A framework for statistical clustering with constant time approximation algorithms for *K*-median and *K*-means clustering

Shai Ben-David

Abstract We consider a framework of *sample-based clustering*. In this setting, the input to a clustering algorithm is a sample generated i.i.d by some unknown arbitrary distribution. Based on such a sample, the algorithm 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 clustering algorithms that approximate the optimal clustering. We show that the *K*-median clustering, as well as *K*-means and the Vector Quantization problems, satisfy these conditions. Our results apply to the combinatorial optimization setting where, assuming that sampling uniformly over an input set can be done in constant time, we get a sampling-based algorithm for the *K*-median and *K*-means clustering problems 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 dependence of the running time of our algorithm on the Euclidean dimension is only linear. Our main technical tool is a uniform convergence result for center based clustering that can be viewed as showing that the effective VC-dimension of k-center clustering equals k.

S. Ben-David (✉)
School of Computer Science, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada
e-mail: shai@ece.cornell.edu

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, Principal 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 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 (Bartlett, Linder and Lugosi, 1998) 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 combinatorial optimization model of clustering as well, and in some cases, yield significant improvements to the best previously known worst-case 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 center-based clusterings, clusterings defined by a choice of a fixed number of centers like the K median problem and vector quantization (or K means), 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, K-mean and vector quantization problems.

The paradigm outlined above has been considered in previous work in the context of sampling based approximate clustering. Buhmann (1998) 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.

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 cases mentioned above.

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 and K means clustering objective functions, as well as the vector quantization cost, enjoy these properties.

Our main technical tool is a uniform convergence result that upper bounds the discrepancy between the empirical cost of clusterings, having their centers in a random sample, to their true cost (as defined by the underlying probability distribution). Convergence results of the empirical estimates of the k-median cost of clusterings were previously obtained for the limiting behavior, as sample sizes go to infinity (see, e.g. Pollard, 1982). Finite-sample convergence bounds where obtained for the k-median problem by Mishra, Oblinger and Pitt (2001), and for the vector quantization problem by Bartlett, Linder and Lugosi (1998), which also provide a discussion of vector quantization in the context of coding theory. Smola, Mika and Scholkopf (1998) 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. Consequently, we are able to prove that sample sizes that are independent of the domain cardinality suffice for obtaining constant-factor approximations to optimal center-based clusterings. Loosely speaking, one can view our results (e.g., Theorems 4 and 5) as showing that the effective VC-dimension of k-center clustering equals k.

1.1 Combinatorial optimization view

Recently there is a growing interest in sampling based algorithms for approximating NP-hard clustering problems (see, e.g, Mishra, Oblinger and Pitt, 2001; de la Vega et al., 2003; Meyerson, O'Callaghan and Plotkin, 2004; Czumaj and Sohler, 2004). 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. A 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, Oblinger and Pitt (2001) where such an algorithm is shown to achieve a sub-linear 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 $O(\frac{k \log(n)}{\epsilon^2})$ suffices to find a set of centers resulting in a clustering with cost (which is defined as the average distance of points to their nearest center) over all the input points) of at most $2Opt + \epsilon$ (where Opt is the 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. In particular, we get an algorithm based on an $\widetilde{O}(\frac{k}{\epsilon^2})$ sample that computes a clustering having cost which is at most ϵ away from a constant approximation for any of the K-means, K-median and Vector Quantization problems. Those sample sizes allow a randomized sampling-based algorithm to find the centers of such clusterings in time quadratic in the sample size, and thus independent of the domain cardinality.

Note that this work considers that task of finding $\alpha Opt + \epsilon$ approximations. In contrast, much of the work on the combinatorial approximation aspects of clustering aim to find constant factor approximations (without the additive ϵ relaxation). In that context, Mettu and Plaxton (2004) prove lower bounds of $\Omega(nk)$ (where n is the input domain cardinality) on the running time of any randomized constant factor approximation algorithm for the K-median problem. The difference between the two versions is mainly due to the fact that the additive ϵ allows us to ignore cases where the domain points are arbitrarily close to each other. Thus, assuming that our input is contained in a bounded set, we can ignore the ratio between the smallest and largest input distances, a ratio that plays a crucial role in the Mettu-Plaxton lower bound.

Meyerson, O'Callaghan and Plotkin (2004) do achieve constant factor approximate clustering in time independent of n, however, they get their results only under the assumption that each cluster is sufficiently large.

The closest results to the implications of our results to the combinatorial optimization setting are similar time complexity upper bounds claimed (without proof) in Czumaj and Sohler (2004). These results were published at the same time and independently of ours (Ben-David, 2004) and apply a different proof technique.

2 The formal setup

We start by providing a definition of our notions of a *statistical clustering problem*. Then, in the "basic tool box" Section 2.2, 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 realization 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 quadruple $(X, \mathcal{T}, R, \mathcal{P})$, 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, and $\mathcal{P}$ is a set of probability distributions over X.[1]

---

[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

–   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.

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 ϵ, $\delta \in (0, 1)$, for every probability distribution P over X and $m \geq f(\epsilon, \delta)$, if an m size sample S is generated i.i.d. by P then with probability exceeding $1 - \delta$,

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

In the combinatorial optimization setting, it is common to consider objective functions that depend on the data set X rather than on a probability distribution over it. That setting can be viewed as a particular case of our framework, obtained by considering only probability distributions that are uniform over some finite domain subset. Alternatively, one could consider a definition in which the clustering problem is defined by a scheme $\{(X_n, \mathcal{T}_n, R_n)\}_{n \in \mathbb{N}}$ (where X_n is an n-size domain) and require that the sample size function $f(\epsilon, \delta)$ is independent of n.

2.1 Concrete clustering problems

The generic examples that we apply our analysis to are center based clusterings, in particular the K-Median, K-means and Vector Quantization problems.

Definition 3 (Center-based clustering problems). Given some domain set X endowed with a metric d and a parameter $k \in \mathbb{N}$, a clustering of X is defined by choosing k points in X as cluster centers and then clustering the points of the domain according to their nearest center points. Formally, $\mathcal{T}$ 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 $\{x_1^T, \dots x_k^T\} \subseteq X$, consisting of the clusters centers.

Center based clustering problems differ by the objective function that each aims to minimize (or maximize). In this paper we shall consider the following popular variants:

The K-median problem: For a probability distribution P over X, and $T \in \mathcal{T}$,

$$R(P, T) = \mathrm{E}_{y \in P} \left(\min_{i \in \{1, \dots k\}} d(y, x_i^T) \right)$$

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.

---

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.

The K-means problem: The aim here is to minimize the sum of squared distances of points to their nearest center. Namely,

$$R(P, T) = \mathrm{E}_{y \in P}\left(\min_{i \in \{1, \dots k\}} d\left(y, x_i^T\right)^2\right)$$

Vector quantization: This problem arises in the context of source coding. On an input set of d-dimensional vectors, one wishes to pick 'code points' $(x_1, \dots x_k) \in \mathbb{R}^d$ and map each input point to one of these code points. The problem is similar to the K-means problem. However, the domain X is restricted to be the Euclidean space $\mathbb{R}^d$, for some d.

Note that we have restricted the range of the risk function, R to the unit interval. This corresponds to assuming that, for the K-center problems described above, the metric d is bounded by 1. This restriction allows simpler formulas for the convergence bounds that we derive. Alternatively, one 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 likely to be arbitrarily off the true cost of a clustering.

2.2 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 4 (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.

An obvious example of such a scheme is the representation of a center based clustering by the set of centers defining it. That is, $G(x_1, \dots, x_k) =$ *the k-clustering whose i'th cluster is* $\{y \in X : d(x_i, y) < d(x_j, y)$ *for all* $j \neq i\}$. (In such a description scheme there is no need to use the extra parameter, so I can be thought of as a singleton set.)

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 $x_1, \dots x_l \in X$ and $i \in I$ such that $G(x_1, \dots x_l, i) = T$.

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 $x_1, \dots x_l \in X$ and $i \in I$,

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

Clearly, for the center-based clustering problems we discussed above, the natural description scheme (describing a clustering by listing its center points) is both complete and local for all of these problems.

Note, that in all the center based clustering problems, 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 time $O(kt_d)$, where t_d denotes the time required to compute the distance between two given points (a point y is assigned to the cluster whose index is $Argmin_{i\in\{1,...k\}}d(y, x_i)$). This time is independent of the data set size.

Locality guarantees that the empirical cost of clusters defined by the description scheme is an unbiased estimation of their true cost. This is the content of the following claim.

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 and $T \in \mathcal{T}$. 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,

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

Proof: By the assumption,

$$R(S, T) = E_{x\in S}(h_T(x)) = \frac{1}{m}\Sigma_{x\in S}h_T(x)$$

therefore,

$$E_{S\in P^m}(R(S, T)) = E_{S\in P^m}\frac{1}{m}\Sigma_{x\in S}h_T(x) = E_{x\in P}h_T(x) = R(P, T)$$

$\square$

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 $m \gg l$ 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 $\{G(x_1, \ldots x_l, i)\}_{x_1, \ldots x_l \in X \ i \in I}$, 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,
   $\forall x_1, \ldots x_l \in S \ \forall i \in I$,

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

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. Note that there are only $|I| \times \binom{m}{l}$ such possible parameter 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: An (l, I)-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 $T \in \mathcal{T}_X$,

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

In particular, 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 $T \in \mathcal{T}_X$,

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

In particular, 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. The formulations of the result for the multiplicative and for the additive schemes, as well as the proofs, are quite similar.

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.*

Furthermore, if $G(x_1 \ldots, x_\ell, i)$ is such a description scheme then for every $\epsilon, \delta > 0$, if

$$m \geq \ell + \max\left\{3\frac{\ell}{\epsilon^2}\ln\left(\frac{\ell}{\epsilon^2}\right), \left(\frac{|I|}{\delta}\right)^{\frac{1}{\ell}}\right\}$$

then, for every probability distribution P (over the domain X), if

$$\mathcal{A}(S) = G\left(\text{argmin}\{R(S, G(x_1, \ldots x_l, i)) : x_1, \ldots x_l \in S, i \in I\}\right)$$

(Namely, $\mathcal{A}$ maps any finite sample, S, to the clustering that optimizes R over all the application of G to members of S.)

then

$$P^m\left[R(P, \mathcal{A}(S)) \leq \alpha R(P, Opt(P)) + \epsilon\right] \geq (1 - \delta)$$

(Where P^m is the probability induced by P over m-size i.i.d. P samples, and $Opt(P)$ is an R minimizing clustering w.r.t. P).

Proof: Let $m \geq \ell + \max\{3\frac{\ell}{\epsilon^2}\ln(\frac{\ell}{\epsilon^2}), (\frac{|I|}{\delta})^{\frac{1}{\ell}}\}$. Note that, for such an m,

$$\sqrt{\frac{\ln(|I|) + \ell\ln m + \ln(1/\delta)}{2(m - \ell)}} \leq \epsilon$$

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 $x_1, \ldots x_\ell \in S$ and $i \in I$,

$$R(S, Opt(S)) \geq \frac{1}{\alpha}R(S, G(x_1, \ldots x_\ell, i))$$

– By Theorem 4, for the above choice of $x_1 \ldots x_\ell, i$,

$$R(S, G(x_1, \ldots x_\ell, i)) \geq R(P, G(x_1, \ldots x_\ell, i)) - \epsilon$$

It therefore follows that

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

$\square$

Theorem 6. *Let $(X, \mathcal{T}, R)$ be a clustering problem and $G(x_1 \ldots, x_\ell, i)$ be a local and complete description scheme which is η-a-covering for $(X, \mathcal{T}, R)$, for some $\eta \in [0, 1]$. Then for every $\epsilon, \delta > 0$, if*

$$m \geq \ell + \max \left\{ 3\frac{\ell}{\epsilon^2} \ln \left(\frac{\ell}{\epsilon^2} \right), \left(\frac{|I|}{\delta} \right)^{\frac{1}{\ell}} \right\}$$

then, for every probability distribution P (over the domain X),

$$P^m \left[R(P, \mathcal{A}(S)) \leq R(P, Opt(P)) + \eta + \epsilon \right] \geq (1 - \delta)$$

(Where, as above, $\mathcal{A}$ maps any finite sample, S, to the clustering that optimizes R over all the application of G to members of S.)

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

4 Applications to K-median, K-means and vector quantization

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

We consider three cases,

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

Euclidean K-median and K-means where X is assumed to be a Euclidean space $\mathbb{R}^d$.

The vector quantization problem where X is a subset of $\mathbb{R}^d$ and, upon seeing a finite sample $S \subset X$, the algorithm has to pick the cluster centers from S.

In the first case there is no extra structure on the underlying domain metric space, whereas in the second and third cases we can exploit the assumption 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.)

The following claims are common knowledge in the clustering community:

Claim. Let X be a metric space and $S \subseteq X$ then,

1. For the K-median risk function, 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).

2. For the K-means risk function, the best clustering with center points from S is at most a factor of 4 away from the optimal clustering for S.

The claim follows by a straightforward application of the triangle inequality.

Corollary 1. *The k-center description scheme, G is 2-m-covering, in the K-median case, is a 4-m-covering scheme for the K-means problem, and it is 1-m-covering for vector quantization.*

Theorem 7. *There exist randomized approximation algorithms for the K-means, K-median and Vector Quantization problem whose running times are $O((\frac{k}{\epsilon^2})^k)$ that compute, with probability $(1 - \delta)$ (the δ enters the constant of the O notation), a clustering which is at most ϵ away from the optimal solution in the case of vector quantization, from a 2-approximation for K-median, and ϵ away from a 4-approximation for the K-means clustering task.*

Furthermore, given any α-approximation algorithm that runs in time $f(m, k)$ (where m is the size of the input data set), our algorithm can run in time $O(f(\frac{k}{\epsilon^2}))$, and output a 2α-approximation.

Proof: Apply Theorem 5 and the above Corollary to obtain the required approximation as the optimal sample based solution for a sample of size $m = O(\frac{k}{\epsilon^2})$, and then run an exhaustive search algorithm to find the optimum set of centers for such a sample.

For the second claim of the theorem, apply the given α-approximation algorithm to a similar sample.   □

For the case of Euclidean, or Hilbert space domain, we can also employ a richer description scheme that results in improved coverage factors. 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). On top of choosing t, we also choose the cardinality, r, of the set of al points of S that participate in these k t-tuples (so r is at most kt, and, for concreteness, one can assume that no point participates in the definition of more than one center, so $r = kt$). 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 $\{1, \ldots, r\}$. For $i \in \{1, \ldots, r^{tk}\}$, let $i(1), \ldots i(tk)$ be a sequence of indices from $\{1, \ldots, r\}$. Let $G_t(x_1 \ldots x_r, i)$ be the clustering defined by the set of centers $\{1/t \sum_{j=th+1}^{t(h+1)} x_{i(j)} : h \in \{0, \ldots, k - 1\}\}$. That is, we take the 'centers of mass' (actually, the averages) 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 $r \geq k$.

We now invoke a lemma of Maurey (Anthony and Bartlett, 1999), to derive better a-covering constants for the above extended description schemes.

Theorem 8 (Maurey (Anthony and Bartlett, 1999)). *Let X be a vector space with a scalar product $(\cdot, \cdot)$ and let $\|x\| \overset{\triangle}{=} \sqrt{(x, x)}$ be the induced norm on X. Suppose $S \subseteq X$ and that, for some $c > 0$, $\|s\| \leq c$ for all $s \in S$. Then for any x from the convex hull of S and any $t \geq 1$ the following holds:*

$$\inf_{s_1, \ldots, s_t \in S} \left\| \frac{1}{t} \sum_{i=1}^{t} s_i - x \right\| \leq \sqrt{\frac{c^2 - \|x\|^2}{t}}$$

Corollary 2. *For any $t \leq r$, the richer description scheme described above enjoys an η-a-coverage, for $\eta = M/\sqrt{t}$, where M is an upper bound on $d(s, s')$ for all $s, s' \in S$.*

The corollary follows by substituting $c = M$ in the Maurey Lemma. Note if we assume that the norm is bounded by 1, we can conclude that the richer description scheme is $\frac{1}{\sqrt{t}}$-a-covering.

Theorem 9. *Consider the K-median problem over a bounded subset of a Hilbert space, X. For every t, consider the clustering algorithm that, on a sample S, outputs*

$$\mathrm{Argmin}\{R(S, G_t(x_1, \ldots, x_{kt}, i)) : x_1, \ldots, x_{kt} \in S, \text{ and } i \leq (kt)^{kt}\}$$

Then, for $|S| \geq (\frac{(kt)^{kt}}{\delta})^{1/\ell}$, the algorithm outputs, with probability exceeding $1 - \delta$, a clustering whose cost is no more then

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

above the cost of the optimal k-centers clustering of the sample-generating distribution (for any sample generating distribution and any $\delta > 0$). Where M is the diameter of the domain set.

Proof: Theorem 6 implies the result once we show that the clustering description scheme G_t is local, complete and $M/\sqrt{t}$-a-covering. Recall that $G_t(x_1, \ldots, x_{kt}, i)$ defines a clustering by picking k subsequences, each of length t, from $x_1, \ldots, x_{kt}$, according to the kt sequence encoded by i, and uses the averages of these subsequences as the cluster centers. The definition of the cost function R for the K-median (as well as the K-means) clustering implies that this scheme is local. Its completeness follows by noting that it expands the simple k-center description scheme (for every k-tuple of cluster centers, the sequence that repeats each of these centers t times is one of the options encoded by the parameters in I). Finally, Corollary 2 implies that this scheme is $M/\sqrt{t}$-a-covering. $\square$

Corollary 3. *There exist a sample based clustering algorithm for the K-median problem that, on samples of size m achieves an additive approximation of*

$$\sqrt{\frac{k \ln(m) + \ln(1/\delta)}{\sqrt{m}}}$$

Proof: Just pick $t = \sqrt{|S|}$ and apply Theorem 9. $\square$

Note that similar results apply as well to Euclidean K-means clustering and to Vector Quantization.

4.1 Implications to the computational complexity of clustering

As we mentioned earlier, clustering tasks have been investigated as combinatorial optimization problems. A problem is defined by a cost function and a number of clusters k. The input of such a problem is a finite domain set with a metric (sometimes a subset of $\mathbb{R}^d$), and the goal of the algorithm is to output a k-clustering of the input domain that minimizes the cost. Many of the common clustering cost functions (such as K-median and K-means) give rise to NP-hard problems. There is a lot of interest in finding efficient approximation algorithms for such problems. One line of attack on these problems is to use randomized algorithms, allowing the algorithm to sample uniformly over the input set (see, e.g., Mishra, Oblinger and Pitt, 2001; de la Vega et al., 2003, Meyerson, O'Callaghan and Plotkin, 2004). 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.

The results of this work yield the first constant time approximation algorithms to both the K-median and the K-means problems.

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 4. *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.

5 Conclusions

The aim of this paper is to set forth a formal framework for analysis of sample-based clustering algorithms. We introduce a notion of clustering description schemes and show that it can be utilized to yield simple sample-based approximation algorithms that have proven finite sample uniform bounds on the rate of their convergence to optimal clusterings.

We offer three basic properties of description schemes, locality, completeness and coverage. We show that schemes that enjoy these properties yield proven guarantees on the quality of the clustering approximations generated by empirical risk minimization over these schemes. We complement these results by proving that the most common center-based clustering problems have such clustering description schemes. We can therefore rip the benefits of our analysis by proving that these simple algorithms provide efficient approximations to the K-median, and K-means (as well as Vector Quantization) clustering tasks. These algorithms provably converge to optimal clusterings with approximation bounds that depend

on the sample sizes and accuracy parameters, but are independent of the generating data distribution and the dimensionality of the data. As far as we know, no such bounds were previously published.

We hope that this work will stimulate further research into the theoretical foundations of statistical clustering. In particular, we are curious about analysis of other clustering algorithms with respect to the three properties of schemes discussed here.

While these properties sound quite basic and general, we also look forward to research into other basic principles that may yield useful clustering algorithms with provable performance guarantees.

References

Anthony, M., & Bartlett, P. L. (1999). Neural network learning: Theoretical foundations. Cambridge University Press.

Bartlett, P., Linder, T., & Lugosi, G. (1998). The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information Theory, 44*, 1802–1813.

Ben-David, S. (2004). A framework for statistical cluatering with a constant time approximation algorithm for K-median clustering. In *Proceedings of the 17th Annual Conference on Learning Theory, COLT'04*, Springer.

Buhmann, J. (1998). Empirical risk approximation: An induction principle for unsupervised learning. Technical Report IAI-TR-98-3, Institut for Informatik III, Universitat Bonn.

Czumaj, A., & Sohler, C. (2004). Sublinear-time approximation for clustering via random samples. In *Proceedings of the 31st International Colloquium on Automata, Language and Programming (ICALP'04),* LNCS 3142:396–407.

Mettu, R. R., & Plaxton, C. G. (2004). Optimal time bounds for approximate clustering. *Machine Learning, 56*, 35–60.

Meyerson, A., O'Callaghan, L., & Plotkin, S. (2004). A k-median algorithm with running time independent of data size. *Journal of Machine Learning*, Special Issue on Theoretical Advances in Data Clustering (MLJ).

Mishra, N., Oblinger, D., & Pitt, L. (2001). Sublinear time approximate clustering. In *Proceedings of Symposium on Discrete Algorithms, SODA*, (pp. 439–447).

Pollard, D. (1982). Quantization and the method of k-means. In *IEEE Transactions on Information Theory, 28*, 199–205.

Smola, A. J., Mika, S., & Scholkopf, B. (1998). Quantization functionals and regularized principal manifolds. Neuro COLT Technical Report Series NC2-TR-1998-028.

de la Vega, F., Karpinski, M., Kenyon, C., & Rabani, Y. (2003). Approximation schemes for clustering problems. In *Proceedings of Symposium on the Theory of Computation*, *STOC'03*.