# A Sober Look at Clustering Stability

Shai Ben-David[1], Ulrike von Luxburg[2], and Dávid Pál[1]

[1] David R. Cheriton School of Computer Science,
University of Waterloo,
Waterloo, Ontario, Canada
{shai,dpal}@cs.uwaterloo.ca
[2] Fraunhofer IPSI, Darmstadt, Germany
ulrike.luxburg@ipsi.fraunhofer.de

**Abstract.** Stability is a common tool to verify the validity of sample based algorithms. In clustering it is widely used to tune the parameters of the algorithm, such as the number $k$ of clusters. In spite of the popularity of stability in practical applications, there has been very little theoretical analysis of this notion. In this paper we provide a formal definition of stability and analyze some of its basic properties. Quite surprisingly, the conclusion of our analysis is that for large sample size, stability is fully determined by the behavior of the objective function which the clustering algorithm is aiming to minimize. If the objective function has a unique global minimizer, the algorithm is stable, otherwise it is unstable. In particular we conclude that stability is not a well-suited tool to determine the number of clusters - it is determined by the symmetries of the data which may be unrelated to clustering parameters. We prove our results for center-based clusterings and for spectral clustering, and support our conclusions by many examples in which the behavior of stability is counter-intuitive.

## 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 (von Luxburg and Ben-David, 2005) In particular, the problem of choosing parameters such as the number $k$ of clusters is still more or less unsolved.

One popular method for model selection in clustering has been the notion of stability, see for instance Ben-Hur et al. (2002), Lange et al. (2004). 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 randomization. As an example, stability measurements are often employed in practice for choosing the number, $k$, of clusters. The rational behind this heuristics is that in a situation where $k$ is too large, the algorithm "randomly" has to split true clusters, and the choice of the cluster it splits might change with the randomness of the sample at hand, resulting in instability. Alternatively, if we choose $k$ too small, then we "randomly" have to merge several true clusters, the choice of which might similarly change with each particular random sample, in which case, once again, instability occurs. For an illustration see Figure 1.

The natural framework for a discussion of stability is that of sample-based algorithms. Much like statistical learning, this framework assumes that there exist some fixed but unknown probability distribution of the data, and the algorithm gets an i.i.d. random sample as input and aims to approximate a solution that is optimal w.r.t. that data distribution. In this paper we focus on clustering algorithms which choose their clustering based on some objective function which they minimize or maximize. The advantage of such cost-based clusterings is that they enjoy an explicit notion of the quality of a clustering. Popular examples in this class are center based and algorithms and spectral clustering.

For such algorithms there are two different sources of instability. The first one is based on the structure of the underlying space and has nothing to do with the sampling process. If there exist several different clusterings which minimize the objective function on the whole data space, then the clustering algorithm cannot decide which one to choose. The clustering algorithm cannot resolve this ambiguity which lies in the structure of the space. This is the kind of instability that is usually expected to occur when stability is applied to detect the correct number of clusters. However, in this article we argue that this intuition is not justified and that stability rarely does what we want in this respect. The reason is that for many clustering algorithms, this kind of ambiguity usually happens only if the data space has some symmetry structure. As soon the space is not perfectly symmetric, the objective function has a unique minimizer (see Figure 1) and stability prevails. Since we believe that most real world data sets are not perfectly symmetric, this leads to the conclusion that for this purpose, stability is not the correct tool to use.

A completely different notion of instability is the one based on the sampling process. As we can only evaluate the objective function on the given sample points, the variance in the sampling process leads to variance in the values of the empirically computed objective function, which in turn results in variance in the choice of the clusterings. This is the kind of stability that has been studied extensively in supervised learning (Bousquet and Elisseeff, 2002, Kutin and Niyogi, 2002, Rakhlin and Caponnetto, 2005). A similar effect happens if we do not have the computational power to exactly compute the global minimum of the objective function, as it for example is the case for the highly non-convex $k$-means objective function. This type of instability typically diminishes as sample sizes grow. Alternatively, one can reduce this type of instability to the previous

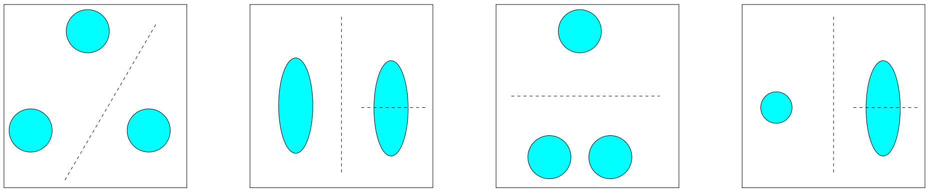

**Fig. 1.** The left two panels show situations where the constructed clustering (depicted by the dashed line) is highly instable, either because the chosen number of clusters is too small or too large. Note that both figures depict very symmetric situations. The right two panels show situations where clustering algorithms return stable results even though they construct a wrong number of clusters. Note that those two figures are not symmetric.

case by considering the set of $\varepsilon$-minimizers of the objective function (Rakhlin and Caponnetto, 2005). The set of $\varepsilon$-minimizers of a function is the set of all clusterings for which the quality function is at most $\varepsilon$ from the minimal value. If we now know that we only have enough sample points to estimate the objective function up to precision $\varepsilon$, then the instability in the algorithm consists in "randomly" picking one of the clusterings in the set of $\varepsilon$-minimizers. In this paper we mainly focus on the first kind of stability. Therefore, we mainly consider the asymptotic behavior of stability as sample sizes grow to infinity.

In this work we analyze the behavior of stability of a large abstract family of clustering algorithms - algorithms that are driven by an objective function (or 'risk') that they aim to minimize. We postulate some basic abstract requirements on such algorithms (such as convergence in probability to a minimum risk solutions as cluster sizes grow to infinity), and show that for algorithms satisfying these requirements, stability is fully determined by the symmetry structure of the underlying data distribution. Specifically, if the risk has a unique minimizer the algorithm is stable, and if there exist a non-trivial symmetry of the set of risk-minimizing solutions, stability fails. Since these symmetry parameters are independent of the number of clusters, we can easily prove that in many cases stability fails to indicate the correct (or even a reasonable) number of clusterings. Our results apply in particular to two large families of clustering algorithms, center based clustering and spectral clustering.

We would like to stress that our findings do not contradict the stability results for supervised learning. The main difference between classification and clustering is that in classification we are only interested in some function which minimizes the risk, but we never explicitly look at this function. In clustering however, we do distinguish between functions even though they have the same risk. It is exactly this fundamental difference which makes clustering so difficult to analyze.

After formulating our basic definitions in Section 2, we formulate an intuitive notion of risk minimizing clustering in Section 3. Section 4 presents our first central result, namely that existence of a unique risk-minimizer implies stability, and Section 5 present the complementary, instability result, for symmetric data structures. We end in section 6 by showing that two popular versions of spectral clustering display similar characterizations of stability in terms of basic data symmetry structure. Throughout the paper, we demonstrate the impact of our results by describing simple examples of data structures for which stability fails to meet 'common knowledge' expectations.

## 2 Definitions

In the rest of the paper we use the following standard notation. We consider a data space $X$ endowed with probability measure $P$. If $X$ happens to be a metric space, we denote by $\ell$ its metric. A sample $S = \{x_1, ..., x_m\}$ is drawn i.i.d from $(X, P)$.

**Definition 1 (Clustering).** *A* clustering $\mathcal{C}$ *of a set* $X$ *is a finite partition* $\mathcal{C} : X \to \mathbb{N}$*. The sets* $C_i := \{x \in X;\ \mathcal{C}(x) = i\}$ *are called clusters. We introduce the notation* $x \sim_C y$ *if* $C(x) = C(y)$ *and* $x \nsim_C y$ *otherwise. In case the clustering is clear from context we drop the subscript and simply write* $x \sim y$ *or* $x \nsim y$*.*

**Definition 2 (Clustering algorithm).** *Any function* $A$*, that for any given finite sample* $S \subset X$ *computes a clustering of* $X$*, is called a clustering algorithm.*

Note that by default, the clustering constructed by an algorithm is only defined on the sample points. However, many algorithms such as center-based clusterings or spectral clustering have natural extensions of the clustering constructed on the sample to the whole data space $X$. For details see section 6.

**Notation 1** *For a finite sample (a multiset),* $S$*, let* $P_S$ *be the uniform probability distribution over* $S$*.*

**Definition 3 (Clustering distance).** *Let* $\mathcal{P}$ *be family of probability distributions over some domain* $X$*. Let* $\mathcal{S}$ *be a family of clusterings of* $X$*. A* clustering distance *is function* $d : \mathcal{P} \times \mathcal{S} \times \mathcal{S} \to [0, 1]$ *satisfying for any* $P \in \mathcal{P}$ *and any* $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3 \in \mathcal{S}$

1. $d_P(\mathcal{C}_1, \mathcal{C}_1) = 0$
2. $d_P(\mathcal{C}_1, \mathcal{C}_2) = d_P(\mathcal{C}_2, \mathcal{C}_1)$ *(symmetry)*
3. $d_P(\mathcal{C}_1, \mathcal{C}_3) \leq d_P(\mathcal{C}_1, \mathcal{C}_2) + d_P(\mathcal{C}_2, \mathcal{C}_3)$ *(triangle inequality)*

We do not require that a clustering distance satisfies that if $d_P(\mathcal{C}_1, \mathcal{C}_2) = 0$ then $\mathcal{C}_1 = \mathcal{C}_2$. As a prototypic example we consider the Hamming distance (or pair-counting distance):

**Definition 4 (Hamming distance).** *For two clusterings $\mathcal{C}_1, \mathcal{C}_2$ of $(X, P)$, the Hamming distance is defined as*

$$d_P(\mathcal{C}_1, \mathcal{C}_2) = \Pr_{\substack{x \sim P \\ y \sim P}} \left[ (x \sim_{\mathcal{C}_1} y) \oplus (x \sim_{\mathcal{C}_2} y) \right],$$

*where $\oplus$ denotes the logical XOR operation.*

It can easily be checked that $d_P$ indeed is a clustering distance. The first two properties trivially hold, and the triangle inequality follows from

$$\begin{aligned} d_P(\mathcal{C}_1, \mathcal{C}_3) &= \Pr_{\substack{x \sim P \\ y \sim P}} \left[ (x \sim_{\mathcal{C}_1} y) \oplus (x \sim_{\mathcal{C}_3} y) \right] \\ &= \Pr_{\substack{x \sim P \\ y \sim P}} \left[ ((x \sim_{\mathcal{C}_1} y) \oplus (x \sim_{\mathcal{C}_2} y)) \oplus ((x \sim_{\mathcal{C}_2} y) \oplus (x \sim_{\mathcal{C}_3} y)) \right] \\ &\leq \Pr_{\substack{x \sim P \\ y \sim P}} \left[ (x \sim_{\mathcal{C}_1} y) \oplus (x \sim_{\mathcal{C}_2} y) \right] + \Pr_{\substack{x \sim P \\ y \sim P}} \left[ (x \sim_{\mathcal{C}_2} y) \oplus (x \sim_{\mathcal{C}_3} y) \right] \\ &= d_P(\mathcal{C}_1, \mathcal{C}_2) + d_P(\mathcal{C}_2, \mathcal{C}_3). \end{aligned}$$

**Proposition 5.** *The Hamming distance $d_P$ satisfies*

$$d_P(\mathcal{C}, \mathcal{D}) \leq 1 - \sum_i \sum_j \left( \Pr[C_i \cap D_j] \right)^2$$

*Proof.* This follows by straight forward transformations:

$$\begin{aligned} d_P(\mathcal{C}, \mathcal{D}) &= 1 - \Pr_{\substack{x \sim P \\ y \sim P}} [(x \sim_{\mathcal{C}} y) \wedge (x \sim_{\mathcal{D}} y)] - \Pr_{\substack{x \sim P \\ y \sim P}} [(x \nsim_{\mathcal{C}} y) \wedge (x \nsim_{\mathcal{D}} y)] \\ &\leq 1 - \Pr_{\substack{x \sim P \\ y \sim P}} [(x \sim_{\mathcal{C}} y) \wedge (x \sim_{\mathcal{D}} y)] \\ &= 1 - \sum_i \sum_j \Pr_{\substack{x \sim P \\ y \sim P}} [(x, y \in C_i) \wedge (x, y \in D_j)] \\ &= 1 - \sum_i \sum_j \left( \Pr[C_i \cap D_j] \right)^2 \qquad \square \end{aligned}$$

Now we define the fundamental notion of this paper:

**Definition 6.** *Let $P$ be probability distribution over $X$. Let $d$ be a clustering distance. Let $A$ be clustering algorithm. The* stability *of the algorithm $A$ for the sample size $m$ with respect to the probability distribution $P$ is*

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

*The* stability *of the algorithm $A$ with respect to the probability distribution $P$ is*

$$\mathit{stab}(A, P) = \limsup_{m \to \infty} \mathit{stab}(A, P, m).$$

*We say that algorithm $A$ is* stable *for $P$, if $\mathit{stab}(A, P) = 0$.*

Note that the algorithm $A$ which for any input only produces the clustering consisting of one cluster $X$, is stable on any probability distribution $P$. More generally, any $A$ which is a constant function is stable.

## 3 Risk optimizing clustering algorithms

A large class of clustering algorithms choose the clustering by optimizing some risk function. The large class of center based algorithms falls into this category, and spectral clustering can also be interpreted in this way.

**Definition 7 (Risk optimization scheme).** *A* risk optimization scheme *is defined by a quadruple* $(X, \mathcal{S}, \mathcal{P}, R)$*, where* $X$ *is some domain set,* $\mathcal{S}$ *is a set of legal clusterings of* $X$*, and* $\mathcal{P}$ *is a set of probability distributions over* $X$*, and* $R : \mathcal{P} \times \mathcal{S} \to \mathbb{R}_0^+$ *is an objective function (or risk) that the clustering algorithm aims to minimize.*

*Denote* $opt(P) := \inf_{\mathcal{C} \in \mathcal{S}} R(P, \mathcal{C})$*. For a sample* $S \subseteq X$*, we call* $R(P_S, \mathcal{C})$ *the* empirical risk *of* $\mathcal{C}$*. A clustering algorithm* $A$ *is called* $R$-minimizing*, if* $R(P_S, A(S)) = opt(P_S)$*, for any sample* $S$*.*

Generic examples are *center based algorithms* such as $k$-means and $k$-medians. Those clusterings pick a set of $k$ center points $c_1, ..., c_k$ and then assign each point in the metric space to the closest center point. Such a clustering is a $k$-cell Voronoi diagram over $(X, \ell)$. To choose the centers, $k$-means minimizes the risk function

$$R(P, \mathcal{C}) = \mathop{\mathbb{E}}_{x \sim P} \min_{1 \le i \le k} (\ell(x, c_i))^2 \mid \mathrm{Vor}(c_1, c_2, \ldots, c_k) = \mathcal{C}$$

while $k$-medians algorithm minimizes

$$R(P, \mathcal{C}) = \mathop{\mathbb{E}}_{x \sim P} \min_{1 \le i \le k} \ell(x, c_i) \mid \mathrm{Vor}(c_1, c_2, \ldots, c_k) = \mathcal{C}$$

Usually, risk based algorithms are meant to converge to the true risk as sample sizes grow to infinity.

**Definition 8 (Risk convergence).** *Let* $A$ *be an* $R$*-minimizing clustering algorithm. We say that* $A$ *is* risk converging*, if for every* $\epsilon > 0$ *and every* $\delta \in (0, 1)$ *there is* $m_0$ *such that for all* $m > m_0$

$$\Pr_{S \sim P^m} [R(P, A(S)) < opt(P) + \epsilon] > 1 - \delta$$

*for any probability distribution* $P \in \mathcal{P}$*.*

For example, in the case of $k$-mean and $k$-medians on bounded subset of $\mathbb{R}^d$ with Euclidean metric, Ben-David (2004) has shown that they both minimize risk from samples.

## 4 Stability of risk minimizing algorithms

In this section we investigate the stability of risk optimizing clustering algorithms. We will see that their stability solely depends on the existence of a unique minimizer of the risk function. In this section we fix a risk minimization scheme $(X, \mathcal{S}, \mathcal{P}, R)$.

**Definition 9.** *Let $d$ be a clustering distance. We say that a probability distribution $P$ has* unique minimizer $\mathcal{C}^*$ *if*

$$(\forall \eta > 0)\ (\exists \epsilon > 0)\ \left(R(P, \mathcal{C}) < opt(P) + \epsilon \implies d_P(\mathcal{C}^*, \mathcal{C}) < \eta\right).$$

*More generally, we say a probability distribution $P$* has $n$ distinct minimizers, *if there exists $\mathcal{C}_1^*, \mathcal{C}_2^*, \dots, \mathcal{C}_n^*$ such that $d_P(\mathcal{C}_i^*, \mathcal{C}_j^*) > 0$ for all $i \neq j$, and*

$$(\forall \eta > 0)\ (\exists \epsilon > 0)\ \left(R(P, \mathcal{C}) < opt(P) + \epsilon \implies (\exists\ 1 \leq i \leq n)\ d_P(\mathcal{C}_i^*, \mathcal{C}) < \eta\right).$$

Note that there is a technical subtlety here; the definition does not require that there is only a single clustering with the minimal cost, but rather that for any two optima $\mathcal{C}_1^*, \mathcal{C}_2^*$, $d_P(\mathcal{C}_1^*, \mathcal{C}_2^*) = 0$. Technically, we can overcome this difference by forming equivalence classes of clusterings, saying that two clusterings are equivalent if their clustering distance is zero. Similarly, $n$ distinct optima correspond $n$ such equivalence classes of optimal clusterings.

**Theorem 10 (Stability theorem).** *If $P$ has unique minimizer $\mathcal{C}^*$, then any $R$-minimizing clustering algorithm which is risk converging is stable on $P$.*

*Proof.* Let $A$ be a risk converging $R$-minimizing clustering algorithm. Suppose we are given $\zeta > 0$ and want to show that for large enough $m$ is $stab(A, P, m) < \zeta$. Let us pick $\delta \in (0, 1)$ and $\eta > 0$, both small enough so that

$$2(\eta + \delta) < \zeta. \tag{1}$$

Let $\mathcal{C}^*$ be the unique minimizer, then for $\eta$ there is some $\epsilon > 0$ such that

$$R(P, \mathcal{C}) < opt(P) + \epsilon \implies d_P(\mathcal{C}, \mathcal{C}^*) < \eta. \tag{2}$$

Since $A$ is risk converging, there is $m_0$ such that for all $m > m_0$

$$\Pr_{S \sim P^m}[R(P, A(S)) \geq opt(P) + \epsilon] < \delta. \tag{3}$$

Combining (2) and (3), for $m > m_0$ we have

$$\Pr_{S \sim P^m}[d_P(A(S), \mathcal{C}^*) \geq \eta] \leq \Pr_{S \sim P^m}[R(P, A(S)) \geq opt(P) + \epsilon] < \delta. \tag{4}$$

Finally, for $m > m_0$ we bound the stability as

$$\begin{aligned}
stab(A, P, m) &= \mathop{\mathbb{E}}_{\substack{S_1 \sim P^m \\ S_2 \sim P^m}} d_P(A(S_1), A(S_2)) \\
&\leq \mathop{\mathbb{E}}_{\substack{S_1 \sim P^m \\ S_2 \sim P^m}} [d_P(A(S_1), \mathcal{C}^*) + d_P(\mathcal{C}^*, A(S_2))] \\
&= 2 \mathop{\mathbb{E}}_{S \sim P^m} d_P(A(S), \mathcal{C}^*) \\
&\leq 2 \left(\eta \cdot \Pr_{S \sim P^m}[d_P(A(S), \mathcal{C}^*) < \eta] + 1 \cdot \Pr_{S \sim P^m}[d_P(A(S), \mathcal{C}^*) \geq \eta]\right) \\
&\leq 2 \left(\eta + \Pr_{S \sim P^m}[R(P, A(S)) \geq opt(P) + \epsilon]\right) \\
&\leq 2(\eta + \delta) \\
&< \zeta.
\end{aligned}$$

□

Note that this result applies in particular to the $k$-means and the $k$-median clustering paradigms (namely, to clustering algorithms that minimize any of these common risk functions).

### 4.1 Unexpected behaviors of stability

As a first example to the surprising consequences of Theorem 10, consider the uniform distribution over the unit interval $[0,1]$. It is not hard to figure out that, for any number of clusters, $k$, both $k$-medians and $k$-means have exactly one risk minimizer—the clustering

$$C(x) = i, \qquad x \in \left[\frac{i-1}{k}, \frac{i}{k}\right) .$$

Therefore, from the stability theorem, it follows that both $k$-medians and $k$-means clustering are stable on the interval uniform distribution *for any value of* $k$. Similarly consider the stability of $k$-means and $k$-medians on the two rightmost examples on the Figure 1. The rightmost example on the picture has for $k = 3$ unique minimizer as shown and therefore is stable, although the correct choice of $k$ should be 2. The second from right example has, for $k = 2$, a unique minimizer as shown, and therefore is again stable, although the correct choice of $k$ should be 3 in that case. Note also that in both cases, the uniqueness of minimizer is implied by the asymmetry of the data distributions. It seems that the number of optimal solutions is the key to instability. For the important case of Euclidean space $\mathbb{R}^d$ we are not aware of any example such that the existence of two optimal sets of centers does not lead to instability. We therefore conjecture:

*Conjecture 11 (Instability).* If $P$ has multiple minimizers then any $R$-minimizing algorithm which is risk converging is unstable on $P$.

While we cannot, at this stage, prove the above conjecture in the generality, we can prove that a stronger condition, *symmetry*, does imply instability for center based algorithms and spectral clustering algorithms.

## 5 Symmetry and instability

In this subsection we define a formal notion of symmetry for metric spaces with a probability distribution. We prove that if there are several risk minimizers which are "symmetric" to each other, then risk minimizing algorithms are bound to be unstable on this distribution. Before we can formulate claim precisely we need introduce some further notation and definitions.

**Definition 12 (Measure-preserving symmetry).** *Let $P$ be a probability distribution over $(X, \ell)$. A function $g : X \to X$, is called* $P$-preserving symmetry *of $(X, \ell)$ if,*

1. *For any $P$-measurable set $A \subseteq X$,* $\Pr[A] = \Pr[g(A)]$.

*2.* $\Pr_{\substack{x\sim P\\ y\sim P}}[\ell(x,y) = \ell(g(x), g(y))] = 1$.

**Note 1:** For any finite sample $S$ (a multi-set), if $g$ is an isometry on $S$ then $g$ is also an $\hat{S}$-preserving symmetry, where $\hat{S}$ is any discrete distribution on $S$. In what follows we adopt the following notation: If $g : X \to X$, then for set $A \subset X$ by $g[A] = \{g(x) \mid x \in A\}$. For a probability distribution $P$ let $P_g$ be defined by $P_g[A] = P[g^{-1}(A)]$ for every set $A$ whose pre-image is measurable. If $g$ is one-to-one then for a clustering $\mathcal{C} : X \to \mathbb{N}$ we define $g[\mathcal{C}]$ by $(g[\mathcal{C}])(x) = \mathcal{C}(g^{-1}(x))$, or in other words that the clusters of $g[\mathcal{C}]$ are images of clusters of $\mathcal{C}$ under $g$.[1]

**Definition 13 (Distance-Distribution dependent risk).** *We say that a risk function $R$ is ODD if it depends only on distances and distribution. Formally, $R$ is ODD if for every probability distribution $P$, every $P$-preserving symmetry $g$, and every clustering $\mathcal{C}$*

$$R(P, \mathcal{C}) = R(P, g(\mathcal{C})).$$

**Note 2:** For any finite sample $S$, if $g$ is an isometry on $S$ and $R$ is ODD, then for every clustering $\mathcal{C}$, $R(P_S, \mathcal{C}) = R(P_S, g(\mathcal{C})) = R(g(P_S), g(\mathcal{C}))$. This follows from Note 1 and the definition of $R$ being ODD.

**Definition 14 (Distance-Distribution dependent clustering distance).** *We say that a clustering distance $d$ is ODD if it depends only on distances and distribution. Formally, $d$ is ODD if for every probability distribution $P$, every $P$-preserving symmetry $g$, and any two clusterings $\mathcal{C}_1, \mathcal{C}_2$*

$$d_P(\mathcal{C}_1, \mathcal{C}_2) = d_P(g(\mathcal{C}_1), g(\mathcal{C}_2)).$$

Note that every natural notion of distance (in particular the Hamming distance and information based distances) is ODD.

**Theorem 15 (Instability from symmetry).** *Let $R$ be an ODD risk function, and $d$ an ODD clustering distance. Let $P$ be probability distribution so that for some $n$, $P$ has $n$ distinct minimizers, and let $g$ be a $P$-symmetry such that for every $R$-minimizer $\mathcal{C}^*$, $d_P(\mathcal{C}^*, g(\mathcal{C}^*)) > 0$, then any $R$-minimizing clustering algorithm which is risk convergent is unstable on $P$.*

*Proof.* Let the optimal solutions minimizing the risk be $\{\mathcal{C}_1^*, \mathcal{C}_2^*, \ldots, \mathcal{C}_n^*\}$. Let $r = \min_{1\le i\le n} d_P(\mathcal{C}_i^*, g(\mathcal{C}_i^*))$. Let $\epsilon > 0$ be such that

$$R(P, \mathcal{C}) < opt(P) + \epsilon \implies (\exists\ 1 \le i \le n)\ d_P(\mathcal{C}_i^*, \mathcal{C}) < r/4$$

(the existence of such an $\epsilon$ is implied by having $n$ distinct minimizers for $P$). Let $T = \{S \in X^m \mid R(P, A(S)) < opt(P) + \epsilon\}$. By the risk-convergence of $A$, there exist some $m_0$ such that for all $m > m_0$, $P(T) > 0.9$.

[1] We can also handle the case where $g$ fails to be one-to-one on a set of probability of zero. For the sake of clarity we omit this technicality.

For $1 \leq i \leq n$, let $T_i = \{S \in T \mid \ d_p(\mathfrak{C}_i^*, A(S)) \leq r/4\}$. Clearly, there exist some $i_0$ for which $P(T_{i_0}) \geq 0.9/n$. Since $g$ is a symmetry, and $R$ is ODD, $g(S) \in T$ for every sample $S \in T$.

Since $d_P(\mathfrak{C}_{i_0}^*, g(\mathfrak{C}_{i_0}^*)) \geq r$, and, for all $S \in T_{i_0}$, $d_P(\mathfrak{C}_{i_0}^*, A(S)) \leq r/4$, and $d_P$ is ODD, we get that for all $S \in T_{i_0}$, $d_P(g(\mathfrak{C}_{i_0}^*), A(g(S))) \leq r/4$. The triangle inequality for $d_P$ implies now that for every $S \in T_{i_0}$ and every $S' \in g[T_{i_0}]$, $d_P(A(S), A(S')) \geq r/2$. Finally, since $g$ is a $P$-symmetry, one gets $P(g[T_{i_0}]) \geq 0.9/n$.

We are now in a position to lower-bound the stability for all $m \geq m_0$:

$$\begin{aligned} stab(A, P, m) &= \mathop{\mathbb{E}}_{\substack{S \sim P^m \\ S' \sim P^m}} d_P(A(S), A(S')) \\ &\geq \frac{r}{2} \Pr_{\substack{S \sim P^m \\ S' \sim P^m}} \left[ d_P(A(S), A(S')) \geq \frac{r}{2} \right] \\ &\geq \frac{r}{2} \Pr_{\substack{S \sim P^m \\ S' \sim P^m}} \left[ S \in T_{i_0} \wedge S' \in g[T_{i_0}] \right] \\ &= \frac{r}{2} \Pr_{S \sim P^m} \left[ S \in T_{i_0} \right] \Pr_{S' \sim P^m} \left[ S' \in g[T_{i_0}] \right] \\ &\geq \frac{r(0.9)^2}{2n^2} \end{aligned}$$

Therefore the stability at infinity, $stab(A, P)$, is positive as well, and hence $A$ is unstable on $P$. □

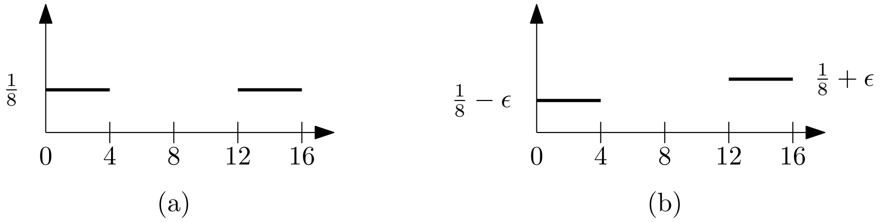

**Fig. 2.** The densities of two "almost the same" probability distributions over $\mathbb{R}$ are shown. (a) For $k = 3$, are the $k$-means and $k$-medians unstable. (b) For $k = 3$, are the $k$-means and $k$-medians stable.

For example, if $X$ is the real line $\mathbb{R}$ with the standard metric $\ell(x, y) = |x - y|$ and $P$ is the uniform distribution over $[0, 4] \cup [12, 16]$ (see Figure 2a), then $g(x) = 16 - x$ is a $P$-preserving symmetry. For $k = 3$, both $k$-means and $k$-median have exactly two optimal triples of centers $(2, 13, 15)$ and $(1, 3, 14)$. Hence, for $k = 3$, both $k$-means and $k$-medians are unstable on $P$.

However, if we change the distribution slightly, such that the weight of the first interval is little bit less than $1/2$ and the weight of the second interval is accordingly a little bit above $1/2$, while retaining uniformity on each individual interval (see Figure 2b), there will be only one optimal triple of centers, namely, $(2, 13, 15)$. Hence, for the same value, $k = 3$, $k$-means and $k$-medians become stable. This illustrates again how unreliable is stability as an indicator of a meaningful number of clusters.

## 6 Stability of spectral clustering

In this section we show similar stability results for spectral clustering. Namely, we show that the existence of a unique minimizer for the associated risk implies stability, and that the existence of non-trivial symmetries implies instability. We consider two variants of spectral clustering, the standard one, and a less standard version related to kernel k-means.

Assume we are given $n$ data points $x_1, ..., x_m$ and their pairwise similarities $s(x_i, x_j)$. Let $W$ denote the similarity matrix, $D$ the corresponding degree matrix, and $L$ the normalized graph Laplacian $L = D^{-1}(D - W)$. One of the standard derivations of normalized spectral clustering is by the normalized cut criterion (Shi and Malik, 2000). The ultimate goal is to construct $k$ indicator vectors $v_i = (v_i^1, ..., v_i^m)^t$ with $v_i^j \in \{0, 1\}$ such that the normalized cut $Ncut = \mathrm{tr}(V^t LV)$ is minimized. Here $V$ denotes the $m \times k$ matrix containing the indicator vectors $v_i$ as columns. As it is NP hard to solve this discrete optimization problem exactly, we have to resort to relaxations. In the next two subsections we investigate the stability of two different spectral clustering algorithms based on two different relaxations.

### 6.1 Stability of the standard spectral clustering algorithm

The "standard relaxation" as used in Shi and Malik (2000) is to relax the integer condition $v_i^j \in \{0, 1\}$ to $v_i^j \in \mathbb{R}$. It can be seen that the solution of the relaxed problem is then given by the first $k$ eigenvectors $v_1, ..., v_k$ of the matrix $L$. To construct a clustering from those eigenvectors we then embed the data points $x_i$ into the $k$-dimensional Euclidean space by $T_v : x_i \mapsto z_i := (v_1^{(i)}, ..., v_k^{(i)})$. Then we apply the standard $k$-means clustering algorithm to the embedded points $z_1, ..., z_m$ to obtain the final clustering $\mathcal{C}$ into $k$ clusters. This algorithm cannot easily be cast into a problem where we minimize one single cost function. Instead we proceed in two stages. In the first one we minimize the eigenvector cost function $\mathrm{tr}(V^t LV)$, and in the second one the standard $k$-means objective function on the embedded points $z_i$.

To discuss the distance between spectral clusterings based on different samples, we first have to extend a clustering constructed on each sample to the whole data space $X$. For spectral clustering there exists a natural extension operator as follows (see von Luxburg et al. (2004) for details). We extend an eigenvector $v_i$ of eigenvalue $\lambda_i$ to a function $\hat{f}_i : X \to \mathbb{R}$ by defining $\hat{f}_i(x) =$

$(\sum_{j=1}^{m} s(x, x_j) v_i^{(j)})/(m(1-\lambda_i))$. Next we extend the embedding $T_v : \{x_1, .., x_m\} \to \mathbb{R}^k$ to an embedding $T_{\hat{f}} : X \to \mathbb{R}^k$ by $T_{\hat{f}} : x \mapsto z := (\hat{f}_1^{(i)}, ..., \hat{f}_k^{(i)})$. Note that $T_{\hat{f}}(x_i) = T_v(x_i)$. Now we perform $k$-means clustering on the images of the sample points $z_i$ in $\mathbb{R}^k$. Finally, this clustering is extended by the standard extension operator for $k$-means, that is we assign all points $z$ to the closest center $c_i$, where $c_1, ..., c_k \in \mathbb{R}^k$ are the centers constructed by $k$-means on the embedded data points $z_1, ..., z_m$. Then we define the exended clustering on $X$ by setting $x \sim_{\mathcal{C}} y$ if the images of $x$ and $y$ are in the same cluster in $\mathbb{R}^k$.

**Theorem 16 (Stability of normalized spectral clustering).** *Let the data space $X$ be compact, and the similarity function $s$ be non-negative, symmetric, continuous, and bounded away from 0. Assume that the limit clustering based on $\mathcal{L}$ is unique (that is, the first $k$ eigenfunctions $f_1, ..., f_k$ of the limit operator $\mathcal{L}$ are unique and the $k$-means objective function applied to $T_f(X)$ has a unique minimizer). Let $\mathcal{C}$ and $\mathcal{D}$ be the extensions of the spectral clusterings computed from two independent samples $x_1, ..., x_m$ and $x'_1, ..., x'_m$. Then* $\lim_{m\to\infty} d_P(\mathcal{C}, \mathcal{D}) = 0$ *in probability.*

*Proof.* (Sketch) The proof is based on techniques developed in von Luxburg et al. (2004), to where we refer for all details. The uniqueness of the first eigenfunctions implies that for large enough $m$, the first $k$ eigenvalues of $\mathcal{L}$ have multiplicity one. Denote the eigenfunctions of $\mathcal{L}$ by $f_i$, the eigenvectors based on the first sample by $v_i$, and the ones based on the second sample by $w_i$. Let $\hat{f}_i$ and $\hat{g}_i$ the extensions of those eigenvectors. In von Luxburg et al. (2004) it has been proved that $\|\hat{f}_i - f_i\|_\infty \to 0$ and $\|\hat{g}_i - f_i\|_\infty \to 0$ almost surely. Now denote by $T_{\hat{f}}$ the embedding of $X$ to $\mathbb{R}^k$ based on the functions $(\hat{f}_i)_{i=1,...,k}$, by $T_{\hat{g}}$ the one based on $(\hat{g}_i)_{i=1,...,k}$, and by $T_f$ the one based on $(f_i)_{i=1,...,k}$. Assume that we are given a fixed set of centers $c_1, ..., c_k \in \mathbb{R}^k$. By the convergence of the eigenfunctions we can conclude that $\sup_{s=1,...,k} | \, \|T_{\hat{f}}(x) - c_s\| - \|(T_f(x) - c_s\| \, | \to 0$ a.s.. In particular, this implies that if we fix a set of centers $c_1, ..., c_k$ and cluster the space $X$ based on the embeddings $T_{\hat{f}}$ and $T_f$, then the two resulting clusterings $\mathcal{C}$ and $\mathcal{D}$ of $X$ will be very similar if the sample size $m$ is large. In particular, $\sup_{i=1,...,k} P(C_i \triangle D_i) \to 0$ a.s., where $\triangle$ denotes the symmetric difference between sets. Together with Proposition 5, for a fixed set of centers this implies $d_P(\mathcal{C}, \mathcal{D}) \to 0$ almost surely. Finally we have to deal with the fact that the centers used by spectral clustering are not fixed, but are the ones computed by minimizing the $k$-means objective function on the embedded sample. Note that the convergence of the eigenvectors also implies that the $k$-means objective functions based $\hat{z}_i$ and $z_i$, respectively, are uniformly close to each other. As a consequence, the minimizers of both functions are uniformly close to each other, which by the stability results proved above leads to the desired result. □

### 6.2 Stability of the kernel-$k$-means version of spectral clustering

In this subsection we would like to consider another spectral relaxation. It can be seen that minimizing Ncut is equivalent to solving a weighted kernel-$k$-means problem with weight matrix $1/nD$ and the kernel matrix $D^{-1}WD^{-1}$ (cf. I. Dhillon, 2005). The solution of this problem can also be interpreted as a relaxation of the original problem, as we can only compute a local instead of the global minimum of the kernel-$k$-means objective function. This approximate solution usually does not coincide with the solution of the standard relaxation presented in the last section.

**Theorem 17 (Stability of kernel-$k$-means spectral clustering).** *Let the data space $X$ be compact, and the similarity function $s$ be non-negative, symmetric, continuous, and bounded away from 0. If there exists a unique optimizer of the kernel-$k$-means objective function, then the kernel-$k$-means relaxation of spectral clustering is stable.*

*Proof.* First we need to show that the sample based objective function converges to the true objective function. This is a combination of the results of von Luxburg et al. (2004) and those above. In von Luxburg et al. (2004) it has been proved that the sample based degree function converges to a continuous function $d$ on the space $X$. This implies that the weights used in the weight matrix $W = D$ converge. Then we can apply the same techniques as in the standard $k$-means setting to show the convergence of the weighted kernel-$k$-means objective function and the stability of the algorithm. □

### 6.3 Symmetry leads to instability of spectral clustering

As it is the case for center based clustering, symmetry is one of the main reasons why standard spectral clustering can be instable. In this section we would like to briefly sketch how this can be seen. Symmetry of graphs is usually described in terms of their automorphism groups (see Chan and Godsil (1997) for an overview). An automorphism of an undirected graph $G$ with vertices $x_1, ..., x_m$ and edge weights $w(x_i, x_j)$ is a surjective mapping $\phi : \{1, ..., m\} \to \{1, ..., m\}$ such that $w(x_{\phi(i)}, x_{\phi(j)}) = w(x_i, x_j)$ for all $i, j$. The set of all automorphisms of a graph forms a group, the automorphism group $Aut(G)$. It is a subgroup of the symmetric group $S_m$. It is easy to see that if $v = (v_1, ..., v_m)^t$ is an eigenvector of $L$ with eigenvalue $\lambda$, and $\phi$ a graph automorphism, then $\phi(v) := (v_{\phi(1)}, ..., v_{\phi(m)})$ is also an eigenvector of $L$ with eigenvalue $\lambda$. If $v$ and $\phi(v)$ are linearly independent, then the eigenvalue $\lambda$ will have geometric multiplicity larger than 1. This immediately leads to ambiguity: from the point of view of spectral clustering, all vectors in the eigenspace of $\lambda$ are equally suitable, as all of them have the same Rayleigh coefficient. But different eigenvectors can lead to different clusterings. As a very simple example consider the graph with 4 vertices connected as a square. The Laplacian $L$ of this graph has the eigenvalues $0, 1, 1, 2$ and the eigenvectors $v_1 = (1, 1, 1, 1)$, $v_2 = (1, 0, -1, 0)$, $v_3 = (0, 1, 0, -1)$, and

$v_4 = (-1, 1, -1, 1)$. The eigenspace of the second eigenvalue thus consists of all vectors of the form $(a, b, -a, -b)$. The spectral embedding based on this eigenvector maps the data points to $\mathbb{R}$ by $x_1 \mapsto a$, $x_2 \mapsto b$, $x_3 \mapsto -a$, and $x_4 \mapsto -b$. The centers construcuted by $k$-means are then either $\pm(a + b)/2$ or $\pm(a - b)/2$, depending on whether $a$ and $b$ have the same sign or not. In the first case, the resulting clustering is $\{x_1, x_2\}, \{x_3, x_4\}$, in the second case it is $\{x_1, x_3\}, \{x_2, x_4\}$. Thus we obtain the two completely symmetric solutions of spectral clustering which we would expect from the square symmetry of the data points.

Now let us consider the underlying data space $X$. The role of automorphisms is now played by measure preserving symmetries as defined above. Assume that $(X, P)$ possesses such a symmetry. Of course, even if $X$ is symmetric, the similarity graph based on a finite sample drawn from $X$ usually will not be perfectly symmetric. However, if the sample size is large enough, it will be "nearly symmetric". It can be seen by perturbation theory that the resulting eigenvalues and eigenvectors will be "nearly" the same ones as resulting from a perfectly symmetric graph. In particular, which eigenvectors exactly will be used by the spectral embedding will only depend on small perturbations in the sample. This will exactly lead to the unstable situation sketched above.

## 7 Conclusions

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. In this work, we have set forward formal definitions for stability and some related clustering notions and used this framework to provide theoretical analysis of stability. Our results show that stability is determined by the structure of the set of optimal solutions to the risk minimization objective. Namely, the existence of a unique minimizer implies stability, and the existence of a symmetry permuting such minimizers implies instability. These results indicate that, contrary to common belief (and practice), stability does 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. Furthermore, our results reduce the problem of stability estimation to concrete geometric and optimization properties of the data distribution. In this paper we prove these results for a wide class of center based and spectral clustering algorithms.

It would be interesting to investigate similar questions with respect to other popular clustering paradigms. Another intriguing issue is to try to figure out what features of real life data make stability successful as a clustering validation tool in practice. As shown in this paper, by our results and examples, stability is not the right tool for such purposes. The success of stability in choosing number of clusters should be viewed as an exception rather than the rule.

# Bibliography

S. Ben-David. A framework for statistical clustering with constant time approximation algorithms for k-median clustering. In J. Shawe-Taylor and Y. Singer, editors, *Proceedings of the 17th Annual Conference on Learning Theory (COLT)*, pages 415–426. Springer, 2004.

A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing*, 2002.

O. Bousquet and A. Elisseeff. Stability and generalization. *JMLR*, 2(3):499–526, 2002.

A. Chan and C. Godsil. Symmetry and eigenvectors. In G. Hahn and G. Sabidussi, editors, *Graph Symmetry, Algebraic Methods and Applications*. Kluwer, 1997.

B. Kulis I. Dhillon, Y. Guan. A unified view of kernel $k$-means, spectral clustering, and graph partitioning. Technical Report TR-04-25, UTCS Technical Report, 2005.

S. Kutin and P. Niyogi. Almost-everywhere algorithmic stability and generalization error. Technical report, TR-2002-03, University of of Chicago, 2002.

T. Lange, V. Roth, M. Braun, and J.Buhmann. Stability-based validation of clustering solutions. *Neural Computation*, 2004.

A. Rakhlin and A. Caponnetto. Stability properties of empirical risk minimization over donsker classes. Technical report, MIT AI Memo 2005-018, 2005.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. Technical Report 134, Max Planck Institute for Biological Cybernetics, 2004.

U. von Luxburg and S. Ben-David. Towards a statistical theory of clustering. In *PASCAL workshop on Statistics and Optimization of Clustering*, 2005.