Stability of k-Means Clustering

Shai Ben-David[1], Dávid Pál[1], and Hans Ulrich Simon[2*]

[1] David R. Cheriton School of Computer Science,
University of Waterloo,
Waterloo, Ontario, Canada
{shai,dpal}@cs.uwaterloo.ca
[2] Ruhr-Universität Bochum, Germany
simon@lmi.rub.de

Abstract. We consider the stability of k-means clustering problems. Clustering stability is a common heuristics used to determine the number of clusters in a wide variety of clustering applications. We continue the theoretical analysis of clustering stability by establishing a complete characterization of clustering stability in terms of the number of optimal solutions to the clustering optimization problem. Our results complement earlier work of Ben-David, von Luxburg and Pál, by settling the main problem left open there. Our analysis shows that, for probability distributions with finite support, the stability of k-means clusterings depends solely on the number of optimal solutions to the underlying optimization problem for the data distribution. These results challenge the common belief and practice that view stability as an indicator of the validity, or meaningfulness, of the choice of a clustering algorithm and number of clusters.

1 Introduction

Clustering is one of the most widely used techniques for exploratory data analysis. Across all disciplines, from social sciences over biology to computer science, people try to get a first intuition about their data by identifying meaningful groups among the data points. Despite this popularity of clustering, distressingly little is known about theoretical properties of clustering [11]. In particular, two central issues, the problem of assessing the meaningfulness of a certain cluster structure found in the data and the problem of choosing k—the number of clusters—which best fits a given data set are basically unsolved.

A common approach to provide answers to these questions has been the notion of *clustering stability*. The intuitive idea behind that method is that if we repeatedly sample data points and apply the clustering algorithm, then a "good" algorithm should produce clusterings that do not vary much from one sample to another. In other words, the algorithm is stable with respect to input

* This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

randomization. In particular, stability is viewed as an indication whether the model proposed by some algorithm fits the data or not. For example, if our data contains three "true" clusters, but we use a clustering algorithm which looks for four clusters, the algorithm wrongly needs to split one of the clusters into two clusters. Which of the three true clusters are split might change from sample to sample, and thus the result will not be very stable. Based on such intuitive considerations, stability is being widely used in practical applications as a heuristics for tuning parameters of clustering algorithms, like the number of clusters, or various stopping criteria, see for example [7], [4], [5], [6].

Aiming to provide theoretical foundations to such applications of stability, Ben-David et al. [3] have set forward formal definitions for stability and some related clustering notions and used this framework to embark on mathematical analysis of stability. Their results challenge these heuristics by showing that stability is determined by the structure of the set of optimal solutions to the risk minimization objective. They postulate that stability is fully determined by the number of distinct clusterings that minimize the risk objective function. They show that the existence of a unique minimizer implies stability. As for the reverse implication, they show that if the probability distribution generating the data has multiple minimizing clusterings, *and is symmetric with respect to these clusterings* then it is unstable. They conjecture that their symmetry condition is not necessary, and that the mere existence of multiple minimizers already implies instability. The main result in this paper is proving this conjecture for k-means clustering over finite-support probability distributions. We believe that our proofs, and therefore our main result, can be generalized to other risk minimization clustering problems.

These results indicate that, contrary to common belief and practice, stability may *not* reflect the validity or meaningfulness of the choice of the number of clusters. Instead, the parameters it measures are rather independent of clustering parameters. Our results reduce the problem of stability estimation to concrete geometric properties of the data distribution.

Using our characterization of stability, one can readily construct many example data distributions in which bad choices of the number of clusters result in stability while, on the other hand, domain partitions reflecting the true basic structure of a data set result in instability. As an illustration of these phenomena, consider the following simple data probability distribution P over the unit interval: For some large enough N, the support of P consists of $2N$ equally weighted points, N of which are equally spaced over the sub-interval $A = [0, 2a]$ and N points are equally spaced over the sub-interval $B = [1 - a, 1]$, where $a < 1/3$. First, let us consider k, the number of target centers, to be 2. It is not hard to see that for some value of $a < 1/3$ two partitions, one having the points in A as one cluster and the points in B as its second cluster, and the other having as one cluster only the points in some sub-interval $[0, 2a - \epsilon]$ and the remaining points as its second cluster, must have the same 2-means cost. It follows from our result that although 2 is the 'right' number of clusters for this distribution, the choice of $k = 2$ induces instability (note that the value of ϵ and a remain practically unchanged for all large enough N). On the other hand, if one considers, for the same value of a, $k = 3$, 3-means will have a unique minimizing solution (having the points in the intervals $[0, a]$, $[a, 2a]$ and $[1 - a, 1]$ as its clusters) and therefore be stable, leading the common heuristics to the conclusion that 3 is a good choice as the number of clusters for our data distributions (and, in particular, a better choice than 2). Note that, in this example, the data distribution is not symmetric, therefore, its instability for 2-means could not be detected by the previously known stability analysis.

The question of the practical value of stability as a clustering evaluation paradigm is intriguing and complex, we shall discuss it some more (without claiming to resolve it) in the Conclusion (Section 6 below).

The term "stability" is used for a variety of meanings in the clustering literature, not all of which are equivalent to our use of the term. In particular, note that the recent work of Rakhlin et al [10], considers a different notion of stability (examining the effect of replacing a small fraction of a clustering sample, as opposed to considering a pair of independent samples, as we do here). They investigate the relative size of a sub-sample that may be replaced without resulting in a big change of the sample clustering and show a bound to that size. Smaller sub-samples are proven to have small effect on the resulting clustering, and for larger fractions, they show an example of "instability".

Here, we analyze the expected distance between clusterings resulting form two independent samples. We define stability as having this expected distance converge to zero as the sample sizes grow to infinity.

Our main result is Theorem 4, in which we state that the existence of multiple optimal-cost clusterings implies instability. We formally state it in Section 3. Since its proof is lengthy, we first outline it, in Section 4. The technical lemmas are stated formally, and some of them are proved in Section 5. Proofs of the rest of the lemmas can be found in the extended version [1] available online. Section 2 is devoted to setting the ground in terms of definitions notation and basic observations.

2 Definitions

In the rest of the paper we use the following standard notation. We consider a data space X endowed with probability measure P. A finite multi-set $S = \{x_1, x_2, \ldots, x_m\}$ of X is called a *sample*. When relevant, we shall assume that samples are drawn i.i.d from (X, P). We denote by $\hat{S}$ the uniform probability distribution over the sample S.

A *clustering* $\mathcal{C}$ of a set X is a finite partition $\mathcal{C}$, of X (namely, an equivalence relation over X with a finite number of equivalence classes). The equivalence classes of a clustering are called clusters. We introduce the notation $x \sim_{\mathcal{C}} y$ whenever x and y lie in the same cluster of $\mathcal{C}$, and $x \nsim_{\mathcal{C}} y$ otherwise. If the clustering is clear from the context we drop the subscript and simply write $x \sim y$ or $x \nsim y$.

A function A, that for any given finite sample $S \subset X$ computes a clustering of X, is called a *clustering algorithm* (in spite of the word 'algorithm', we ignore any computability considerations). Note that this definition differs slightly from some common usage of "clustering algorithms" in which it is assumed that the algorithm outputs only a partition of the input sample.

In order to define the stability of a clustering algorithm we wish to measure by how much two clusterings differ. Given a probability distribution P over X, we define the *P-Hamming clustering distance* between two clusterings $\mathcal{C}$ and $\mathcal{D}$ as

$$d_P(\mathcal{C}, \mathcal{D}) = \Pr_{\substack{x \sim P \\ y \sim P}} \left[(x \sim_{\mathcal{C}} y) \oplus (x \sim_{\mathcal{D}} y) \right],$$

where $\oplus$ denotes the logical XOR operation. In other words, $d_P(\mathcal{C}, \mathcal{D})$ is the P-probability of drawing a pair of points on which the equivalence relation defined by $\mathcal{C}$ differs from the one defined by $\mathcal{D}$. Other definitions of clustering distance may also be used, see [3] and [8]. However, the Hamming clustering distance is conceptually the simplest, universal, and easy to work with. For a probability distribution P with a finite support, the Hamming distance has the additional property that two clusterings have zero distance if and only if they induce the same partitions of the support of P. We shall thus treat clusterings with zero Hamming clustering distance as equal.

The central notion of this paper is *instability*:

Definition 1 (Instability). *The* instability *of a clustering algorithm A with respect to a sample size m and a probability distribution P is*

$$\text{Instability}(A, P, m) = \mathop{\mathbb{E}}_{\substack{S_1 \sim P^m \\ S_2 \sim P^m}} d_P(A(S_1), A(S_2)).$$

The instability *of A with respect to P is*

$$\text{Instability}(A, P) = \lim_{m \to \infty} \text{Instability}(A, P, m).$$

We say that an algorithm A is stable *on P, if* $\text{Instability}(A, P) = 0$*, otherwise we say that A is* unstable*.*

A large class of clustering problems aim to choose the clustering by minimizing some risk function. We call these clustering optimization problems.

Definition 2 (Risk Minimization Clustering Problems).

– *A clustering* risk minimization problem *is a quadruple $(X, \mathcal{L}, \mathcal{P}, R)$, where X is some domain set, $\mathcal{L}$ is a set of legal clusterings of X, $\mathcal{P}$ is a set of probability distributions over X, and $R : \mathcal{P} \times \mathcal{L} \to \mathbb{R}_+$ [3] is an objective function (or risk) that the clustering algorithm aims to minimize.*

[3] We denote by $\mathbb{R}_+$ the set of non-negative real numbers, and by $\mathbb{R}_{++}$ the set of positive real numbers.

– *An* instance *of the risk minimizing problem is a concrete probability distribution P from $\mathcal{P}$. The optimal cost $opt(P)$ for an instance P, is defined as $opt(P) = \inf_{\mathcal{C}\in\mathcal{L}} R(P,\mathcal{C})$.*
– *For a sample $S \subseteq X$, we call $R(\hat{S},\mathcal{C})$ the* empirical risk *of $\mathcal{C}$ with respect to the sample S.*
– *A* risk-minimizing *(or R-*minimizing*) clustering algorithm is an algorithm that for any sample S, has $R(\hat{S}, A(S)) = opt(\hat{S})$. For all practical purposes this requirement defines A uniquely. We shall therefore refer to* the *risk-minimizing algorithm.*

Given a probability distribution P over some Euclidean space $X \subseteq \mathbb{R}^d$ and a clustering $\mathcal{C}$ of X with clusters $C_1, C_2, \ldots, C_k$, let $c_1, c_2, \ldots, c_k$ be the P-centers of mass of the clusters C_i. Namely, $c_i = \mathbb{E}_{x\sim P}[x|x \in C_i]$, and, for every $x \in X$, let c_x denote the center of mass of the class to which x belongs. The k-means risk R is defined as

$$R(P,\mathcal{C}) = \mathop{\mathbb{E}}_{x\sim P} \|x - c_x\|_2^2 \, . \tag{1}$$

In many cases, risk minimizing algorithms converge to the true risk as sample sizes grow to infinity. For the case of k-mean and k-medians on bounded subset of $\mathbb{R}^d$ with the Euclidean metric, such convergence was proved by Pollard [9], and uniform, finite-sample rates of convergence were shown in [2].

Definition 3 (Uniform Convergence). *Let P be a probability distribution. The risk function R* converges uniformly *if for any positive ϵ and δ, there exists sample size m_0 such that for all $m > m_0$*

$$\Pr_{S\sim P^m}\left[\forall\mathcal{C} \in \mathcal{S} \quad |R(\hat{S},\mathcal{C}) - R(P,\mathcal{C})| < \epsilon\right] > 1 - \delta \, .^4$$

3 Stability of Risk Optimizing Clustering Algorithms

Informally speaking, our main claim is that the stability of the risk minimizing algorithm with a uniformly converging risk function is fully determined by the number of risk optimal clusterings. More concretely, a risk-minimizing algorithm is stable on an input data distribution P, if and only if P has a unique risk minimizing clustering. We prove such a result for the k-means clustering problem

The first step towards such a characterization follows from Pollard [9]. He proves that the existence of a unique k-means minimizing clustering (for a P's with bounded support over Euclidean spaces) implies stability. Ben-David et al, [3] extended this result to a wider class of clustering problems.

As for the reverse implication, [3] shows that if P has multiple risk-minimizing clusterings, and is *symmetric* with respect to these clusterings, then it is unstable. Where symmetry is defined as an isometry $g : X \to X$ of the underlying metric space (X, ℓ) which preserves P (that is, for any measurable set A,

[4] Here m_0 can also depend on P, and not only on ϵ and δ. The uniform convergence bound proved in [2] is a stronger in this sense, since it expresses m_0 as a function of ϵ and δ only and holds for any P.

$\Pr_{x\sim P}[x \in A] = \Pr_{x\sim P}[g(x) \in A]$), and the clustering distance and the risk function are invariant under g. Note that the k-means risk function and the Hamming clustering distance are invariant under any such symmetry. See [3] for details.

Ben-David et al [3] conjecture that symmetry is not a necessary condition. Namely, that the mere existence of multiple risk-minimizing clusterings suffices for instability. In this paper we prove that this conjecture holds for k-means clustering over finite-support probability distributions.

Theorem 4. *Let P be a probability distribution over the Euclidean space $\mathbb{R}^d$ with a finite support. Then, the k-means risk-minimizing algorithm is stable on P if and only if there exist unique clustering minimizing the k-means risk function $R(P,\cdot)$.*

The next section outlines the proof. In Section 5 we follow that outline with precise statements of the needed technical lemmas and some proofs. Some of the proofs are omitted and can be found in the extended version [1] available online.

4 Proof Outline

A finite-support probability distribution may be viewed as a vector of weights. Similarly, any finite sample over such a domain can be also described by a similar relative frequency vector. We view the clustering problem as a function from such vectors to partitions of the domain set. Loosely speaking, having multiple optimal clusterings for some input distribution, P, say, $\mathcal{C}_1, \mathcal{C}_2, \ldots \mathcal{C}_h$, we consider the decision function that assigns each sample-representing vector to the index $i \in \{1, \ldots h\}$ of its optimal solution. (Note that due to the uniform convergence property, for large enough samples, with high probability, these sample based partitions are among the actual input optimal clusterings.) We analyze this decision function and show that, for large enough sample sizes, none of its values is obtained with probability 1. This implies instability, since having two different partitions, each with non-zero probability, implies a non-zero expectation of the distance between sample-generated clustering solutions.

To allow a more detailed discussion we need some further notation.

Let $F = \{x_1, x_2, \ldots, x_n\}$ be the support of P with $P(\{x_i\}) = \mu_i > 0$ for all $i = 1, 2, \ldots, n$ and $\mu_1 + \mu_2 + \cdots + \mu_n = 1$. Let

$$\mu = (\mu_1, \mu_2, \ldots, \mu_n).$$

If $n \leq k$ or $k \leq 1$ there is a trivial unique minimizer. Hence we assume that $n > k \geq 2$.

For a sample S of size m, we denote the number of occurrences of the point x_i in S by m_i, and use $w_i = m_i/m$ to denote the empirical frequency (*weight*) of the point x_i in the sample. The sample is completely determined by the vector of weights

$$w = (w_1, w_2, \ldots, w_n)\ .$$

Since the support of P is finite, there are only finitely many partitions of F.

A partition, $\mathcal{C}$, is called *optimal* if its risk, $R(P, \mathcal{C})$ equals $opt(P)$. A partition is called *empirically optimal* for a sample S, if its empirical risk, $R(\hat{S}, \mathcal{C})$ equals $opt(\hat{S})$. We shall freely replace $\hat{S}$ with its weight vector w, in particular, we overload the notation and write $R(w, \mathcal{C}) = R(\hat{S}, \mathcal{C})$.

Consider a pair of distinct optimal partitions $\mathcal{C}$ and $\mathcal{D}$. For weights w consider the empirical risk, $R(w, \mathcal{C})$, of the partition $\mathcal{C}$ on a sample with weights w. Likewise, consider the empirical risk $R(w, \mathcal{D})$. The k-means risk minimizing algorithm "prefers" $\mathcal{C}$ over $\mathcal{D}$ when $R(w, \mathcal{C}) < R(w, \mathcal{D})$. We consider the set of weights

$$Q = \{w \in \mathbb{R}^n_{++} \mid R(w, \mathcal{C}) < R(w, \mathcal{D})\} \ ,$$

where, $\mathbb{R}_{++}$ denotes the set of (strictly) positive real numbers. We allow Q to contain weight vectors w having arbitrary positive sum of weights, $w_1 + w_2 + \cdots + w_n$, not necessarily equal to one. Due to the homogeneity of the k-means risk as a function of the weights, weight vectors of arbitrary total weight can be rescaled to probability weights without effecting the risk preference between two clusterings (for details see the proof Lemma 13). This relaxation simplifies the analysis.

Step 1: We analyze the set Q in a small neighborhood of μ. In Lemma 12, we show that Q contains an *open cone T with peak at μ*. The proof of the Lemma consists of several smaller steps.

(a) We first define the function $f : \mathbb{R}^n \to \mathbb{R}$, $f(w) = R(w, \mathcal{D}) - R(w, \mathcal{C})$. In this notation $Q = \{w \mid f(w) > 0\}$. Note the important fact that $f(\mu) = 0$. We analyze the behavior of f near μ.

(b) From Observation 5 it follows that $R(w, \mathcal{C})$ is a rational function of w. Then, in Lemma 6, we compute the Taylor expansion of $R(w, \mathcal{C})$ at the point μ.

(c) In Lemma 10 we show that the first non-zero term in the Taylor expansion of f attains both positive and negative values, and thus f itself attains both positive and negative values arbitrarily close to μ.

(d) We show that, since f is rational and hence analytic in the neighborhood of μ, it follows that Q contains a cone T whose peak is at μ.

Step 2: Consider the hyperplane

$$H = \{w \in \mathbb{R}^n \mid w_1 + w_2 + \cdots + w_n = 1\}$$

in which the weights actually lie. In Lemma 13 we show that $Q \cap H$ contains an $(n-1)$-dimensional open cone Y.

Step 3: The distribution of the random vector w describing the sample is a multinomial distribution with m trials. From central limit theorem it follows that as the sample size m approaches infinity the probability distribution of w can be approximated by a multivariate Gaussian distribution lying in H. The Gaussian distribution concentrates near its mean value μ as the sample size increases. The shape of Q near μ determines the probability that the algorithm prefers partition $\mathcal{C}$ over $\mathcal{D}$. Formally, in Lemma 14 we show that $\lim_{m\to\infty} \Pr[w \in Y] > 0$; hence $\lim_{m\to\infty} \Pr[w \in Q] > 0$.

Step 4: For sufficiently large sample sizes the partition of F output by the algorithm is, with high probability, one of the optimal partitions. From the previous step it follows that with non-zero probability any optimal partition has lower empirical risk than any other optimal partition. Hence, there exist at least two optimal partitions of F, such that each of them is empirically optimal for a sample with non-zero probability. These two partitions cause instability of the algorithm. A precise argument is presented in Lemma 15.

5 The Technical Lemmas

Observation 5 (Explicit Risk Formula). *For a partition $\mathcal{C}$ and a weight vector w,*

$$R(w, \mathcal{C}) = \sum_{i=1}^{k} \sum_{x_t \in C_i} w_t \left\| x_t - \frac{\sum\limits_{x_s \in C_i} w_s x_s}{\sum\limits_{x_s \in C_i} w_s} \right\|_2^2 , \tag{2}$$

where $C_1, C_2, \ldots, C_k$ are the clusters of $\mathcal{C}$. Therefore $R(w, \mathcal{C})$ is a rational function of w.

Proof. This is just a rewriting of the definition of the k-means cost function for the case of a finite domain. We use weighted sums expressions for the expectations and

$$c_i = \frac{\sum_{x_s \in C_i} w_s x_s}{\sum_{x_s \in C_i} w_s}$$

to calculate the centers of mass of the clusters. □

Lemma 6 (Derivatives of f). *Let $\mathcal{C}$ be a partition of the support of P. The first two derivatives of the risk function $R(w, \mathcal{C})$ with respect to w at μ are as follows.*

1. *The gradient is*

$$(\nabla R(\mu, \mathcal{C}))_p = \left.\frac{\partial R(w, \mathcal{C})}{\partial w_p}\right|_{w=\mu} = \|c_\ell - x_p\|_2^2 ,$$

   assuming that x_p lies in the cluster C_ℓ.
2. *The (p, q)-th entry of the Hessian matrix*

$$\left(\nabla^2 R(\mu)\right)_{p,q} = \left.\frac{\partial^2 R(w, \mathcal{C})}{\partial w_p \partial w_q}\right|_{w=\mu}$$

   equals to

$$-2 \frac{(c_\ell - x_p)^T (c_\ell - x_q)}{\sum\limits_{x_s \in C_\ell} \mu_s}$$

   if x_p, x_q lie in a common cluster C_ℓ, and is zero otherwise.

Here, $c_1, c_2, \ldots, c_k$ are the optimal centers $c_i = \left(\sum_{x_s \in C_i} \mu_s x_s\right) / \left(\sum_{x_s \in C_i} \mu_s\right)$, and $C_1, C_2, \ldots, C_k \subseteq F$ are the clusters of $\mathcal{C}$.

Proof. Straightforward but long calculation, starting with formula (2). See the extended version paper [1] available online. □

Lemma 7 (Weights of Clusters). *Let $\mathcal{C}$ and $\mathcal{D}$ be two partitions of F. Consider the weights μ assigned to points in F. Then, either for every point in F the weight of its cluster in $\mathcal{C}$ is the same as the weight of its cluster in $\mathcal{D}$. Or, there are two points in F, such that the weight of the cluster of the first point in $\mathcal{C}$ is strictly larger than in $\mathcal{D}$, and the weight of the cluster of the second point in $\mathcal{C}$ is strictly smaller than in $\mathcal{D}$.*

Proof. For any point $x_t \in F$ let $a_t = \sum_{x_s \in C_i} \mu_s$ be the weight of the cluster C_i in which x_t lies in the partition $\mathcal{C}$. Likewise, let $b_t = \sum_{x_s \in D_j} \mu_s$ be the weight of the cluster D_j in which x_t lies in the clustering $\mathcal{D}$. Consider the two sums $\sum_{t=1}^{n} \frac{\mu_t}{a_t}$ and $\sum_{t=1}^{n} \frac{\mu_t}{b_t}$. It easy to see that the sums are equal, $\sum_{t=1}^{n} \frac{\mu_t}{a_t} = \sum_{i=1}^{k} \sum_{x_t \in C_i} \frac{\mu_t}{a_t} = k = \sum_{i=1}^{k} \sum_{x_t \in D_i} \frac{\mu_t}{b_t} = \sum_{t=1}^{n} \frac{\mu_t}{b_t}$. Either all the corresponding summands μ_t/a_t and μ_t/b_t in the two sums are equal and hence $a_t = b_t$ for all t. Or, there exist points x_t and x_s such that $\mu_t/a_t < \mu_t/b_t$ and $\mu_s/a_s > \mu_s/b_s$, and hence $a_t > b_t$ and $a_s < b_s$. □

Lemma 8 (No Ties). *Let $\mathcal{C}$ be an optimal partition and let $c_1, c_2, \ldots, c_k$ be the centers of mass of the clusters of $\mathcal{C}$ computed with respect to the weight vector μ. Then, for a point x of the support lying in a cluster C_i of $\mathcal{C}$, the center of mass c_i is strictly closer to x than any other center.*

Proof. Suppose that the distance $\|c_j - x\|_2$, $j \neq i$, is smaller or equal to the distance $\|c_i - x\|$. Then, we claim that moving the point x to the cluster C_j decreases the risk. After the move of x, recompute the center of C_i. As a result the risk strictly decreases. Then recompute the center of mass of C_j, the risk decreases even more. □

Lemma 9 (Hessian determines Clustering). *For partitions $\mathcal{C}, \mathcal{D}$ of the support of P, the following holds. If the Hesse matrices of the risk functions $R(w, \mathcal{C})$ and $R(w, \mathcal{D})$, respectively, coincide at μ, then $\mathcal{C} = \mathcal{D}$.*

Proof. For sake of brevity, let

$$A_{p,q} := \left. \frac{\partial^2 R(w, \mathcal{C})}{\partial w_p \partial w_q} \right|_{w=\mu} .$$

It suffices to show that centers $c_1, c_2, \ldots, c_k$ of partition $\mathcal{C}$ are uniquely determined by matrix A. To this end, we view A as the adjacency matrix of a graph G with nodes $x_1, x_2, \ldots, x_n$, where nodes x_p, x_q are connected by an edge if and only if $A_{p,q} \neq 0$. Let $K_1, K_2, \ldots, K_\ell$ be the connected components of G. Note that there is an edge between x_p and x_q only if p and q belong to the same cluster in $\mathcal{C}$. Thus, the connected components of G represent a refinement of partition $\mathcal{C}$. Consider a fixed cluster C_j in $\mathcal{C}$ with center c_j. Recall that

$$c_j = \sum_{x_i \in C_j} \mu_i x_i \ . \tag{3}$$

Let $K \subseteq C_j$ be any connected component of G that is contained in C_j and define, for sake of brevity, $\mu(K) := \sum_{x_i \in C} \mu_i$ and $K' = C_j \setminus K$. We claim that

$$c_j = \frac{1}{\mu(K)} \sum_{x_i \in K} \mu_i x_i \ , \tag{4}$$

that is, c_j is determined by any component $K \subseteq C_j$. Since this is obvious for $K = C_j$, we assume that $K \subsetneq C_j$. We can rewrite (3) as

$$0 = \left(\sum_{x_i \in K} \mu_i (x_i - c_j) \right) + \left(\sum_{x_{i'} \in K'} \mu_{i'} (x_{i'} - c_j) \right) \ . \tag{5}$$

Pick any pair i, i' such that $x_i \in K$ and $x_{i'} \in K'$. Since x_i and $x_{i'}$ are not neighbors in G, $A_{i,i'} = 0$, which means that $x_i - c_j$ is orthogonal to $x_{i'} - c_j$. Thus the vector represented by the first sum in (5) is orthogonal on the vector represented by the second sum. It follows that both sums yield zero, respectively. Rewriting this for the first sum, we obtain (4). □

Lemma 10 (Indefinitness). *Let $\mathcal{C}$ and $\mathcal{D}$ be any two optimal partitions. Let $f(w) = R(w, \mathcal{D}) - R(w, \mathcal{C})$. Consider the Taylor expansion of f around μ. Then, $\nabla f(\mu) \neq 0$ or the Hessian, $\nabla^2 f(\mu)$, is indefinite.*[5]

Proof. We denote by $C_1, C_2, \ldots, C_k \subseteq F$ the clusters of $\mathcal{C}$ and by $D_1, D_2, \ldots, D_k \subseteq F$ the clusters of $\mathcal{D}$. We denote by $c_1, c_2, \ldots, c_k$ the optimal centers for $\mathcal{C}$, and by $d_1, d_2, \ldots, d_k$ the optimal centers for $\mathcal{D}$. That is, the center c_i is the center of mass of C_i, and d_j is the center of mass of D_j.

Consider the Taylor expansion of f at μ. Lemma 9 implies that the Hessian, $\nabla^2 f(\mu)$, is not zero. Assuming $\nabla f(\mu) = 0$ i.e. $\nabla R(\mu, \mathcal{C}) = \nabla R(\mu, \mathcal{D})$, we need to show that $\nabla^2 f(\mu)$ is indefinite.

For any point $x_p \in F$ we define three numbers e_p, a_p and b_p as follows. Suppose $x_p \in C_\ell$ and $x_p \in D_{\ell'}$. The first part of the Lemma 6 and $\nabla R(\mu, \mathcal{C}) = \nabla R(\mu, \mathcal{D})$ imply that the distance between x_p and c_ℓ equals to the distance between x_p and $d_{\ell'}$; denote this distance by e_p. Denote by a_p the weight of the cluster C_ℓ, that is, $a_p = \sum_{x_t \in C_\ell} \mu_t$. Likewise, let b_p be the weight of the cluster $D_{\ell'}$, that is, $b_p = \sum_{x_t \in D_{\ell'}} \mu_t$.

Consider the diagonal entries of Hessian matrix of f. Using the notation we had just introduced, by the second part of the Lemma 6 the (p,p)-th entry is

$$(\nabla^2 f(\mu))_{p,p} = \left(\frac{\partial^2 R(w, \mathcal{D})}{\partial w_p^2} - \frac{\partial^2 R(w, \mathcal{C})}{\partial w_p^2} \right) \bigg|_{w=\mu} = 2 e_p^2 \left(\frac{1}{a_p} - \frac{1}{b_p} \right) \ .$$

[5] A matrix is *indefinite* if it is neither positively semi-definite, nor negatively semi-definite.

We claim that if $e_p = 0$, then $a_p = b_p$. Let $x_p \in C_\ell \cap D_{\ell'}$, and suppose without loss of generality that $a_p > b_p$. Since $e_p = 0$ it is $x_p = c_\ell = d_{\ell'}$. Since $a_p > b_p$ there is another point x_q that causes the decrease of the weight the cluster C_ℓ. Formally, $x_q \in C_\ell$, $x_q \notin D_{\ell'}$, but $x_q \in D_{\ell''}$. This means that in $\mathcal{D}$ the point x_q is closest to both $d_{\ell'}$ and $d_{\ell''}$. By Lemma 8, a tie can not happen in an optimal partition, which is a contradiction.

By Lemma 7, either (a) for all indices p, $a_p = b_p$, or (b) there are indices i, j such that $a_i > b_i$ and $a_j < b_j$. In the subcase (a), all the diagonal entries of Hessian matrix are zero. Since the Hessian matrix is non-zero, there must exist a non-zero entry off the diagonal making the matrix is indefinite. In the subcase (b), the above claim implies that the indices i, j for which $a_i > b_i$ and $a_j < b_j$ are such that $e_i, e_j > 0$. Hence, the (i,i)-th diagonal entry of the Hessian matrix is negative, and the (j,j)-the diagonal entry of the Hessian matrix is positive. Therefore the Hessian matrix is indefinite. □

Corollary 11. *There exists arbitrarily small $\delta \in \mathbb{R}^n$, $f(\mu + \delta) > 0$ (and similarly, there exists arbitrarily small δ', $f(\mu + \delta') < 0$).*

Proof. Consider the Taylor expansion of f at μ and its lowest order term $T(x-\mu)$. that does not vanish (according to Lemma 10, either the gradient or the Hessian). Since T can take values of positive and of negative sign (obvious for the gradient, and obvious from Lemma 10 for the Hessian), we can pick a vector $x = \mu + \delta$ such that $T(x - \mu) = T(\delta) > 0$. Since T is homogeneous in δ, $T(\lambda\delta) > 0$ for every $\lambda > 0$. If λ is chosen sufficiently small, then $f(\mu + \lambda\delta)$ has the same sign as $T(\lambda\delta)$. The considerations for negative sign are symmetric. □

Lemma 12 (Existence of a Positive Open Cone). *There exist positive real numbers ϵ and δ, and a unit vector $u \in \mathbb{R}^n$ such that the open cone*

$$T = \left\{ w \in \mathbb{R}^n_{++} \;\middle|\; 0 < \|w - \mu\|_2 < \epsilon,\ \frac{u^T(w - \mu)}{\|w - \mu\|_2} > 1 - \delta \right\}$$

is contained in Q, the set of weights for which $R(w, \mathcal{C}) < R(w, \mathcal{D})$.

Proof. Let h be the order of the first non-zero term in the Taylor expansion of $f(\mu + u)$ around μ (as a multi-variate polynomial in u). Using Corollary 11, pick a vector u so that $f(\mu + u) > 0$ and, for some $\eta > 0$, for every v in the η-ball around u, that term of the Taylor expansion of $f(\mu + u)$ dominates the higher order terms. The existence of such a ball follows from the smoothness properties of f. Note that this domination holds as well for any $f(\mu + \lambda v)$ such that $0 < \lambda \leq 1$.

Let δ be the supremum, over the vectors v in the η-ball, of the expression $1 - (u^T v)/\|v\|_2$. (That is, $1 - \delta$ is the infimum of the cosines of the angles between u and v's varying over the η-ball.) And let $\epsilon = \eta$.

For the vectors v, by the Taylor expansion formula, λv, $\mathrm{sign}(f(\mu + \lambda v)) = \mathrm{sign}(f(\mu + v)) = \mathrm{sign}(f(\mu + u)) > 0$. Hence, all the points of the form $w = \mu + \lambda v$ contain the cone sought. □

Lemma 13 (Existence of a Positive Open Cone II). *There exists positive real numbers ϵ, δ and a unit vector $u \in \mathbb{R}^n$ with sum of coordinates, $u_1 + u_2 + \cdots + u_n$, equal to zero, such that the $(n-1)$-dimensional open cone*

$$Y = \left\{ w \in H \cap \mathbb{R}^n_{++} \;\middle|\; 0 < \|w - \mu\|_2 < \epsilon,\ \frac{u^T(w-\mu)}{\|w-\mu\|_2} > 1 - \delta \right\}$$

is contained in $Q \cap H$.

Proof. We use the projection $\phi : \mathbb{R}^n_{++} \to H$, $\phi(w) = w/(w_1 + w_2 + \cdots + w_n)$. Note that for the k-means cost function, for every clustering $\mathcal{C}$ and every positive constant λ, $R(\lambda w, \mathcal{C}) = \lambda R(w, \mathcal{C})$. It follows that the projection ϕ does not affect the sign of f. That is, $\operatorname{sign}(f(w)) = \operatorname{sign}(f(\phi(w)))$. Therefore $Q \cap H = \phi(Q) \subset Q$. The projection $\phi(T)$ clearly contains an $(n-1)$-dimensional open cone Y of the form as stated in the Lemma. More precisely, there exists positive numbers ϵ, δ and unit vector u (the direction of the axis of the cone), such that the cone

$$Y := Y_{\epsilon,\delta,u} = \left\{ w \in H \cap \mathbb{R}^n_{++} \;\middle|\; 0 < \|w - \mu\|_2 < \epsilon,\ \frac{u^T(w-\mu)}{\|w-\mu\|_2} > 1 - \delta \right\}$$

is contained in $\phi(T)$. Since the cone Y lies in H, the direction of the axis, u, can be picked in such way that the sum of its coordinates $u_1 + u_2 + \cdots + u_n$ is zero. Since $T \subseteq Q$, we get $Y \subset \phi(T) \subset \phi(Q) = Q \cap H$. $\square$

Lemma 14 (Instability). *Let $\mathcal{C}$ and $\mathcal{D}$ be distinct optimal partitions. Let Q be the set of weights where the k-means clustering algorithm prefers $\mathcal{C}$ over $\mathcal{D}$. Then,* $\lim_{m\to\infty} \Pr[w \in Q] > 0$.

Proof. Let $Y \subset (Q \cap H)$ be an $(n-1)$-dimensional open cone (as implied by lemma 13) lying in the hyperplane H defined by the equation $w_1 + w_2 + \cdots + w_n = 1$. We show that,

$$\lim_{m\to\infty} \Pr[w \in Y] > 0 \,,$$

which implies the claim.

We have

$$\begin{aligned}
\Pr[w \in Y] &= \Pr\left[\frac{u^T(w-\mu)}{\|w-\mu\|_2} > 1 - \delta,\ 0 < \|w-\mu\|_2 < \epsilon \right] \\
&= \Pr\left[\frac{u^T(\sqrt{m}(w-\mu))}{\sqrt{m}\|w-\mu\|_2} > 1 - \delta,\ 0 < \sqrt{m}\|w-\mu\|_2 < \epsilon\sqrt{m} \right] .
\end{aligned}$$

By the central limit theorem $\sqrt{m}(w - \mu)$ weakly converges to a normally distributed random variable $Z \sim N(0, \Sigma)$, where Σ is the covariance matrix.[6] In

[6] $\Sigma = \operatorname{diag}(\mu_1, \mu_2, \ldots, \mu_n) - \mu\mu^T$, the rank of Σ is $n-1$, and its rows (or columns) span the $(n-1)$-dimensional vector space $\{u \in \mathbb{R}^n \mid u_1 + u_2 + \cdots + u_n = 0\}$.

particular this means that there is a sequence $\{\zeta_m\}_{m=1}^{\infty}$, $\zeta_m \to 0$, such that

$$\left| \Pr\left[\frac{u^T(\sqrt{m}(w-\mu))}{\sqrt{m}\|w-\mu\|_2} > 1-\delta,\ 0 < \sqrt{m}\|w-\mu\|_2 < \epsilon\sqrt{m}\right] \right.$$
$$\left. - \Pr\left[\frac{u^T Z}{\|Z\|_2} > 1-\delta,\ 0 < \|Z\|_2 < \epsilon\sqrt{m}\right] \right| < \zeta_m$$

Consequently, we can bound the probability $\Pr[w \in Y]$ as

$$\begin{aligned}\Pr[w \in Y] &\geq \Pr\left[\frac{u^T Z}{\|Z\|_2} > 1-\delta,\ 0 < \|Z\|_2 < \epsilon\sqrt{m}\right] - \zeta_m \\ &\geq 1 - \Pr\left[\frac{u^T Z}{\|Z\|_2} < 1-\delta\right] - \Pr\left[\|Z\|_2 \geq \epsilon\sqrt{m}\right] - \Pr\left[\|Z\|_2 = 0\right] - \zeta_m \ .\end{aligned}$$

Take the limit $m \to \infty$. The last three terms in the last expression vanish. Since u has sum of its coordinates zero and $Z \sim N(0, \Sigma)$ is normally distributed, the term $\lim_{m\to\infty} \Pr\left[\frac{u^T Z}{\|Z\|_2} < 1-\delta\right]$ lies in $(0, 1)$. □

Lemma 15 (Multiple Optimal Partitions). *If there are at least two optimal partitions of the support F, then the k-means algorithm is unstable.*

Proof. Let $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_h$, $h \geq 2$, be the optimal partitions. Suppose that

$$\lim_{m\to\infty} \Pr[A(S) = \mathcal{C}_i] = \pi_i \ ,$$

where by the event $A(S) = \mathcal{C}_i$ we mean that the k-means algorithm on the sample S outputs the partition $\mathcal{C}_i$ of the support.

Claim: Each number π_i is strictly less than one.

Proof of the Claim:

$$\begin{aligned}\Pr_{S\sim P^m}[A(S) = \mathcal{C}_i] &\leq \Pr\left[R(w, \mathcal{C}_i) \leq \min_{\substack{\ell=1,2,\ldots,h \\ \ell\neq i}} R(w, \mathcal{C}_\ell)\right] \\ &\leq \Pr[R(w, \mathcal{C}_i) \leq R(w, \mathcal{C}_j)] \\ &= 1 - \Pr[R(w, \mathcal{C}_i) > R(w, \mathcal{C}_j)]\end{aligned}$$

Taking limit $m \to \infty$ on both sides of the inequality and applying Lemma 14, $\lim_{m\to\infty} \Pr[R(w, \mathcal{C}_i) > R(w, \mathcal{C}_j)] > 0$ the claim follows.

Since k-means is risk converging, as sample size increases with probability approaching one, $A(S)$ outputs an optimal partition, and hence $\pi_1 + \pi_2 + \cdots + \pi_h = 1$. Necessarily at least two numbers π_i, π_j are strictly positive. That is, the algorithm outputs two different partitions $\mathcal{C}_i, \mathcal{C}_j$ with non-zero probability for arbitrarily large sample size. The algorithm will be switching between these two partitions. Formally, $\mathrm{Instability}(A, P) \geq d_P(\mathcal{C}_i, \mathcal{C}_j)\pi_i\pi_j$ is strictly positive. □

6 Conclusions and Discussion

Stability reflects a relation between clustering algorithms and the data sets (or data generating probability distributions) they are applied to. Stability is commonly viewed as a necessary condition for the suitability of the clustering algorithm, and its parameter setting, to the input data, as well as to the meaningfulness of the clustering the algorithm outputs. As such, stability is often used for model selection purposes, in particular for choosing the number of clusters for a given data. While a lot of published work demonstrates the success of this approach, the stability paradigm is mainly a heuristic and is not supported by clear theoretical guarantees. We embarked on the task of providing theoretical analysis of clustering stability. The results of Ben-David el al [3] and this paper challenge the common interpretation of stability described above. We show that the stability of risk-minimizing clustering algorithms over data generating distributions is just an indicator of weather the objective function (the risk) that the algorithm is set to minimize has one or more optimal solutions over the given input. This characterization is orthogonal to the issues of model selection to which stability is commonly applied. Based on our characterization, it is fairly simple to come up with examples of data sets (or data generating distributions) for which a 'wrong' choice of the number of clusters results in stability, whereas a 'correct' number of clusters results in instability (as well as examples for any of the other combinations of 'wrong/correct number of clusters' and 'stable/unstable'). The results of this paper apply to k-means over finite domains, but we believe that they are extendable to wider classes of clustering tasks.

How can that be? How can a paradigm that works in many practical applications be doomed to failure when analyzed theoretically? The answers should probably reside in the differences between what is actually done in practice and what our theory analyzes. The first suspect in that domain is the fact that, while in practice every stability procedure is based on some finite sample, our definition of stability refers to the limit behavior, as sample sizes grow to infinity. In fact, it should be pretty clear that, for any reasonable clustering risk function, an overwhelming majority of realistic data sets should have a unique optimal clustering solution. It is unlikely that for a real data set two different partitions will result in *exactly* the same k-means cost. It therefore follows that for large enough samples, these differences in the costs of solutions will be detected by the samples and the k means clustering will stabilize. On the other hand, sufficiently small samples may fail to detect small cost differences, and therefore look stable. It may very well be the case that the practical success will breakdown if stability tests would take into account larger and larger samples. If that is the case, it is a rather unusual occasion where working with larger samples obscures the 'truth' rather than crystalizes it.

At this point, this is just a speculation. The most obvious open questions that we see ahead is determining whether this is indeed the case by coming up with a useful non-asymptotic characterization of stability. Can our work be extended to predicting the behavior of stability over finite sample sizes?

Other natural questions to answered include extending the results of this paper to arbitrary probability distributions (doing away with our finite support assumption), as well as extending our analysis to other risk-minimizing clustering tasks.

7 Acknowledgments

We are happy to express our gratitude to Shalev Ben-David (Shai's son) for his elegant proof of Lemma 7.

References

1. Extended version of this paper. Availabe at `http://www.cs.uwaterloo.ca/~dpal/papers/stability/stability.pdf` or at `http://www.cs.uwaterloo.ca/~shai/publications/stability.pdf`.
2. S. Ben-David. A framework for statistical clustering with a constant time approximation algorithms for k-median clustering. In *Proceedings of the Conference on Computational Learning Theory*, pages 415–426, 2004.
3. S. Ben-David, U. von Luxburg, and D. Pál. A sober look at clustering stability. In *Proceedings of the Conference on Computational Learning Theory*, pages 5–19, 2006.
4. A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing*, volume 7, pages 6–17, 2002.
5. S. Dudoit and J. Fridlyand. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 3(7), 2002.
6. T. Lange, M. L. Braun, V. Roth, and J. Buhmann. Stability-based model selection. In *Advances in Neural Information Processing Systems*, volume 15, pages 617–624, 2003.
7. E. Levine and E. Domany. Resampling method for unsupervised estimation of cluster validity. *Neural Computation*, 13(11):2573–2593, 2001.
8. M. Meila. Comparing clusterings. In *Proceedings of the Conference on Computational Learning Theory*, pages 173–187, 2003.
9. D. Pollard. Strong consistency of k-means clustering. *The Annals of Statistics*, 9(1):135–140, 1981.
10. Alexander Rakhlin and Andrea Caponnetto. Stability of k-means clustering. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.
11. U. von Luxburg and S. Ben-David. Towards a statistical theory of clustering. In *PASCAL workshop on Statistics and Optimization of Clustering*, 2005.