mathbb{R}^d$. Each $p \in P$ defines the function

$$\operatorname{dist}(p, X) = \min_{q \in X} \operatorname{dist}(p, q),$$

where X is a finite set of points in $\mathbb{R}^d$.

Then for any positive integer k and $\epsilon > 0$, there exists an ϵ-approximation $A \subseteq P$ such that for any set X of k points in $\mathbb{R}^d$ and $r \in \mathbb{R}^+$,

$$\left| \frac{\sum_{p \in P_{X,r}} \operatorname{dist}(p, X)}{n} - \frac{\sum_{p \in A_{X,r}} \operatorname{dist}(p, X)}{|A|} \right| \leq 3\epsilon r, \tag{15.4}$$

where $P_{X,r} = \{p \in P \colon \operatorname{dist}(p, X) \leq r\}$ and $A_{X,r} = A \cap P_{X,r}$.

Note that the role of each $p \in P$ is captured by the function $\operatorname{dist}(p, \cdot)$. We now prove Equation (15.4) in an abstract setting where $\operatorname{dist}(p, \cdot)$ is replaced by an arbitrary function $g_p \colon \mathcal{X} \to \mathbb{R}^+$, where $\mathcal{X}$ is a given domain. That is,

$$\begin{aligned} \text{set of all } k\text{-tuples of points in } \mathbb{R}^d \quad &\longrightarrow \quad \text{a domain } \mathcal{X}, \\ \text{set of } n \text{ functions } \operatorname{dist}(p, \cdot),\ p \in P \quad &\longrightarrow \quad \text{set } G \text{ of } n \text{ functions from } \mathcal{X} \text{ to } \mathbb{R}^+, \\ P_{X,r} = \{p \in P \colon \operatorname{dist}(p, X) \leq r\} \quad &\longrightarrow \quad G_{X,r} = \{g \in G \colon g(X) \leq r\}. \end{aligned}$$

The main theorem of this section states and proves the analog of Equation (15.4) in this abstract setting for a set G of n functions.

THEOREM 15.5. *Let $G = \{g_1, \ldots, g_n\}$ be a set of n functions over a domain $\mathcal{X}$[1], where $g_i \colon \mathcal{X} \to \mathbb{R}^+$. Define the set system $(G, \mathcal{F})$, with*

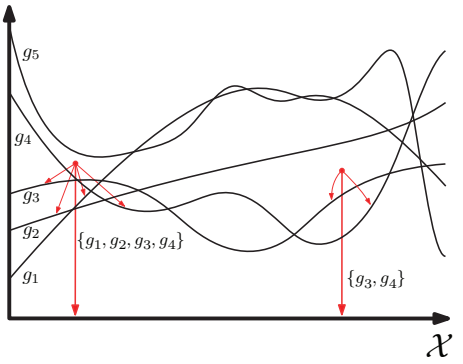
$$\mathcal{F} = \{G_{X,r} \colon X \in \mathcal{X} \text{ and } r \in \mathbb{R}^+\}, \quad \text{where} \quad G_{X,r} = \{g \in G \colon g(X) \leq r\}.$$

Let $A \subseteq G$ be an ϵ-approximation of $\mathcal{F}$, for a given parameter $\epsilon > 0$. Then for any $X \in \mathcal{X}$ and $r \in \mathbb{R}^+$, setting $A_{X,r} = A \cap G_{X,r}$, we have

$$\left| \frac{\sum_{g \in G_{X,r}} g(X)}{|G|} - \frac{\sum_{g \in A_{X,r}} g(X)}{|A|} \right| \leq 3\epsilon r.$$

To visualize Theorem 15.5, consider the case when $\mathcal{X} = \mathbb{R}$. The figure illustrates an example of five functions $g_1, \ldots, g_5$. In this example, the set $G_{X,r}$ is simply the set of functions lying below the point (X, r).

[1] $\mathcal{X}$ need not be finite. In the previous example, $\mathcal{X}$ was the set of all k-tuples of points in $\mathbb{R}^d$.



Before we proceed to the proof, we illustrate the versatility of Theorem 15.5 by showing a specific consequence.

Let P be a set of n points in $\mathbb{R}^d$ and $\mathcal{X}$ the set of all k-tuples of points in $\mathbb{R}^d$. Additionally, for each $p \in P$ we are given a function $f_p \colon \mathcal{X} \to \mathbb{R}^+$. Set

$$G = \{f_p \colon p \in P\}.$$

For each $X \in \mathcal{X}$, let r_X be the smallest value for which $G_{X,r} = G$. That is,

$$r_X = \max_{p \in P} f_p(X).$$

Applying Theorem 15.5 to P and G, we arrive at the following.

COROLLARY 15.6. *Let P be a set of n points in $\mathbb{R}^d$ and k a positive integer. Further each $p \in P$ has an associated function*

$$f_p \colon \left(\mathbb{R}^d\right)^k \to \mathbb{R}^+.$$

These functions define a set system $(P, \mathcal{R})$, with

$$\mathcal{R} = \left\{ P_{X,r} \colon X \in \left(\mathbb{R}^d\right)^k \quad \text{and} \quad r \in \mathbb{R}^+ \right\},$$
$$\text{where } P_{X,r} = \left\{ p \in P \colon f_p(X) \le r \right\}.$$

Let A be an ϵ-approximation of $\mathcal{R}$. Then for any $X \subseteq \mathbb{R}^d$ with $|X| = k$, we have

$$\left| \frac{\sum_{p \in P} f_p(X)}{|P|} - \frac{\sum_{p \in A} f_p(X)}{|A|} \right| \le 3\epsilon r_X = 3\epsilon \max_{p \in P} f_p(X).$$

Note that for the case where $f_p(X) = \operatorname{dist}(p, X)$, $\mathcal{R}$ is precisely the primal set system induced on P by the union of k equal-radius balls in $\mathbb{R}^d$.

Overview of ideas. For a fixed $X \in \mathcal{X}$ and $r \in \mathbb{R}^+$, we need to relate the quantities

$$\sum_{g \in G_{X,r}} g(X) \qquad \text{and} \qquad \sum_{g \in A_{X,r}} g(X).$$

As earlier, one way to proceed is to partition $G_{X,r}$ into disjoint sets $G_0, G_1, \ldots$, where

$$G_i = \left\{ g \in G \,:\, g(X) \in \left((1-\epsilon)^{i+1} \, r, (1-\epsilon)^i \, r \right] \right\}.$$

Then all the functions in G_i have approximately the same value on X, and so one can approximately bound the summation of functions in G_i in terms of $|G_i|$, and the summation of functions in $A \cap G_i$ in terms of $|A \cap G_i|$.

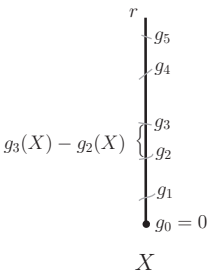
However, we present a different proof based on an elegant trick: we sort the functions in $G_{X,r}$ by increasing $g(X)$ values, and rewrite each of the above two summations as sums of the interval lengths between two consecutive function values in the sorted order. Concretely,

let $G_{X,r} = \{g_1, \ldots, g_t\}$, sorted by increasing $g(X)$ values.

Then any interval

$$C_i = g_i(X) - g_{i-1}(X)$$

is 'contributed' in $\sum_{g \in G_{X,r}} g(X)$ by precisely the functions $\{g_i, \ldots, g_t\}$ (see figure). Summing over all consecutive intervals, and using the fact that A is an 2ϵ-approximation of each $\{g_i, \ldots, g_t\}$, we get the required bound.



In the formal proof one has to be a little careful though, as multiple functions might have the same value on X.

PROOF OF THEOREM 15.5. Fix any $X \in \mathcal{X}$ and $r \in \mathbb{R}^+$. Sort the functions in $G_{X,r}$ by increasing $g(X)$ values, and partition $G_{X,r}$ into groups along this order:

$$G_{X,r} = G_1 \cup \cdots \cup G_m,$$

where all the functions in G_i, $i \in [m]$, have the same value on X. Note that $m \leq |G_{X,r}|$. Set

$$G_{\geq i} = G_i \cup \cdots \cup G_m.$$

As A is an ϵ-approximation of the sets $G_1 \cup \cdots \cup G_i$ for all $i \in [m]$, Claim 13.6 (2) implies the following.

CLAIM 15.7. For each $i \in [m]$, A is a 2ϵ-approximation of $G_{\geq i}$.

Now we sum up the functions in $G_{X,r}$ and $A_{X,r}$ by summing up over the differences between values of adjacent functions. For each $i \in [m]$, fix an arbitrary function $g_i \in G_i$ and let $g_0 = 0$. Then

$$\sum_{g \in G_{X,r}} g(X) = \sum_{i=1}^{m} \big(g_i(X) - g_{i-1}(X)\big) \cdot \big|G_{\geq i}\big|, \qquad \text{and}$$

$$\sum_{g \in A_{X,r}} g(X) = \sum_{i=1}^{m} \big(g_i(X) - g_{i-1}(X)\big) \cdot \big|A \cap G_{\geq i}\big|.$$

Thus the required expression

$$\left| \frac{\sum_{g \in G_{X,r}} g(X)}{n} - \frac{\sum_{g \in A_{X,r}} g(X)}{|A|} \right|$$

is upper bounded by

$$\begin{aligned}
&\left| \left(\sum_{i=1}^{m} \big(g_i(X) - g_{i-1}(X)\big) \cdot \frac{\big|G_{\geq i}\big|}{n} \right) - \left(\sum_{i=1}^{m} \big(g_i(X) - g_{i-1}(X)\big) \cdot \frac{\big|A \cap G_{\geq i}\big|}{|A|} \right) \right| \\
&\leq \sum_{i=1}^{m} \big(g_i(X) - g_{i-1}(X)\big) \cdot \left| \frac{|G_{\geq i}|}{n} - \frac{|A \cap G_{\geq i}|}{|A|} \right| \\
&\leq \sum_{i=1}^{m} \big(g_i(X) - g_{i-1}(X)\big) \cdot 2\epsilon \leq 2\epsilon r,
\end{aligned}$$

where the second-to-last step used Claim 15.7. □

Bibliography and discussion. The material in this section is from [**FL11**] (with some simplifications). The size of the ϵ-approximation in Theorem 15.5 depends on a parameter called the *pseudo-dimension*, which is a generalization of the notion of VC-dimension for general functions (see [**HP11**, Chapter 7] for details).

[FL11] D. Feldman and M. Langberg, *A unified framework for approximating and clustering data*, STOC'11—Proceedings of the 43rd ACM Symposium on Theory of Computing, ACM, New York, 2011, pp. 569–578, DOI 10.1145/1993636.1993712. MR2932007

[HP11] S. Har-Peled, *Geometric approximation algorithms*, Mathematical Surveys and Monographs, vol. 173, American Mathematical Society, Providence, RI, 2011, DOI 10.1090/surv/173. MR2760023

2. Application: Sensitivity and Coresets for Clustering

> *Elegant algorithms are easy to program correctly, as well as being efficient. A clever algorithm that is clean and elegant is much more likely to be used than a messy one. When people understand how an algorithm works, which is much more likely with an elegant algorithm, they are more likely to have confidence in the results it produces.*
>
> *Also, elegant solutions are much easier to generalize, to extend to other problems. My goal is to find general approaches and solutions, not ad hoc tricks.*
>
> Robert Tarjan

Given a set P of n points in $\mathbb{R}^d$, the k-median problem asks to compute a set X of k points that minimizes the cost function[2]

$$\mathrm{Cost}(P,k) = \min_{\substack{X\subseteq\mathbb{R}^d \\ |X|=k}} \mathrm{Cost}(P,X), \quad \text{where} \quad \mathrm{Cost}(P,X) = \sum_{p\in P} \mathrm{dist}(p,X).$$

One approach towards solving this problem is to first compute a smaller set A that 'approximates' P with respect to $\mathrm{Cost}(P,X)$. That is, for *every* set X of k points in $\mathbb{R}^d$ we would like $\mathrm{Cost}(P,X)$ to be approximately equal to $\mathrm{Cost}(A,X)$ (scaled up appropriately). Then the original problem on P is reduced to finding an X minimizing $\mathrm{Cost}(A,X)$—an easier problem if $|A|$ is much smaller than $|P|$.

This leads to the following definition.

Definition 15.8. Given a set P of n points in $\mathbb{R}^d$ and a parameter $\epsilon > 0$, a set $A \subseteq \mathbb{R}^d$ together with a weight function $w\colon A \to \mathbb{R}$ is an ϵ-coreset for the k-median problem on P if for every $X \subseteq \mathbb{R}^d$ of k points,

$$\sum_{p\in A} \mathrm{dist}\,(p,X)\cdot w\,(p) = (1\pm\epsilon)\cdot \sum_{p\in P} \mathrm{dist}\,(p,X)\,. \tag{15.9}$$

Our goal then is to construct an ϵ-coreset for the k-median problem. We will prove two main theorems, the first of which is the following.

Theorem 15.10. *Let P be a set of n points in $\mathbb{R}^d$ and $k \in \mathbb{Z}^+$, $\epsilon > 0$ be two given parameters. Define*

$$S = \sum_{p\in P} \sup_{\substack{Y\subseteq\mathbb{R}^d \\ |Y|=k}} \frac{\mathrm{dist}(p,Y)}{\sum_{q\in P}\mathrm{dist}(q,Y)}.$$

Then there exists an ϵ-coreset $A \subseteq P$ of size $O\left(\frac{S^2\, d\, k\log k}{\epsilon^2}\right)$ for the k-median problem on P.

The size of the ϵ-coreset in Theorem 15.10 relies on the seemingly mysterious quantity S; however its proof will demonstrate that S 'falls out' naturally when constructing coresets using ϵ-approximations. There do exist good upper bounds on S but using techniques and ideas outside the scope of this text (see discussion).

[2]Recall that $\mathrm{dist}\,(p,X) = \min_{q\in X}\mathrm{dist}\,(p,X)$.

Our second main result shows that the dependency on S can be removed if one is also given an approximate solution B.

THEOREM 15.11. *Let P be a set of n points in $\mathbb{R}^d$ and $k \in \mathbb{Z}^+, \epsilon > 0$ be two given parameters. Further let $B \subseteq \mathbb{R}^d$ be a set of points and $C \geq 1$ such that*

$$\mathrm{Cost}(P, B) \leq C \cdot \mathrm{Cost}(P, k).$$

Then there exists an ϵ-coreset for the k-median problem on P of size

$$O\left(\frac{C^2\, d\, k \log k}{\epsilon^2} + |B|\right).$$

Note that $|B|$ could be larger than k. Furthermore B is only a C-approximation to $\mathrm{Cost}(P, k)$, where C can be large, even a function of k and n. Theorem 15.11 shows that this approximate solution is already sufficient to get a small ϵ-coreset for P.

Overview of ideas. The proof of Theorems 15.10 and 15.11 rely on the following two insights. Let X be any set of k points in $\mathbb{R}^d$.

Relation to ϵ-approximations: Rewrite Equation (15.9) to get

$$\left|\sum_{p\in P} \mathrm{dist}\,(p, X) - \sum_{p\in A} \mathrm{dist}\,(p, X) \cdot w\,(p)\right| \leq \epsilon \cdot \sum_{p\in P} \mathrm{dist}\,(p, X)\,.$$

This resembles the notion of ϵ-approximations. Indeed, applying Corollary 15.6 to P with functions $f_p\,(X) = \mathrm{dist}\,(p, X)$ for each $p \in P$, an ϵ-approximation $A \subseteq P$ of the set system induced on P by the union of k balls in $\mathbb{R}^d$ satisfies

$$\left|\sum_{p\in P} \mathrm{dist}(p, X) - \sum_{p\in A} \mathrm{dist}(p, X)\frac{|P|}{|A|}\right| \leq 3\,\epsilon \cdot |P| \cdot \max_{p\in P} \mathrm{dist}(p, X).$$

The set A would be an $O\,(\epsilon)$-coreset, with weight function $w(p) = \frac{|P|}{|A|}$, if for each X,

$$|P| \cdot \max_{p\in P} \mathrm{dist}\,(p, X) = O\left(\sum_{q\in P} \mathrm{dist}\,(q, X)\right),$$

or equivalently, if for each p and each X,

$$\mathrm{dist}\,(p, X) = O\left(\frac{\sum_{q\in P} \mathrm{dist}\,(q, X)}{|P|}\right).$$

This is not the case, of course—each distance cannot be upper bounded by the average distance for *all* $X \subseteq \mathbb{R}^d$.

Weighted ϵ-approximations: The condition $\frac{\mathrm{dist}(p,X)}{\sum_{q\in P} \mathrm{dist}(q,X)} = O(\frac{1}{|P|})$ suggests that one should construct an ϵ-approximation according to a weight distribution, which can then be set depending on the relative values of $\mathrm{dist}\,(p, \cdot)$. This idea, sometimes called *importance sampling*, is thematically very similar to the idea in Theorem 8.20. Specifically, consider the following weighted version of Corollary 15.6.

LEMMA 15.12. *Let P be a set of n points in $\mathbb{R}^d$ and k a positive integer. For each $p \in P$ we are given a rational weight m_p and a function $f_p \colon \left(\mathbb{R}^d\right)^k \to \mathbb{R}^+$. These functions define a set system $(P, \mathcal{R})$ with*

$$\mathcal{R} = \left\{P_{X,r} \colon X \subseteq \mathbb{R}^d, |X| = k \text{ and } r \in \mathbb{R}\right\}, \text{ where } P_{X,r} = \left\{p \in P \colon f_p(X) \leq r\right\}.$$

Then given $\epsilon > 0$ there exists a multiset $A \subseteq P$ of size $O\left(\frac{\text{VC-dim}(\mathcal{R})}{\epsilon^2}\right)$ such that for any $X \in \left(\mathbb{R}^d\right)^k$,

$$\left| \frac{\sum_{p \in P} f_p(X)}{\sum_{p \in P} m_p} - \frac{\sum_{p \in A} \frac{f_p(X)}{m_p}}{|A|} \right| \leq 3\epsilon \left(\max_{p \in P} \frac{f_p(X)}{m_p} \right). \tag{15.13}$$

PROOF. By scaling up, we can assume that each m_p is an integer. Let P' be the set constructed by adding m_p copies of each $p \in P$ to P', where each copy of p is assigned the function $\frac{f_p(X)}{m_p}$. By applying Corollary 15.6 to P', there exists a set A such that for all $X \in \left(\mathbb{R}^d\right)^k$,

$$\left| \frac{\sum_{p' \in P'} f_{p'}(X)}{|P'|} - \frac{\sum_{p' \in A} f_{p'}(X)}{|A|} \right| \leq 3\epsilon \left(\max_{p' \in P'} f_{p'(X)} \right). \tag{15.14}$$

Noting that $\sum_{p' \in P'} f_{p'}(X) = \sum_{p \in P} f_p(X)$, Equation (15.14) is equivalent to Equation (15.13).
Finally, as the VC-dimension is unchanged by adding duplicate elements, Theorem 13.2 implies that $|A| = O\left(\frac{\text{VC-dim}(\mathcal{R})}{\epsilon^2}\right)$. □

In the proof of Theorems 15.10 and 15.11 we will set the parameters m_p, $f_p(\cdot)$ and ϵ' such that an ϵ'-approximation A given by Lemma 15.12 can be used to construct the required ϵ-coreset.

Given our preparation, the proof of our first main result is immediate.

THEOREM 15.10. *Let P be a set of n points in $\mathbb{R}^d$ and $k \in \mathbb{Z}^+$, $\epsilon > 0$ be two given parameters. Define*

$$S = \sum_{p \in P} \sup_{\substack{Y \subseteq \mathbb{R}^d \\ |Y| = k}} \frac{\text{dist}(p, Y)}{\sum_{q \in P} \text{dist}(q, Y)}.$$

Then there exists an ϵ-coreset $A \subseteq P$ of size $O\left(\frac{S^2 d k \log k}{\epsilon^2}\right)$ for the k-median problem on P.

PROOF. Set $f_p(X) = \text{dist}(p, X)$ for each $p \in P$ and let m_p be the weight of $p \in P$. These weights will be set later and normalized so that $\sum_{p \in P} m_p = 1$. Further let $\epsilon' > 0$ be a parameter to be set later.
Let A be an ϵ'-approximation given by Lemma 15.12 applied to P with weights m_p and functions $f_p(\cdot)$. That is, for each set X of k points in $\mathbb{R}^d$, A satisfies

$$\left| \sum_{p \in P} \text{dist}(p, X) - \sum_{p \in A} \text{dist}(p, X) \cdot \frac{1}{|A| \, m_p} \right| \leq 3\epsilon' \left(\max_{p \in P} \frac{\text{dist}(p, X)}{m_p} \right). \tag{15.15}$$

For A to be an ϵ-coreset, the right-hand side of the above inequality must be upper bounded by $\epsilon \cdot \sum_{p\in P} \operatorname{dist}(p, X)$. The natural choice is to set $m_p = \frac{\operatorname{dist}(p,X)}{\sum_{q\in P} \operatorname{dist}(q,X)}$ for each $p \in P$ and $\epsilon' = \frac{\epsilon}{3}$. However m_p must be independent of X, as A must work for all choices of X! Considering the worst-case bound for m_p over all choices of X leads to the notion of the *sensitivity* of a point.

DEFINITION 15.16. The sensitivity of each $p \in P$ with respect to $\{f_p : p \in P\}$ is defined to be

$$s(p) = \sup_{\substack{Y\subseteq\mathbb{R}^d \\ |Y|=k}} \frac{f_p(Y)}{\sum_{q\in P} f_q(Y)}.$$

Let $S = \sum_{p\in P} s(p)$ and set

$$f_p(X) = \operatorname{dist}(p, X) \qquad \text{and} \qquad m_p = \frac{s(p)}{S}^3.$$

From Equation (15.15),

$$\left| \sum_{p\in P} \operatorname{dist}(p, X) - \sum_{p\in A} \operatorname{dist}(p, X) \cdot \frac{1}{|A|\, m_p} \right| \leq 3\epsilon' \cdot S \cdot \left(\max_{p\in P} \frac{\operatorname{dist}(p, X)}{s(p)} \right)$$

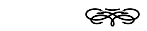
The R.H.S., after substituting for $s(p)$ and multiplying/dividing by $\sum_{q\in P} \operatorname{dist}(q, X)$:

$$3\epsilon' S \sum_{q\in P} \operatorname{dist}(q, X) \cdot \left(\max_{p\in P} \frac{\frac{\operatorname{dist}(p,X)}{\sum_{q\in P} \operatorname{dist}(q,X)}}{\sup_{\substack{Y\subseteq\mathbb{R}^d \\ |Y|=k}} \frac{\operatorname{dist}(p,Y)}{\sum_{q\in P} \operatorname{dist}(q,Y)}} \right) \leq 3\epsilon' S \left(\sum_{q\in P} \operatorname{dist}(q, X) \right).$$

Setting $\epsilon' = \frac{\epsilon}{3S}$ implies that A is an ϵ-coreset where each $p \in A$ is assigned the weight $\frac{1}{|A|\, m_p}$.

Note that A is an ϵ-approximation of the set system induced on P by the union of k balls in $\mathbb{R}^d$; that is, the set system induced by the k-fold union of balls in $\mathbb{R}^d$. Lemma 10.3 and Theorem 11.6 implies that the VC-dimension of this set system is $\Theta(dk \log k)$ and thus $|A| = O\left(\frac{S^2 dk \log k}{\epsilon^2}\right)$ by Lemma 15.12. □

We remark here that to compute A above, we need to compute the weights m_p for each $p \in P$. This, together with the problem of deriving good upper bounds on the total sensitivity S, is a non-trivial algorithmic problem by itself (see discussion).



[3]The division by S is just to get $\sum_p m_p = 1$.

THEOREM 15.11. *Let P be a set of n points in $\mathbb{R}^d$ and $k \in \mathbb{Z}^+, \epsilon > 0$ be two given parameters. Further let $B \subseteq \mathbb{R}^d$ be a set of points and $C \geq 1$ such that*

$$\operatorname{Cost}(P, B) \leq C \cdot \operatorname{Cost}(P, k).$$

Then there exists an ϵ-coreset for the k-median problem on P of size

$$O\left(\frac{C^2\, d\, k \log k}{\epsilon^2} + |B|\right).$$

PROOF. Given B, set the parameters for each $p \in P$ as follows:

$$f_p(X) = \operatorname{dist}(p, X) - \operatorname{dist}(\operatorname{closest}(p, B), X) + \operatorname{dist}(p, B),$$
$$m_p = \frac{\operatorname{dist}(p, B)}{\sum_{q \in P} \operatorname{dist}(q, B)},$$

where $\operatorname{closest}(p, Q) = \arg\min_{q \in Q} \operatorname{dist}(p, q)$ denotes the closest point in Q to p. Note that $f_p(X)$ is non-negative[4] due to triangle inequality:

$$\operatorname{dist}(\operatorname{closest}(p, B), X) \leq \operatorname{dist}(\operatorname{closest}(p, B), p) + \operatorname{dist}(p, X).$$

Applying Lemma 15.12 with f_p and m_p set above and noting that $\sum_{p \in P} m_p = 1$, we get an ϵ'-approximation A such that for any $X \subseteq \mathbb{R}^d$ of k points,

$$\left| \sum_{p \in P} \Big(\operatorname{dist}(p, X) - \operatorname{dist}(\operatorname{closest}(p, B), X) + \operatorname{dist}(p, B) \Big) - \sum_{p \in A} \Big(\operatorname{dist}(p, X) - \operatorname{dist}(\operatorname{closest}(p, B), X) + \operatorname{dist}(p, B) \Big) \cdot \frac{1}{|A|\, m_p} \right| \leq 3\epsilon' \left(\max_{p \in P} \frac{\operatorname{dist}(p, X) - \operatorname{dist}(\operatorname{closest}(p, B), X) + \operatorname{dist}(p, B)}{m_p} \right).$$

Using the fact that

$$\sum_{p \in P} \operatorname{dist}(p, B) = \sum_{p \in A} \left(\operatorname{dist}(p, B) \frac{\sum_{q \in P} \operatorname{dist}(q, B)}{|A|\, \operatorname{dist}(p, B)} \right) = \sum_{p \in A} \operatorname{dist}(p, B) \frac{1}{|A|\, m_p},$$

we arrive at

$$\left| \left(\sum_{p \in P} \operatorname{dist}(p, X) \right) - \left(\sum_{p \in P} \operatorname{dist}(\operatorname{closest}(p, B), X) \right) - \left(\sum_{p \in A} \operatorname{dist}(p, X) \frac{1}{|A|\, m_p} \right) + \left(\sum_{p \in A} \operatorname{dist}(\operatorname{closest}(p, B), X) \frac{1}{|A|\, m_p} \right) \right|$$
$$\leq 3\epsilon' \left(\max_{p \in P} \frac{\operatorname{dist}(p, X) - \operatorname{dist}(\operatorname{closest}(p, B), X) + \operatorname{dist}(p, B)}{m_p} \right). \tag{15.17}$$

R.H.S.: Using a consequence of triangle inequality, that

$$\operatorname{dist}(p, X) - \operatorname{dist}(\operatorname{closest}(p, B), X) \leq \operatorname{dist}(p, \operatorname{closest}(p, B)) = \operatorname{dist}(p, B),$$

[4]Indeed, the additive term $\operatorname{dist}(p, B)$ is present in $f_p(X)$ just to make $f_p(X)$ non-negative so that one can apply Lemma 15.12. *Conceptually* we only need $f_p(X) = \operatorname{dist}(p, X) - \operatorname{dist}(\operatorname{closest}(p, B), X)$.

as well as substituting the value of m_p, the R.H.S. of Equation (15.17) is at most

$$3\epsilon' \max_{p\in P} \frac{2\,\mathrm{dist}(p,B)}{\frac{\mathrm{dist}(p,B)}{\sum_{q\in P}\mathrm{dist}(q,B)}} = 6\epsilon' \sum_{p\in P} \mathrm{dist}(p,B)$$
$$\leq 6\epsilon' \cdot C \cdot \mathrm{Cost}\,(P,k) \leq 6\epsilon' C \sum_{p\in P} \mathrm{dist}(p,X).$$

L.H.S.: For each $b \in B$, let P_b be the set of points of P whose closest point in B is b. Then the L.H.S. of Equation (15.17) becomes

$$\left| \left(\sum_{p\in P} \mathrm{dist}(p,X)\right) - \left(\sum_{b\in B} |P_b| \cdot \mathrm{dist}(b,X)\right) - \left(\sum_{p\in A} \mathrm{dist}(p,X)\frac{1}{|A|\,m_p}\right) + \left(\sum_{b\in B}\sum_{p\in P_b\cap A} \mathrm{dist}(b,X)\frac{1}{|A|\,m_p}\right)\right|. \tag{15.18}$$

We're done—set $\epsilon' = \frac{\epsilon}{6C}$ and return $A \cup B$ as our ϵ-coreset, with weights dictated by Equation (15.18):

$$p \in A \;:\; w(p) = \frac{1}{|A|\,m_p}.$$
$$b \in B \;:\; w(b) = |P_b| - \sum_{p\in A\cap P_b} \frac{1}{|A|\,m_p}.$$

□

We remark that to compute the set A one again needs to compute the weights m_p for all $p \in P$. However this time it is easier as we are also given the set B.

Bibliography and discussion. The beautiful application of this section is from [**FL11**]. See [**VX12**] for upper bounds on sensitivity for a variety of optimization problems. There are many interesting variations, improvements and applications of the basic ideas presented in this section (e.g., see [**HV20**]). We refer the reader to the surveys [**AHPV07**, **Phi18**] for more information on coresets.

[AHPV07] P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan, *Geometric approximation via coresets*, Combinatorial and computational geometry, Math. Sci. Res. Inst. Publ., vol. 52, Cambridge Univ. Press, Cambridge, 2005, pp. 1–30, DOI 10.4171/PRIMS/172. MR2178310

[FL11] D. Feldman and M. Langberg, *A unified framework for approximating and clustering data*, STOC'11—Proceedings of the 43rd ACM Symposium on Theory of Computing, ACM, New York, 2011, pp. 569–578, DOI 10.1145/1993636.1993712. MR2932007

[HV20] L. Huang and N. K. Vishnoi, *Coresets for clustering in euclidean spaces: importance sampling is nearly optimal*, STOC '20—Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, ACM, New York, 2020, pp. 1416–1429. MR4141850

[Phi18] J. M. Phillips. *Coresets and sketches. Handbook of Discrete and Computational Geometry.* CRC Press, 2018. pp. 1269–1286.

[VX12] K. R. Varadarajan and X. Xiao. *On the sensitivity of shape fitting problems. Proceedings of the IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS).* Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2012. pp. 486–497.