A Framework for Statistical Clustering with a Constant Time Approximation Algorithms for K-Median Clustering

Shai Ben-David

Department of Computer Science
Technion, Haifa 32000, Israel
and
School of ECE**
Cornell university, Ithaca 14853, NY
shai@ece.cornell.edu

Abstract. We consider a framework in which the clustering algorithm gets as input a sample generated i.i.d by some unknown arbitrary distribution, and has to output a clustering of the full domain set, that is evaluated with respect to the underlying distribution. We provide general conditions on clustering problems that imply the existence of sampling based clusterings that approximate the optimal clustering. We show that the K-median clustering, as well as the Vector Quantization problem, satisfy these conditions. In particular our results apply to the *sampling - based approximate clustering* scenario. As a corollary, we get a sampling-based algorithm for the K-median clustering problem that finds an almost optimal set of centers in time depending only on the confidence and accuracy parameters of the approximation, but independent of the input size. Furthermore, in the Euclidean input case, the running time of our algorithm is independent of the Euclidean dimension.

1 Introduction

We consider the following fundamental problem:

> Some unknown probability distribution, over some large (possibly infinite) domain set, generates an i.i.d. sample. Upon observing such a sample, a learner wishes to generate some simple, yet meaningful, description of the underlying distribution.

The above scenario can be viewed as a high level definition of *unsupervised learning*. Many well established statistical tasks, such as Linear Regression, Principle Component Analysis and Principal Curves, can be viewed in this light. In this work, we restrict our attention to *clustering* tasks. That is, the description that the learner outputs is in the form of a finite collection of subsets (or a

** This work is supported in part by the Multidisciplinary University Research Initiative (MURI) under the Office of Naval Research Contract N00014-00-1-0564.

partition) of the domain set. As a measure of the quality of the output of the clustering algorithm, we consider objective functions defined over the underlying domain set and distribution.

This formalization is relevant to many realistic scenarios, in which it is natural to assume that the information we collect is only a sample of a larger body which is our object of interest. One such example is the problem of Quantizer Design [2] in coding theory, where one has to pick a small number of vectors, 'code words', to best represent the transmission of some unknown random source.

Results in this general framework can be applied to the worst-case model of clustering as well, and in some cases, yield significant improvements to the best previously known complexity upper bounds. We elaborate on this application in the subsection on *worst-case complexity view* below.

The paradigm that we analyze is the simplest sampling-based meta-algorithm. Namely,

1. `Draw an i.i.d random sample of the underlying probability distribution.`
2. `Find a good clustering of the sample.`
3. `Extend the clustering of the sample to a clustering of the full domain set.`

A key issue in translating the above paradigm into a concrete algorithm is the implementation of step 3; How should a clustering of a subset be extended to a clustering of a full set? For clusterings defined by a choice of a fixed number if centers, like the K median problem and vector quantization, there is a straightforward answer; namely, use the cluster centers that the algorithm found for the sample, as the cluster centers for the full set. While there are ways to extend clusterings of subsets for other types of clustering, in this paper we focus on the K-median and vector quantization problems.

The focus of this paper is an analysis of the approximation quality of sampling based clustering. We set the ground for a systematic discussion of this issue in the general context of statistical clustering, and demonstrate the usefulness of our approach by considering the concrete case of K-median clustering.

We prove that certain properties of clustering objective functions suffice to guarantee that an implicit description of an almost optimal clustering can be found in time depending on the confidence and accuracy parameters of the approximation, but independent of the input size. We show that the K-median clustering objective function, as well as the vector quantization cost, enjoy these properties. We are therefore able to demonstrate the first known constant-time approximation algorithm for the K-median problem.

The paradigm outlined above has been considered in previous work in the context of sampling based approximate clustering. Buhmann [3] describes a similar meta-algorithm under the title "Empirical Risk Approximation". Buhmann suggests to add an intermediate step of averaging over a set of empirically good clusterings, before extending the result to the full data set. Such a step helps reduce the variance of the output clustering. However, Buhmann's analysis is under the assumption that the data- generating distribution is *known* to the

learner. We address the distribution free (or, worst case) scenario, where the only information available to the learner is the input sample and the underlying metric space.

Our main technical tool is a uniform convergence result that upper bounds, as a function of the sample sizes, the discrepancy between the empirical cost of certain families of clusterings to their true cost (as defined by the underlying probability distribution). Convergence results of the empirical estimates of the k-median cost of clusterings where previously obtained for the limiting behavior, as sample sizes go to infinity (see, e.g. Pollard [6]). Finite-sample convergence bounds where obtained for the k-median problem by Mishra et al [5], and for the vector quantization problem by Bartlett et al [2], which also provide a discussion of vector quantization in the context of coding theory see [2]. Smola et al [7] provide a framework for more general quantization problems, as well as convergence results for a regularized versions of these problems. However, the families of cluster centers that our method covers are much richer than the families of centers considered in these papers.

1.1 Worst-Case Complexity view

Recently there is a growing interest in sampling based algorithms for approximating NP-hard clustering problems (see, e.g, Mishra et al [5], de la Vega et al [8] and Meyerson et al [4]). In these problems, the input to an algorithm is a finite set X in a metric space, and the task is to come up with a clustering of X that minimizes some objective function. The sampling based algorithm performs this task by considering a relatively small $S \subseteq X$ that is sampled uniformly at random from X, and applying a (deterministic) clustering algorithm to S. The motivating idea behind such an algorithm is the hope that relatively small sample sizes may suffice to induce good clusterings, and thus result in computational efficiency. In these works one usually assumes that a point can be sampled uniformly at random over X in constant time. Consequently, using this approach, the running time of such algorithms is reduced to a function of the size of the sample (rather than of the full input set X) and the computational complexity analysis boils down to the statistical analysis of sufficient sample sizes.

The analysis of the model proposed here is relevant to these settings too. By taking the underlying distribution to be the uniform distribution over the input set X, results that hold for our general scenario readily apply to the sampling based approximate clustering as well.

The worst case complexity of sampling based K-median clustering is addressed in Mishra et al [5] where such an algorithm is shown to achieve a sublinear upper bound on the computational complexity for the approximate K-median problem. They prove their result by showing that with high probability, a sample of size $O\left(\left(\frac{k\log(n)}{\epsilon^2}\right)\right.$ suffices to achieve a clustering with average cost (over all the input points) of at most $2Opt + \epsilon$ (where Opt is the average cost of an optimal k clustering). By proving a stronger upper bound on sufficient sample sizes, we are able to improve these results. We prove upper bounds on

the sufficient sample sizes (and consequently on the computational complexity) that are independent of the input size n.

2 The Formal Setup

We start by providing a definition of our notions of a *statistical clustering problem*. Then, in the "basic tool box' subsection, we define the central tool for this work, the notion of a *clustering description scheme*, as well as the properties of these notions that are required for the performance analysis of our algorithm. Since the generic example that this paper addresses is that of K-median clustering, we shall follow each definition with its concrete manifestation for the K-median problem.

Our definition of clustering problems is in the spirit of combinatorial optimization. That is, we consider problems in which the quality of a solution (i.e. clustering) is defined in terms of a precise objective function. One should note that often, in practical applications of clustering, there is no such well defined objective function, and many useful clustering algorithms cannot be cast in such terms.

Definition 1 (Statistical clustering problems).

- *A clustering problem is defined by a triple $(X, \mathcal{T}, R)$, where X is some domain set (possibly infinite), $\mathcal{T}$ is a set of legal clusterings (or partitions) of X, and $R : \mathcal{P} \times \mathcal{T} \mapsto [0, 1]$ is the objective function (or risk) the clustering algorithm aims to minimize, where $\mathcal{P}$ is a set of probability distributions over X* [1].
- *For a finite $S \subseteq X$, the* empirical risk *of a clustering T on a sample S, $R(S, T)$, is the risk of the clustering T with respect to the uniform distribution over S.*
- For the K-median problem, the domain set X is endowed with a metric d and $\mathcal{T}$ is the set of all k-cell Voronoi diagrams over X that have points of X as centers. Clearly each $T \in \mathcal{T}$ is determined by a set $\{x_1^T, \ldots x_k^T\} \subseteq X$, consisting of the cell's centers. Finally, for a probability distribution P over X, and $T \in \mathcal{T}$, $R(P, T) = \mathrm{E}_{y \in P} \left(\min_{i \in \{1,\ldots k\}} d(y, x_i^T) \right)$. That is, the risk of a partition defined by a set of k centers is the expected distance of a P-random point from its closest center.

Note that we have restricted the range of the risk function, R to the unit interval. This corresponds to assuming that, for the K-median and vector quantization problems, the data points are all in the unit ball . This restriction allows simpler formulas for the convergence bounds that we derive. Alternatively, one

[1] In this paper, we shall always take $\mathcal{P}$ to be the class of all probability distributions over the domain set, therefore we do not specify it explicitly in our notation. There are cases in which one may wish to consider only a restricted set of distributions (e.g., distributions that are uniform over some finite subset of X) and such a restriction may allow for sharper sample size bounds.

could assume that the metric spaces are bounded by some constant and adjust the bounds accordingly. On the other extreme, if one allows unbounded metrics, then it is easy to construct examples for which, for any given sample size, the empirical estimates are arbitrarily off the true cost of a clustering.

Having defined the setting for the problems we wish to investigate, we move on to introduce the corresponding notion of desirable solution. The definition of a clustering problem being 'approximable from samples' resembles the definition of learnability for classification tasks.

Definition 2 (Approximable from samples). *A clustering problem* $(X, \mathcal{T}, R)$ *is* α - approximable from samples*, for some* $\alpha \geq 1$*, if there exist an algorithm* $\mathcal{A}$ *mapping finite subsets of* X *to clusterings in* $\mathcal{T}$*, and a function* $f : (0,1)^2 \mapsto \mathbb{N}$*, such that for every probability distribution* P *over* X *and every* ϵ*,* $\delta \in (0,1)$*, if a sample* S *of size* $\geq f(\epsilon, \delta)$ *is generated i.i.d. by* P *then with probability exceeding* $1 - \delta$*,*

$$R(P, \mathcal{A}(S)) \leq \min_{T \in \mathcal{T}} \alpha R(P, T) + \epsilon$$

.

Note that formally, the above definition is trivially met for any fixed finite size domain X. We have in mind the setting where X is some infinite universal domain, and one can embed in it finite domains of interest by choosing the underlying distribution P so that it has that set of interest as its support. Alternatively, one could consider a definition in which the clustering problem is defined by a scheme $\{(X_n, \mathcal{T}_n, R_n)\}_{n \in \mathbb{N}}$ and require that the sample size function $f(\epsilon, \delta)$ is independent of n.

2.1 Our basic tool box

Next, we define our notion of an implicit representation of a clustering. We call it a *clustering description scheme*. Such a scheme can be thought of as a compact representation of clusterings in terms of sets of l elements of X, and maybe some additional parameters.

Definition 3 (Clustering description scheme).
Let $(X, \mathcal{T}, R)$ *be a clustering problem. An* (l, I) clustering description scheme *for* $(X, \mathcal{T}, R)$ *is a function,* $G : X^l \times I \mapsto \mathcal{T}$*, where* l *is the number of points a description depends on, and* I *is a set of possible values for an extra parameter.*

We shall consider three properties of description schemes. The first two can, in most cases, be readily checked from the definition of a description scheme. The third property has a statistical nature, which makes it harder to check. We shall first introduce the first two properties, *completeness* and *localization*, and discuss some of their consequences. The third property, *coverage*, will be discussed in Section 3 .

Completeness: A description scheme, G, is *Complete* for a clustering problem $(X, \mathcal{T}, R)$, if for every $T \in \mathcal{T}$ there exist $x_1, \ldots x_l \in X$ and $i \in I$ such that $G(x_1, \ldots x_l, i) = T$.

Localization: A description scheme, G, is *Local* for a clustering problem $(X, \mathcal{T}, R)$, if there exist a functions $f : X^{l+1} \times I \mapsto \mathbb{R}$ such that for any probability distribution P, for all $x_1, \ldots x_l \in X$ and $i \in I$,

$$R(P, G(x_1, \ldots x_l, i)) = E_{y \in P} f(y, x_1, \ldots x_l, i)$$

Examples:

The K-median problem endowed with the natural description scheme: in this case,$l = k$ (the number of clusters), there is no extra parameter i, and $G(x_1, \ldots, x_k)$ is the clustering assigning any point $y \in X$ its closest neighbor among $\{x_1, \ldots, x_k\}$. So, given a clustering T, if $\{x_1^T, \ldots x_k^T\}$ are the centers of T's clusters, then $T = G(x_1^T, \ldots, x_k^T)$. Clearly, this is a complete and local description scheme (with $f(y, x_1 \ldots, x_k) = \min_{i \in \{1, \ldots k\}} d(y, x_i)$ and F being the identity function).

Vector Quantization: this problem arises in the context of source coding. The problem is very similar to the K-median problem. The domain X is the Euclidean space $\mathbb{R}^d$, for some d, and one is given a fixed parameter l. On an input set of d-dimensional vectors, one wishes to pick 'code points' $(x_1, \ldots x_l) \in \mathbb{R}^d$ and map each input point to one of these code points. The only difference between this and the K-median problem is the objective function that one aims to minimize. Here it is $R(P, T_{x_1, \ldots x_l}) = \mathrm{E}_{y \in P} \left[\min_{i \in \{1, \ldots l\}} d(y, x_i)^2\right]$. The natural description scheme in this case is the same one as in the K-median problem - describe a quantizer T by the set of code point (or centers) it uses. It is clear that, in this case as well, the description scheme is both complete and local.

Note, that in both the K-median clustering and the vector quantization task, once such an implicit representation of the clustering is available, the cluster to which any given domain point is assigned can be found from the description in constant time (a point y is assigned to the cluster whose index is $Argmin_{i \in \{1, \ldots k\}} d(y, x_i)$).

The next claim addresses the cost function. Let us fix a sample size m. Given a probability distribution P over our domain space, let P^m be the distribution over i.i.d. m- samples induced by P. For a random variable $f(S)$, let $E_{S \in P^m}(f)$ denote the expectation of f over this distribution.

Claim 1 *Let $(X, \mathcal{T}, R)$ be a clustering problem. For $T \in \mathcal{T}$, if there exists a function $h_T : X \mapsto \mathbb{R}^+$ such that for any probability distribution P, $R(P, T) = E_{x \in P}(h_T(x))$, then for every such P and every integer m,*

$$E_{S \in P^m}(R(S, T)) = R(P, T)$$

Corollary 2 *If a clustering problem $(X, \mathcal{T}, R)$ has a local and complete description scheme then, for every probability distribution P over X, every $m \geq 1$ and every $T \in \mathcal{T}$,*

$$E_{S \in P^m}(R(S, T)) = R(P, T)$$

Lemma 1. *If a clustering problem $(X, \mathcal{T}, R)$ has a local and complete description scheme then, for every probability distribution P over X, every $m \geq 1$ and every $T \in \mathcal{T}$,*

$$P^m\{|R(P,T) - R(S,T)| \geq \epsilon\} \leq 2e^{-2\epsilon^2 m}$$

The proof of this Lemma is a straightforward application of Hoeffding inequality to the above corollary (recall that we consider the case where the risk R is in the range $[0,1]$).

Corollary 3 *If a clustering problem $(X, \mathcal{T}, R)$ has a local and complete description scheme then, for every probability distribution P over X, and every clustering $T \in \mathcal{T}$, if a sample $S \subseteq X$ of size $m \geq \frac{\ln 2/\delta}{2\epsilon^2}$ is picked i.i.d. via P then, with probability $> 1 - \delta$ (over the choice of S),*

$$|R(S,T) - R(P,T)| \leq \epsilon$$

In fact, the proofs of the sample-based approximation results in this paper require only the one-sided inequality, $R(S,T) \leq R(P,T) + \epsilon$.

So far, we have not really needed description schemes. In the next theorem, claiming that the convergence of sample clustering costs to the true probability costs, we heavily rely on the finite nature of description schemes. Indeed, clustering description schemes play a role similar to that played by compression schemes in classification learning.

Theorem 4. *Let G be a local description scheme for a clustering problem $(X, \mathcal{T}, R)$. Then for every probability distribution P over X, if a sample $S \subseteq X$ of size $m >> l$ is picked i.i.d. by P then, with probability $> 1 - \delta$ (over the choice of S), for every $x_1, \ldots x_l \in S$ and every $i \in I$,*

$$|R(S, G(x_1, \ldots x_l, i)) - R(P, G(x_1, \ldots x_l, i))| \leq \sqrt{\frac{\ln(|I|) + l \ln m + \ln(1/\delta)}{2(m-l)}}$$

Proof. Corollary 3 implies that for every clustering of the form $G(x_1, \ldots x_l, i)$, if a large enough sample S is picked i.i.d. by P, then with high probability, the empirical risk of this clustering over S is close to its true risk. It remains to show that, with high probability, for S sampled as above, this conclusion holds simultaneously for all choices of $x_1, \ldots x_l \in S$ and all $i \in I$.

To prove this claim we employ the following uniform convergence result:

Lemma 2. *Given a family of clusterings $\{G(x_1, \ldots x_l, i)\}_{x_1,\ldots \in X \ i \in I}$, let $\epsilon(m, \delta)$ be a function such that, for every choice of $x_1, \ldots x_l, i$ and every choice of m and $\delta > 0$, if a sample S is picked by choosing i.i.d uniformly over X, m times, then with probability $\geq 1 - \delta$*

$$|R(S, G(x_1, \ldots x_l, i)) - R(P, G(x_1, \ldots x_l, i))| < \epsilon(m, \delta)$$

then, with probability $\geq 1-\delta$ *over the choice of* S*,*
$\forall x_1, \ldots x_l \in S \ \forall i \in I$*,*

$$|R(S, G(x_1, \ldots x_l, i)) - R(P, G(x_1, \ldots x_l, i))| < \epsilon(m-l, \frac{\delta}{|I| \times \binom{m}{l}})$$

One should note that the point of this lemma is the change of order of quantification. While in the assumption one first fixes $x_1, \ldots x_l, i$ and then randomly picks the samples S, in the conclusion we wish to have a claim that allows to pick S first and then guarantee that, no matter which $x_1, \ldots x_l, i$ is chosen, the S-cost of the clustering is close to its true P-cost. Since such a strong statement is too much to hope for, we invoke the sample compression idea, and restrict the choice of the x_i's by requiring that they are members of the sample S.

Proof (Sketch). The proof follows the lines of the uniform convergence results for sample compression bounds for classification learning. Given a sample S of size m, for every choice of l indices, $i_1, \ldots, i_l \in \{1, \ldots, m\}$, and $i \in I$, we use the bound of Corollary 3 to bound the difference between the empirical and true risk of the clustering $G(x_1, \ldots x_l, i)$. We then apply the union bound to 'uniformize' over all possible such choices.

In fact, the one-sided inequality,

$$R(P, G(x_1, \ldots x_l, i)) \leq R(S, G(x_1, \ldots x_l, i)) + \epsilon$$

suffices for proving the sample-based approximation results of this paper.

3 Sample based approximation results for clustering in the general setting

Next we apply the convergence results of the previous section to obtain guarantees on the approximation quality of sample based clustering. Before we can do that, we have to address yet another component of our paradigm. The convergence results that we have so far suffice to show that the empirical risk of a description scheme clustering that is based on sample points is close to its true risk. However, there may be cases in which any such clustering fails to approximate the optimal clustering of a given input sample. To guard against such cases, we introduce our third property of clustering description schemes, the *coverage* property.

The Coverage property: We consider two versions of this property:

Multiplicative coverage: A description scheme is α*-m-covering* for a clustering problem $(X, \mathcal{T}, R)$ if for every $S \subset X$ s.t. $|S| \geq l$, there exist $\{x_1, \ldots x_l\} \subseteq S$ and $i \in I$ such that for every $T \in \mathcal{T}_X$,

$$R(S, G(x_1, \ldots x_l, i)) \leq \alpha R(S, T)$$

Namely, an optimal clustering of S can be α-approximated by applying the description scheme G to an l-tuple of members of S.

Additive coverage: A description scheme is *η-a-covering* for a clustering problem $(X, \mathcal{T}, R)$ if for every $S \subset X$ s.t. $|S| \geq l$, there exist $\{x_1, \ldots x_l\} \subseteq S$ and $i \in I$ such that for every $T \in \mathcal{T}_X$,

$$R(S, G(x_1, \ldots x_l, i)) \leq R(S, T) + \eta$$

Namely, an optimal clustering of S can be approximated to within (additive) η by applying the description scheme G to an l-tuple of members of S.

We are now ready to prove our central result. We formulate it for the case of multiplicative covering schemes. However, it is straightforward to obtain an analogous result for additive coverage.

Theorem 5. *Let $(X, \mathcal{T}, R)$ be a clustering problem that has a local and complete description scheme which is α-m-covering, for some $\alpha \geq 1$. Then $(X, \mathcal{P}, R)$ is α-approximable from samples.*

Proof. Let $m = O\left(\frac{\ln\left(\frac{|I|}{\delta\epsilon^l}\right)}{\epsilon^2}\right)$. Let $T^* \in \mathcal{T}$ be a clustering of X that minimizes $R(P, T)$, and let $S \subset X$ be an i.i.d. P-random sample of size m.

Now, with probability $\geq 1 - \delta$, S satisfies the following chain of inequalities:

- By Corollary 3,
$$R(P, T^*) + \epsilon \geq R(S, T^*)$$
- Let $Opt(S)$ be a clustering of S that minimizes $R(S, T)$. Clearly,
$$R(S, T^*)) \geq R(S, (Opt(S))$$
- Since G is α covering, for some $x_1, \ldots x_l \in S$ and $i \in I$,
$$R(S, Opt(S)) \geq \frac{1}{\alpha} R(S, G(x_1, \ldots x_l, i))$$
- By Theorem 4, for the above choice of $x_1 \ldots x_l, i$,
$$R(S, G(x_1, \ldots x_l, i)) \geq R(P, G(x_1, \ldots x_l, i)) - \epsilon$$

It therefore follows that

$$R(P, G(x_1, \ldots x_l, i)) \leq \alpha(R(P, T^*) + \epsilon) + \epsilon$$

□

Theorem 6. *Let $(X, \mathcal{T}, R)$ be a clustering problem and let $G(x_1 \ldots, x_l, i)$ be a local and complete description scheme which is η-a-covering, for some $\eta \in [0, 1]$. Then for every probability distribution P over X and $m >> l$, if a sample, S, of size m is generated i.i.d by P, then with probability exceeding $1 - \delta$,*

$$\min\{R(P, G(x_1, \ldots, x_l, i)) : \ x_1, \ldots, x_l \in S, \ i \in I\} \leq$$

$$\min\{R(P, T) : T \in \mathcal{T}\} + \eta + \sqrt{\frac{\ln(|I|) + l \ln m + \ln(1/\delta)}{2(m - l)}}$$

The proof is similar to the proof of Theorem 5 above.

4 *K*-Median Clustering and Vector Quantization

In this section we show how to apply our general results to the specific cases of K-median clustering and vector quantization. We have already discussed the natural clustering description schemes for these cases, and argued that they are both complete and local. The only missing component is therefore the analysis of the *coverage* properties of these description schemes.

We consider two cases,

Metric K-median problem where X can be any metric space.

Euclidean K-median where X is assumed to be a Euclidean space $\mathbb{R}^d$. This is also the context for the vector quantization problem.

In the first case there is no extra structure on the underlying domain metric space, whereas in the second we assume that it is a Euclidean space (it turns out that the assumption that the domain a Hilbert space suffices for our results).

For the case of general metric spaces, we let $G(x_1, \ldots, x_k)$ be the basic description scheme that assigns each point y to the x_i closest to it. (So, in this case we do not use the extra parameter i).

It is well known, (see e.g., [5]) that for any sample S, the best clustering with center points from S is at most a factor of 2 away from the optimal clustering for S (when centers can be any points in the underlying metric space). We therefore get that is that case G is a 2-m-covering.

For the case of Euclidean, or Hilbert space domain, we can also employ a richer description scheme. For a parameter t, we wish to consider clustering centers that are the centers of mass of t-tuples of sample points (rather than just the sample points themselves). Fixing parameters t and r, let our index set I be r^{tk}, that is, the set of all vectors of length k whose entries are t-tuples of indices in $\{1, \ldots, r\}$. Let $G_t(x_1, \ldots, x_r, i) \stackrel{\triangle}{=} G(x_{i,1}, \ldots x_{i,tk})$, where $i \in r^{tk}$ indexes a sequence $(x_{i,1}, \ldots x_{i,kt})$ of points in $\{x_1, \ldots, x_r\}$, and $G(x_{i,1}, \ldots x_{i,tl})$ is the clustering defined by the set of centers $\{1/t\Sigma_{j=th+1}^{t(h+1)} x_{i,j} : h \in \{0, \ldots, k-1\}\}$. That is, we take the 'centers of mass' of t tuples of points of S, where i is the index of the sequence of kt points that defines or centers. It is easy to see that such G_t is complete iff $r \geq k$.

The following lemma of Maurey, [1], implies that, for $t \leq r$, this description scheme enjoys an η-a-coverage, for $\eta = 1/\sqrt{t}$.

Theorem 7 (Maurey, [1]). *Let F be a vector space with a scalar product $(\cdot, \cdot)$ and let $\|f\| \stackrel{\triangle}{=} \sqrt{(f, f)}$ be the induced norm on F. Suppose $G \subseteq F$ and that, for some $c > 0$, $\|g\| \leq c$ for all $g \in G$. Then for all f from the convex hull of G and all $k \geq 1$ the following holds:*

$$\inf_{g_1, \ldots, g_k \in G} \left\| \frac{1}{k} \sum_{i=1}^{k} g_i - f \right\| \leq \sqrt{\frac{c^2 - \|f\|^2}{k}}.$$

Corollary 8 *Consider the K median problem over a Hilbert space, X. For every t and $r \geq \max\{k, t\}$, the clustering algorithm that, on a sample S, outputs $Argmin\{R(G_t(x_1, \ldots, x_r, i)) : x_1, \ldots, x_r \in S, \text{ and } i \leq r^{tk}\}$ produces, with probability exceeding $1 - \delta$ a clustering whose cost is no more then*

$$\frac{1}{\sqrt{t}} + \sqrt{\frac{k(t \ln r + \ln |S|) + \ln(1/\delta)}{2(|S| - r)}}$$

above the cost of the optimal k-centers clustering of the sample generating distribution (for any sample generating distribution and any $\delta > 0$).

4.1 Implications to worst case complexity

As we mentioned earlier, worst case complexity models of clustering can be naturally viewed as a special case of the statistical clustering framework. The computational model in which there is access to random uniform sampling from a finite input set, can be viewed as a statistical clustering problem with P being the uniform distribution over that input set.

Let (X, d) be a metric space, $\mathcal{T}$ a set of legal clusterings of X and R an objective function. A worst case sampling-based clustering algorithm for $(X, \mathcal{T}, R)$ is an algorithm that gets as input finite subsets $Y \subseteq X$, has access to uniform random sampling over Y, and outputs a clustering of Y.

Corollary 9 *Let $(X, \mathcal{T}, R)$ be a clustering problem. If, for some $\alpha \geq 1$, there exist a clustering description scheme for $(X, \mathcal{T}, R)$ which is both complete and α-m-covering, then there exists a worst case sampling-based clustering algorithm for $(X, \mathcal{T}, R)$ that runs in constant time depending only of the approximation and confidence parameters, ϵ and δ (and independent of the input size $|Y|$) and outputs an $\alpha Opt + \epsilon$ approximations of the optimal clustering for Y, with probability exceeding $1 - \delta$.*

Note that the output of such an algorithm is an *implicit* description of a clustering of Y. It outputs the parameters from which the description scheme determines. For natural description schemes (such as describing a Voronoi diagram by listing its center points) the computation needed to figure out the cluster membership of any given $y \in Y$ requires constant time.

Acknowledgments:

I would like to express warm thanks to Aharon Bar-Hillel for insightful discussions that paved the way to this research.

References

1. Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.
2. Peter Bartlett, Tamas Linder and gabor Lugosi "the minimax distortion Redundancy in empirical Quantizer Design" *IEEE Transactions on Information theory*, vol. 44, 1802–1813, 1998.
3. Joachim Buhmann, "Empirical Risk Approximation: An Induction Principle for Unsupervised Learning" Technical Report IAI-TR-98-3, Institut for Informatik III, Universitat Bonn. 1998.
4. Adam Meyerson, Liadan O'Callaghan, and Serge Plotkin "A k-median Algorithm with Running Time Independent of Data Size" *Journal of Machine Learning*, Special Issue on Theoretical Advances in Data Clustering (MLJ) 2004.
5. Nina Mishra, Dan Oblinger and Leonard Pitt "Sublinear Time Approximate Clustering" in *Proceedings of Syposium on Discrete Algorithms, SODA 2001* pp. 439-447.
6. D. Pollard "Quantization and the method of k-means" in *IEEE Transactions on Information theory* 28:199-205, 1982.
7. Alex J. Smola, Sebastian Mika, and Bernhard Scholkopf "Quantization Finctionals and Regularized Principal Manifolds" *NeuroCOLT Technical Report Series* NC2-TR-1998-028.
8. Fernandes de la Vega, Marek Karpinski, Calire Kenyon and Yuval Rabani "Approximation Schemes for Clustering Problems" *Proceedings of Symposium on the Theory of computation, STOC'03*, 2003.