Relating clustering stability to properties of cluster boundaries

Shai Ben-David
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
shai@cs.uwaterloo.ca

Ulrike von Luxburg
Max Planck Institute for Biological Cybernetics
Tübingen, Germany
ulrike.luxburg@tuebingen.mpg.de

Abstract

In this paper, we investigate stability-based methods for cluster model selection, in particular to select the number K of clusters. The scenario under consideration is that clustering is performed by minimizing a certain clustering quality function, and that a unique global minimizer exists. On the one hand we show that stability can be upper bounded by certain properties of the optimal clustering, namely by the mass in a small tube around the cluster boundaries. On the other hand, we provide counterexamples which show that a reverse statement is not true in general. Finally, we give some examples and arguments why, from a theoretic point of view, using clustering stability in a high sample setting can be problematic. It can be seen that distribution-free guarantees bounding the difference between the finite sample stability and the "true stability" cannot exist, unless one makes strong assumptions on the underlying distribution.

1 Introduction

In the domain of data clustering, the problem of model selection is one of the most difficult challenges. In particular the question of selecting the number of clusters has drawn a lot of attention in the literature. A very popular method to solve this problem is to use a stability-based approach. The overall idea is that a clustering algorithm with a certain setting of parameters is meaningful for a given input data if it produces "stable" results, that is, inputs similar to that data lead to similar clustering results. The other way round, an algorithm which is unstable cannot be trusted. This argument is then turned into a model selection criterion: to determine a "good" number K of clusters on a particular data set, one runs a clustering algorithm with different choices of K on many perturbed versions of that data set and selects the parameter K where the algorithm gives the most stable result.

This stability approach has been implemented in various different ways (e.g., Levine and Domany, 2001, Ben-Hur et al., 2002, Lange et al., 2004, Smolkin and Ghosh, 2003) and gains more and more influence in applications, for example in the domain of bioinformatics (Bittner et al., 2000, Fridlyand and Dudoit, 2001, Kerr and Churchill, 2001, Bertoni and Valentini, 2007). However, its theoretical foundations are not yet well understood. While it is a reasonable requirement that an algorithm should demonstrate stability in general, it is not obvious that, among several stable algorithms, the one which is most stable leads to the best performance.

One important situation has been analyzed in Ben-David et al. (2006) and Ben-David et al. (2007). There it has been proved that in a setting where clustering is performed by globally minimizing an objective function, clustering stability can be characterized by simple properties of the underlying objective function. Namely, given a data set from some particular input distribution, a clustering algorithm is stable for this distribution for large sample sizes if and only if its objective function has a unique global minimizer for that input. As several counter-examples have shown, the latter property is not necessarily related to the fact that the algorithm constructs the correct number of clusters. Some examples for this behavior have also been given in Krieger and Green (1999) (but without rigorous analysis). The dilemma worked out by Ben-David et al. (2006) and Ben-David et al. (2007) is not so much that algorithms get unstable in case of multiple global optima, but the fact that all algorithms with unique global optima are stable. That is, for large sample size (in)stability converges to the same value 0, no matter what parameter K we choose. This result suggests that for large sample size, stability criteria are unsuitable for model selection.

While this looks like a very negative result on the first glance, recent follow-up work by Shamir and Tishby (2008b) and Shamir and Tishby (2008a) indicates a possible way out of this trap. In a simple situation where the data is distributed according to well separated, univariate Gaussians, the authors show that even though the K-means algorithm is stable for many values of K, the rate of convergence of a rescaled measure of stability behaves differently for different numbers of clusters. In this example, the authors show that a model selection criterion based on stability can be used to select the correct number of clusters. The difference to the approach considered in Ben-David et al. (2006) and Ben-David et al. (2007) is that the scaling constant in the definition of stability is chosen as $1/\sqrt{n}$ rather than $1/n$. Hence, the authors consider a central limit theorem setting rather

than a law of large numbers. In the central limit theorem setting, they show that stability does not necessarily converge to 0, but to some normal distribution with particular parameters. Intuitively this means that stability behaves like $c(K)/\sqrt{n}$ where the constant $c(K)$ depends (in some complicated way) on the number K of clusters. In the simple univariate mixture of Gaussian settings studied in Shamir and Tishby (2008b) and Shamir and Tishby (2008a), this constant is higher for the "incorrect" parameter choice. This work indicates that even for large sample size, stability criteria might be useful for model selection after all. It remains to be seen whether this approach can successfully extended to more complex data scenarios reflecting real world data.

The work of Shamir and Tishby (2008b) shows how stability might be used to select the number of clusters in the setting of large sample size and unique global optimizer. However, one crucial question still remains unanswered: what is it really that stability reflects, how will stable clusterings look like in general, and what properties will they have? This is the direction we want to take in our current paper. The general setup is similar to the one discussed above, that is we study clustering algorithms which minimize a certain clustering quality function. As the other case has already been treated completely in Ben-David et al. (2006) and Ben-David et al. (2007), we are now solely concerned with the setting where the clustering quality function has one unique global optimizer. Our goal is to relate the stability of clustering algorithms (on finite sample sizes) to properties of the optimal data clustering itself.

One candidate for such a relation is the conjecture that in the large sample regime, differences in stability of clustering algorithms can be explained by whether the cluster boundaries of the optimal clustering of the underlying space lie in a low or a high density areas of the underlying space. The conjecture is that if the boundaries are in low density areas of the space, an algorithm which constructs clusterings sufficiently close to the optimal clustering will be stable. The other way round, we expect it to be more unstable if the decision boundaries of the optimal clustering are in a high density area. The intuition behind this conjecture is simple: if the decision boundary is in a low density area of the space, small perturbations of the samples might move the boundary a bit, but this movement of the boundary will only affect the cluster labels of very few points (as there are not many points close to the boundary). On the other hand, if the boundary is in a high density area, even small perturbations in the samples will change the cluster assignments of many data points. If this conjecture were true, it would have a very large impact on understanding the mechanism of stability-based model selection.

In this paper, we first prove one direction of this conjecture: the quantitative value of stability can be upper bounded by the mass in a small tube around the optimal clustering boundary. Such a statement has already been implicitly used in Shamir and Tishby (2008b), but only in a very simple one-dimensional setting where the cluster boundary just consists of one single point. The challenge is to prove this statement in a more general, multidimensional setting.

Unfortunately, it turns out that the opposite direction of the conjecture does not hold. In general, there can be clusterings whose decision boundary lies in a high density area, but we have high stability. We demonstrate this fact with counterexamples which also shed light on the reasons for the failure of this direction of the conjecture.

Finally, we end our paper with a few cautionary thoughts about using stability in large sample scenarios. Essentially, we argue that even if one found satisfactory reasons which explain why a certain clustering tends to be more stable than an other one, such statements are not very useful for drawing conclusions about stability measures of any given *finite* sample size. The reason is that as opposed to the standard statistical learning theory settings, there cannot exist uniform convergence bounds for stability. Thus there is no way one can state any theoretical guarantees on the decisions based on stability for any fixed sample size, unless one makes very strong assumptions on the underlying data distributions.

2 Notation and ingredients

2.1 General setup

Let $(\mathcal{X}, d)$ denote an arbitrary metric space. For convenience, in the following we will always assume that $\mathcal{X}$ is compact. By $\operatorname{diam} \mathcal{X} := \max_{x,y \in \mathcal{X}} d(x, y)$ we denote the diameter of the space. The space of all probability measures on $\mathcal{X}$ (with respect to the Borel σ-algebra) is denoted by $M_1(\mathcal{X})$. Let P be a fixed probability measure on $\mathcal{X}$, and $X_1, ..., X_n$ a sample of points drawn i.i.d. from $\mathcal{X}$ according to P. The empirical measure of this sample will be denoted by P_n.

Let F be a set of admissible clustering functions of the form $f : \mathcal{X} \to \{1, ..., K\}$, where $K \in \mathbb{N}$ denotes the number of clusters. In the following, we will consider clusterings with respect to the equivalence relation of renaming the cluster labels. Namely, define the equivalence relation $\sim$ on F by

$$f \sim g : \Longleftrightarrow \exists \pi : \ f(x) = \pi(g(x))$$

where π is a permutation of the set $\{1, ..., K\}$. Denote by $\mathcal{F} := F/\sim$ the space of equivalence classes of this relation. This will be the space of clusterings we will work with. To perform clustering, we will rely on a clustering quality function $Q : \mathcal{F} \times M_1(\mathcal{X}) \to \mathbb{R}$. The optimal "true" clustering of $\mathcal{X}$ with respect to P is defined as

$$f^* := \operatorname*{argmin}_{f \in \mathcal{F}} Q(f, P).$$

Throughout this paper we will assume that f^* is the unique global optimizer of Q. If this is not the case, it has already been proved that the corresponding clustering algorithm is not stable anyway (Ben-David et al., 2006, 2007).

When working on a finite sample, we will use an empirical quality function $Q_n : \mathcal{F} \times M_1(\mathcal{X}) \to \mathbb{R}$. We consider the clustering algorithm which, on any given sample, selects the clustering f_n by

$$f_n := \operatorname*{argmin}_{f \in \mathcal{F}} Q_n(f, P_n).$$

Note that implicit in this formulation, one makes the assumption that the clustering algorithm is able to detect the global minimum of Q_n. Of course, this is not the case for many commonly used clustering algorithms. For example, the standard K-means algorithm is not guaranteed to do so. Even though in applications, experience shows that the K-means algorithm is reasonably successful on "well-clustered" data sets, to get provable guarantees one has to revert to other algorithms, such as the nearest neighbor clustering introduced in von Luxburg et al. (2008) or approximation schemes such as the one introduced in Ostrovsky et al. (2006).

In the following, we will only deal with clustering algorithms which are statistically consistent, that is $Q(f_n, P) \to Q(f^*, P)$ in probability. It has been proved that minimizing well-known objective functions such as the one used by K-means or the normalized cut used in spectral clustering can be performed consistently (von Luxburg et al., 2008).

For two independent samples $\{X_1, ..., X_n\}$ and $\{X'_1, ..., X'_n\}$ denote the clustering solutions based on minimizing a quality function Q_n by f_n and f'_n, respectively. For a given distance function $D : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ which measures some kind of distance between clusterings, the instability of the clustering algorithm minimizing the quality function Q based on sample size n is defined as

$$\mathrm{InStab}_D(Q, n, P) := \mathbb{E}\left(D(f_n, f'_n)\right)$$

where the expectation is over the random drawing of the two samples. So, the stability (or instability) is a function of several quantities: the input data distribution P, the clustering algorithm (defined by the quality function Q that the algorithm optimizes), the sample size n, and the clustering distance measure used. Unless otherwise mentioned, we shall be using the minimal matching distance (see below) for the definition of instability and drop the subscript D in the instability notation. Also, if it is clear which objective function Q we refer to, we drop the dependence on Q, too, and simply write $\mathrm{InStab}(n, P)$ for instability.

2.2 Distance functions between clusterings

Various measures of clustering distances have been used and analyzed in the literature (see for example Meila, 2005). We define below two measures that our most relevant to our discussion.

Minimal matching distance. This is perhaps the most widely used distance between clusterings. For two clusterings defined on a finite point set $X_1, ..., X_n$, this distance is defined as

$$D_{\mathrm{MinMatch}}(f_n, f'_n) := \min_{\pi} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{f(X_i) \neq \pi(g(X_i))}$$

where the minimum is taken over all permutations π of the set $\{1, ..., K\}$. This distance is close in spirit to the 0-1-loss used in classification. It is well known that D_{MinMatch} is a metric, and that it can be computed efficiently using a minimal bipartite matching algorithm.

A distance based on cluster boundaries. For our current work, we need to introduce a completely new distance between clusterings. Intuitively, this distance measures how far the class boundaries of two clusterings are away from each other. Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^s$, d a metric on $\mathbb{R}^s$ such as the Euclidean one, and $\mathcal{F}$ the space of all clustering functions $f : \mathcal{X} \to \{1, ..., K\}$, up to the equivalence relation $\sim$. For a given $f \in \mathcal{F}$, we define the **boundary** $B(f)$ of f as the set

$$B(f) := \{x \in \mathcal{X} \mid f \text{ discontinuous at } x\}.$$

The distance of a point x to the boundary $B(f)$ is defined as usual by

$$d(x, B(f)) := \inf\{d(x, y) \mid y \in B(f)\}.$$

For $\gamma > 0$, we then we define the **tube** $T_\gamma(f)$ as the set

$$T_\gamma(f) := \{x \in \mathcal{X} \mid d(x, B(f)) \leq \gamma\}.$$

For $\gamma = 0$ we set $T_0(f) = B(f)$.

We say that a clustering function g is in the γ-tube of f, written $g \lhd T_\gamma(f)$, if

$$\forall x, y \notin T_\gamma(f) : f(x) = f(y) \iff g(x) = g(y).$$

Finally, we define the distance function D_{boundary} on $\mathcal{F}$ as

$$D_{\text{boundary}}(f, g) := \inf_{\gamma > 0}\{f \lhd T_\gamma(g) \text{ and } g \lhd T_\gamma(f)\}.$$

The distance D_{boundary} satisfies several nice properties:

Proposition 1 (Properties of D_{boundary}) *Assume that the metric space $\mathcal{X} \subset \mathbb{R}^s$ is compact. Let $\mathcal{F}$ be the set of equivalence classes of clustering functions $f : \mathcal{X} \to \{1, ..., K\}$ as defined above. Then the following technical properties hold:*

1. *$D_{boundary}$ is well-defined on the equivalence classes.*
2. *Let $f, g \in \mathcal{F}$. Then: $g \lhd T_\gamma(f)$ implies that $B(g) \subset T_\gamma(f)$.*
3. *Let f, g two clusterings with $D_{boundary}(f, g) \leq \gamma$. Then there exists a permutation π such that for all $x \in \mathcal{X}$,*

$$f(x) \neq \pi(g(x)) \implies x \in T_\gamma(g).$$

Furthermore, the following fundamental properties hold:

5. *The distance function $D_{boundary}$ is a metric on $\mathcal{F}$.*
6. *$\mathcal{F}$ is relatively compact under the topology induced by $D_{boundary}$.*

Proof.

1. The definitions of all quantities above do not depend on the particular labeling of the clusters, but only on the positions of the cluster boundaries.

2. Let $g \lhd T_\gamma(f)$, but assume that $B(g) \not\subseteq T_\gamma(f)$. That is, there exists a point $x \in B(g)$ with $x \notin T_\gamma(f)$. By definition of $B(g)$, x is a point of discontinuity of g, thus the clustering g changes its label at x. On the other hand, by the definition of $T_\gamma(f)$, f does not change its label at x (otherwise, x would be in $B(f) \subset T_\gamma(f)$). But the latter contradicts the definition of $g \lhd T_\gamma(f)$ which requires that f and g only change their labels at the same points outside of $T_\gamma(f)$. Contradiction.

3. Similar to Part 2.

4. $D_{\text{boundary}}(f,g) \leq \operatorname{diam}\mathcal{X} < \infty$: As $\mathcal{X}$ is compact, it has a finite diameter $\operatorname{diam}\mathcal{X}$. Then for all $f, g \in \mathcal{F}$ we have $T_{\operatorname{diam}\mathcal{X}}(f) = \mathcal{X}$ and $T_{\operatorname{diam}\mathcal{X}}(g) = \mathcal{X}$. Thus, trivially $f \triangleleft T_{\operatorname{diam}\mathcal{X}}(g)$ and vice versa, that is $D_{\text{boundary}}(f,g) \leq \operatorname{diam}\mathcal{X}$.

   $D_{\text{boundary}}(f,g) \geq 0$: clear.

   $D_{\text{boundary}}(f,f) = 0$: clear.

   $D_{\text{boundary}}(f,g) = 0 \implies f = g$: $D_{\text{boundary}}(f,g) = 0$ implies that $B(f) \subset T_0(g) = B(g)$ and vice versa, thus we have $B(f) = B(g)$. So the class boundaries of both clusterings coincide. Moreover, we have that for all $x, y \notin B(g)$, $f(x) = f(y) \iff g(x) = g(y)$. Thus there exists a permutation of the labeling of g such that $f(x) = \pi(g(x))$ for all $x \notin B(g)$. Thus f and g are in the same equivalence class with respect to $\sim$, that is $f = g$ in the space $\mathcal{F}$.

   Triangle inequality: assume that $D_{\text{boundary}}(f,g) = \gamma_1$ and $D_{\text{boundary}}(g,h) = \gamma_2$, that is

$$\begin{aligned}
&\forall x, y \notin T_{\gamma_1}(f) : [f(x) = f(y) \iff g(x) = g(y)] \\
&\forall x, y \notin T_{\gamma_1}(g) : [f(x) = f(y) \iff g(x) = g(y)] \\
&\forall x, y \notin T_{\gamma_2}(g) : [h(x) = h(y) \iff g(x) = g(y)] \\
&\forall x, y \notin T_{\gamma_2}(h) : [h(x) = h(y) \iff g(x) = g(y)].
\end{aligned} \tag{1}$$

   Now define $\gamma := \gamma_1 + \gamma_2$. We first need to prove a small sub-statement, namely that

$$x \notin T_\gamma(f) \implies x \notin T_{\gamma_2}(g). \tag{2}$$

   To this end, let $x \in T_{\gamma_2}(g)$, that is there exists some point $y \in B(g)$ with $d(x,y) \leq \gamma_2$. As we know that $g \triangleleft T_{\gamma_1}(f)$, we also have $B(g) \subset T_{\gamma_1}(f)$, that is for all $y \in B(g)$ exists $z \in B(f)$ such that $d(y,z) \leq \varepsilon_1$. Combining those two statements and using the triangle inequality for the metric d on the original space $\mathcal{X}$, we can conclude that $d(x,z) \leq d(x,y) + d(y,z) = \gamma_1 + \gamma_2 = \gamma$, that is $x \in T_\gamma(f)$. This shows statement (2) by its contra-position. Now we can go ahead and prove the triangle inequality for D_{boundary}. Using the property (2) and the equations (1) we get that

$$\begin{aligned}
x, y \notin T_\gamma(f) &\implies x, y \notin T_{\gamma_2}(g) \\
&\implies [g(x) = g(y) \iff h(x) = h(y)].
\end{aligned}$$

   Moreover, by the definition of $T_\gamma(f)$ and the fact that $\gamma \geq \gamma_1$ we trivially have that $x, y \notin T_\gamma(f)$ implies $x, y \notin T_{\gamma_1}(f)$. Together with equations (1) this leads to

$$\begin{aligned}
x, y \notin T_\gamma(f) &\implies x, y \notin T_{\gamma_1}(f) \\
&\implies [g(x) = g(y) \iff f(x) = f(y)].
\end{aligned}$$

   Combining those two statements we get

$$x, y \notin T_\gamma(f) \implies [f(x) = f(y) \iff h(x) = h(y)],$$

   that is $h \triangleleft T_\gamma(f)$. Similarly we can prove that $f \triangleleft T_\gamma(h)$, that is we get $D_{\text{boundary}}(f,h) \leq \gamma$. This proves the triangle inequality.

   All statements together prove that D_{boundary} is a metric.

5. By the theorem of Heine-Borel, a metric space is relatively compact if it is totally bounded, that is for any $\gamma > 0$ it can be covered with finitely many γ-balls. By assumption, we know that $\mathcal{X}$ is compact. Thus we can construct a finite covering of balls of size γ of $\mathcal{X}$ (in the metric d). Denote the centers of the covering balls as $x_1, ..., x_s$. We want to use this covering to construct a finite covering of $\mathcal{F}$. To this end, let $f \in \mathcal{F}$ be an arbitrary function (for now let us fix a labeling, we will go over to the equivalence class in the end). Given f, we reorder the centers of the covering balls such that all centers x_i with $x_i \notin T_{2\gamma}(f)$ come in the ordering before the points x_j with $x_j \in T_{2\gamma}(f)$, that is:

$$x_i \notin T_{2\gamma}(f) \text{ and } x_j \in T_{2\gamma}(f) \implies i < j.$$

   Now we construct a clustering $\tilde{f}$ as follows: one after the other, in the ordering determined before, we color the balls of the covering according to the color $f(x_i)$ of its center, that is we set:

   - $x \in B(x_1) \implies \tilde{f}(x) := f(x_1)$
   - $x \in B(x_2) \setminus B(x_1) \implies \tilde{f}(x) := f(x_2)$

   ...

   - $x \in B(x_i) \setminus \cup_{t=1,...,i-1} B(x_t) : \tilde{f}(x) := f(x_i)$

   By construction, for all points $x \notin T_\gamma(f)$ we have $\tilde{f}(x) = f(x)$. Consequently, $\tilde{f} \triangleleft T_\gamma(f)$. Similarly, the other way round we have $f \triangleleft T_\gamma(\tilde{f})$. Thus, $D_{\text{boundary}}(f,\tilde{f}) \leq \gamma$. Note that given two representatives f, g of the same clustering in $\mathcal{F}$ (that is, two functions such that $f = \pi(g)$ for some permutation π), the corresponding functions $\tilde{f}$ and $\tilde{g}$ are also representatives of the same clustering, that is $\tilde{f} = \pi(\tilde{g})$. Thus the whole construction is well-defined on $\mathcal{F}$.

   Finally, it is clear that the set $\tilde{\mathcal{F}} := \{\tilde{f} \mid f \in \mathcal{F}\}$ has finitely many elements: there only exist finitely many orderings of the s center points $x_1, ..., x_s$ and finitely many labelings of those center points using K labels. Hence, the set $\tilde{\mathcal{F}}$ forms a finite γ-covering of $\mathcal{F}$.

   ☺

In the current paper, we will only use the distance D_{boundary} for clusterings of $\mathbb{R}^s$, but its construction is very general. The distance D_{boundary} can also be defined on more general metric spaces, and even discrete spaces. One just has to give up defining $B(f)$ and directly define the set $T_\gamma(f)$ as the set $\{x \in \mathcal{X} \mid \exists y \in \mathcal{X} : f(x) \neq f(y) \text{ and } d(x,y) \leq \varepsilon\}$. However, in that case, some care has to be taken when dealing with "empty regions" of the space.

3 Upper bounding stability by the mass in γ-tubes

In this section we want to establish a simple, but potentially powerful insight: given any input data distribution, P, for large enough n, the stability of a quality-optimizing consistent clustering algorithm can be described in terms of the P-mass of along the decision boundaries of the optimal clustering. The intuition is as follows. The distance D_{MinMatch} counts the number of points for which two clusterings do not coincide, that is it counts the number of points which lie "between" the decision boundaries of the two clusterings. Stability is the expectation over D_{MinMatch}, computed on different random samples.

3.1 Relation between stability and tubes

Let us first assume that we know that with high probability over the random drawing of samples, we have that $D_{\text{boundary}}(f_n, f) \leq \gamma$ for some constant γ. Then the following proposition holds:

Proposition 2 (Relating stability and mass in tubes) *Let f be any fixed clustering, and f_n the clustering computed from a random sample of size n. Assume that with probability at least $1-\delta$ over the random samples, we have that $D_{boundary}(f_n, f) \leq \gamma$. Then the instability (based on distance $D_{MinMatch}$) satisfies*

$$\text{InStab}(n, P) \leq 2\delta + 2P(T_\gamma(f)).$$

Proof. Denote the set of samples on which the event $D_{\text{boundary}}(f_n, f) \leq \gamma$ is true by M. W.l.o.g. assume that for all n, the labels of the clustering f_n are chosen such that they already coincide with the ones of f, that is the permutation for which the minimum in $D_{\text{MinMatch}}(f_n, f)$ is attained is the identity. Then we have:

$$\begin{aligned}
&\text{InStab}(n, P) = \mathbb{E}(D_{\text{MinMatch}}(f_n, f'_n)) \\
&\leq \mathbb{E}(D_{\text{MinMatch}}(f_n, f) + D_{\text{MinMatch}}(f'_n, f)) \\
&= 2\mathbb{E}D_{\text{MinMatch}}(f_n, f) \\
&= 2\int_M \mathbf{1}_{f_n(X)\neq f(X)}\, dP(X) + 2\int_{M^c} \mathbf{1}_{f_n(X)\neq f(X)}\, dP(X) \\
&\quad (\text{ on M, } f_n(x) \neq f(x) \implies x \in T_\gamma(f), \text{ see Prop. 1}) \\
&\leq 2\int_M \mathbf{1}_{X\in T_\gamma(f)}\, dP(X) \quad + 2P(M^c) \\
&= 2P(T_\gamma(f) + 2\delta
\end{aligned}$$

☺

Proposition 2 gives several very plausible reasons for why a clustering can be unstable:

- The decision boundaries themselves vary a lot (i.e., γ is large). This case is pretty obvious.
- The decision boundaries do not vary so much (i.e., γ is small), but lie in an area of high density. This is a more subtle reason, but a very valuable one. It suggests that if we compare two clusterings, one of them has its cluster boundary in a high density area and the other one in a low density area, then the first one tends to be more unstable than the second one. However, to formally analyze such a comparison between stability values of different algorithms, one also has to prove a lower bound on stability, see later.
- The decision boundaries do not vary so much (i.e., γ is small), are in a region of moderate density, but they are very long, so significant mass accumulates along the boundary.

3.2 Determining the width γ in terms of the limit clustering

Now we want to apply the insight from the last subsection to relate properties of the optimal clustering to stability. In this section, we still want to work in an abstract setting, without fixing a particular clustering objective function. In order to prove our results, we will have to make a few crucial assumptions:

- The objective function Q has a unique global minimum. Otherwise we know by Ben-David et al. (2006) and Ben-David et al. (2007) that the algorithm will not be stable anyway.
- The clustering algorithm is consistent, that is $Q(f_n, P) \to Q(f^*, P)$ in probability. If this assumption is not true, any statement about the stability on a finite sample is pretty meaningless, as the algorithm can change its mind with the sample size. For example, consider the trivial algorithm which returns a fixed function f_1 if the sample size n is even, and another fixed function f_2 if the sample size is odd. This algorithm is perfectly stable for every n, but since the results do not converge, it is completely meaningless.
- The sample size n is sufficiently large so that $Q(f_n) - Q(f^*)$ is sufficiently small: f_n is inside the region of attraction of the global minimum. With this assumption we want to exclude trivial cases where instability is induced due to too high sample fluctuations. See also Section 5 for discussion.

To state the following proposition, we recall the definition of a quasi-inverse of a function. The quasi-inverse of a function is a generalization of the inverse of a function to cases where the function is not injective. Let $f : \mathcal{X} \to \mathcal{Y}$ be a function with range $\text{rg}(f) \subset \mathcal{Y}$. A function $g : \text{rg}(f) \to \mathcal{X}$ which satisfies $f \circ g \circ f = f$ is called a quasi-inverse of f. Note that quasi-inverses are not unique, unless the function f is injective.

Proposition 3 (Consequences of unique global optimum) *Let $(\mathcal{X}, d)$ a compact metric space with probability distribution P, and $\mathcal{F}$ the space of P-measurable clusterings with K clusters on $\mathcal{X}$. As a topology on $\mathcal{F}$, consider the one induced by the distance $D_{boundary}$. Let $Q := Q(\cdot, P) : \mathcal{F} \to \mathbb{R}$ be continuous and assume that it has a unique global minimizer f^*. Then, every quasi-inverse $Q^{-1} : \text{rg}(Q) \subset \mathbb{R} \to \mathcal{F}$ is continuous at $Q(f^*)$. In particular, for all $\gamma > 0$ there exists some $\varepsilon(\gamma, f^*, P) > 0$ such that for all $f \in \mathcal{F}$,*

$$|Q(f, P) - Q(f^*, P)| \leq \varepsilon \implies D_{boundary}(f, f^*) \leq \gamma. \tag{3}$$

Proof. Assume Q^{-1} is not continuous at $Q(f^*)$, that is there exists a sequence of functions $(g_n)_n \subset \mathcal{F}$ such that $Q(g_n) \to Q(f^*)$ but $g_n \not\to f^*$. By the compactness assumption, the sequence $(g_n)_n$ has a convergent subsequence $(f_{n_k})_k$ with $f_{n_k} \to \tilde{f}$ for some $\tilde{f} \in \mathcal{F}$. Also by assumption, we can find such a subsequence such that $\tilde{f} \neq f^*$. By the continuity of Q we know that $Q(f_{n_k}) \to Q(\tilde{f})$, and by the definition of $(g_n)_n$ we know also that $Q(f_{n_k}) \to Q(f^*)$. So we know that $Q(f^*) = Q(\tilde{f})$, and by the uniqueness of the optimum f^* this leads to $f^* = \tilde{f}$. Contradiction. ☺

Note that the "geometry of Q" plays an important role in this proposition. In particular, the size of the constant ε heavily depends on the "steepness" of Q in a neighborhood of the global optimum and on "how unique" the global optimum is. We formalize this by introducing the following quantity:

$$U_P^Q(\gamma) := \sup\Big\{\varepsilon > 0 : |Q(f,P) - Q(f^*,P)| \leq \varepsilon \implies D_{\text{boundary}}(f, f^*) \leq \gamma\Big\}.$$

One can think of U_P^Q as indicating how unique is the optimal clustering f^* of P is.

The following theorem bounds the stability of a clustering algorithm on a given input data distribution by the mass it has in the tube around the decision boundary. It replaces the assumption of uniform convergence of the empirical clusterings under the D_{boundary} metric of Proposition 2 by the more intuitive assumption that the underlying clustering algorithm is *uniformly consistent*. That is, $Q(f_n, P) \to Q(f^*, P)$ in probability, uniformly over all probability distributions P:

$$\forall \varepsilon > 0\ \forall \delta > 0\ \exists n \in \mathbb{N}\ \forall P : P(|Q(f_n,P) - Q(f^*,P)| > \varepsilon) \leq \delta.$$

In particular, for any positive ε and δ, the required sample size n does not depend on P. Such an assumption holds, for example, for the algorithm constructing the global minimum of the K-means objective function, as shown by Ben-David (2007). For background reading on consistency of clustering algorithms and bounds for many types of objective function see von Luxburg et al. (2008). When such uniform consistency holds for Q, let us quantify the sample size by defining

$$C_Q(\varepsilon, \delta) := \min\Big\{m \in \mathbb{N} : \forall P\ \forall n \geq m\ \ P(|Q(f_n,P) - Q(f^*,P)| > \varepsilon) \leq \delta\Big\}.$$

We can now provide a bound on stability which refers to the following quantities: the uniqueness U_P^Q of the optimal clustering, the consistency C_Q of the quality measure, and the P-weight of the tubes around the optimal clustering of the input data distribution.

Theorem 4 (High instability implies cut in high density region) *Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^s$, Assume that the cluster quality function $Q(\cdot, P) : \mathcal{F} \to \mathbb{R}$ is continuous with respect to the topology on $\mathcal{F}$ induced by $D_{boundary}$. Let $Q(\cdot, P)$ have a unique global minimizer f^*, and assume that $Q(\cdot, P)$ can be minimized uniformly consistently, Then, for all $\gamma > 0$ and for all $\delta > 0$,*
if

$$n \geq C_Q(U_P^Q(\gamma), \delta)$$

then

$$\text{InStab}(n, P) \leq 2\delta + 2P(T_\gamma(f^*)).$$

Proof. By definition of C_Q we know that if $n \geq C_Q(U_P^Q(\gamma), \delta)$ then we have that

$$P(|Q(f_n,P) - Q(f^*,P)| \leq U_P^Q(\gamma)) > 1 - \delta.$$

By definition of $U_P^Q(\gamma)$ we know that if $|Q(f_n,P) - Q(f^*,P)| \leq U_P^Q(\gamma)$, then we have that $D_{\text{boundary}}(f_n, f^*) \leq \gamma$. Together this means that whenever $n \geq C_Q(U_P^Q(\gamma), \delta)$ then with probability at least $1 - \delta$ we have that $D_{\text{boundary}}(f_n, f^*) \leq \gamma$. Now the statement of the theorem follows by Proposition 2. ☺

3.3 Application to particular objective functions

In this subsection we briefly want to show that the conditions in Theorem 4 are satisfied for many of the commonly used clustering quality functions. The major conditions to investigate are the consistency condition and the condition that Q is continuous with respect to D_{boundary} on $\mathcal{F}$.

K-means objective function. The empirical K-means objective function Q_n on a finite sample of n points is defined as

$$Q_n(f) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \mathbf{1}_{f(X_i)=k}\|X_i - c_k\|^2$$

where c_i denote the cluster centers. Its continuous counterpart is the quality function

$$Q(f) = \int \sum_{k=1}^{K} \mathbf{1}_{f(X)=k}\|X - c_k\|^2\, dP(X).$$

Assume that on any finite sample, the clustering algorithm returns the global optimizer of the empirical K-means function. Then it is known that this empirical optimizer converges to the true optimum uniformly over all probability distributions (e.g., Corollary 8 in Ben-David, 2007). (However, note that this guarantee does not apply to the standard K-means algorithm, which only constructs local optima of the empirical quality function.)
Moreover, the K-means objective function is continuous with respect to D_{boundary}, as can be seen by the following proposition:

Proposition 5 (Continuity of K-means wrt. D_{boundary}) *Let $\mathcal{X} \subset \mathbb{R}^s$ compact, and P a probability distribution on $\mathcal{X}$ with a density with respect to the Lebesgue measure. Then the K-means quality function Q is continuous with respect to $D_{boundary}$.*

Proof. Assume f and g are two K-means clusterings with distance $D_{\text{boundary}}(f,g) \leq \gamma$. W.l.o.g. assume that the labeling of g is permuted such that outside of the γ-tubes, the labels of f and g coincide. Denote the complement of a set T by T^c. Then we can compute:

$$\begin{aligned}
Q(g) &= \int \sum_{k=1}^{K} \mathbf{1}_{g(X)=k} \|X - c_k(g)\|^2 \, dP(X) \\
&\leq \int \sum_{k=1}^{K} \mathbf{1}_{g(X)=k} \|X - c_k(f)\|^2 \, dP(X) \\
&= \int_{T_\gamma(f)^c} \sum_{k=1}^{K} \mathbf{1}_{g(X)=k} \|X - c_k(f)\|^2 \, dP(X) \\
&\quad + \int_{T_\gamma(f)} \sum_{k=1}^{K} \mathbf{1}_{g(X)=k} \|X - c_k(f)\|^2 \, dP(X) \\
&\qquad \text{(now on } T_\gamma(f)^c : f(X) = k \iff g(X) = k) \\
&= \int_{T_\gamma(f)^c} \sum_{k=1}^{K} \mathbf{1}_{f(X)=k} \|X - c_k(f)\|^2 \, dP(X) \\
&\quad + \int_{T_\gamma(f)} \sum_{k=1}^{K} \mathbf{1}_{g(X)=k} \|X - c_k(f)\|^2 \, dP(X) \\
&\leq Q(f) + \operatorname{diam}(\mathcal{X})^2 \cdot P(T_\gamma(f)).
\end{aligned}$$

By the symmetry in f and g this leads to

$$|Q(g) - Q(f)| \leq \operatorname{diam}(\mathcal{X})^2 \cdot \max\{P(T_\gamma(f)), P(T_\gamma(g))\}.$$

Finally, the assumption $g \triangleleft T_\gamma(f)$ implies that $T_\gamma(g) \subset T_{2\gamma}(f)$. Thus we finally get that

$$|Q(f) - Q(g)| \leq \operatorname{diam}(\mathcal{X})^2 \cdot P(T_{2\gamma}(f)),$$

which shows the continuity of Q at function f, that is

$$\forall f \; \forall \gamma \; \exists \delta \; \forall g : D_{\text{boundary}}(f,g) \leq \delta \implies |Q(f) - Q(g)| \leq \gamma.$$

☺

Case of graph cut objective functions. As an example, consider the normalized cut objective function, which is defined as follows. Let $s : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ be a similarity function which is upper bounded by a constant C. For a given cluster described by the cluster indicator function $f_k : \mathbb{R}^d \to \{0,1\}$, we set

$$\operatorname{cut}(f_k) := \operatorname{cut}(f_k, P) := \mathbb{E} f_k(X_1)(1 - f_k(X_2)) s(X_1, X_2)$$
$$\operatorname{vol}(f_k) := \operatorname{vol}(f_k, P) := \mathbb{E} f_k(X_1) s(X_1, X_2)$$

For a clustering function $f \in \mathcal{F}$ we can then define the normalized cut by

$$\operatorname{Ncut}(f) := \operatorname{Ncut}(f, P) := \sum_{k=1}^{K} \frac{\operatorname{cut}(f_k)}{\operatorname{vol}(f_k)}.$$

In Bubeck and von Luxburg (2007) it has been proved that there exists an algorithm such that Ncut can be minimized uniformly consistently. So it remains to be shown that Ncut is continuous with respect to D_{boundary}.

Proposition 6 (Continuity of Ncut wrt. D_{boundary}) *Let $\mathcal{X} \subset \mathbb{R}^s$ compact, and P a probability distribution on $\mathcal{X}$ with a density with respect to the Lebesgue measure. For a fixed constant $C > 0$, let $\mathcal{F}_C$ be the space of all clusterings $f : \mathcal{X} \to \{1, ..., K\}$ such that all clusters have a minimal P-mass C. Then the* Ncut *objective function is continuous with respect to D_{boundary} on $\mathcal{F}_C$.*

Proof. The proof is very similar to the one for the K-means case, thus we just provide a sketch. We consider the enumerator and denominator of Ncut separately. As for the K-means case, one splits the integrals over $\mathcal{X}$ in a sum of the integrals over $T_\gamma(f)$ and $T_\gamma(f)^c$. Both parts are dominated by the contributions from points in T_γ^c, and the contributions from inside the tubes can be bounded by some constant times the mass in the tubes. This leads to a similar argument as in the K-means case. ☺

Explicit form of the constant γ. We have seen that Theorem 4 can be applied to several of the standard clustering objective functions, such as the K-means one and the normalized cut. What remains a bit vague is the exact functional form of the constant γ in this theorem. Essentially, this constant is the result of an existence statement in Proposition 3. For the case of K-means, it is possible to upper bound this constant by using the tools and methods from Meila (2006) and Meila (2007). There it has been proved in a finite sample setting that under certain conditions, if $|Q(f) - Q(g)|$ is small, then also the $D_{\text{MinMatch}}(f,g)$ is small. For K-means, one can show that small $D_{\text{MinMatch}}(f,g)$ implies small D_{boundary}. Furthermore, all quantities used in the finite sample results of Meila (2006) need to be carried over to the limit setting. For example, the eigenvalues of the similarity matrices have to be replaced by eigenvalues of the corresponding limit operators, for example by using results from Blanchard et al. (2007). Combining all those arguments leads to an explicit upper bound for the constant γ in Theorem 4 for the K-means objective function. However, this upper bound became so technical that we refrain from deriving it in this paper. A similar argument might be possible for the normalized cut, as the results of Meila (2007) also cover this case. However, we have not worked out this case in detail, so we do not know whether it really goes through. If it does, the result is likely to look even more complicated than in the K-means case.

3.4 High-density boundaries do *not* imply instability

In the following example we demonstrate that, in some sense, the converse of Theorem 4 fails. We construct a data distribution over the two-dimensional plane for which the 2-means clustering has high probability mass in a narrow tube around the optimal clustering boundary, and yet the instability levels converge to zero fast (as a function of the sample sizes).

Example 1 *Let P_η^ν be a mixture distribution consisting of the following components (see Figure 1 for illustration). Define the sets $A = \{-1\} \times [-1,1]$, $B = \{1\} \times [-1,1]$, $C = \{(-\eta, 0)\}$, and $D = \{(\eta, 0)\}$. Let U_A and U_B be the uniform distributions on A and B, and δ_C and δ_D the probability distributions giving weight*

1 to the point C and D, respectively. Define $P_\eta^\nu = \frac{1}{2}\left((1-\nu)(U_A+U_B)+\nu(\delta_C+\delta_D)\right)$. Namely, the distribution that allocates weight $\nu/2$ to each of the singleton points C and D, and the rest of its weight is uniformly spread over the two vertical intervals at $x=-1$ and at $x=1$.

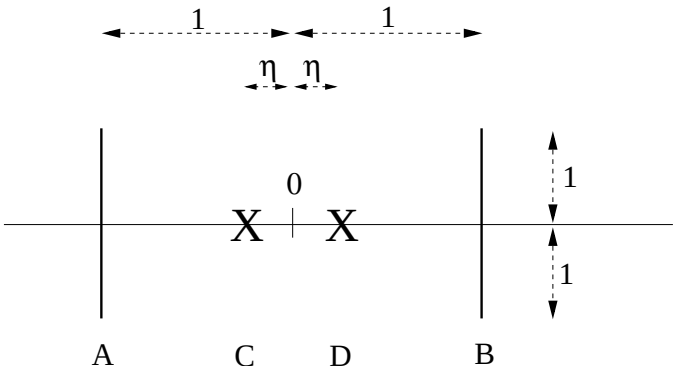

Figure 1: *Illustration of Example 1*

Clearly, the optimal 2-means clustering, f^*, divides the plane along the y axis. It is straight forward to see that if the parameters η and ν are, say, $\eta = 0.01$ and $\nu = 0.2$, then the following statements hold:

1. For γ comparable to the variance of $D_{\text{boundary}}(f_n, f^*)$, the γ-tube around this optimal boundary includes the points C and D and therefore has significant weight, namely $P(T_\gamma(f^*)) = \nu$.

2. $\text{InStab}(n, P_\eta^\nu)$ goes to zero exponentially fast (as a function of n).
   To see this, note that as long as both of the cluster centers are outside the interval, $[-1+\eta, 1-\eta]$, the clustering will be fixed (cutting along the y axis). This condition, in turn, holds whenever the sample S satisfies
   $$|S \cap A| > 20|S \cap C|$$
   and
   $$|S \cap B| > 20|S \cap D|$$
   ('20'here just stands for 'many times'). Note that these conditions are implied by having, for every $T \in \{A, B, C, D\}$,
   $$\left|\frac{|S \cap T|}{|S|} - P(T)\right| < 0.01$$
   Finally, note that, by the Chernoff Bound, the probability (over samples S of size n) that this condition fails is bounded by $c'e^{-cn}$ for some constants $c,\ c'$. Consequently, for every sample size, n,
   $$\text{InStab}(n, P_\eta^\nu) \leq c'e^{-cn}$$

3. For any $\varepsilon > 0$, $P(|Q(f_n) - Q(f^*)| > \varepsilon)$ goes to zero exponentially fast with n.

Thus, while the preconditions of Theorem 4 hold, in spite of having γ-tubes with significant P-weight, the instability values are going to zero at an extremely fast rate. The reason is that the sample fluctuations will move the cluster centers up and down in a rather narrow tube around the two vertical intervals. The resulting fluctuations of the empirical clustering boundary will (with overwhelming probability) keep the boundary *between* the points C and D. Therefore the instability will practically be zero (no points change cluster membership). On the other hand, those up and down sample-based fluctuations of cluster centers cause the boundary between the two empirical clusters to rotate around the origin point (for example, if the cluster center corresponding to A sits above the x-axis, and the center corresponding to B sits below the x-axis). Such rotations result in relatively high expected value for the D_{boundary} distance between the sample based empirical clusterings and the optimal clustering. These fluctuations could even be be made larger by concentrating the probability weight of the two vertical intervals at the end points of these intervals.
Furthermore, the phenomena of having significant weight in $T_\gamma(f^*)$, for small γ (i.e., comparable to the variance of the cluster centers) and yet retaining negligible instability can be shown for arbitrarily large sample sizes. Given any sample size n, one can choose η small enough so that, in spite of the decrease in the expected D_{boundary} empirical-to-optimal distances (due to having large samples), the points C and D will remain inside the $T_\gamma(f^*)$, for γ equal the variance of that D_{boundary} distance. Such a choice of parameters can be done while retaining the property that empirical clusterings are unlikely to move these points between clusters, and hence the stability.

High boundary density version: Example 1 has large weight on the γ-tube around the boundary of its optimal clustering partition. Yet, the value of the probability density function on the boundary is zero. One can construct a similar example, in which the probability density along the boundary itself is high, and yet the data has close-to-zero instability.

Example 2 *Similar to the example above, we consider a mixture distribution made up of three parts: S and T are the vertical intervals $S = \{-1\} \times [-1/2, 1/2]$ and $T = \{1\} \times [-1/2, 1/2]$. However, now the third component is a the rectangle $R = [-\eta, \eta] \times [-1, 1]$. Our data space is then defined as $\mathcal{X} := R \cup S \cup T$, and as probability distribution we choose $D_\eta^\nu = (1-\nu)/2 \cdot (U_S + U_T) + \nu \cdot U_R$. Finally, we define a distance $d_\mathcal{X}$ on this space by letting $d_\mathcal{X}(a, b)$ be the usual Euclidean distance whenever a and b belong to the same component of $\mathcal{X}$, and $d_\mathcal{X}(a, b)$ is defined as the distance between the projections of a and b on the x-axis whenever a and b belong to different components. Note that this metric is not Euclidean and that $S \cup R \cup T$ is our full domain space, not the real plane.*

Once again the optimal 2-means clustering splits the space along the y-axis. However, now this boundary has significantly high density. Yet, we claim that D_η^ν instability goes to zero exponentially fast with the sample size. Intuitively, this is because the up and down fluctuations of the centers

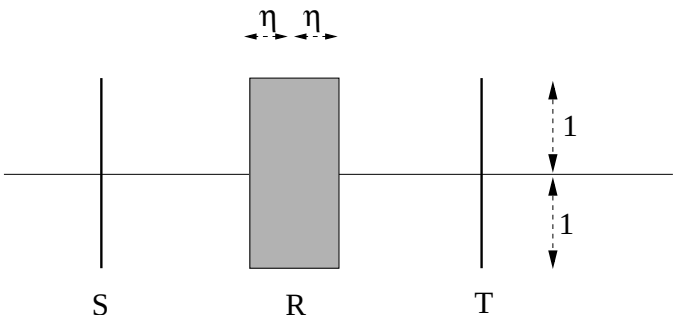

Figure 2: *Illustration of Example 2*

of the two clusters do not perturb the boundary between the two clusters.

More concretely, say we pick $\eta < 0.01$ and $\nu = 0.1$. we wish to show that, with high probability over the choice of samples, the 2-means optimal sample clustering has its cluster centers in the sets S and T. Note that by our choice of distance function, if one center is in S and the other is in T then the clustering cuts our domain set along the y axis (regardless of the y coordinates of the centers).

Since our domain set equals $S \cup R \cup T$ (there are no other points in our domain space), it suffices to show that it is unlikely that a sample based clustering will have a cluster center in the set R. To see that, note that if for some sample W, the 2-means cost clustering based on W has a cluster center, say of the left-hand cluster, is in R then the 2-means cost of that clustering is at least $|S \cap W|0.99$. On the other hand, if the center of that cluster is in S then the 2-means cost of that cluster is at most $|S \cap W|0.25 + |W \cap R|(1.01)^2$. It follows that, as long as $|W \cap R| < 0.11|W|$ the optimal 2-means clustering of the sample W will have one cluster center in S and the other cluster center in T. We can now apply a similar argument to the one used for example 1. Namely, note that as long as the empirical weight of each of the three data components is within 0.01 of its true weight it will indeed be the case that $|W \cap R| < 0.11|W|$. It therefore follows, by the Chernoff Bound, that the probability of having a sample W violate this condition is bounded by $c'e^{-c|W|}$ for some constants c, c'. Consequently, except for such minuscule probability, the clustering always splits our domain set along the y axis. Consequently the 2-means instability of our data distribution is exponentially small (in the sample size).

4 Some inherent limitations of the stability approach in the large sample regime

We consider a setting in which one tries to gain insight into the structure of some unknown data set (or probability distribution over such a set) by sampling i.i.d. from that set. A major question is when can such samples be considered a reliable reflection of structure of that unknown domain. This is the typical setting in which notions of stability are applied. The most common use of stability is as a model selection tool. In that context stability is viewed as an indication that a clustering algorithm does the "right thing" and, in particular, that its choice of number of clusters is "correct". The work of Shamir and Tishby (2008b) as well as the analysis in this paper claim that stability can be viewed as an indication that the clusters output by an algorithm are "correct" in the sense of having their boundaries pass through low-density data regions.

However, all such results relate the desired clustering properties to the eventual values of stability when the sample sizes grow unboundedly. Since in applications a user always examines finite size samples, the reliability of stability as a model selection tool requires the bound on the rate by which stabilities over n-size samples converge to their limit values to be uniform over the class of potential data distributions. We show below that no such bounds hold. Arbitrarily large sample sizes can have arbitrarily misleading stability values. The implications of stability values discussed in these papers kick in for sample sizes that depend upon the data distribution, and are therefore not available to the user in most practical applications. We are going to analyze this behavior based on the following example.

Example 3 *Consider the following probability distribution over the two dimensional plane (see Figure 3). Let B be the disk $\{(x,y) : (x-1)^2 + y^2 \leq 1/2\}$, let C be the disk $\{(x,y) : (x+1)^2 + y^2 \leq 1/2\}$. Let x_0 be the point $(0, M)$ for some large positive M (say, $M = 100$). Given $\varepsilon > 0$, let P_ε^M be the probability distribution defined as $P_\varepsilon^M = \varepsilon\delta_{x_0} + (1-\varepsilon)/2(U_B + U_C)$ (in the notation of the example in Section 3.4), where ε is some small number, say $\varepsilon = 0.01$.*

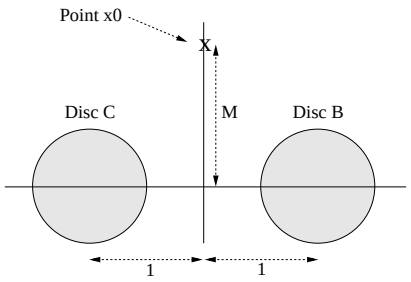

Figure 3: *Illustration of Example 3*

No distribution-free stability convergence rates possible. Consider the distribution of Example 3, and let $\mathcal{A}$ be an algorithm that finds an optimal 2-means solution for every input data set. For n rather small, a sample of n points is rather unlikely to contain the point x_0 as it has a very small mass on it. In those cases, the algorithm $\mathcal{A}$ will cluster the data by vertically splitting between the disks B and C. Thus, $\mathrm{InStab}(n, P_\varepsilon^M)$ computed on such a data set is very low. However, as the sample size grows, the probability that a sample will contain the point x_0 becomes significant. Now observe that as we chose M to be very large, then whenever x_0 is a member of a sample S the optimal 2-clustering of S will have one of its center points at x_0. Consequently, as long as n is such that a significant fraction of the n-samples pick x_0 and a significant fractions of the samples miss it, $\mathrm{InStab}(n, P_\varepsilon^M)$ is very high. Finally, when the sample size are large enough to guarantee that hardly any sample misses

x_0, stability is regained.

All in all we have constructed an example of a probability distribution where the 2-means optimizing algorithm is very stable for sample size n, is very unstable for some sample size $n' > n$ and converges to perfect stability as the sample sizes go to infinity. By playing with the parameters of P_ε^M one can in particular adjust the sample size n' for which the instable regime holds. As a consequence, there cannot be a distribution-free convergence rate for stability.

It is also worth while to note that throughout the above family of distributions (for all non-degenerate values of M and ε), the optimal clustering has a wide tube of zero-density around it. Just the same, for arbitrarily large values of n', n'-size samples display large instability. In particular, this example shows that the assumption "$D_{\text{boundary}}(f_n, f) \leq \gamma$" in Proposition 2, is indeed necessary.

Stability does not imply close-to-optimal clustering cost. Proposition 2 states that when sample sizes are such that the sample based clustering quality is close to its optimal value, and if that optimum is achieved with low-density tubes, then the value of instability is low. Example 3 shows that the converse of this statement does not always hold. For data distributions of the form P_ε^M, due to having a far outlier, x_0, when a sample misses that outlier point, the cost of the sample-based clusterings is at least $M^2\varepsilon$. On the other hand, the cost of the optimal clustering (that allocates a center to cover the outlier point) is less than $3(1-\varepsilon)$. As long as $\varepsilon \leq 1/n^2$, samples are unlikely to hit x_0 and therefore $\mathrm{InStab}(n, P_\varepsilon^M)$ is very low. However, if M is picked to be greater than, say $10/\varepsilon$ we get a large gap between the cost of the sample based clustering and the cost of the distribution-optimal clustering.

Stability does not imply proximity between the sample based clustering and the optimal clustering. Again, it can be readily seen that the above family of P_ε^M data distributions demonstrates this point as well.

Stability is not monotone as a function of the sample sizes. Clearly Example 3 demonstrates such non-monotonicity. The values of $\mathrm{InStab}(n, P_\varepsilon^M)$ decrease with n for $n < 1/\sqrt{\varepsilon}$, they increase with n for values of n round $1/\varepsilon$ and the they decrease to zero for $n \geq 1/\varepsilon^2$.

We end this section with a few further observations demonstrating the somewhat "erratic behavior" of stability.

No *uniform* convergence of cluster centers to a normal distribution. Although Pollard (1982) has proved that as the sample sizes grow to infinity, the distribution of the empirical cluster centers converges to a normal distribution, there is no uniform bound on the rate of this convergence. For example, consider a two-mode probability distribution over the real line that has high peeks of its density function at the points $(0, -\varepsilon)$ and $(0, \varepsilon)$, has 0 density for $x = 0$, and then tails off smoothly as $|x|$ goes to infinity. Obviously, for every sample size, n, by choosing small enough ε, the distribution of each of the cluster centers for 2-means of random n-samples drawn from this distribution is highly non-symmetric (it has higher variance in the direction away from the 0 than its variance towards 0), and therefore far from being a normal distribution.

Arbitrarily slow convergence of stability for 'nice' data. Even when data is stable and has a rather regular structure (no outliers like in the example discussed above), and the optimal boundaries pass through wide low-density data regions, the convergence to this stability, although asymptotically fast, is not uniformly bounded over different (well structured) data distributions. For every n there exists a data distribution D_n that enjoys the above properties, and yet $\mathrm{InStab}(n, D_n)$ is large. As an example of this type of non-uniformity, consider a planar distribution having its support on four small (say, of radius 0.1) discs centered on the four corners of the unit square. Assume the distribution is uniform over each of the discs, is symmetric around the x axis, but gives slightly more weight to the left hand side two disks than to the right hand side disks. For such a distribution, the optimal 2-means clustering is a unique partition along the x axis, and has wide 0-density margins around its boundary. Just the same, as long that the sample sizes are not big enough to detect the asymmetry of the distribution (around the y axis), a significant fraction of the sample based 2-means clustering will pick a partition along the y axis and a significant fraction of samples will pick a partition along the x axis, resulting in high instability. This instability can be made to occur for arbitrarily large sample sizes, by just making the asymmetry of the data sufficiently small.

5 Discussion

In this paper, we discuss the mechanism of stability-based model selection for clustering. The first part of the paper investigates a promising conjecture: in the large sample regime, the stability of a clustering algorithm can be described in terms of properties of the cluster boundary, particularly whether the boundary lies in a small or high density area. In the case of K-means, this would explain the success of stability-based methods by demonstrating that stability adds the "the missing piece" to the algorithm. As the K-means clustering criterion is only concerned by within-cluster similarity, but not with between-cluster dissimilarity, a model selection criterion based on low density areas would add a valuable aspect to the algorithm.

In parts, our results are promising: the conjecture holds at least in one direction. However, it is pretty discouraging that the conjecture does not hold the other way round, as we can show by a simple counterexample. This counterexample also indicates that a simple mechanism such as "low density" vs. "high density" does not exist. So, after all, the question which are the underlying geometric principles of stability-based model selection in the large sample regime remains unanswered.

On the other hand, we also provide some reasons why using stability-based methods in the large sample setting might be problematic in general. The reason is that it is impossible to

give global convergence guarantees for stability. Thus, while one can use stability criteria in practice, it is impossible to give distribution-free performance guarantees on any of its results. No matter how large our sample size n is, we can always find distributions where the stability evaluated on that particular sample size is misleading, in the sense that it is far from the "true stability"

Finally, we would like to put our results in a broader context and point out future research directions for investigating stability. In general, there are different reasons why cluster instability can arise:

Instability due to multiple global optima. If the global optimizer of the clustering objective function is not unique, this always leads to instability. However, this kind of instability is usually not related to the correct number of clusters, as has been proved in Ben-David et al. (2006), Ben-David et al. (2007). Instead, it might depend on completely unrelated criteria, for example symmetries in the data. In this situation, stability criteria are not useful for selecting the number of clusters.

Geometric instability in the large sample setting. This is the kind of instability we considered in this paper. Here one assumes that no issues with local optima exist, that is the algorithm always ends up in the global optimum, and that a unique global optimum exists (for all values of K under consideration). In this paper, we made an attempt to connect the mechanism behind stability-based model selection to geometric properties of the underlying distribution and clustering, but with moderate success only. On the other hand, we can demonstrate that using stability in the large sample setting has problems in general. While it might be possible that future work shows a tighter connection between geometric properties of the data space and stability issues, we are doubtful whether those methods can be applied successfully in practice, unless one makes strong assumptions on the underlying distributions.

Instability due to too small sample size. If the sample size is too small, and the cluster structure is not sufficiently well pronounced in the data set, we will observe instability. Here, clustering stability can be a useful criterion to detect whether the number of clusters is much too high. If this is the case, the algorithm will construct clusters which are mainly based on sampling artifacts, and those clusters will be rather unstable. Here, stability tells us whether we have enough data to support a given cluster structure. This is of course a useful thing to know. However, it is still not obvious whether stability can be used to detect the "best" number of clusters, as there might be several values of K which lead to stable results. We believe that it is a very important direction to investigate what guarantees can be given on stability-based methods in this scenario.

Algorithmic instability. This kind of instability occurs if the algorithm itself can converge to very different solutions, for example it ends up in different local optima, depending on starting conditions. Note that algorithmic instability is rather a property of an algorithm than of an underlying distribution or sample. If we had a perfect algorithm which always found the global optimum, then this kind of instability would not occur. In our opinion, in a setting of algorithmic instability it is not clear that stability selects the "best" or "correct" number of clusters. Essentially, in this case stability simply detects whether there is a well-pronounced local optimum where the objective function has the shape of a "wide bowl" such that the algorithm gets trapped in this local optimum all the time. However, we find it unlikely that the conclusion "local optimum in wide bowl implies good K" is true. It has been argued that the conclusion the other way round is true: "distribution with well-pronounced cluster structure implies global optimum in wide bowl" (e.g., Meila, 2006 or Srebro et al., 2006). However, this is not the direction which is needed to show that clustering stability is a good criterion to select the number of clusters. We conclude that in the "algorithmic instability" scenario, stability is not very well understood, and it would be very interesting to give conditions on distributions and algorithms in which this kind of stability can provably be useful for model selection.

In all settings discussed above, stability is useful in one respect: high instability can be used as an alarm sign to distrust the clustering result, be it for sampling, algorithmic or other reasons. However, the other way round, namely that the most stable algorithm leads to the best clustering result, so far has not been established for any of the settings above in a satisfactory way.

Acknowledgments

We are grateful to Markus Maier who pointed out an error in an earlier version of this manuscript, and to Nati Srebro and David Pal for insightful discussions.

References

S. Ben-David. A framework for statistical clustering with constant time approximation algorithms for k-median and k-means clustering. *Machine Learning*, 66:243 – 257, 2007.

S. Ben-David, U. von Luxburg, and D. Pál. A sober look on clustering stability. In G. Lugosi and H. Simon, editors, *Proceedings of the 19th Annual Conference on Learning Theory (COLT)*, pages 5 – 19. Springer, Berlin, 2006.

S. Ben-David, D. Pál, and H.-U. Simon. Stability of k -means clustering. In N. Bshouty and C. Gentile, editors, *Conference on Learning Theory (COLT)*, pages 20–34. Springer, 2007.

A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing*, pages 6 – 17, 2002.

A. Bertoni and G. Valentini. Model order selection for bio-molecular data clustering. *BMC Bioinformatics*, 8(Suppl 2):S7, 2007.

M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, V. Sondak, M. Hayward, and J. Trent. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406: 536 – 540, 2000.

G. Blanchard, O. Bousquet, and L. Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2-3):259–294, 2007.

S. Bubeck and U. von Luxburg. Overfitting of clustering and how to avoid it. Preprint, 2007.

J. Fridlyand and S. Dudoit. Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method. Technical Report 600, Department of Statistics, University of California, Berkeley, 2001.

M. K. Kerr and G. A. Churchill. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *PNAS*, 98(16):8961 – 8965, 2001.

A. Krieger and P. Green. A cautionary note on using internal cross validation to select the number of clusters. *Psychometrika*, 64(3):341 – 353, 1999.

T. Lange, V. Roth, M. Braun, and J. Buhmann. Stability-based validation of clustering solutions. *Neural Computation*, 16(6):1299 – 1323, 2004.

E. Levine and E. Domany. Resampling Method for Unsupervised Estimation of Cluster Validity. *Neural Computation*, 13(11):2573 – 2593, 2001.

M. Meila. Comparing clusterings: an axiomatic view. In *Proceedings of the International Conference of Machine Learning (ICML)*, pages 577–584, 2005.

M. Meila. The uniqueness of a good optimum for K-means. In W. Cohen and A. Moore, editors, *Proceedings of the Twenty-Third International Conference on Machine Learning (ICML)*, pages 625–632. ACM, 2006.

M. Meila. The stability of a good clustering. Manuscript in preparation, 2007.

R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy. The effectiveness of Lloyd-type methods for the k-means problem. In *FOCS*, pages 165–176. IEEE Computer Society, 2006.

D. Pollard. A central limit theorem for k-means clustering. *Annals of Probability*, 10(4):919 – 926, 1982.

O. Shamir and N. Tishby. Model selection and stability in k-means clustering. In *Conference on Learning Theory (COLT), to appear*, 2008a.

O. Shamir and T. Tishby. Cluster stability for finite samples. In J.C. Platt, D. Koller, Y. Singer, and S. Rowseis, editors, *Advances in Neural Information Processing Systems (NIPS) 21*. MIT Press, Cambridge, MA, 2008b.

M. Smolkin and D. Ghosh. Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics*, 36 (4), 2003.

N. Srebro, G. Shakhnarovich, and S. Roweis. An investigation of computational and informational limits in Gaussian mixture clustering. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 865 – 872. ACM Press, New York, 2006.

U. von Luxburg, S. Bubeck, S. Jegelka, and M. Kaufmann. Consistent minimization of clustering objective functions. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems (NIPS) 21*, Cambridge, MA, 2008. MIT Press.