Clustering over Multi-Objective Samples: The one2all Sample

Edith Cohen
Google Research, USA
Tel Aviv University, Israel

Shiri Chechik
Tel Aviv University, Israel

Haim Kaplan
Tel Aviv University, Israel

June 13, 2017

Abstract

Clustering is a fundamental technique in data analysis. Consider data points $X$ that lie in a (relaxed) metric space (where the triangle inequality can be relaxed by a constant factor). Each set of points $Q$ (centers) defines a clustering of $X$ according to the closest center with cost $V(Q) = \sum_{x \in X} d_{xQ}$. This formulation generalizes classic $k$-means clustering, which uses squared distances. Two basic tasks, parametrized by $k \geq 1$, are cost estimation, which returns (approximate) $V(Q)$ for queries $Q$ such that $|Q| = k$, and clustering, which returns an (approximate) minimizer of $V(Q)$ of size $|Q| = k$. With very large data sets $X$, we seek efficient constructions of small summaries that allow us to efficiently approximate clustering costs over the full data.

We present a novel data reduction tool based on multi-objective probability-proportional-to-size (pps) sampling: Our one2all construction inputs any set of centers $M$ and efficiently computes a sample of size $O(|M|)$ from which we can tightly estimate the clustering cost $V(Q)$ for any $Q$ that has at least a fraction of the clustering cost of $M$. For cost queries, we apply one2all to a bicriteria approximation to obtain a sample of size $O(k\epsilon^{-2})$ for all $|Q| = k$. For clustering, we propose a wrapper that applies a black-box algorithm to a sample and tests clustering quality over $X$, adaptively increasing the sample size. Our approach exploits the structure of the data to provide quality guarantees through small samples, without the use of typically much larger worst-case-size summaries.

1 Introduction

Clustering is a fundamental and prevalent tool in data analysis. We have a set $X$ of data points that lie in a (relaxed) metric space $M$, where distances satisfy a relaxed triangle inequality: For some constant $\rho \geq 1$, for any three points $x, y, z$, $d_{xy} \leq \rho(d_{xz} + d_{zy})$. Note that any metric space with distances replaced by their $p$th power satisfies this relaxation: For $p \leq 1$ it remains a metric and otherwise we have $\rho = 2^{p-1}$. In particular, for squared distances, commonly used for clustering, we have $\rho = 2$.

Each set $Q \subset M$ of points (centers) defines a clustering, which is a partition of $X$ into $|Q|$ clusters, which we denote by $X_q$ for $q \in Q$, so that a point $x \in X$ is in $X_q$ if and only if it is in the Voronoi region of $q$, that is $q = \arg\min_{y \in Q} d_{xy}$. We allow points $x \in X$ to have optional weights $w_x > 0$, and define the cost of clustering $X$ by $Q$ to be

$$V(Q \mid X, w) = \sum_{x \in X} w_x d_{xQ},$$

where $d_{xQ} = \min_{y \in Q} d_{xy}$ is the distance from point $x$ to the set $Q$.

Two fundamental computational tasks are cost queries and clustering (cost minimization). The clustering cost $[\Pi]$ of query $Q$ can be computed using $n|Q|$ pairwise distance computations, where $n = |X|$ is the number of points in $X$. With multiple queries, it is useful to pre-process $X$ and return fast approximate answers. Clustering amounts to finding $Q$ of size $|Q| \leq k$ with minimum cost:

$$\arg\min_{Q, |Q| \leq k} V(Q \mid X, w).$$

1

arXiv:1706.03607v1 [cs.LG] 12 Jun 2017
Clustering is computationally hard \[2\] even on Euclidean spaces and even to tightly approximate \[4\]. There is a local search polynomial algorithm with \(9 + \epsilon\) approximation ratio \[20\]. In practice, clustering is solved using heuristics, most notably Lloyd’s algorithm \[22\], and scalable approximation algorithms such as kmeans++ \[3\], which obtains a \(\log k\) factor approximation using \(k\) centers. A recent result \[27\] established that kmeans++ provides constant factors bi-criteria guarantees: When applied to obtain \(\beta k\) centers for some constant \(\beta > 1\), the clustering cost is within a constant factor of the optimum \(k\)-means cost.

When the set of points \(X\) is very large, we seek an efficient procedure that computes a small summary structure from which we can approximate clustering costs. These structures are commonly in the form of subsets \(S \subset X\) with weights \(w'\) so that \(V(Q | S, w')\) approximates \(V(Q | X, w)\) for each \(Q\) of size \(k\). The term \textit{coresets} for such structures was introduced in the computational geometry literature \[1\] \[18\], building on the theory of \(\epsilon\)-nets. Some notable constructions include \[23\] \[9\] \[15\] \[16\]. Initial coresets constructions had bounds with high (exponential or high polynomial) dependence on some parameters (dimension, \(\epsilon\), \(k\)) and poly logarithmic dependence on \(n\), with the best current asymptotic bound of \(O(k\epsilon^{-2}\log k\log n)\) claimed in \[5\].

The bulk of these randomized summary structures are aimed to provide strong “\texttt{ForAll}” statistical guarantees, which bound the distribution of the maximum approximation error over all \(Q\) of size \(k\). The constructions are worst-case and based on general (VC) dimension bounds. Moreover, they do not exploit typical structure present in the data and must use size that meets the worst-case bounds in order to obtain guarantees on quality of the results. The \texttt{ForAll} requirement, however, is a costly overkill for the two tasks we have at hand: For clustering cost queries, weaker per-query “\texttt{ForEach}\" typically suffice, which for each \(Q\), with very high probability over the structure distribution, bound the error of the estimate of \(V(Q)\). For clustering, it suffices to guarantee that the (approximate) minimizers of \(V(Q | S, w')\) are approximate minimizers of \(V(Q | X, w)\). In particular, sets \(Q\) with high clustering costs require only very coarse approximation. Moreover, our data typically has structure that includes a natural clustering (which is what we seek) and lower dimensionality than the ambient space. While a notion of “weak coresets” that aim to only support optimization was considered in the coreset literature \[15\], the constructions are also worst-case.

### Contribution Overview

We propose summary structures for clustering costs based on multi-objective probability-proportional-to-size (pps) samples \[13\] \[11\], which build on the classic notion of sample coordination \[21\] \[6\] \[24\] \[10\].

Consider a particular set \(Q\) of centers. The theory of weighted sampling \[25\] \[26\] tells us that to estimate the sum \(V(Q | X, w)\) it suffices to sample \(\epsilon^{-2}\) points with probabilities \(p_x \propto w_x d_{Qx}\) proportional to their contribution to the sum \[17\]. The inverse-probability \[19\] estimate obtained from the sample \(S\),

\[
\hat{V}(Q | X, w) \equiv V(Q | S, \{w_x/p_x\})
\]

is an unbiased estimate of \(V(Q | X, w)\) with well-concentrated normalized squared error of \(\epsilon\). The challenge here for us is that we are interested in quality guarantees for \textit{all} subsets \(Q\) of size \(k\) whereas the estimate \(V(Q' | S, \{w_x/p_x\})\) when \(S\) is taken for a particular \(Q \neq Q'\) will not provide quality guarantees for the set \(Q'\). To obtain these quality guarantees for all \(Q\) by a single sample, we apply the machinery of multi-objective sampling: We use multi-objective pps sampling probabilities, where each point \(x \in X\) is sampled with the maximum pps probability over all \(Q\) of size \(k\).

Priori, however, it seems that the size of such a multi-objective sample can be very large. Surprisingly, we show that on any (relaxed) metric space, the multi-objective pps sample size is \(O(k\epsilon^{-2})\). Note that the size does not depend on the dimensionality of the space or on the size of the data. Our result generalizes previous work \[5\] that only applied to the case where \(k = 1\), where clustering cost reduces to inverse classic closeness centrality (sum of distances from a single point \(Q\)).

For our applications, we also need to efficiently perform the sampling. Clearly, a straightforward computation of these multi-objective pps probabilities is not feasible.

Our main technical contribution, which is the basis of both the existential and algorithmic results, is an efficient general construction which we refer to as \textit{one2all}: For any given set \(M\) of centers, and any \(\alpha \leq 1\),
we compute using \(|M|n^{d} fast computation, multi-objective sampling probabilities for all subsets \(Q\) with clustering cost \(V(Q) \geq \alpha V(M)\). The size of the sample is \(O(\frac{\alpha^{-1}|M|}{\epsilon^{-2}})\).

By applying our one2all construction to an optimal clustering \(M\) of size \(k\) and \(\alpha = 1\), we establish existentially that a multi-objective pps sample for all sets \(Q\) of size \(k\) has size \(O(k\epsilon^{-2})\). To obtain an efficient construction, we can apply kmeans++ [3] or another efficient bi-criteria approximation algorithm to compute \(M\) of size \(2k\) that has cost that is at most the optimum divided by some constant \(\alpha\) [27].

We then apply our one2all construction to \(M\) with \(\alpha\) to obtain a multi-objective sample for all \(k\)-subsets of size \(O(\frac{\alpha^{-1}|M|}{\epsilon^{-2}}) = O(k\epsilon^{-2})\).

For the task of approximate cost queries, we pre-process the data as above to obtain a multi-objective sample of size \(O(k\epsilon^{-2})\). We then process cost queries \(Q\) by computing and returning the clustering cost of \(S\) by \(Q: V(Q | S, \{w_{x}/p_{x}\})\). Our sample provide statistical guarantees that apply to each \(Q\) of size \(k\), or more generally, for each \(Q\) with \(V(Q | X, w) \geq \alpha V(M | X, w)\) over the distribution of the sample \(S\).

For approximate clustering tasks, we propose a wrapper algorithm which uses as a black-box (approximate, bicriteria) clustering algorithm \(A\). The algorithm \(A\) is applied to the smaller sample to obtain a respective approximate minimizer of the clustering cost over the sample. The \textbf{ForEach} guarantee, however, in-and-off itself, does not guarantee us that the solution over the sample has the respective quality over the full data set. A larger sample size may be required and can be achieved by controlling the size parameter of the multi-objective sample. We use an optimization framework over multi-objective samples [15]. This framework relies on a critical property of \textbf{ForEach} samples: While not guaranteeing optimization, they do facilitate testing of the quality of the sample approximate optimizer \(Q\) returned by \(A\): If the clustering cost of \(V(Q | X, w)\) agrees with the estimate \(V(Q | S, w')\) then we can certify that \(Q\) is an approximate optimizer (in the sense of \(A\)) over the full data \(X\). Otherwise, \(V(Q | S, w')\) serves as an approximate lower bound for what we could do over \(X\). Conveniently, this test can be performed with high confidence using another independent validation \textbf{ForEach} multi-objective pps sample. Our wrapper algorithm iteratively doubles the multi-objective pps sample size \(S\) and applies \(A\) to cluster \(S\) until the sample (approximate) minimizer satisfies the test. This approach allows us to work with the smallest sample that suffices to meet the specified quality.

The paper is organized as follows. The relevant components of the machinery of multi-objective pps sampling and its application to clustering are reviewed in Section 2. Section 3 presents a statement of our one2all theorem and some implications. Section 4 provides a full proof of our one2all Theorem.

2 Multi-objective pps samples for clustering

We review the framework of weighted and multi-objective weighted sampling in our context of clustering costs. Consider approximating the clustering cost \(V(Q | X, w)\) from a sample \(S\) of \(X\). For probabilities \(p_{x} > 0\) for \(x \in X\) and a sample \(S\) drawn according to these probabilities, we have the unbiased inverse probability estimator [19] of \(V(Q | X, w)\):

\[
\hat{V}(Q | X, w) = \sum_{x \in S} \frac{w_{x}d_{xQ}}{p_{x}} = V(Q | S, \{w_{x}/p_{x}\}) .
\]

Note that the estimate is equal to the clustering cost of \(S\) with weights \(w_{x}/p_{x}\) by \(Q\).

2.1 Probability proportional to size (pps) sampling

[17] To obtain guarantees on the estimate quality of the clustering cost by \(Q\), we need to use weighted sampling. The pps base probabilities for \(x \in X\) are

\[
\psi_{x}(Q | X, w) = \frac{w_{x}d_{xQ}}{\sum_{y \in X} w_{y}d_{yQ}} .
\]

The pps probabilities for a sample with size parameter \(r > 1\) are

\[
\psi_{x}^{(r)}(Q | X, w) = \min\{1, r\psi_{x}(Q | X, w)\} .
\]
With pps sampling we obtain the following guarantees:

**Theorem 2.1** Consider a sample $S$ where each $x \in X$ is included (independently or using VarOpt dependent sampling \[13, 12\]) with probability $p_x \geq \psi_x^{(r)}(Q \mid X, w)$. Then the estimate \[3\] has the following statistical guarantees:

- The coefficient of variation (CV), defined as the ratio of the standard deviation to the mean, (measure of the “relative error”) is at most $\epsilon = 1/\sqrt{r}$.
- The estimate is well concentrated in the Chernoff-Bernstein sense: The probability of relative error larger than $c\epsilon$ decreases exponentially with $c$.

Note that the (expected) sample size is $\sum_x p_x$. When $p_x = \psi_x^{(r)}(Q \mid X, w)$, the size is at most $r$.

### 2.2 Multi-objective pps sampling

When we seek estimates with statistical guarantees for a set $Q$ of queries (for example, all sets of $k$ points in the metric space $M$), we use multi-objective samples \[13, 11\]. The multi-objective (MO) pps base sampling probabilities are defined as the maximum of the pps base probabilities over $Q \in Q$:

$$
\psi_x(Q \mid X, w) = \max_{Q \in Q} \psi_x(Q \mid X, w).
$$

For a size parameter $r$, the multi-objective pps probabilities are

$$
\psi_x^{(r)}(Q \mid X, w) = \min\{1, r\psi_x(Q \mid X, w)\} = \max_{Q \in Q} \psi_x^{(r)}(Q \mid X, w).
$$

As a corollary of Theorem 2.1, noting that $\forall Q \in Q, \psi_x(Q \mid X, w) \geq \psi_x(Q \mid X, w)$, we obtain the following.

**Corollary 2.1** Consider a sample of $X$ where each $x \in X$ is included (independently or using VarOpt dependent sampling) with probability $p_x \geq \min\{1, r\psi_x(Q \mid X)\}$. Then for each $Q \in Q$ the inverse probability estimator \[3\] has the statistical guarantees stated in Theorem 2.1.

With some notation abuse, we define the overhead of multi-objective sampling $Q$ to be:

$$
h(Q \mid X, w) \equiv h(\psi(Q \mid X, w)) = \sum_{x \in X} \psi_x(Q \mid X, w).
$$

Note that with size parameter $r$, the multi-objective pps sample size is at most $rh(Q \mid X, w)$, so $h(Q \mid X, w)$ bounds the factor-increase in sample size due to the “multi-objectiveness” of the sample.

Sometimes we can not compute $\psi$ exactly but can instead efficiently obtain upper bounds $\pi_x \geq \psi_x(Q \mid X, w)$. Accordingly, we use sampling probabilities $\pi_x^{(r)} = \min\{1, r\pi_x\}$. The use of upper bounds increases the sample size, and we refer to $h(\pi) = \sum_x \pi_x$ as the overhead of $\pi$. We therefore seek upper-bounds $\pi$ with a overhead not much larger than $h(Q \mid X, w)$. Note that when sampling with the upper-bounds $\pi$, we retain the statistical guarantees on estimation quality as stated in Theorem 5.

### 2.3 Optimization over multi-objective pps samples

We now consider clustering, with an approximate or bicriteria-approximate objective. We would like to perform this optimization by running a clustering algorithm $A$ on the sample $S$ with weights $w_x/p_x$. 4
A multi-objective pps sample for \( Q \) with size parameter \( r = \epsilon^{-2} \) provides ForEach guarantees. This guarantee is not sufficient to ensure that \( Q \) returned by \( A \) satisfies the quality requirements of \( A \), even approximately, over \( X \), but it does allow us to test that using:

\[
V(Q \mid X, w) \leq (1 + \epsilon)V(Q \mid S, \{w_x/p_x\}) .
\]  

(6)

To perform the test, we can compute the exact cost \( V(Q \mid X, w) \). Alternatively, we can simply draw a different, independent, ForEach sample \( S' \), using the same distribution, and compute the estimate \( V(Q \mid S', \{w_x/p_x\}) \).

If the test is satisfied, we know (within the ForEach guarantees and the approximation quality guarantee provided by \( A \)) that \( Q \) is an approximate solution over \( X \). If the condition (6) does not hold, it still provides us with an approximate lower bound of the solution over \( X \). We can increase the sample size and apply \( A \) to the larger sample. A pseudocode for this wrapper is provided as Algorithm 1.

Algorithm 1 Wrapper for clustering over samples

\textbf{Input:} points \( X \), weights \( w > 0 \), (upper bounds on) base pps probabilities \( \pi \) for \( Q \), \( \epsilon > 0 \), an approximate clustering algorithm \( A \) that inputs a weighted set of points and returns \( Q \in Q \) that satisfies some (bi-criteria) quality guarantees.

\begin{algorithm}
\begin{algorithmic}
\STATE // Initialization
\FOREACH \( x \in X \) \DO // for sampling
\STATE \( u_x \sim U[0,1] \)
\STATE \( r \leftarrow \epsilon^{-2} \) \hfill \( \text{// Start with ForEach guarantee} \)
\STATE // Main Loop
\REPEAT
\STATE \( S \leftarrow \perp \) \hfill \( \text{// Initialize sample} \)
\FOREACH \( x \in X \) \such \that \( u_x \leq r\pi_x \) \DO \hfill \( \text{// sample points} \)
\STATE \( S \leftarrow S \cup \{x\} \)
\STATE // Cluster the sample \( S \)
\FOREACH \( x \in S \) \DO \hfill \( \text{// weights for sampled points} \)
\STATE \( w'_x \leftarrow w_x/\min\{1, r\pi_x\} \)
\STATE \( Q \leftarrow A(S, w) \) \hfill \( \text{// Apply approximate clustering algorithm} \ A \text{ to sample} \)
\STATE \( r \leftarrow 2r \) \hfill \( \text{// Double the sample size parameter} \)
\UNTIL \( V(Q \mid X, w) \leq (1 + \epsilon)V(Q \mid S, w') \) \hfill \( \text{// Exact or approx using a validation sample} \)
\RETURN \( Q \)
\end{algorithmic}
\end{algorithm}

3 The one2all Theorem statement and implications

Consider a relaxed metric space \( M \) where distances satisfy all properties of a metric space except that the triangle inequality is relaxed using a parameter \( \rho \geq 1 \):

\[
\forall x, y, z \in M, \ d_{xy} \leq \rho(d_{xz} + d_{zy}) .
\]  

(7)

For \( v > 0 \), we use the notation

\[ Q^{(v)} = \{Q \mid V(Q \mid X, w) \geq v\} \]

for the set of all subsets \( Q \subset M \) that cluster \( X \) with cost at least \( v \).

For a set \( M \), we denote by

\[ X_q = \{x \in X \mid d_{xq} = d_{xM}\} \]

the points in \( X \) that are closest to \( q \in M \). In case of ties we apply arbitrary tie breaking to ensure that \( X_q \) \( q \in M \) is a partition of \( X \). We will assume that \( X_q \) is not empty for all \( q \in M \), since otherwise, we can remove the point \( q \) from \( M \) without affecting the clustering cost of \( X \) by \( M \).
For the set $M$, we define the probabilities $\pi^{(M)} \equiv \pi^{(M|X,w)}$ as follows:

$$\forall m \in M, \forall x \in X_m, \pi_x^{(M|X,w)} = \min \left\{ 1, \max \left\{ 2\rho \frac{d_{xM}}{V(M|X,w)}, \frac{8\rho^2}{w(X_m)} \right\} \right\}.$$ (8)

We are now ready to state our one2all Theorem: We show that $\frac{1}{\alpha} \pi^{(M|X,w)}$ upper bound the multi-objective pps base probabilities for $Q^{(v)}$ for $v = \alpha V(M|X,w)$. The full proof of the Theorem is provided in the next section.

**Theorem 3.1** Let $M$ be a relaxed metric space with parameter $\rho$, $X \subset M$ be finite set of points with weights $w$, and $M \subset M$ be a finite set of points. Let $v = V(M|X,w)$ Then for all $\alpha \leq 1$,

$$\min\{ 1, \frac{1}{\alpha} \pi^{(M|X,w)} \} \geq \psi(Q^{(\alpha v)}|X,w).$$

**Corollary 3.1** The multi-objective overhead of the set of all sets of centers $Q$ with clustering cost at least $v = \alpha V(M|X,w)$ is at most $h(Q^{(v)}) \leq \frac{1}{\alpha}(8\rho^2|M| + 2\rho)$.

**Proof**  Note that the sum $\sum_x \pi_x^{(M|X,w)} \leq 8\rho^2|M| + 2\rho$.

**Corollary 3.2** For $k \geq 1$, let $Q$ be the set of all $k$-subsets of points in a relaxed metric space $M$ with parameter $\rho$. The multi-objective pps overhead of $Q$ satisfies

$$h(Q) \leq 8\rho^2k + 2\rho.$$  

**Proof**  We apply Theorem 3.1 with $M$ being the $k$-means optimum and $\alpha = 1$.

### 4 Proof of the one2all Theorem

To prove Theorem 3.1, we need to show that $\forall Q$ such that $V(Q|X,w) \geq \alpha V(M|X,w)$ and $\forall x \in X$,

$$\frac{d_{xQ}}{V(Q|X,w)} \leq \frac{1}{\alpha} \pi_x^{(M|X,w)}.$$ (9)

We will do a case analysis, as illustrated in Figure 1. We first consider $Q$ and points $x$ such that the distance of $x$ to $Q$ is not much larger than the distance of $x$ to $M$. Property 9 follows using the first term of the maximum in (8) by applying the following Lemma with $c = 2\rho$.

**Lemma 4.1** Let $Q$ and $x$ be such that $d_{xQ} \leq c d_{xM}$. Then

$$\frac{d_{xQ}}{V(Q|X,w)} \leq \frac{c}{\alpha} \frac{d_{xM}}{V(M|X,w)}.$$  

**Proof**  Since $Q \in Q^{(V(M))}$, $V(Q|X,w) \geq \alpha V(M|X,w)$. Therefore

$$\frac{d_{xQ}}{V(Q|X,w)} \leq \frac{1}{\alpha} \frac{d_{xQ}}{V(M|X,w)} \leq \frac{c}{\alpha} \frac{d_{xM}}{V(M|X,w)}.$$  


Figure 1: Illustration of the one2all construction proof with $\rho = 1$. The data points $X$ are in black. The points in $M$ are colored red. We show the respective Voronoi partition and for each cluster, we show circles centered at the respective $m \in M$ (red) point with radius $\Delta_m$. The points in blue are a set $Q$. The points $x \in X$ are labeled $A$ if $d_{xQ} < 2d_{xM}$ (and we apply Lemma 4.1). Otherwise, when there is a point $m$ such that $d_{xQ} > d_{xm}$, the point is labeled $B$ when $d_{mQ} \geq 2\Delta_m$ (Lemma 4.2) and is labeled $C$ otherwise (Lemma 4.3).

It remains to consider the complementary case where point $x$ is much closer to $M$ than to $Q$:

$$d_{xQ} \geq 2\rho d_{xM}.$$  \hfill (10)

We first introduce a useful definition: For a point $q \in M$, we denote by $\Delta_q$ the weighted median of the distances $d_{qy}$ for $y \in X_q$, weighted by $w_y$. The median $\Delta_q$ is a value that satisfies the following two conditions:

$$\sum_{x \in X_q \mid d_{xq} \leq \Delta_q} w_x \geq \frac{1}{2} w(X_q) \quad \text{and} \quad \sum_{x \in X_q \mid d_{xq} \geq \Delta_q} w_x \geq \frac{1}{2} w(X_m). \quad \text{(11, 12)}$$

It follows from (12) that for all $q \in M$,

$$V(M \mid X_q, w) = \sum_{x \in X_q} w_x d_{qx} \geq \sum_{x \in X_q \mid d_{xq} \geq \Delta_q} w_x d_{xq} \geq \Delta_q \sum_{x \in X_q \mid d_{xq} \geq \Delta_q} w_x \geq \frac{1}{2} w(X_m) \Delta_q.$$  

Therefore,

$$V(M \mid X, w) = \sum_{q \in M} V(M \mid X_q, w) \geq \frac{1}{2} \sum_{q \in M} w(X_m) \Delta_q.$$  \hfill (13)

We now return to our proof for $x$ that satisfies (10). We will show that property (9) holds using the second term in the max operation in the definition (8). Specifically, let $m$ be the closest $M$ point to $x$. We will show that

$$\frac{d_{xQ}}{V(Q \mid X, w)} \leq \frac{8\rho^2}{\alpha \cdot w(X_m)}.$$  \hfill (14)

We divide the proof to two subcases, in the two following Lemmas, each covering the complement of the other: When $d_{mQ} \geq 2\rho \Delta_m$ and when $d_{mQ} \leq 2\rho \Delta_m.$
Lemma 4.2 Let $Q$ and point $x$ be such that

$$\exists m \in M, \ d_{mx} < \frac{1}{2\rho}d_{xQ} \text{ and } d_{mQ} \geq 2\rho \Delta_m .$$

Then

$$\frac{d_{xQ}}{V(Q \mid X, w)} \leq \frac{8\rho^2}{w(X_m)} .$$

Proof Let $q = \arg\min_{z \in Q} d_{mz}$ be the closest $Q$ point to $m$. From (relaxed) triangle inequality and our assumptions:

$$d_{xQ} \leq d_{xq} \leq \rho(d_{mq} + d_{mx}) = \rho(d_{mQ} + d_{mx}) \leq \rho d_{mQ} + \frac{1}{2}d_{xQ} .$$

Rearranging, we get

$$d_{xQ} \leq 2\rho d_{mQ} . \quad (15)$$

Consider a point $y$ such that $d_{my} \leq \Delta_m$. Let $q' = \arg\min_{z \in Q} d_{yz}$ be the closest $Q$ point to $y$. From relaxed triangle inequality we have $d_{mq'} \leq \rho(d_{yy'} + d_{ym})$ and therefore

$$d_{yQ} = d_{yq'} \geq \frac{1}{\rho}d_{mq'} - d_{ym} \geq \frac{1}{\rho}d_{mQ} - \Delta_m \geq \frac{1}{\rho}d_{mQ} - \frac{1}{2\rho}d_{mQ} \geq \frac{1}{2\rho}d_{mQ} .$$

Thus, using the definition of $\Delta_m$:

$$V(Q \mid X, w) \geq \sum_{y \mid d_{yQ} \leq \Delta_m} w_y d_{yQ} \geq \frac{1}{2\rho} \sum_{y \mid d_{yQ} \leq \Delta_m} w_y d_{mQ} \geq \frac{1}{2\rho} d_{mQ} \sum_{y \in X_m \mid d_{yQ} \leq \Delta_m} w_y \geq \frac{1}{2\rho} d_{mQ} \frac{w(X_m)}{2} = \frac{1}{4\rho} d_{mQ} w(X_m) . \quad (16)$$

Combining (15) and (16) we obtain:

$$\frac{d_{xQ}}{V(Q \mid X, w)} \leq \frac{2\rho d_{mQ}}{4\rho w(X_m) d_{mQ}} = \frac{8\rho^2}{w(X_m)} \frac{1}{d_{mQ}} .$$

Lemma 4.3 Let $Q$ and point $x$ be such that

$$\exists m \in M, \ d_{xm} < \frac{1}{2\rho}d_{xQ} \text{ and } d_{mQ} \leq 2\rho \Delta_m .$$

Then

$$\frac{d_{xQ}}{V(Q \mid X, w)} \leq \frac{8\rho^2}{\alpha w(X_m)} .$$

Proof Let $q = \arg\min_{z \in Q} d_{zm}$ be the closest $Q$ point to $m$. We have

$$d_{xQ} \leq d_{xq} \leq \rho(d_{xm} + d_{mq}) \leq \rho d_{mQ} \leq \frac{1}{2}d_{xQ} + 2\rho^2 \Delta_m .$$

Therefore,

$$d_{xQ} \leq 4\rho^2 \Delta_m . \quad (17)$$
Using (13) we obtain

\[ V(Q | X, w) \geq \alpha V(M | X, w) \geq \alpha \frac{1}{2} \sum_{y \in M} w(X_y) \Delta_y \geq \alpha \frac{1}{2} w(X_m) \Delta_m . \]  

(18)

Combining (17) and (18) we obtain

\[ \frac{d_{xQ}}{V(Q | X, w)} \leq \frac{4\rho^2 \Delta_m}{\frac{1}{2} \alpha w(X_m) \Delta_m} \leq \frac{8\rho^2}{\alpha} \frac{1}{w(X_m)} . \]

5 Conclusion

We presented the one2all construction for clustering costs: From any set of centers \( M \) with cost \( V(M) \) we obtain a summary structure of size \( O(|M|) \), in the form of a multi-objective sample, with which we can estimate the clustering cost by any set \( Q \) with cost that is at least a fraction of \( V(M) \).

Looking forward, we point on some potential applications beyond estimation and optimization of clustering cost. First, we note that the set of distances of \( Q \) to the one2all sample \( S \) is essentially a sketch of the full (weighted) distance vector of \( Q \) to \( X \) [13]. Sketches of different sets \( Q \) allow us to estimate relations between the respective full vectors, such as distance norms, weighted Jaccard similarity, quantile aggregates, and more, which can be useful building blocks in other applications. Second, recent work casted Euclidean \( k \)-means clustering as a constrained rank-\( k \) approximation problem [14]. This connection facilitated interesting feedback between techniques designed for low-rank approximation and for clustering. We thus hope that our technique and insights might lead to further progress on other low-rank approximation problems.

References

[1] P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan. Geometric approximation via coresets. In *Combinatorial and computational geometry, MSRI*. University Press, 2005.

[2] D. Aloise, A. Deshpande, P. Hansen, and P. Popat. NP-hardness of Euclidean sum-of-squares clustering. *Mach. Learn.*, 75(2), 2009.

[3] D. Arthur and S. Vassilvitskii. K-means++: The advantages of careful seeding. In *SODA*, 2007.

[4] P. Awasthi, M. Charikar, R. Krishnaswamy, and A. K. Sinop. The hardness of approximation of Euclidean \( k \)-means. In *SoCG*, 2015.

[5] V. Braverman, D. Feldman, and H. Lang. New frameworks for offline and streaming coreset constructions. *CoRR*, abs/1612.00889, 2016.

[6] K. R. W. Brewer, L. J. Early, and S. F. Joyce. Selecting several samples from a single population. *Australian Journal of Statistics*, 14(3):231–239, 1972.

[7] M. T. Chao. A general purpose unequal probability sampling plan. *Biometrika*, 69(3):653–656, 1982.

[8] S. Chechik, E. Cohen, and H. Kaplan. Average distance queries through weighted samples in graphs and metric spaces: High scalability with tight statistical guarantees. In *RANDOM*. ACM, 2015.

[9] K. Chen. On coresets for \( k \)-median and \( k \)-means clustering in metric and Euclidean spaces and their applications. *SIAM J. Comput.*, 39(3), 2009.
[10] E. Cohen. Size-estimation framework with applications to transitive closure and reachability. *J. Comput. System Sci.*, 55:441–453, 1997.

[11] E. Cohen. Multi-objective weighted sampling. In *HotWeb*. IEEE, 2015. full version: http://arxiv.org/abs/1509.07445.

[12] E. Cohen, N. Duffield, C. Lund, M. Thorup, and H. Kaplan. Efficient stream sampling for variance-optimal estimation of subset sums. *SIAM J. Comput.*, 40(5), 2011.

[13] E. Cohen, H. Kaplan, and S. Sen. Coordinated weighted sampling for estimating aggregates over multiple weight assignments. *VLDB*, 2(1–2), 2009. full: http://arxiv.org/abs/0906.4560.

[14] M. B. Cohen, S. Elder, C. Musco, C. Musco, and M. Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *STOC*. ACM, 2015.

[15] D. Feldman and M. Langberg. A unified framework for approximating and clustering data. In *STOC*. ACM, 2011.

[16] D. Feldman, M. Schmidt, and C. Sohler. Turning big data into tiny data: Constant-size coresets for k-means, PCA and projective clustering. In *SODA*. ACM-SIAM, 2013.

[17] M. H. Hansen and W. N. Hurwitz. On the theory of sampling from finite populations. *Ann. Math. Statist.*, 14(4), 1943.

[18] S. Har-Peled and S. Mazumdar. On coresets for k-means and k-median clustering. In *STOC*. ACM, 2004.

[19] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[20] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. A local search approximation algorithm for k-means clustering. *Computational Geometry*, 28(2), 2004.

[21] L. Kish and A. Scott. Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66(335):pp. 461–470, 1971.

[22] S. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theor.*, 1982.

[23] R. R. Mettu and C. G. Plaxton. Optimal time bounds for approximate clustering. *Mach. Learn.*, 56(1-3), 2004.

[24] P. J. Saavedra. Fixed sample size pps approximations with a permanent random number. In *Proc. of the Section on Survey Research Methods*, pages 697–700, Alexandria, VA, 1995. American Statistical Association.

[25] C-E. Särndal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling*. Springer, 1992.

[26] Y. Tillé. *Sampling Algorithms*. Springer-Verlag, New York, 2006.

[27] D. Wei. A constant-factor bi-criteria approximation guarantee for k-means++. In *NIPS*, 2016.