A formal analysis of stability - lessons and challenges

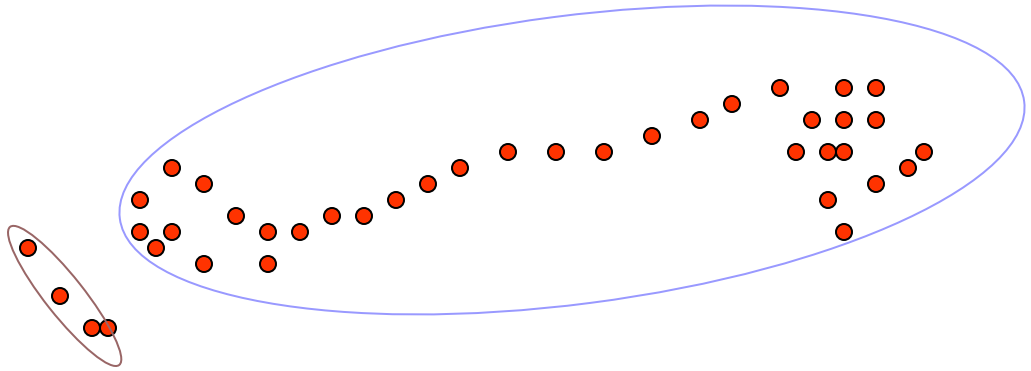
Shai Ben-David

PASCAL Workshop on Stability and Resampling Methods for Clustering

What is a good clustering???

“Clustering” is an ill defined problem

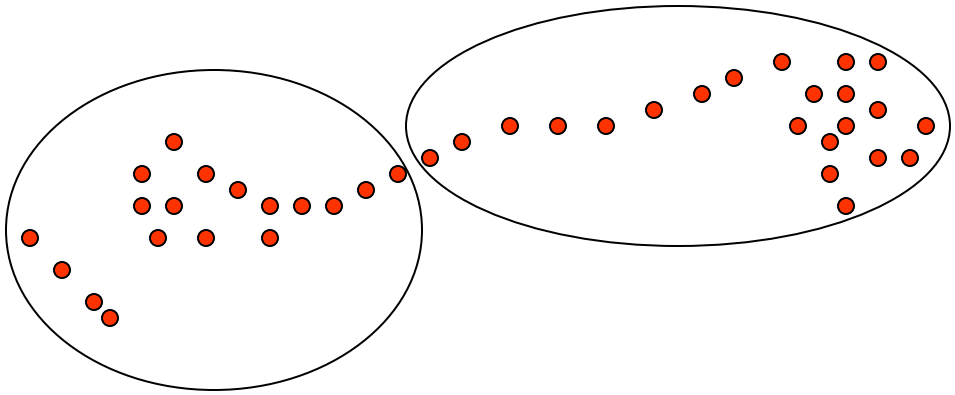
There are many different clustering tasks, leading to different clustering paradigms:



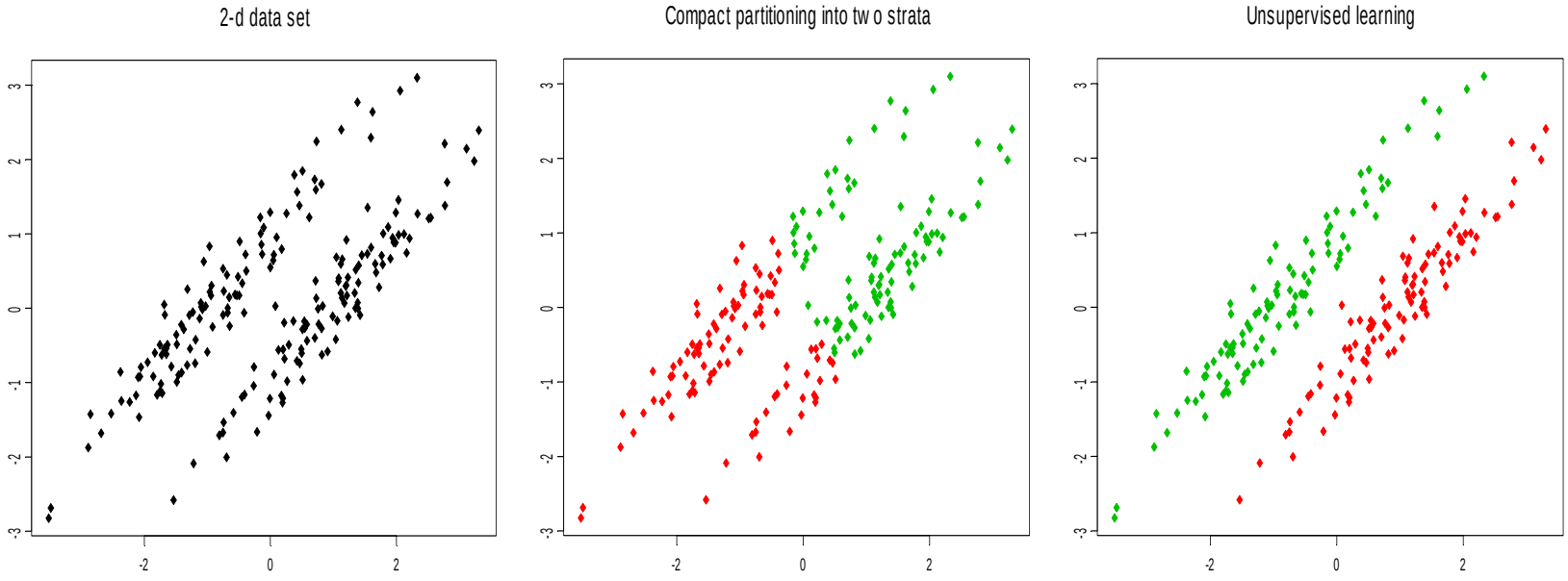
What is a good clustering???

“Clustering” is an ill defined problem

There are many different clustering tasks, leading to different clustering paradigms:



In what sense is the leftmost clustering better than the middle one?



Even if we commit to a fixed cost function

- You get a data set.
- Run your 5-means clustering algorithm, and get a clustering ***C***.
- You compute its 5-means cost and its 0.7.

- Can you conclude that ***C*** is a good clustering?
- How can we verify that structure described by ***C*** is not just “noise”?

Even harder questions

- How can we tell if a given data set *has* a good k-clustering solution (for a given k)?
- Can we have an efficient algorithm for the above task (say, running in time sub-linear in the size of the input data)?
- Note that even ***approximating*** the ***cost*** of the optimal k-clustering is NP- hard.

Quest for a general theory

Can we find answers that are *independent* of any

particular algorithm,

particular objective function

or specific generative data model

?

A more modest approach

Formulate conditions that should be satisfied by any conceivably good clustering function.

(Sidestepping the issue of “what is a good clustering clustering?”)

In other words –
find ***necessary*** conditions for good clustering

Stability - the basic idea

1) *Cluster independent samples of the data.*

2) *Compare the resulting clusterings.*

Meaningful clusterings should not change much

from one independent sample to another.

This idea has been employed as a tool for choosing the number of clusters in several empirical studies ([Ben-Hur et al'02], [Lange, Brown, Roth, Buhmann '03] and many more).

However, currently there is very limited theoretical support.

Stability - the formal definition

Given,

- Probability dist. P over some domain X.
- Clustering function A defined on $\{S : S \subseteq X\}$.
- Similarity measure over clusterings, d.
- Sample size m.

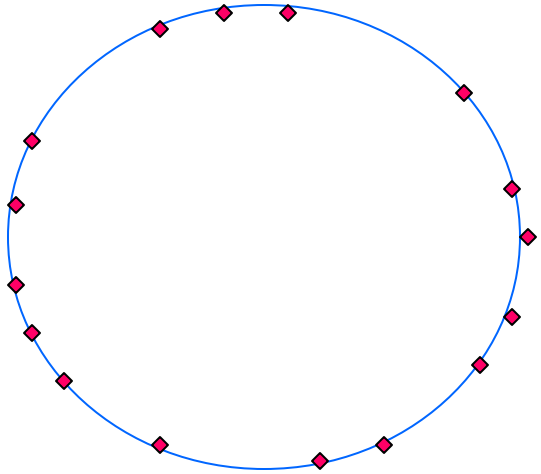
$$\mathrm{InStab}_m(A, P) = \mathrm{E}_{S, S' \in P^m} d(A(S), A(S'))$$

Namely, the expected distance between the clusterings generated by two P-random i.i.d. samples of size m.

(In)Stability detects non-clusterability:

There is no distribution-free stability guarantee.

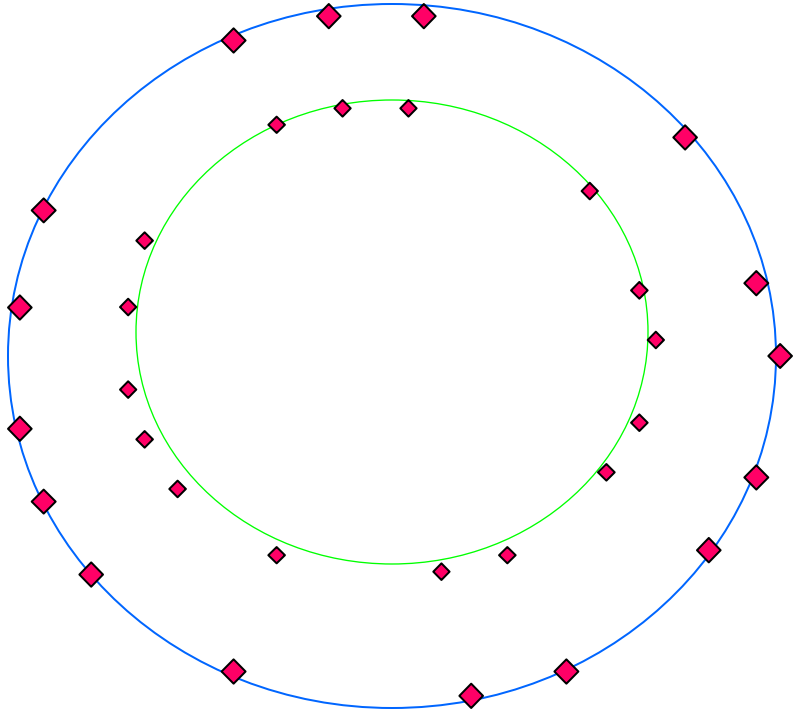
Example 1: The uniform distribution over a circle



InStab(C,P) will be large for any non-trivial clustering function.

Stability distinguishes relevant from irrelevant clustering paradigms:

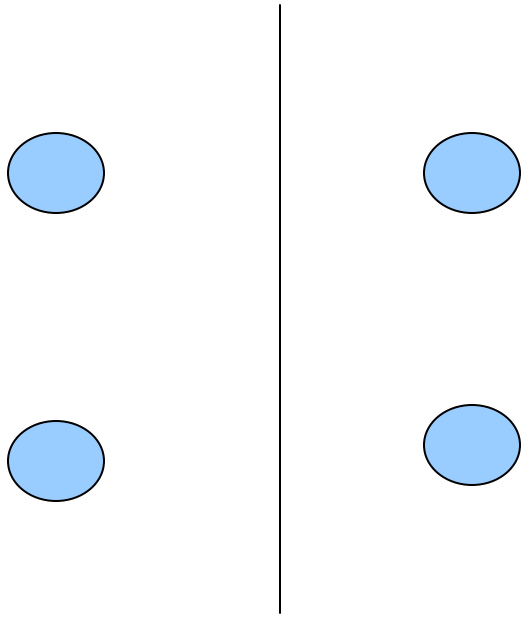
Example 2: A mixture of two uniform distribution over circles



InStab(C,P) will be large for any center-based clustering function, but single-linkage turns out to be stable (for some choice of parameters)

Stability detects correct k:

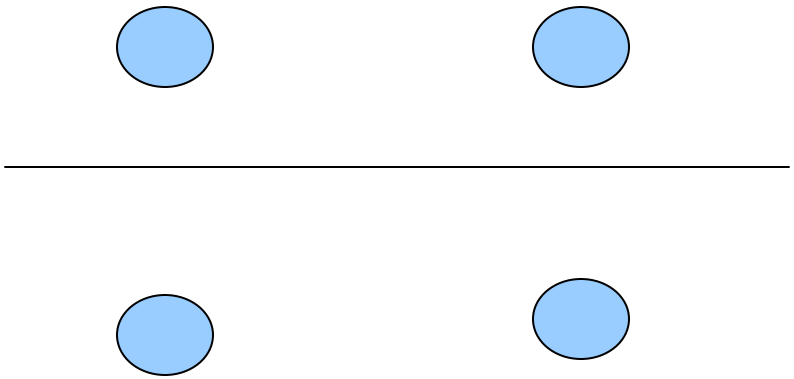
Example 3: A mismatch of # of clusters



UnStab(C,P) will be large for any center-based 2-cluster function.

Stability detects correct k:

Example 3: A mismatch of # of clusters



UnStab(C,P) will be large for any center-based 2-cluster function.

Conclusions (as of Dec. 2005)

- ✓ We formally define a measure of statistical generalization for sampling-based clustering –***stability***.
- ✓ Stability is a necessary property for any clustering method to be considered ‘meaningful’.
- ✓ Stability is viewed as a measure of the fit between a clustering function and an input data.
- ✓ We show that this measure can be reliably estimated from finite samples.

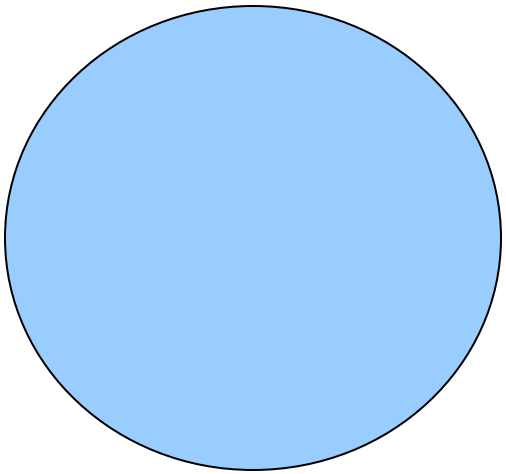
Have we found a good answer?

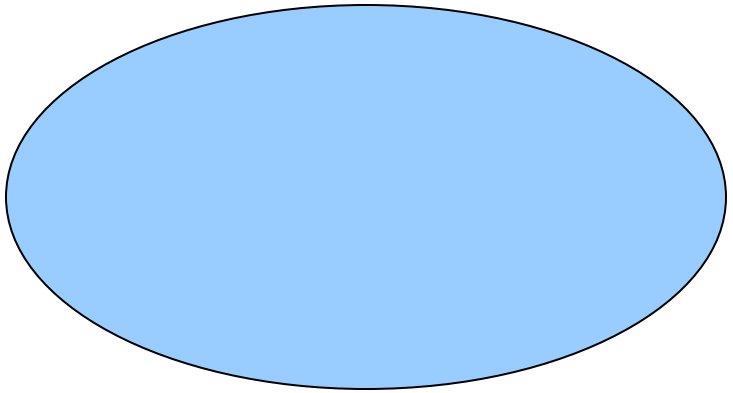
This is what we thought in January 2006, when we (with Ule and David Pal) set up to prove that

"stability is a reliable model-selection tool".

Its another interesting question how can one provide a mathematical formulation of such a statement.

Some bothersome examples





- A perfect ‘circular’ data is unstable for every k>1,
- Once the symmetry is broken, it becomes stable for every choice of k.

The bottom line of a formal analysis

- **Stability does a nice model-selection job on *simple synthetic distributions*, (as well as in many practical applications).**
- **We *characterize* it for the k-means optimization algorithms (BD-Luxburg-Pal, COLT06, BD-Pal-Simon, COLT07).**
- **We conclude that that success should be considered a lucky coincidence rather than a reliable rule.**

The formal results

We consider cost-minimizing clustering algorithms.

(E.g. K-Means, minimizing the sum of square distances of points to their respective centers).

We say that an algorithm ***A is stable on a data set D*** if

$\mathbf{Lim_{m->\infty} InStab_m(A,D)=0}$

Theorem (BD-Pal-Luxburg 06, DB-Pal-Simon 07):

A cost minimizing algorithm, *A*, is **stable** on data set *D*

if and only if

there is a **unique** clustering solution to the cost minimization problem.

Proof Idea 1: Uniqueness implies stability

This was proved in Shai Ben-David, Ulrike von Luxburg, Dávid Pál, COLT'06 [2]. It essentially follows from the uniform convergence:

Theorem (Uniform Convergence, Shai Ben-David, COLT'04 [1])

For any $\epsilon > 0$, for large enough m, for $S \sim P^m$ with high probability, for all k-tuples of centers $c_1, c_2, \ldots, c_k$ simultaneously

$$|R(P; c_1, c_2, \ldots, c_k) - R(S; c_1, c_2, \ldots, c_k)| < \epsilon .$$

Proof idea (2): Multiple solutions imply instability

- Support of P is finite $\Longrightarrow$ finitely many different clusterings. (Note that $d_P(\mathcal{C}, \mathcal{D}) > 0$ for any two distinct clusterings $\mathcal{C}, \mathcal{D}$.)
- Some of them are P-optimal: $\mathbf{OPT} = \{\mathcal{C}, \mathcal{D}, \mathcal{E}, \dots\}$.
- Let m be large enough, so that, with high probability a sample $S \sim P^m$ is such that $A_k(S) \in \mathbf{OPT}$. This follows from the uniform convergence theorem. (We can ignore S's not having this property.)
- Pick two clusterings $\mathcal{C}, \mathcal{D} \in \mathbf{OPT}$, $d_P(\mathcal{C}, \mathcal{D}) > 0$. We will show that

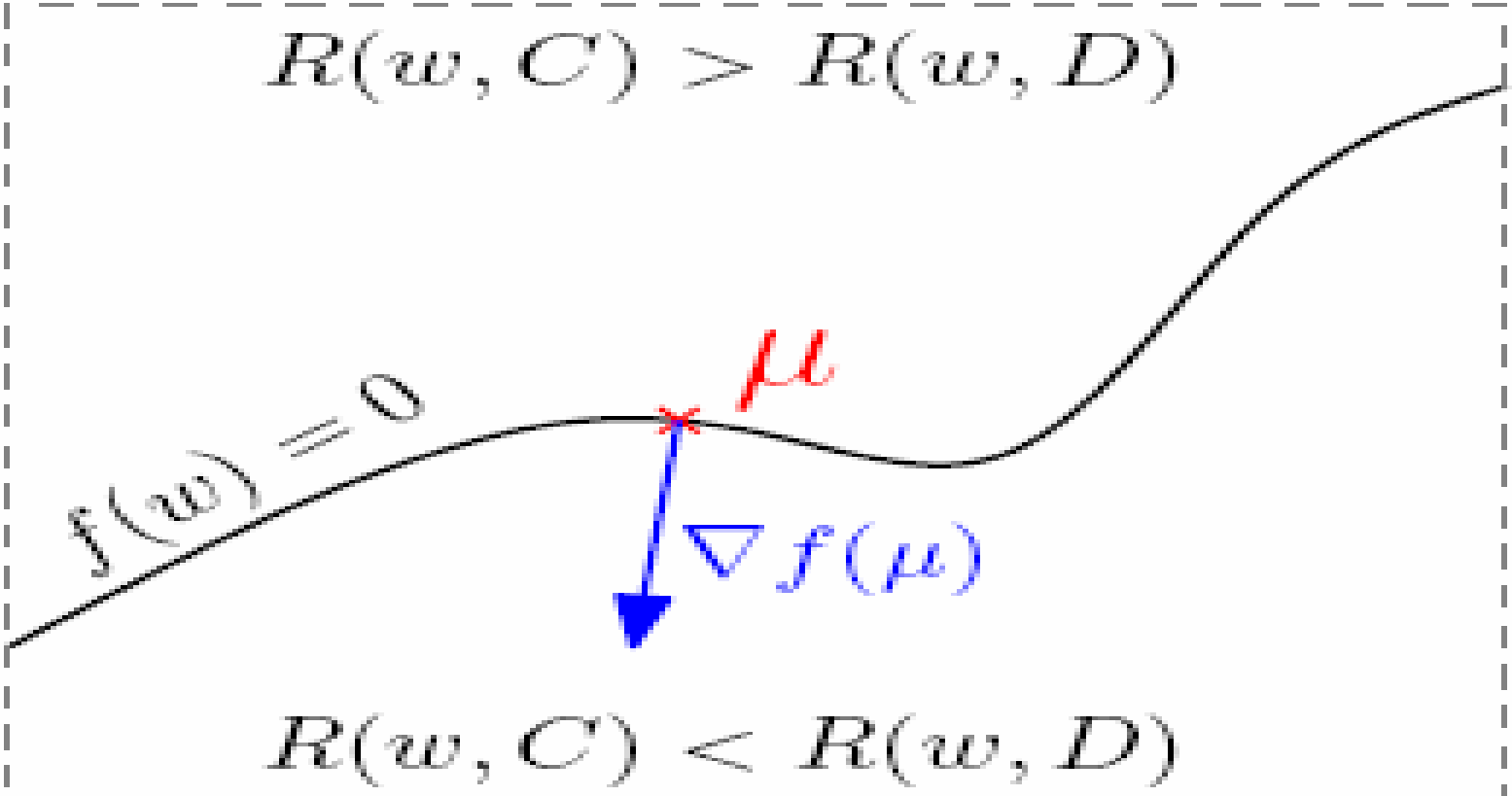
$$\lim_{m \to \infty} \Pr[R(S, \mathcal{C}) < R(S, \mathcal{D})] \quad \notin \quad \{0, 1\} .$$

- In other words, in the limit, there will be at least two P-optimal clusterings which are S-optimal with non-zero probability.

Proof idea (continued)

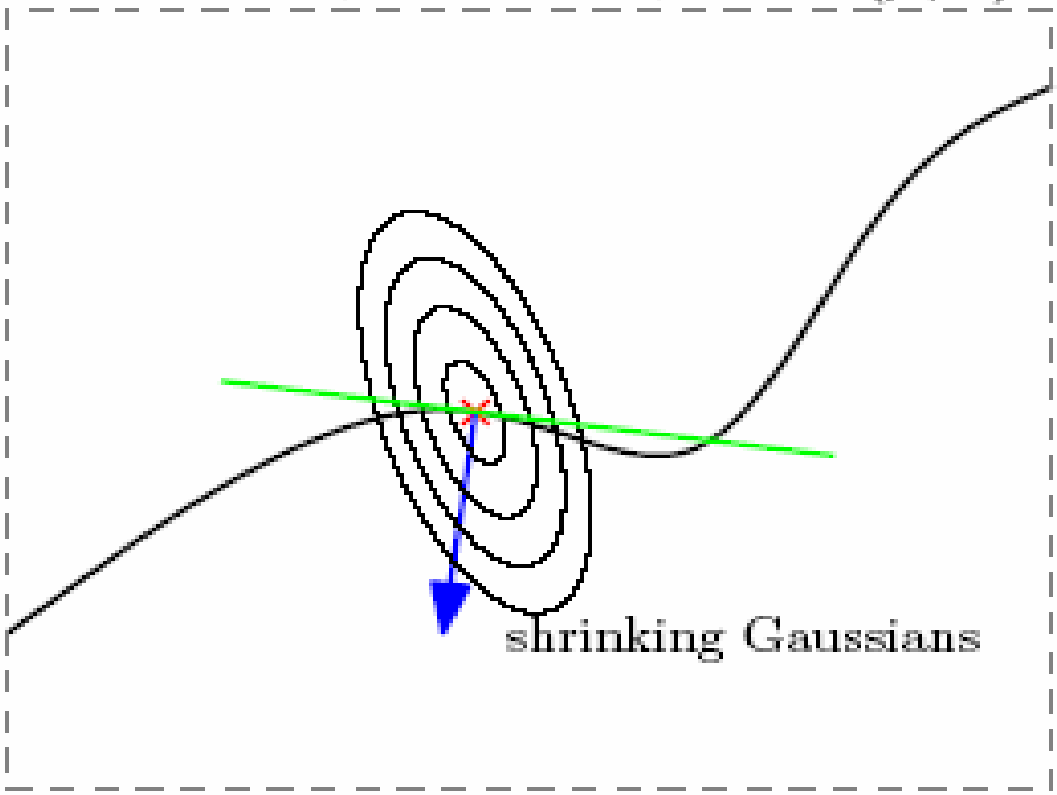
Consider the decision function:

$f : \mathbb{R}^n \longrightarrow \mathbb{R},\ f(w) = R(w, \mathcal{D}) - R(w, \mathcal{C})$



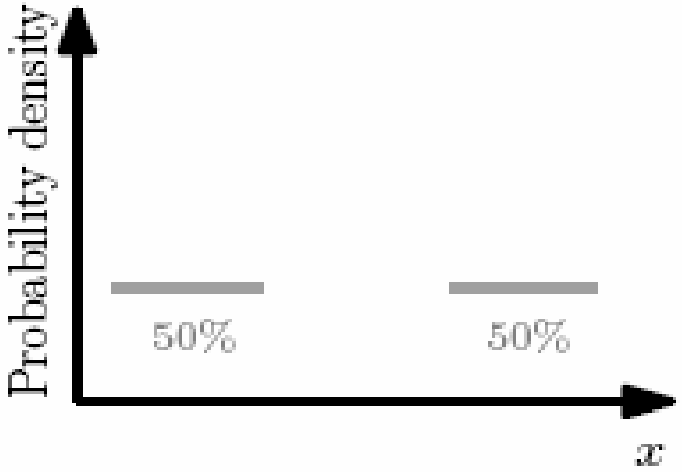
Proof Idea (continued)

By central limit theorem, as $m \to \infty$, $w \sim N(\mu, \Sigma)$ with $\Sigma \to 0$:

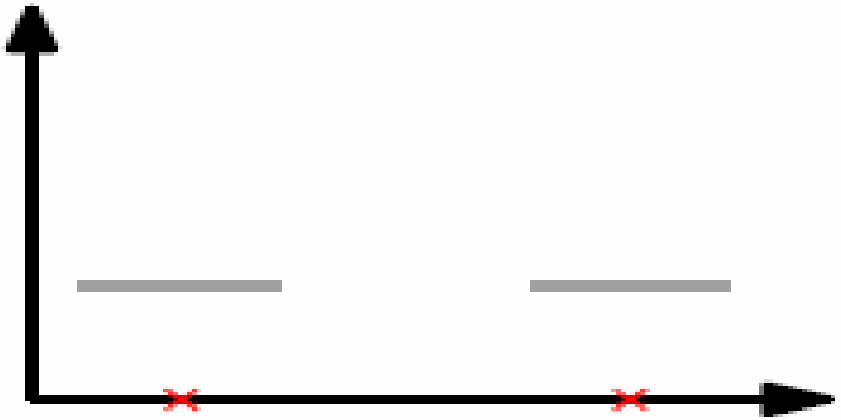


Some Examples

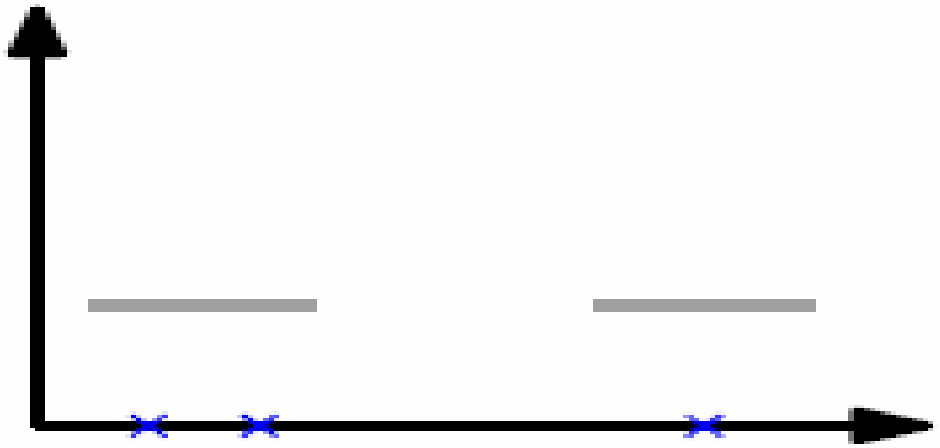
1D probability distribution



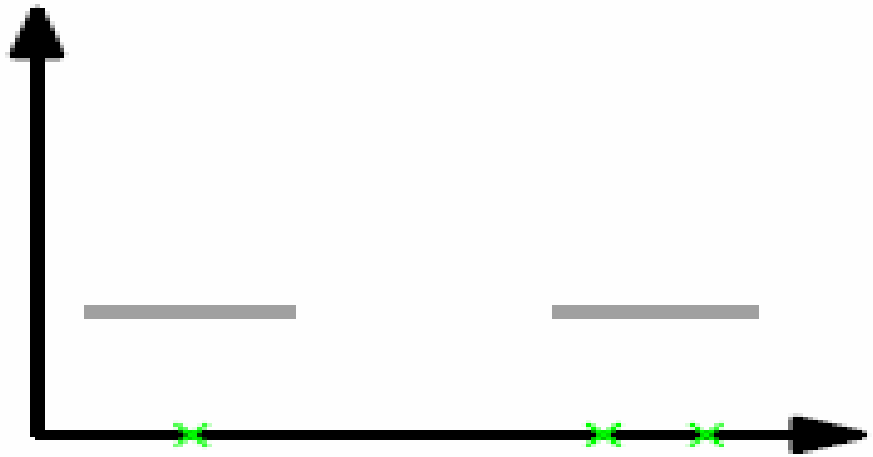
2 centers – stable



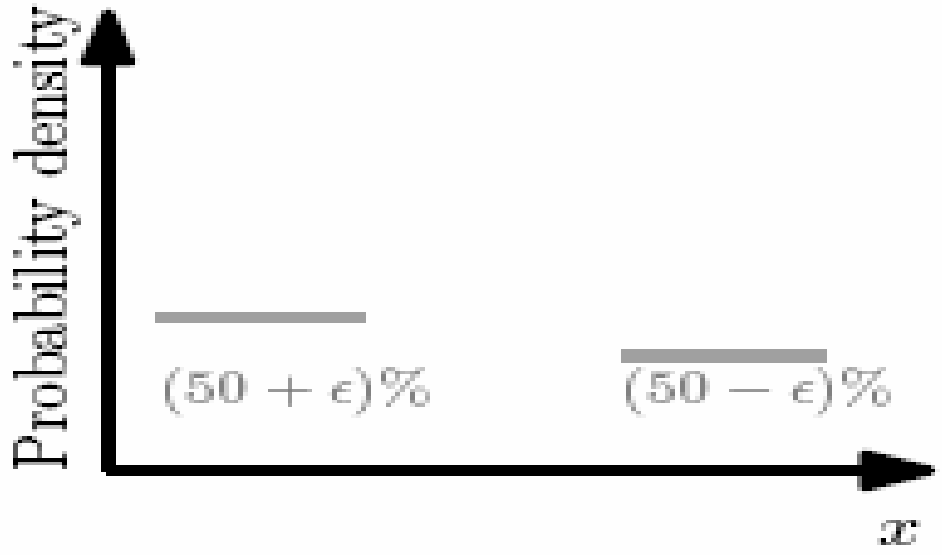
3 centers — solution #1



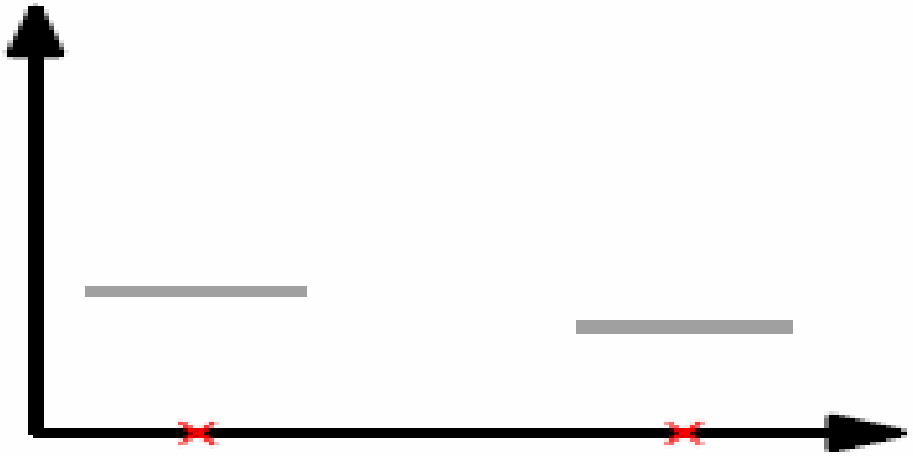
3 centers – solution #2 $\Longrightarrow$ unstable



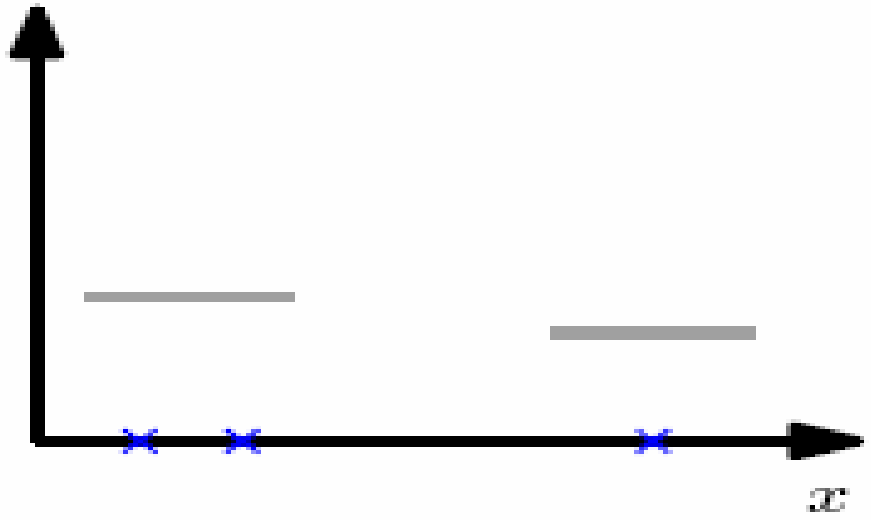
slightly asymmetric distribution



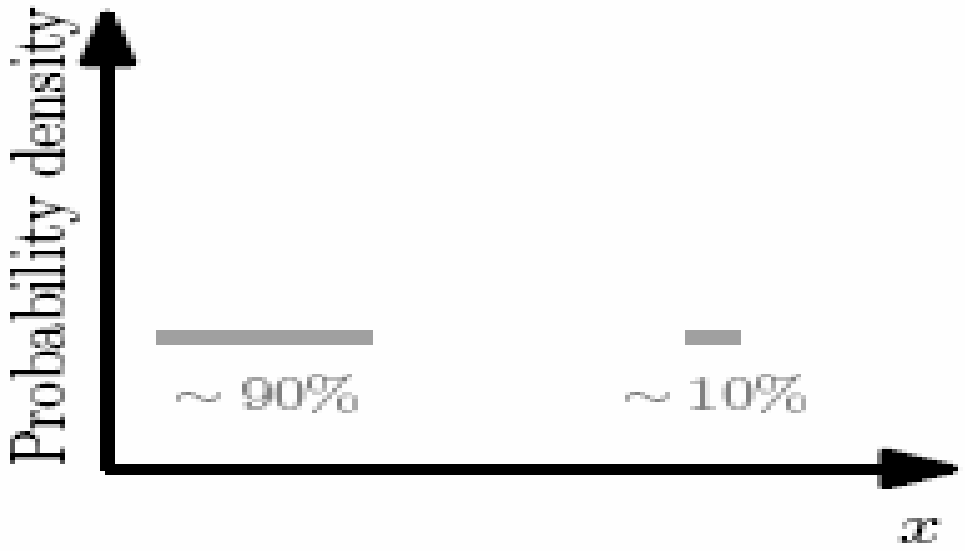
2 centers – stable



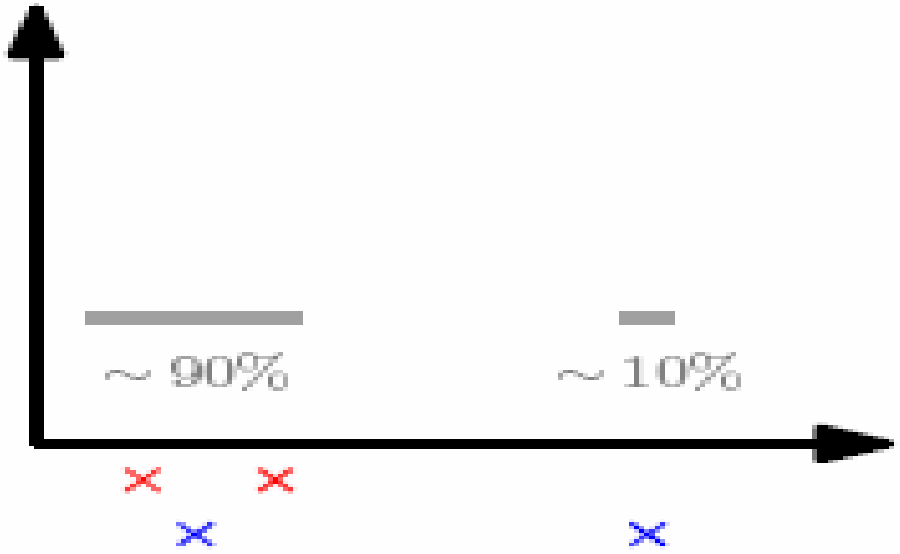
3 centers – stable



1D probability distribution



2 centers — unstable



The bottom line

- In practice, since no real data set is nicely symmetric, there will always be a unique cost-minimizing solution.

- Consequently, *Any choice of clustering parameters, on any real data set,*

  will always end up being stable.

Stability does not do the job we thought it did!

Other notions of stability

- There are several different reasonable notions of clustering stability.
- For example, one could consider *data-perturbations* rather than random *re-sampling*.
- While our proof applies only to the re-sampling stability, we believe that or results apply to such other notions as well.

Two different topics for discussion

- Is *uniqueness of optimal solution*
  a reasonable measure of data clusterability?

- Can the *asymptotic nature* of our results
  explain the “theory-practice gap” concerning clustering stability?

Some thoughts on the 'finite samples' issue

- Clearly, the claim that in practice any data set has a unique clustering cost minimizer, may fail if relaxed to 'almost minimal' solution'. It follows that sample sizes that are not large enough may 'view' the data as having multiple minima, and show the expected instability. But how large will 'not large enough' be?

- To make a practically useful contribution,

  we'd rather have a sample size estimate that can be derived from random samples (without any prior information about the structure of data).

  I doubt if this may be possible.

Alternative notions of clusterability

In forthcoming work, we investigate a variety of notions of clusterability and of clustering quality:

- Clustering Separability – the ratio between the k-means cost and the (k-1)-means cost.
- (Variance within clusters)/(Variance between clusters).
- Clustering robustness to data pertubations.