The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Clustering Small Samples with Quality Guarantees: Adaptivity with One2all pps

Edith Cohen
Google Research, USA
Tel Aviv University, Israel

Shiri Chechik
Tel Aviv University, Israel

Haim Kaplan
Tel Aviv University, Israel

Abstract

Clustering of data points is a fundamental tool in data analysis. We consider points X in a relaxed metric space, where the triangle inequality holds within a constant factor. A clustering of X is a partition of X defined by a set of points Q (*centroids*), according to the closest centroid. The *cost* of clustering X by Q is $V(Q) = \sum_{x \in X} d_{xQ}$. 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 $|Q| = k$. When the data set X is very large, we seek efficient constructions of small samples that can act as surrogates for performing these tasks. Existing constructions that provide quality guarantees, however, are either worst-case, and unable to benefit from structure of real data sets, or make explicit strong assumptions on the structure. We show here how to avoid both these pitfalls using adaptive designs.

The core of our design are the novel *one2all* probabilities, computed for a set M of centroids and $\alpha \geq 1$: The clustering cost of *each* Q with cost $V(Q) \geq V(M)/\alpha$ can be estimated well from a sample of size $O(\alpha |M| \epsilon^{-2})$. For cost estimation, we apply one2all with a bicriteria approximate M, while adaptively balancing $|M|$ and α to optimize sample size per quality. For clustering, we present a wrapper that adaptively applies a base clustering algorithm to a sample S, using the smallest sample that provides the desired statistical guarantees on quality. We demonstrate experimentally the huge gains of using our adaptive instead of worst-case methods.

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 $(\mathcal{M}, d)$, 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 pth 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 ($p = 2$), commonly used for clustering, we have $\rho = 2$.

Each set $Q \subset \mathcal{M}$ of points (*centroids*) 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

$$V(Q \mid X, \boldsymbol{w}) = \sum_{x \in X} w_x d_{xQ} \ , \tag{1}$$

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 (1) 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:

$$\arg \min_{Q \mid |Q| \leq k} V(Q \mid X, \boldsymbol{w}) \ . \tag{2}$$

Optimal clustering is computationally hard (Aloise et al. 2009) even on Euclidean spaces and even to tightly approximate (Awasthi et al. 2015). There is a local search polynomial algorithm with $9 + \epsilon$ approximation ratio (Kanungo et al. 2004). In practice, clustering is solved using heuristics, most notably Lloyd's algorithm (EM) for squared Euclidean distances (Lloyd 1982) and scalable approximation algorithms such as KMEANS++ (Arthur and Vassilvitskii 2007) for general metrics. EM iterates allocating points to clusters defined by the nearest centroid, and replacing each centroid with the center of mass $\sum_x w_x x$ of its cluster. Each iteration uses $|X|k$ pairwise distance computations. It is a heutistic because although each iteration reduces the clustering cost, the algorithm can terminates in a local minima. KMEANS++ produces a sequence of points $\{m_i\}$: The first point m_1 is selected randomly with probability $\propto w_x$ and a point m_i us selected with probability $\propto w_x d_{\{m_1,\ldots,m_{i-1}\}x}$. Each iteration requires $O(|X|)$ pairwise distance computations. KMEANS++ guarantees that the expected clustering cost of the first k points is within an $O(\log k)$ factor of the optimum k-means cost. Moreover, KMEANS++ provides bicriteria guarantees (Aggarwal, Deshpande, and Kannan 2009; Wei 2016): The first βk points selected (for some constant $\beta > 1$) have expected clustering cost is within a constant factor of the optimum k-means cost. In practice, *kmeans++* is often used to initiallize Lloyd's algorithm.

With very large point sets X, we seek a summary structure that can be efficiently computed and act as a surrogate of the full data for the purpose of approximating clustering costs. When designing such strutcures we seek to optimize a tradeoff between its size and the quality guarantees it provides. *Coresets* that build on the theory of ϵ-nets were studied in the computational geometry literature (Agarwal, Har-Peled, and Varadarajan 2005; Har-Peled and Mazumdar 2004) and notable constructions include (Mettu and Plaxton 2004; Chen 2009; Feldman and Langberg 2011; Feldman, Schmidt, and Sohler 2013). A common coreset form is a subset $S \subset X$ with weights $\boldsymbol{w}'$ so that $V(Q \mid S, \boldsymbol{w}') \approx V(Q \mid X, \boldsymbol{w})$ for each Q of size k.

The bulk of coreset constructions are aimed to provide strong "ForAll" statistical guarantees, which bound the distribution of the maximum approximation error of all Q of size k. The ForAll requirement, however, comes with a hefty increase in size that is generally unnecessary for the two tasks we have at hand: For clustering cost queries, weaker per-query "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 $V(Q \mid S, \boldsymbol{w}')$ are approximate minimizers of $V(Q \mid X, \boldsymbol{w})$. Moreover, previous constructions use coreset sizes that are *worst-case*, based on general (VC) dimension or union bounds. A state-of-the-art asymptotic worst-case bound of $O(k\epsilon^{-2} \log k \log n)$ is claimed in (Braverman, Feldman, and Lang 2016).

Even when a worst-case bound is tight up to constants, which typically it is not (constants are not even specified in state of the art coreset constructions), it only means it is tight for pathological data sets of the particular size and dimension. A much smaller summary structure might suffice in the presence of structure typical in data such as natural clusterability (which is what we seek) and lower dimensionality than the ambient space.

It seems on the surface, however, that in order to achieve statistical guarantees on quality of the results one must either make explicit assumptions on the data or use the worst-case size. We show here how to avoid both these pitfalls via elegant adaptive designs.

Contribution Overview

At the heart of our approach are novel summary structures for clustering costs based on multi-objective probability-proportional-to-size (pps) samples (Cohen, Kaplan, and Sen 2009; Cohen 2015), which build on the classic notion of sample coordination (Kish and Scott 1971; Brewer, Early, and Joyce 1972; Saavedra 1995; Cohen 1997).

Consider a particular set Q of centroids. The theory of weighted sampling (Särndal, Swensson, and Wretman 1992; Tillé 2006) tells us that to estimate the sum $V(Q \mid X, \boldsymbol{w})$ it suffices to compute a sample S of ϵ^{-2} points included with probabilities $p_x \propto w_x d_{Qx}$ proportional to their contribution to the sum (Hansen and Hurwitz 1943). The inverse-probability (Horvitz and Thompson 1952) estimator of $V(Q \mid X, \boldsymbol{w})$

$$V(\widehat{Q \mid X}, \boldsymbol{w}) := V(Q \mid S, \{w_x/p_x\}) \;,$$

is unbiased with normalized root mean squared error that is bounded by ϵ and with Bernstein-Chernoff concentration. Note that any set of nonnegative probabilities yields an unbiased estimate but weighted sampling is necessary for providing concentration and variance bounds. A challenge here for us is that we are interested in simultaneously having pps-like quality guarantees for *any* subset Q of size k whereas the estimate $V(Q' \mid S, \{w_x/p_x\})$ when S is taken from a sample distribution according to $Q \neq Q'$ will not provide these guarantees for Q'. To obtain these quality guarantees for all Q by a single sample, we use *multi-objective* pps sampling probabilities, where the sampling probability of each point $x \in X$ is the maximum pps probability over all Q of size k. Evidently, a multi-objective sample can be larger than a dedicated sample providing the same guarantees. We refer to the increase factor in sample size as the *overhead*.

For our applications, we need not only to bound these probabilities but also to efficiently compute such bounds – evidently it is not feasible to enumerate the infinite number of subsets Q. The basis of our work is a novel general construction of the *one2all* probabilities assigned to points in X and defined with respect to a set M of points and $\alpha \geq 1$. The one2all probabilities upper bound the pps probabilities for any Q with clustering cost $V(Q) \geq V(M)/\alpha$. The overhead is only $O(\alpha|M|)$ and the computation uses $|M|n$ pairwise distance computations. Note that this holds for any (relaxed) metric space and in particular, does not depend on the dimension or the size of the data.

The one2all probabilities of an optimal clustering Q^* of size k and $\alpha = 1$ upper bound the pps probabilities for all Q of size k and *existentially* establishes that the overhead is $O(k)$. This generalizes previous work (Chechik, Cohen, and Kaplan 2015) for the case where $k = 1$, where clustering cost reduces to inverse classic closeness centrality (sum of distances from a single point Q).

We use one2all to obtain a clustering cost oracle for queries with cost at least a specified C: We pre-process the data by applying KMEANS++ (Arthur and Vassilvitskii 2007) to obtain a sequence M of centroids. We can then apply one2all with M and $\alpha = \max\{1, V(M \mid X, \boldsymbol{w})/C\}$. In a further adaptive optimization, we consider all prefixes $M' \subset M$ and corresponding $\alpha' \leftarrow \alpha V(M')/V(M)$ and select the sweet-spot prefix M' of the centroids sequence returned by KMEANS++ that minimizes the sample size. Finally, we sample X to obtain S.

Our oracle then processes cost query Q by computing and returning the clustering cost of S by Q: $V(Q \mid S, \{w_x/p_x\})$. Each computation performs $O(|S||Q|)$ pairwise distance computations instead of the $O(n|Q|)$ that would have been required over the full data.

To obtain an oracle for all Q of size k, we use the bicriteria approximation (Wei 2016) guarantee of KMEANS++: For some small constants β and $\alpha > 1$, M of size βk is likely to have cost $V(M) \leq \alpha V(Q^*)$. One2all probabilities for these M and α provide us with an oracle of size at most $\beta\alpha$ times larger than if we had used Q^*. In particular, the sample size is $O(\alpha|M|\epsilon^{-2}) = O(k\epsilon^{-2})$. This construction, however, uses the bicriteria bound in a worst-case manner. Instead, we propose a feedback oracle design that adaptively lower the target clustering cost C by increasing sample size when provided with a query with lower clustering cost than C.

For the task of approximate clustering, we adapt an optimization framework over multi-objective samples (Cohen 2015). The meta algorithm is a wrapper that inputs multi-objective pps probabilities, specified error guarantee ϵ, and a black-box (exact, approximate, bicriteria, heuristic) base clustering algorithm $\mathcal{A}$. The wrapper applies $\mathcal{A}$ to a sample to obtain a respective approximate minimizer of the clustering cost over the sample. When the sample is much smaller than the full data set and is yet large enough to support the optimization, we can improve clustering quality with reduced computation. Our initial multi-objective pps sample provides ForEach guarantees that apply to each estimate in isolation but not to the sample optimum. In particular, it does not guarantee us that the solution over the sample has the respective quality over the full data set. A larger sample may or may not be required. One can always increase the sample by a worst-case upper bound (using a union bound or domain-specific dimensionality arguments). Our adaptive approach exploits a critical benefit of ForEach: That is, we are able to *test* the quality of the sample approximate optimizer Q returned by $\mathcal{A}$: If the clustering cost of $V(Q \mid X, \boldsymbol{w})$ agrees with the estimate $V(Q \mid S, \boldsymbol{w}')$ then we can certify that Q has similar (within $(1+\epsilon)$ quality over X as it has over the sample S. Otherwise, the wrapper doubles the sample size S and repeats until the test is satisfied. Since the base algorithm is always at least linear, the total computation is dominated by that last largest sample size we use.

Note that the only computation performed over the full data set are the $O(k)$ iterations of KMEANS++ that produce (M, α) to which we apply one2all. Each such iteration performs $O(|X|)$ distance computations. This is a significant gain, as even with Lloyd's algorithm (EM heuristic), each iteration is $O(k|X|)$.

A further adaptive optimization targets this initial cost: On real-world data it is often the case that much fewer iterations of KMEANS++ bring us to within some reasonable factor α of the optimal k-clustering. We thus propose to adaptively perform additional KMEANS++ iterations as to balance their cost with the size of the sample that we need to work with.

We demonstrate through experiments on both synthetic and real-world data the potentially huge gains of our data-adaptive methods as a replacement to worst-case-bound size samples or coresets.

Multi-objective pps samples for clustering

We review the framework of weighted and multi-objective weighted sampling (Cohen 2015) in our context of clustering costs. Consider approximating the clustering cost $V(Q \mid X, \boldsymbol{w})$ 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 (Horvitz and Thompson 1952) of $V(Q \mid X, \boldsymbol{w})$:

$$V(\widehat{Q \mid X}, \boldsymbol{w}) := \sum_{x \in S} w_x \frac{d_{xQ}}{p_x} = V(Q \mid S, \{w_x/p_x\}) \ . \quad (3)$$

Note that the estimate is equal to the clustering cost of S with weights w_x/p_x by Q.

Probability proportional to size (pps) sampling

To obtain guarantees on the estimate quality of the clustering cost by Q, we need to use weighted sampling (Hansen and Hurwitz 1943). The pps *base* probabilities of Q for $x \in X$ are

$$\psi_x^{(Q|X,\boldsymbol{w})} = \frac{w_x d_{xQ}}{\sum_{y \in X} w_y d_{yQ}} \ . \quad (4)$$

The pps probabilities for a sample with size parameter $r > 1$ are

$$r * \psi_x^{(Q|X,\boldsymbol{w})} = \min\{1, r\psi_x^{(Q|X,\boldsymbol{w})}\} \ .$$

Note that the (expected) sample size is $\sum_x p_x$. When $p_x = r * \psi_x^{(Q|X,\boldsymbol{w})}$, the size is at most r. With pps sampling we obtain the following guarantees:

Theorem 0.1 ((weak) pps sampling) *Consider a sample S where each $x \in X$ is included independently (or using VarOpt dependent sampling (Chao 1982; Cohen et al. 2011)) with probability $p_x \geq \nu\epsilon^{-2} * \psi_x^{(Q|X,\boldsymbol{w})}$, where $\nu \leq 1$. Then the estimate* (3) *has the following statistical guarantees:*

- *Normalized root mean square error (NRMSE) bounded by*

$$\begin{aligned} \text{NRMSE}[V(\widehat{Q \mid X}, \boldsymbol{w})] &= \\ \frac{\sqrt{\mathsf{E}_{S\sim \boldsymbol{p}}[V(Q \mid S, \{w_x/p_x\})^2] - V(Q \mid X, \boldsymbol{w})^2}}{V(Q \mid X, \boldsymbol{w})} &\leq \frac{\epsilon}{\sqrt{\nu}} \ . \end{aligned}$$

- *Chernoff-Bernstein concentration, which bounds the relative error as follows:*

$$\begin{aligned} \Pr[V(Q \mid S) \geq (1+\delta)V(Q \mid X)] &\leq e^{-\frac{\delta \ln(1+\delta)\nu\epsilon^{-2}}{2}} \\ \Pr[V(Q \mid S) \leq (1-\delta)V(Q \mid X)] &\leq e^{-\frac{\delta^2 \nu\epsilon^{-2}}{2}} \end{aligned}$$

For our purposes here, we bound on the probability that when $\nu < 1$ the estimate exceeds $\nu^{-1}V(Q \mid X, \boldsymbol{w})$:

Corollary 0.1 (Overestimation probability)

$$\Pr[V(Q \mid S) \geq \frac{V(Q \mid X)}{\nu}] \leq e^{-(1-\nu)\ln(1/\nu)\epsilon^{-2}/2}$$

Multi-objective pps

When we seek estimates with statistical guarantees for a set $\mathcal{Q}$ of queries (for example, all possible sets of k centroids), we use multi-objective pps (Cohen, Kaplan, and Sen 2009; Cohen 2015). The *multi-objective (MO) pps base sampling probabilities* are defined as the maximum of the pps base probabilities over $Q \in \mathcal{Q}$:

$$\psi_x^{(\mathcal{Q}|X,\boldsymbol{w})} = \max_{Q \in \mathcal{Q}} \psi_x^{(Q|X,\boldsymbol{w})} \ . \quad (5)$$

Accordingly, for a size parameter r, the multi-objective pps probabilities are

$$r * \psi_x^{(\mathcal{Q}|X,\boldsymbol{w})} = \min\{1, r\psi_x^{(\mathcal{Q}|X,\boldsymbol{w})}\} = \max_{Q \in \mathcal{Q}} r * \psi_x^{(Q|X,\boldsymbol{w})} \ .$$

A key property of multi-objective pps is that the error bounds of dedicated pps samples (Theorem 0.1 and Corollary 0.1) hold. We refer to these multi-objective statistical quality guarantees as "ForEach," meaning that they hold for each Q over the distribution of the samples. We define the *overhead* of multi-objective sampling $\mathcal{Q}$ or equivalently of the respective base probabilities as:

$$h(\mathcal{Q} \mid X, \boldsymbol{w}) := |\boldsymbol{\psi}^{(\mathcal{Q}|X,\boldsymbol{w})}|_1 := \sum_{x\in X} \psi_x^{(\mathcal{Q}|X,\boldsymbol{w})} .$$

The overhead bounds the factor-increase in sample size due to "multi-objectiveness:" The multi-objective pps sample size with size parameter r is at most $|r * \psi^{(\mathcal{Q}|X,\boldsymbol{w})}|_1 \leq rh(\mathcal{Q} \mid X, \boldsymbol{w})$.

Often we can not efficeintly compute $\boldsymbol{\psi}$ but only obtain upper bounds $\boldsymbol{\pi} \geq \boldsymbol{\psi}^{(\mathcal{Q}|X,\boldsymbol{w})}$. Accordingly, we use sampling probabilities $r * \boldsymbol{\pi}$. The use of upper bounds increases the sample size. We refer to $h(\boldsymbol{\pi}) = |\boldsymbol{\pi}|_1$ as the overhead of $\boldsymbol{\pi}$. We seek upper-bounds $\boldsymbol{\pi}$ with overhead not much larger than $h(\mathcal{Q} \mid X, \boldsymbol{w})$.

one2all probabilities

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

$$\forall x, y, z \in \mathcal{M},\ d_{xy} \leq \rho(d_{xz} + d_{zy}) . \tag{6}$$

Let $(X, \boldsymbol{w})$ where $X \subset \mathcal{M}$ and $\boldsymbol{w} > 0$ be weighted points in $\mathcal{M}$. For another set of points $M \subset \mathcal{M}$, which we refer to as *centroids*, and $q \in M$, we denote by

$$X_q^{(M)} = \{x \in X \mid d_{xq} = d_{xM}\}$$

the points in X that are closest to centroid q. In case of ties we apply arbitrary tie breaking to ensure that $X_q^{(M)}$ for $q \in M$ forms a partition of X. We will assume that $X_q^{(M)}$ 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.

Our *one2all* construction takes one set of centroids M and computes base probabilities for $x \in X$ such that samples from it allow us to estimate the clustering costs of all Q with estimation quality guarantees that depends on $V(Q \mid X, \boldsymbol{w})$. For a set M we define the *one2all base probabilities* $\boldsymbol{\pi}^{(M|X,\boldsymbol{w})}$ as:

$$\forall m \in M,\ \forall x \in X_m, \tag{7}$$
$$\pi_x^{(M|X,\boldsymbol{w})} \quad = \quad \min\left\{1, \max\left\{2\rho\frac{w_x d_{xM}}{V(M \mid X, \boldsymbol{w})}, \frac{8\rho^2 w_x}{w(X_m)}\right\}\right\} .$$

We omit the superscripts when clear from context.

Theorem 0.2 (one2all) *Consider weighted points* $(X, \boldsymbol{w})$ *in a relaxed metric space with parameter* ρ*, points* M*, and a set* Q *of centroids. Then*

$$\boldsymbol{\pi}^{(M|X,\boldsymbol{w})} \geq \min\{1, \frac{V(Q \mid X, \boldsymbol{w})}{V(M \mid X, \boldsymbol{w})}\}\boldsymbol{\psi}^{(Q|X,\boldsymbol{w})} ,$$

where $\boldsymbol{\pi}^{(M|X,\boldsymbol{w})}$ *are the one2all base probabilities for* M.

The full proof of the Theorem is provided in the next section. As a corollary, we obtain that for $r \geq 1$, we can upper bound the multi-objective base pps probabilities $\boldsymbol{\psi}^{(\mathcal{Q}|X,\boldsymbol{w})}$ and the overhead $h(\mathcal{Q})$ of the set $\mathcal{Q}$ of all Q with at least a fraction $1/r$ of the clustering cost of M:

Corollary 0.2 *Consider* M *and* $r \geq 1$ *and the set* $\mathcal{Q} = \{Q \mid V(Q \mid X, \boldsymbol{w}) \geq V(M \mid X, \boldsymbol{w})/r\}$*. Then,* $r * \boldsymbol{\pi}^{(M|X,\boldsymbol{w})} \geq \boldsymbol{\psi}^{(\mathcal{Q}|X,\boldsymbol{w})}$ *and* $h(\mathcal{Q}) \leq r(8\rho^2|M| + 2\rho)$.

Proof For $Q \in \mathcal{Q}$, $r * \boldsymbol{\pi}^{(M|X,\boldsymbol{w})} \geq r \min\{1, \frac{V(Q|X,\boldsymbol{w})}{V(M|X,\boldsymbol{w})}\} * \boldsymbol{\psi}^{(Q|X,\boldsymbol{w})} \geq \boldsymbol{\psi}^{(Q|X,\boldsymbol{w})}$. Note that $|\boldsymbol{\pi}^{(M|X,\boldsymbol{w})}|_1 \leq 8\rho^2|M| + 2\rho$. ■

We can also upper bound the multi-objective overhead of all sets of centroids of size k:

Corollary 0.3 *For* $k \geq 1$*, let* $\mathcal{Q}$ *be the set of all* k*-subsets of points in a relaxed metric space* $\mathcal{M}$ *with parameter* ρ*. The multi-objective pps overhead of* $\mathcal{Q}$ *satisfies*

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

Proof We apply Corollary 0.2 with M being the k-means optimum and $r = 1$. ■

Proof of the one2all Theorem

Consider a set of points Q and let $\alpha = \max\{1, \frac{V(M|X,\boldsymbol{w})}{V(Q|X,\boldsymbol{w})}\}$. To prove Theorem 0.2, we need to show that $\forall x \in X$,

$$\psi_x^{(Q|X,\boldsymbol{w})} = \frac{w_x d_{xQ}}{V(Q \mid X, \boldsymbol{w})} \leq \alpha\pi_x^{(M|X,\boldsymbol{w})} . \tag{8}$$

We will do a case analysis, as illustrated in Figure 1. We first consider points x such that the distance of x to Q is not much larger than the distance of x to M. Property (8) follows using the first term of the maximum in (7).

Lemma 0.4 *Let* x *be such that* $d_{xQ} \leq 2\rho d_{xM}$*. Then*

$$\frac{w_x d_{xQ}}{V(Q \mid X, \boldsymbol{w})} \leq 2\rho\alpha\frac{w_x d_{xM}}{V(M \mid X, \boldsymbol{w})} .$$

Proof Using $V(Q \mid X, \boldsymbol{w}) \geq V(M \mid X, \boldsymbol{w})/\alpha$ we get

$$\frac{d_{xQ}}{V(Q \mid X, \boldsymbol{w})} \leq \alpha\frac{d_{xQ}}{V(M \mid X, \boldsymbol{w})} \leq 2\rho\alpha\frac{d_{xM}}{V(M \mid X, \boldsymbol{w})} .$$

■

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

$$d_{xQ} \geq 2\rho d_{xM} . \tag{9}$$

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

$$\sum_{x\in X_q | d_{xq} \leq \Delta_q} w_x \quad \geq \quad \frac{1}{2}w(X_q) \tag{10}$$

$$\sum_{x\in X_q | d_{xq} \geq \Delta_q} w_x \quad \geq \quad \frac{1}{2}w(X_m) . \tag{11}$$

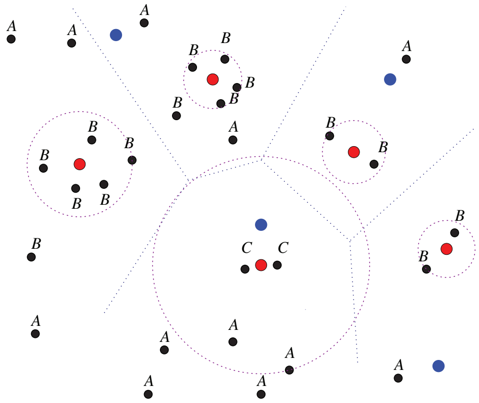

Figure 1: Illustration of the one2all construction proof with $\rho = 1$. The data points X are in black. The points in M are colored red. We show the respective Voronoi partition and for each cluster, we show circles centered at the respective $m \in M$ (red) point with radius Δ_m. The points in blue are a set Q. The points $x \in X$ are labeled A if $d_{xQ} < 2d_{xM}$ (and we apply Lemma 0.4). 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 0.5) and is labeled C otherwise (Lemma 0.6).

It follows from (11) that for all $q \in M$,

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

Therefore,

$$V(M \mid X, \boldsymbol{w}) = \sum_{q\in M} V(M \mid X_q, \boldsymbol{w}) \geq \frac{1}{2} \sum_{q\in M} w(X_m)\Delta_q . \tag{12}$$

We now return to our proof for x that satisfies (9). We will show that property (8) holds using the second term in the max operation in the definition (7). Specifically, let m be the closest M point to x. We will show that

$$\frac{d_{xQ}}{V(Q \mid X, \boldsymbol{w})} \leq 8\rho^2\alpha\frac{1}{w(X_m)} . \tag{13}$$

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 0.5 *Let x be such that*

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

Then

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

Proof Let $q = \arg\min_{z\in Q} d_{mz}$ be the closest Q point to m. From (relaxed) triangle inequality (6) 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

$$d_{xQ} \leq 2\rho d_{mQ} . \tag{14}$$

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 $d_{mq'} \leq \rho(d_{yq'} + d_{ym})$ and therefore

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

Thus, using the definition of Δ_m (10):

$$\begin{aligned} V(Q \mid X, \boldsymbol{w}) &\geq \sum_{y \mid d_{yQ}\leq\Delta_m} w_y d_{yQ} \geq \frac{1}{2\rho} \sum_{y \mid d_{yQ}\leq\Delta_m} w_y d_{mQ} \\ &\geq \frac{1}{2\rho}d_{mQ} \sum_{y\in X_m \mid d_{yQ}\leq\Delta_m} w_y \\ &\geq \frac{1}{2\rho}d_{mQ}\frac{w(X_m)}{2} = \frac{1}{4\rho}d_{mQ}w(X_m) . \end{aligned} \tag{15}$$

Combining (14) and (15) we obtain:

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

■

Lemma 0.6 *Let a point x be such that*

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

Then

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

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

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

Therefore,

$$d_{xQ} \leq 4\rho^2\Delta_m . \tag{16}$$

Using (12) we obtain

$$\begin{aligned} V(Q \mid X, \boldsymbol{w}) &\geq V(M \mid X, \boldsymbol{w})/\alpha \geq \frac{1}{2\alpha}\sum_{y\in M} w(X_y)\Delta_y \\ &\geq \frac{1}{2\alpha}w(X_m)\Delta_m . \end{aligned} \tag{17}$$

Combining (16) and (17) we obtain

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

■

n	d	k	guarantee ϵ	adaptive $\frac{\lvert S\rvert}{n}$	worst-case $\frac{\lvert S\rvert}{n}$	× gain	est. err	$\frac{V(Q\lvert X)}{V_{\text{ground-truth}}}$	$\frac{V(\{m_0,\ldots,m_k\}\lvert X)}{V_{\text{ground-truth}}}$	sweet-spot
Mixture of Gaussians data sets										
5×10^5	10	5	0.10	0.0500	1.00	20.0	0.008	1.07	2.50	2.3
5×10^5	10	5	0.20	0.0136	1.00	73.3	0.012	1.10	2.39	2.6
2.5×10^6	10	5	0.20	0.0025	1.00	90.2	0.0160	1.14	2.14	2.2
1×10^7	10	5	0.20	0.00066	1.00	94.4	0.018	1.12	2.07	2.6
2×10^6	10	20	0.10	0.04839	1.00	20.7	0.0018	1.14	2.27	9.1
2×10^6	10	20	0.20	0.012007	1.00	83.2	0.008	1.18	2.25	9.0
2×10^6	10	50	0.20	0.0298	1.00	33.5	0.0057	1.16	2.24	19.5
2×10^6	10	100	0.20	0.061918	1.00	16.2	0.0058	1.15	2.22	40.8
1×10^6	20	10	0.10	0.05293	1.00	18.9	0.0035	1.17	2.39	5.0
1×10^6	50	10	0.10	0.04726	1.00	21.2	0.0037	1.19	2.65	3.9
1×10^6	100	10	0.10	0.05287	1.00	18.9	0.0035	1.18	2.6	4.9
MNIST data set										
6×10^5	784	10	0.20	0.0371	1.00	26.9	0.018	0.985	1.765	1.0
Fashion data set										
6×10^5	784	10	0.20	0.05720	1.00	17.5	0.021	0.91	1.65	1.0

Table 1: Sample size needed to meet the specified statistical guarantees by our adaptive and worst-case approaches. The reported estimation error is for the estimate $V(Q \mid S)$ of $V(Q \mid X)$ of the returned clustering Q from the final sample S. It is the square root of the normalized squared error, averaged over repetitions. .

Algorithm 1 Clustering Wrapper

```
Input: points $X$, weights $\boldsymbol{w} > \mathbf{0}$, $\epsilon > 0$, a clustering algorithm $\mathcal{A}$ that inputs a weighted set of
       points and returns $Q \in \mathcal{Q}$.
Output: Set $Q$ of $k$ centroids with statistical guarantees on quality over $X$ that match within $(1+\epsilon)$
        those provided by $\mathcal{A}$
// Initialization
$c \leftarrow \infty$ // Apply KMEANS++ to $(X, \boldsymbol{w})$
foreach iteration $i \in [2k]$ of KMEANS++$(X, \boldsymbol{w})$ do
    $m_i \leftarrow$ next centroid
    $v_i \leftarrow V(\{m_1, \ldots, m_i\} \mid X, \boldsymbol{w})$                    // Clustering cost
    if $iv_i < c$ then // sweet-spot one2all prob. (7)
        $\boldsymbol{\pi} \leftarrow \boldsymbol{\pi}^{(M|X,\boldsymbol{w})}$; $V_M \leftarrow v_i$ ; $c \leftarrow iv_i$
$r \leftarrow \frac{v_M}{v_{2k}}$                    // Initial sample size increase factor
$Q^* \leftarrow \{m_1, \ldots, m_k\}; \overline{V} \leftarrow v_k$                    // Best so far and upper bound
foreach $x \in X$ do $u_x \sim U[0,1]$ // Randomization for sampling
$S \leftarrow \{x \mid u_x \leq r\epsilon^{-2}\pi_x\}$ // Initial sample. $O(|S|)$ given preprocessed
  $\boldsymbol{\pi}$
  foreach $x \in S$ do $w'_x \leftarrow w_x / \min\{1, r\pi_x\}$ // weights for sampled points
// Main Loop
repeat
    // Cluster the sample $S$
    $Q \leftarrow \mathcal{A}(S, \boldsymbol{w}')$                    // Apply algorithm $\mathcal{A}$ to sample
    $V_Q \leftarrow V(Q \mid X, \boldsymbol{w})$ // Exact or approx using a validation sample
    if $V_Q < \overline{V}$ then $\overline{V} \leftarrow V_Q$  $Q^* \leftarrow Q$
    if $V_Q \leq (1+\epsilon)V(Q \mid S, \boldsymbol{w}')$ and $V_Q \geq V_M/r$ then
        break
    $r \leftarrow \max\{2, V_Q/V_M\}r$          // Increase the sample size parameter
    repeat// Increase sample size until $Q$ is cleared
        $S \leftarrow \{x \mid u_x \leq r\epsilon^{-2}\pi_x\}$ // Add points to sample
          foreach $x \in S$ do $w'_x \leftarrow w_x / \min\{1, r\pi_x\}$ // weights for sampled
          points
        $r \leftarrow 2r$
    until $V(Q \mid S, \boldsymbol{w}') > \min\{(1+\epsilon)\overline{V}, (1-\epsilon)V_Q\}$
until True
return $Q^*$
```

Clustering wrapper

The input to a clustering problem is a (weighted) set of points $(X, \boldsymbol{w})$ and $k > 0$. The goal is to compute a set Q of k centroids aimed to minimize the clustering cost $V(Q \mid X, \boldsymbol{w})$.

We present a *wrapper,* Algorithm 1, which inputs a clustering problem, a clustering algorithm $\mathcal{A}$, and $\epsilon > 0$, and returns a set of Q of k centroids. The wrapper computes weighted samples $(S, \boldsymbol{w}')$ of the input points $(X, \boldsymbol{w})$ and applies $\mathcal{A}$ to S. It then performs some tests on the clustering Q returned by $\mathcal{A}$, based on which, it either terminates and returns a clustering, or adaptively increases the sample size. The wrapper provides a statistical guarantee that the quality of the clustering Q returned by $\mathcal{A}$ on the sample $(S, \boldsymbol{w}')$ reflects, within $(1+\epsilon)$, its quality on the data.

The first part of the wrapper performs $2k$ iterations of KMEANS++ tor $(X, \boldsymbol{w})$ to compute a list $\{m_i\}$ of centroids and respective clustering costs $v_i = V(\{m_1, \ldots, m_i\} \mid X, \boldsymbol{w})$. While performing this computation, we identify a sweet-spot $M = \{m_1, \ldots, m_i\}$ and retain $\boldsymbol{\pi} : X$, which are the one2all base probabilities for M. Our wrapper separately maintains a size parameter r, that is initially set to $r = V_i/v_{2k}$. From Theorem 0.2, the probabilities $r * \boldsymbol{\pi}$ upper bound the base pps probabilities for all Q with clustering cost $V(Q \mid (X, \boldsymbol{w})) \geq V_M/r$. Initially, $r * \boldsymbol{\pi}$ is set for cost above v_{2k}. We then selects a fixed randomization $\boldsymbol{u}$, that will allow for coordination of samples selected with different size parameters.

The main iteration computes a weighted sample $(S, \boldsymbol{w}')$ selected with probabilities $\epsilon^{-2}r * \pi$. Our algorithm is applied to the sample $Q \leftarrow \mathcal{A}(S, \boldsymbol{w}')$ to obtain a set Q of k centroids. We compute (or estimate from a validation sample) the clustering cost over the full dataset $V_Q = V(Q \mid x, \boldsymbol{w})$. If V_Q is not lower than V_M/r and is also not much higher than the sample clustering cost $V(Q \mid S, \boldsymbol{w}')$, we break and return the best Q we found so far. Otherwise, we increase the size parameter r, augment the sample accordingly, and iterate. The increase in the size parameter at least doubles it and is set so that (i) We have $V_M/r \leq \overline{V}$, where $\overline{V}$ is the smallest clustering cost encountered so far. (ii) The set Q that was underestimated by the sample has estimate that is high enough to clear it from $\overline{V}$ or to comprise an accurate estimate.

Analysis

We show that if our algorithm $\mathcal{A}$ provides a certain approximation ratio on the quality of the clustering, then this ratio would also hold (approximately, with high confidence) over the full data set. A similar argument applies to a bicriteria bound.

The wrapper works with an optimistic initial choice of r, but increases it adaptively as necessary. The basis of the correctness of our algorithm is that we are able to detect when our choice of r is too low.

There are two separate issues that we tackle with adaptivity instead of with a pessimistic worst-case bound. The first is also addressed by our feedback oracle: For accurate estimates we need V_M/r to be lower than $V^* = V(Q^* \mid X, \boldsymbol{w})$ (the optimal clustering cost over X), which we do not know. Initially, $V_M/r = v_{2k}$, which may be higher than V^*. We increase r when we find a clustering Q with $V(Q) < V_M/r$. The potential "bad" event is when the optimum clustering Q^* has $V^* \ll V_M/r$ but is overestimated by a large amount in the sample resulting in the sample optimum V_S^* is much larger than V^*. As a consequence, the clustering algorithm $\mathcal{A}$ applied to the sample can find Q, for which the estimate is correct, and has cost above V_m/r. The approximation ratio over the sample is $V(Q \mid S)/V_S^*$ which can be much better than the true (much weaker) approximation ratio $V(Q \mid S)/V^*$ over the full data.

This bad event happens when $V_S^* \gg V^*$. But note that in expectation, $\mathsf{E}[V_S^*] \leq V^*$. Moreover, the probability of this bad event is bounded by $\exp(-\epsilon^{-2}/6)$ (see Theorem 0.1 and Corollary 0.1). We can make the probability of such bad event smaller by augmenting the wrapper as follows. When the wrapper is ready to return Q, we generate multiple samples of the same size and apply $\mathcal{A}$ to all these samples and take the best clustering generated. If we find a clustering with cost below V_m/r, we continue the algorithm. Otherwise, we return the best Q. The probability that all the repetitions are "bad" drops exponentially with the number of repetitions (samples) we use.

The second issue is inherent with optimization over samples. Suppose now that r is such that $V^* \geq V_M/r$. The statistical guarantees provided by the sample are "ForEach," which assure us that the cost is estimated well for a *given* Q. In particular, $V(Q^* \mid S, \boldsymbol{w}')$ is well concentrated around V^* (Theorem 0.1). This means that V_S^*, the optimal clustering cost over S, can only (essentially - up to concentration) be lower than V^*.

When we consider all Q of size k, potentially an infinite or a very large number of them, it is possible that some Q has clustering cost $V(Q \mid X, \boldsymbol{w}) \gg V^*$ but is grossly underestimated in the sample, having sample-based cost $V(Q \mid S, \{w_x/p_x\}) < V^*$. In this case, $V_S^* \ll V^*$ and our algorithm $\mathcal{A}$ that is applied to the sample will be fooled and can return such a Q. The worst-case approach to this issue is to use a union or a dimensionality bound that drastically increases sample size. We get around it using an adaptive optimization framework (Cohen 2015).

We can identify and handle this scenario, however, by testing Q returned by the base algorithm to determine if our algorithm was "fooled" by the sample:

$$V(Q \mid X, \boldsymbol{w}) \leq (1+\epsilon)V(Q \mid S, \{w_x/p_x\}) \,. \tag{18}$$

by either computing the exact cost $V(Q \mid X, \boldsymbol{w})$ or by drawing another independent *validation* sample S', and using the estimate $V(Q \mid S, \{w_x/p_x\})$. When the test fails, we increase the sample size and repeat. In fact, we at least double the sample size parameter, but otherwise increase it at least to the point that $V(Q \mid S, \{w_x/p_x\})$ can no longer fool the algorithm. The only bad event possible here is that the sample optimum is much larger than $V(Q^*)$. But as noted, when $V^* \geq V_M/r$ the probability of this for a particular sample is bounded by Theorem 0.1. Moreover, note that each increase of the sample size significantly strengthens the concentration of estimates for particular Q. Thus, the worst quality, over iterations, in which Q^* is estimated in the sample is dominated by the first iteration with $V(Q^*) \geq V_M/r$. Therefore, the approximation ratio over the sample is at least (up to the statistical concentration of the estimates of Q^*) the ratio over the full data.

Computation

The computation performed is dominated by two components. The first is the $2k$ iterations of KMEANS++ on the data, which are dominated by $2k|X|$ pairwise distance computations. These is the only component that must be performed over the original data. The second is the application of $\mathcal{A}$ to the sample. When $\mathcal{A}$ is (super)linear, it is dominated by the largest sample we use.

Experiments

We performed illustrative experiments for Euclidean k-means clustering on both synthetic and real-world data sets. We implemented our wrapper Algorithm 1 in numpy with the following base clustering algorithm $\mathcal{A}$: We use 5 applications of KMEANS++ and take the set of k centroids that has the smallest clustering cost. This set is used as an initialization to 20 iterations of Lloyd's algorithm. The use of KMEANS++ to initialize Lloyd's algorithm is a prevalent method in practice.

Synthetic data: We generated synthetic data sets by drawing n points $X \subset R^d$ from a mixture of k Gaussians. The means of the Gaussians are arranged to lie in a line with equal distances. The standard deviations of the Gaussians were drawn from a range equal to the distance to the closest mean. As a reference, we use the means of the Gaussians as the *ground truth* centroids.

MNIST and Fashion MNIST datasets: We use the MNIST data set of images of handwritten digits (LeCun and Cortes 2010) and the Fashion data set of images of clothing items (Xiao, Rasul, and Vollgraf 2017). Both data sets contain $n = 6 \times 10^5$ images coded as $d = 784$ dimensional vectors. There are $k = 10$ natural classes that correspond to the 10 digits or 10 types of clothing items. Our reference ground-truth centroids were taken as the mean of each class.

Worst-case bounds: We also report, for comparison, sizes based on state-of-the-art coresets constructions that provide the same statistical guarantees. The coreset sizes are determined using worst-case upper bounds. When constant factors are not specified, we *underestimate* them. Details are provided in the full version.

Adaptive bounds: Table 1 reports the results of our experiments. The first four columns report the basic parameters of each data set: The number of points n, clusters k, dimension

d, and the specified value of ϵ for the desired statistical guarantee. The middle columns report the final sample size $|S|$ used by the algorithm as a fraction of n, an underestimate on the corresponding coreset size from state of the art worst-case bounds, and the gain factor in sample size by using our adaptive algorithm instead of a worst-case bound. We can observe significant benefit that increases with the size of the data sets. On the MNIST data, the worst-case approach provides no data reduction.

The third set of columns reports the accuracy of the sample-based estimate of the cost of the final clustering Q. We can see that the error is very small (much smaller than ϵ). We also report the quality of the final clustering Q and the quality of the clusters obtained by applying KMEANS++ to X, relative to the cost of the "ground truth" centroids. We can see that the cost of the final clustering is very close (in the case of MNIST, is lower) than the "ground truth" cost. We also observe significant improvement over the cost of the KMEANS++ centroids used for initialization. The last column reports the number of KMEANS++ iterations on the full data set that was eventually used (the sweet spot value).

Conclusion

A salient feature of our methods is that we start with an optimistic small sample and increase it *adaptively* only in the face of evidence that a larger sample is indeed necessary for meeting the specified statistical guarantees on quality. Previous constructions use *worst-case* size summary structures that can be much larger. We demonstrate experimentally the very large potential gain, of orders of magnitude in sample sizes, when using our adaptive versus worst-case methods.

References

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

Aggarwal, A.; Deshpande, A.; and Kannan, R. 2009. Adaptive sampling for k-means clustering. In *RANDOM*.

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

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

Awasthi, P.; Charikar, M.; Krishnaswamy, R.; and Sinop, A. K. 2015. The hardness of approximation of Euclidean k-means. In *SoCG*.

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

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

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

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

Chen, K. 2009. On coresets for k-median and k-means clustering in metric and Euclidean spaces and their applications. *SIAM J. Comput.* 39(3).

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

Cohen, E.; Kaplan, H.; and Sen, S. 2009. Coordinated weighted sampling for estimating aggregates over multiple weight assignments. *VLDB* 2. full: http://arxiv.org/abs/0906.4560.

Cohen, E. 1997. Size-estimation framework with applications to transitive closure and reachability. *J. Comput. System Sci.* 55:441–453.

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

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

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

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

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

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

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

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

LeCun, Y., and Cortes, C. 2010. MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/.

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

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

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

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

Tillé, Y. 2006. *Sampling Algorithms*. Springer-Verlag.

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

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR* abs/1708.07747.