

Data Clustering and the Stability Method

Joachim M. Buhmann

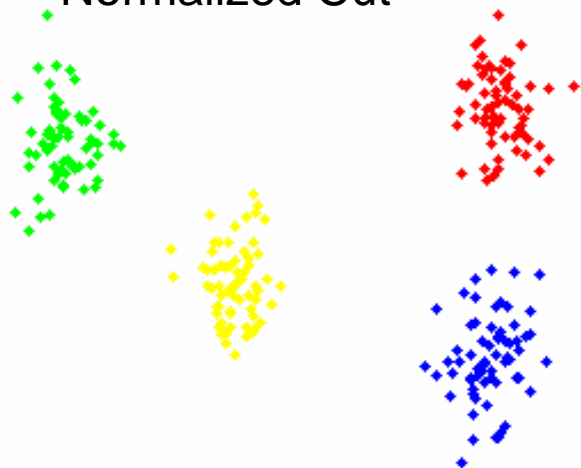
Institute for Computational Science, ETH Zurich



Grouping/Segmentation Principles

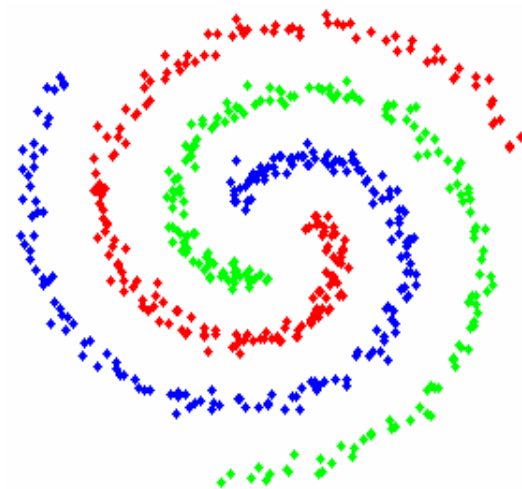
Compactness criterion

- K-Means Clustering
- Pairwise Clustering, Average Association
- Max-Cut, Average Cut
- Normalized Cut



Connectedness criterion

- Single Linkage
- Path Based Clustering



Overview of this Talk

- **My view** on clustering?
- The **stability approach** to cluster validation
(joint work with T. Lange, V. Roth, M. Braun)
- **Empirical Risk Approximation** and its connection to annealing

What is Data Clustering?

- Given are **measurements/data** $\mathbf{X} \in \mathcal{X}$ to characterize **objects** $o \in \mathcal{O}$.
- Clusterings **partition** objects into groups, i.e.,

$$c : \mathcal{O} \rightarrow \{1, \dots, k\}$$

$$o \mapsto c(o) \in \mathcal{C} \quad \text{hypothesis class}$$

- **Clustering quality:** cost function

$$R : \mathcal{X} \times \mathcal{C} \rightarrow \mathbb{R}_+$$

$$(c, \mathbf{X}) \mapsto R(c, \mathbf{X}) = \sum_{o \in \mathcal{O}} R_o(c, \mathbf{X})$$

Example: k-means clustering

- **Cost per object:**

$$R_o(c, \mathbf{X}) = \|x(o) - y_{c(o)}\|^2$$

y_α : centroids

- **Optimal clustering solution**

$$c^{\text{opt}}(o) =$$

$$\arg \min_{c \in \mathcal{C}} \mathbb{E}_{\mathbf{X}} \{ \|x(o) - y_{c(o)}\|^2 \}$$

s.t. y_α : centroids for $c^{\text{opt}}(o)$

- **Hypothesis class**

- Vector quantization

$$\mathcal{C}^{\text{VQ}} = \{c(o) : c(o) = \arg \min_{\alpha} \|x(o) - y_\alpha\|\}$$

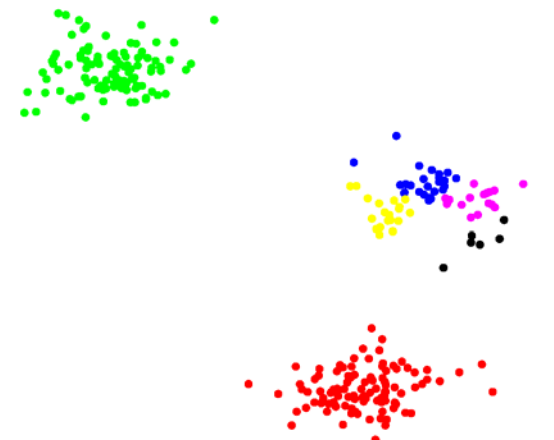
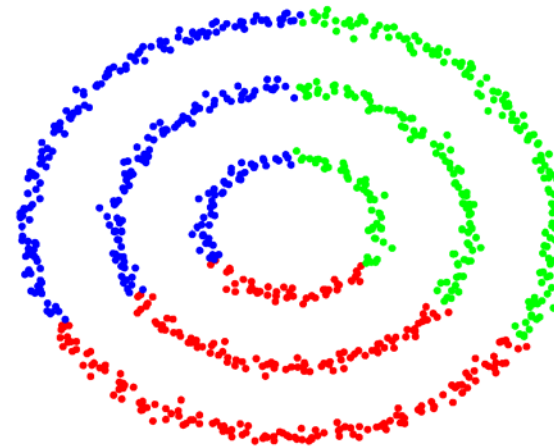
- Mixture models

$$\mathcal{C}^{\text{MM}} = \{c : \text{all partitions of } \mathcal{O}\}$$

$$\dim_{\mathcal{V}\mathcal{C}}(\mathcal{C}^{\text{MM}}) = \infty$$

The Validation Problem in Clustering

- **Modelling problem:** Does the cluster model describe the data? Selection of the costs/hypothesis class!
- **Model order selection problem:** Is the number of clusters and/or features correct?



Stability Based Cluster Validation Method

Two Sample Scenario (Lange, Roth, Braun, JB 2003)

Idea of Stability: *Solutions on two data sets from the same source should be similar!*

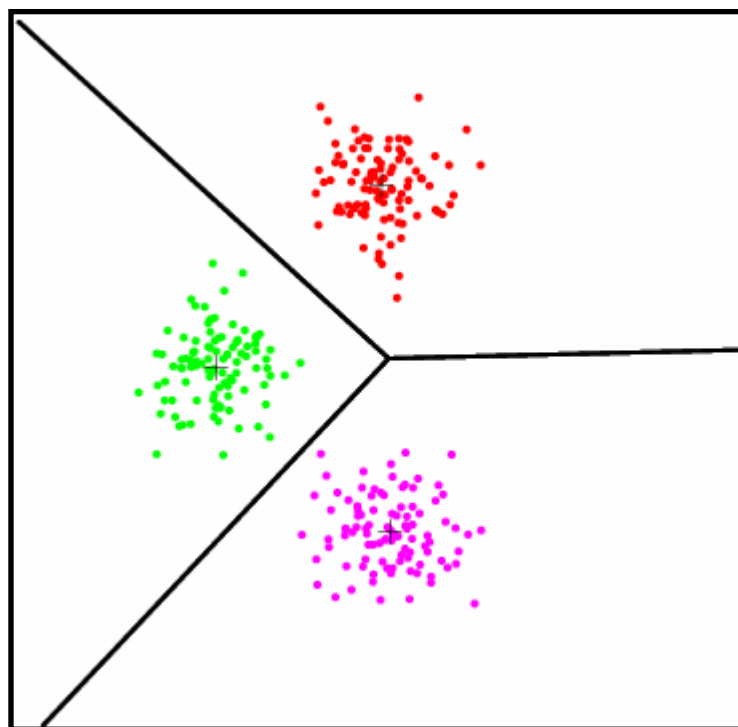
General procedure:

1. Draw two data sets X, X' from the same source.
2. Cluster both data sets using algorithm $\alpha(o, X)$.
3. Compute agreement between both solutions

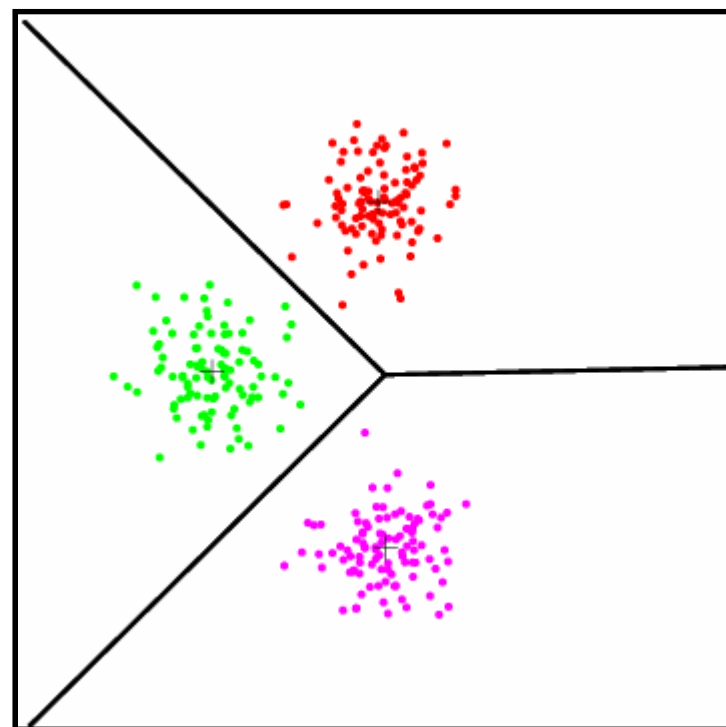
Stability := expected agreement of solutions

Practical problem: X' is not available! => resampling

Stable Solution



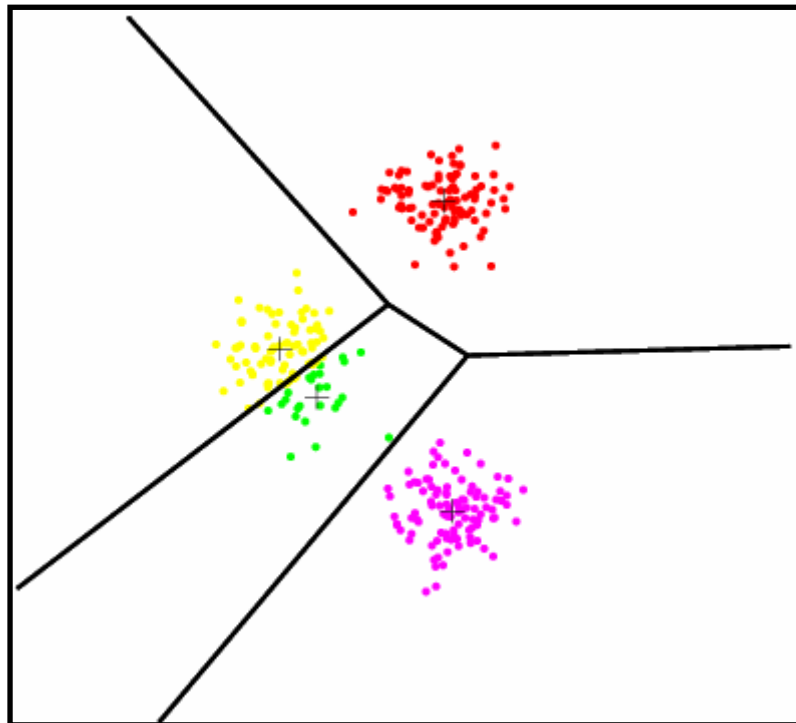
first set



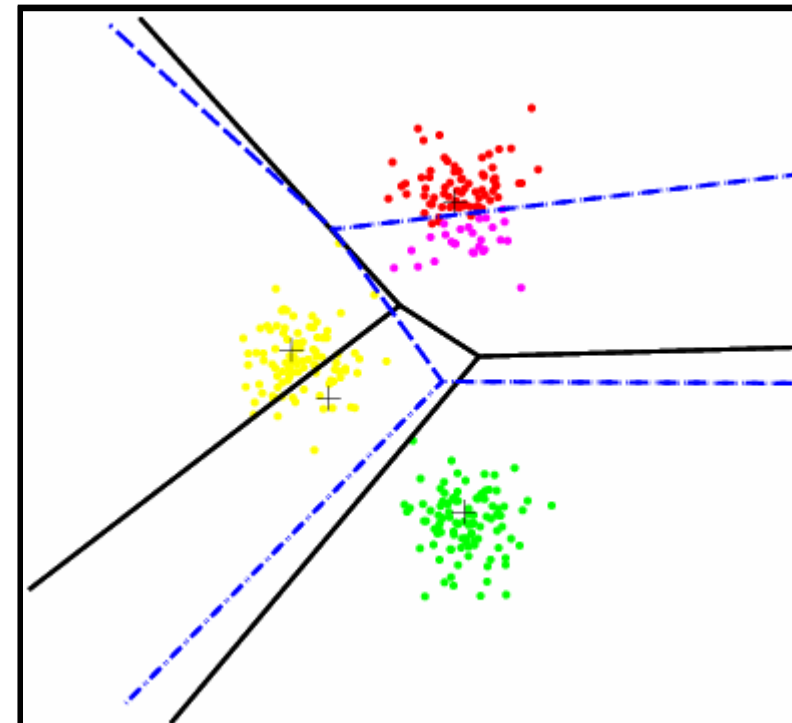
second set

Conclusion: If the model (order) is appropriate for the data then groupings on different data from the same source sets are similar with high probability.

Unstable Solution



first set



second set

Conclusion: If the model (order) is **not** appropriate for the data then groupings on different data from the same source sets are **dissimilar** with high probability.

Measuring Disagreement

Two labelings on one data set:

disagreement := fraction of differently labeled objects

Three conceptual problems:

1. Clustering solutions label disjoint sets of objects.
2. Symmetries: Labeling is unique up to permutation $\pi \in \mathcal{S}_k$.
3. Disagreement is sensitive to model complexity:
50% @ $k = 2$ -> totally random,
50% @ $k = 10$ -> often acceptable.

Extension of Solution from X to X' ...

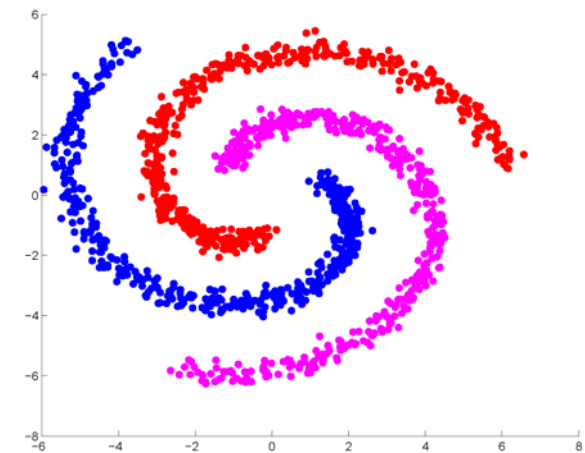
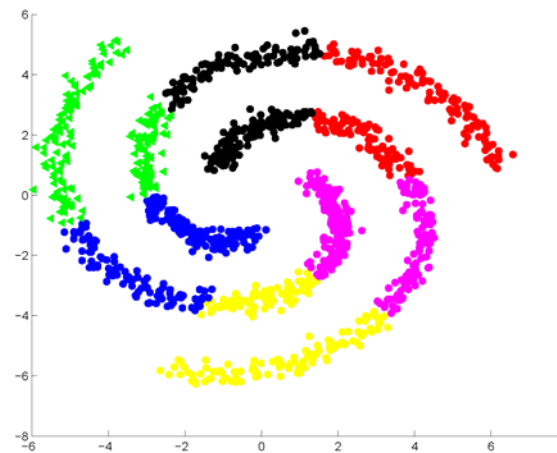
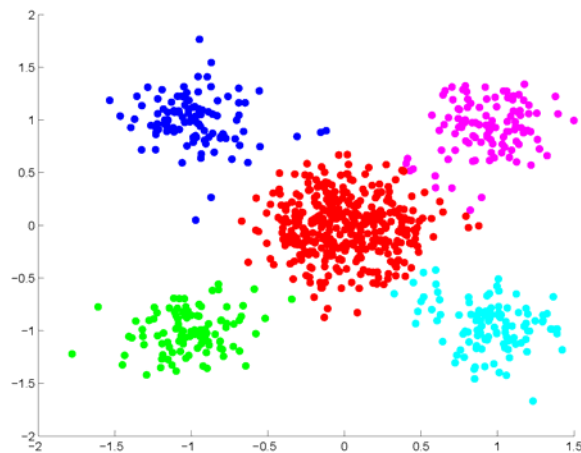
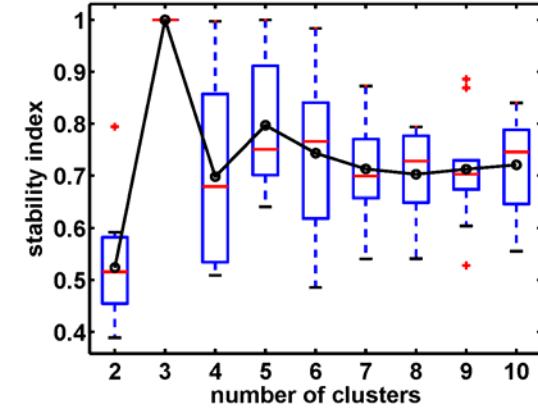
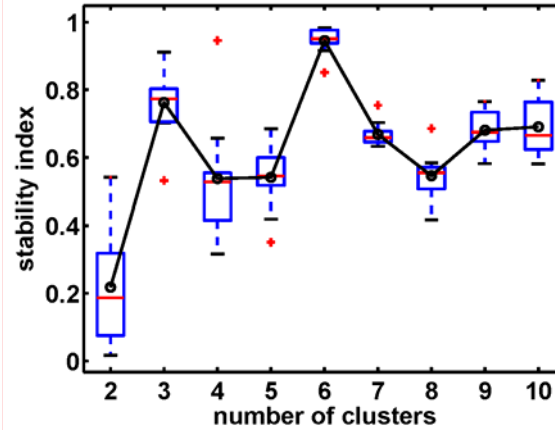
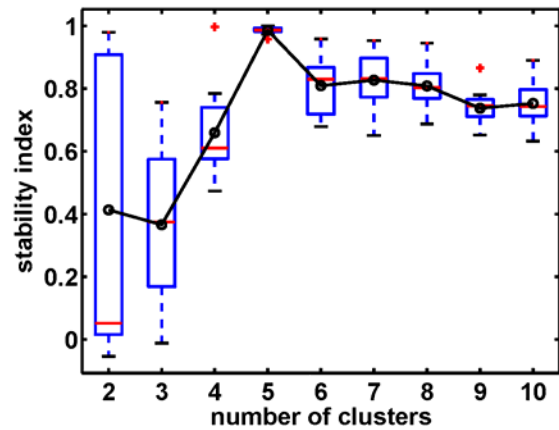
- (1) **training** a predictor on $\mathbf{X}^{(1)}$
- (2) **predicting** labels on $\mathbf{X}^{(2)}$
- (3) **compare** clusterings on $\mathbf{X}^{(2)}$
- **Replacement predictor** φ : predict labeling for object $o^{\text{new}} \in \mathcal{O}'$
 1. Find most similar (in costs) object in $o^* \in \mathcal{O}$
 2. Replace o^* by o^{new}
 3. Optimize labeling for o^{new} by conditioning on all other objects

Stability Measure for Clusterings

$$\inf_{g \in \mathcal{G}} \frac{1}{S(\rho)} \mathbb{E}_{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}} \left(\min_{\pi} \frac{1}{n} \sum_{o \in \mathcal{O}} \#\{\alpha(o, \mathbf{X}^{(2)}) = \pi \circ \varphi(o; X_o^{(2)}, \mathbf{X}^{(1)})\} \right)$$

- Disagreement rate
- Permutation symmetry breaking
- Expectation w.r.t. 2 samples from same source
- Normalize by stability of random labelings $S(\rho)$.
- Estimate $\mathbb{E}_{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}}$ by resampling

Results on Toy Data



Clustering of Microarray Data

(dataset from Golub *et al.*, Science, Oct. 1999, pp.531-537)

Task: Find groups of different Leukemia tumour samples

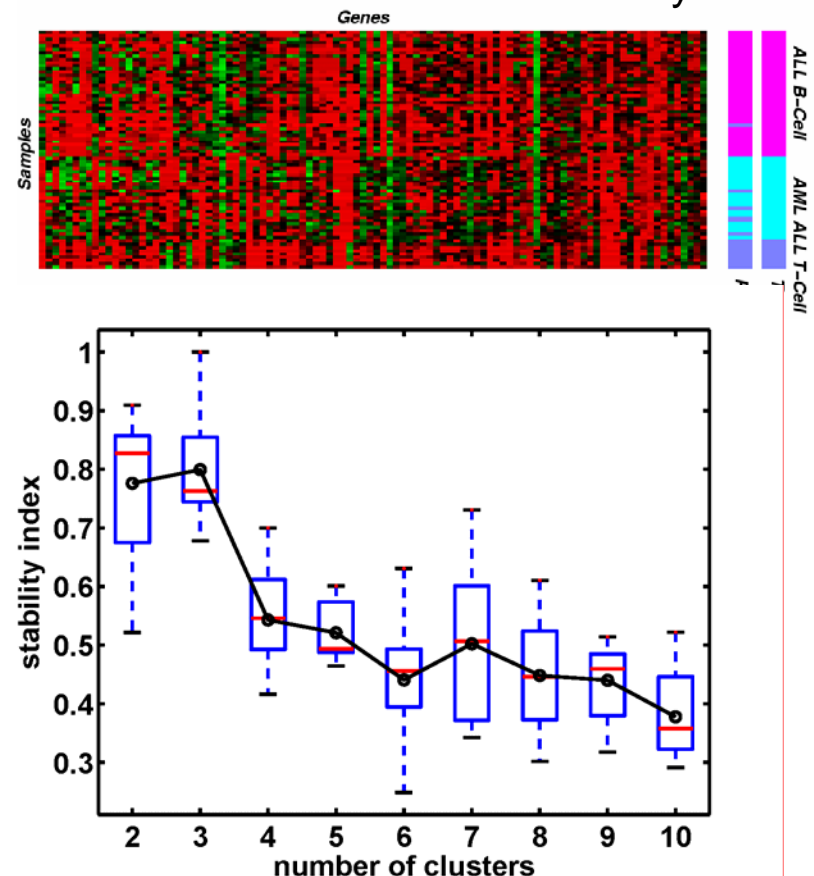
(two- and three class classifications are known).

Problem: Number of groups is unknown a priori.

Via Stability with k -means:
Estimated number of groups is 3.

Result: 3-means solution recovers $\approx 91\%$ of known sample classifications.

3-means grouping of Golub *et al.* data set and estimated instability



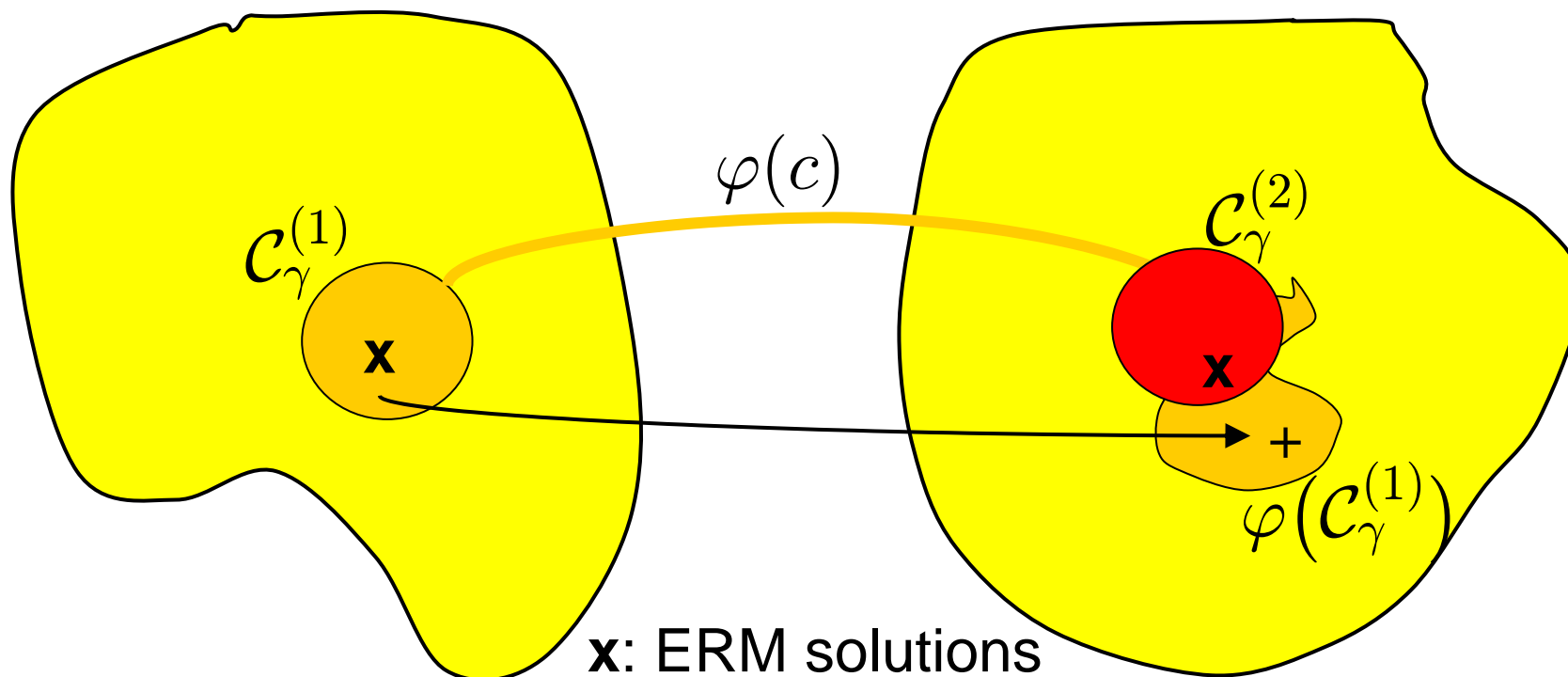
Comments on the Stability Method

- **Plus:**
 - applicable to any clustering algorithm $\alpha(o, \mathbf{X}^{(1)})$
 - success in many different pattern analysis problems (Vision, bioinformatics, ...)
- **Minus:** theoretical understanding is lacking
- **Idea:** Extend stability requirement for empirical risk minimizers of training and test instances $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ to stability of an approximation set.

Prediction of Solutions

Solution space for training instance

Solution space for test instance



x : ERM solutions

$+$: transferred ERM solution

Size of the Approximation Set?

- Optimality condition:
 - **Too small** => intersection empty $\varphi(\mathcal{C}_\gamma^{(1)}) \cap \mathcal{C}_\gamma^{(2)} = \emptyset$
or nearly empty $|\mathcal{C}_\gamma^{(1)} \cap \mathcal{C}_\gamma^{(2)}| \ll \max\{|\mathcal{C}_\gamma^{(1)}|, |\mathcal{C}_\gamma^{(2)}|\}$
=> the training solution has little to do with the test solution => overfitting
 - **Too large** => approximation is not precise enough
- Randomly sample from $\mathcal{C}_\gamma^{(2)}$ and from $\varphi(\mathcal{C}_\gamma^{(2)})$.
“Optimal” Precision: Find the smallest γ for which both sets are maximally overlapping.

Stochastic Approximation

Learning procedure: sample typical solutions
from an approximation set

$$c_\gamma \in \mathcal{C}_\gamma^{(1)} = \{c : R(c, \mathbf{X}^{(1)}) \leq \min_{\tilde{c}} R(\tilde{c}, \mathbf{X}^{(1)}) + \gamma\}$$

Generalization performance: $c^\perp := \arg \min_c R(c, \mathbf{X}^{(2)})$

$$\mathbb{E}_{\mathbf{X}^{(2)}} \{ R(\varphi(c_\gamma), \mathbf{X}^{(2)}) - R(c^\perp, \mathbf{X}^{(2)}) \}$$

$\varphi(c)$ maps solutions from the training instance $\mathbf{X}^{(1)}$ to solutions of the test instance $\mathbf{X}^{(2)}$ by prediction.

Vapnik-Chervonenkis Inequality

Bounding test performance of training solution

$$\begin{aligned} R\left(\varphi(c_\gamma), \mathbf{X}^{(2)}\right) - R\left(c^\perp, \mathbf{X}^{(2)}\right) \leq \\ R\left(\varphi(c_\gamma), \mathbf{X}^{(2)}\right) - R\left(c_\gamma, \mathbf{X}^{(1)}\right) + \\ R\left(\varphi^{-1}(c^\perp), \mathbf{X}^{(1)}\right) - R\left(c^\perp, \mathbf{X}^{(2)}\right) + \gamma \end{aligned}$$

Take expectations w.r.t. test data $\mathbf{X}^{(2)}$

Bound on Expected Performance

Vapnik-Chervonenkis inequality $c^\perp := \arg \min_c R(c, \mathbf{X}^{(2)})$

$$\begin{aligned} E_{\mathbf{X}^{(2)}} \left\{ R\left(\varphi(c_\gamma), \mathbf{X}^{(2)}\right) - R\left(c^\perp, \mathbf{X}^{(2)}\right) \right\} &\leq \gamma + \\ 2 \max \left\{ \mathbb{E} R\left(\varphi(c_\gamma), \mathbf{X}^{(2)}\right) - R\left(c_\gamma, \mathbf{X}^{(1)}\right), \right. \\ &\quad \left. \mathbb{E} R\left(\varphi^{-1}(c^\perp), \mathbf{X}^{(1)}\right) - \mathbb{E} R\left(c^\perp, \mathbf{X}^{(2)}\right) \right\} \end{aligned}$$

Take expectations w.r.t. test data $\mathbf{X}^{(2)}$

Probability of Large Deviation

Estimate probability of large deviations

$$\begin{aligned} & \mathbb{P} \left\{ \mathbb{E} \mathcal{R}^{(2)} (\varphi(c_\gamma)) - \mathbb{E} \mathcal{R}^{(2)} (c^\perp) > \epsilon + \gamma \right\} \leq \\ & \mathbb{P} \left\{ \left| \mathbb{E} \mathcal{R}^{(1)} (\varphi^{-1}(c^\perp)) - \mathbb{E} \mathcal{R}^{(2)} (c^\perp) \right| > \frac{\epsilon}{2} \right\} + \\ & \mathbb{P} \left\{ \left| \mathcal{R}^{(1)} (c_\gamma) - \mathbb{E} \mathcal{R}^{(2)} (\varphi(c_\gamma)) \right| > \frac{\epsilon}{2} \right\} \end{aligned}$$

1st term can be bounded in simple cases by Hoeffding or Bernstein inequality since c^\perp does not depend on training data.

2nd term requires uniform convergence since c_γ is data dependent.

Union Bound / Uniform Convergence

Estimate probability of large deviations

$$\mathbb{P} \left\{ \left| \mathbb{E} \mathcal{R}^{(1)} (\varphi^{-1}(c^\perp)) - \mathbb{E} \mathcal{R}^{(2)} (c^\perp) \right| > \frac{\epsilon}{2} \right\} \lesssim 2 \exp(-\lambda n \epsilon^2)$$

$$\mathbb{P} \left\{ \left| \mathcal{R}^{(1)} (c_\gamma) - \mathbb{E} \mathcal{R}^{(2)} (\varphi(c_\gamma)) \right| > \frac{\epsilon}{2} \right\} \lesssim 2 \frac{|\mathcal{C}|}{|C_\gamma|} \exp(-\lambda n \epsilon^2)$$

Bound on expected risk

$$\mathbb{E} \mathcal{R}^{(2)} (\varphi(c_\gamma)) \lesssim \mathbb{E} \min_{c \in \mathcal{C}} \mathcal{R}^{(2)} (c) + \gamma + c \sqrt{\log(1 + \frac{|\mathcal{C}|}{c_\gamma}) + \log \frac{2}{\delta}}$$

Relation to Gibbs Sampling

Relation to statistical mechanics of learning:

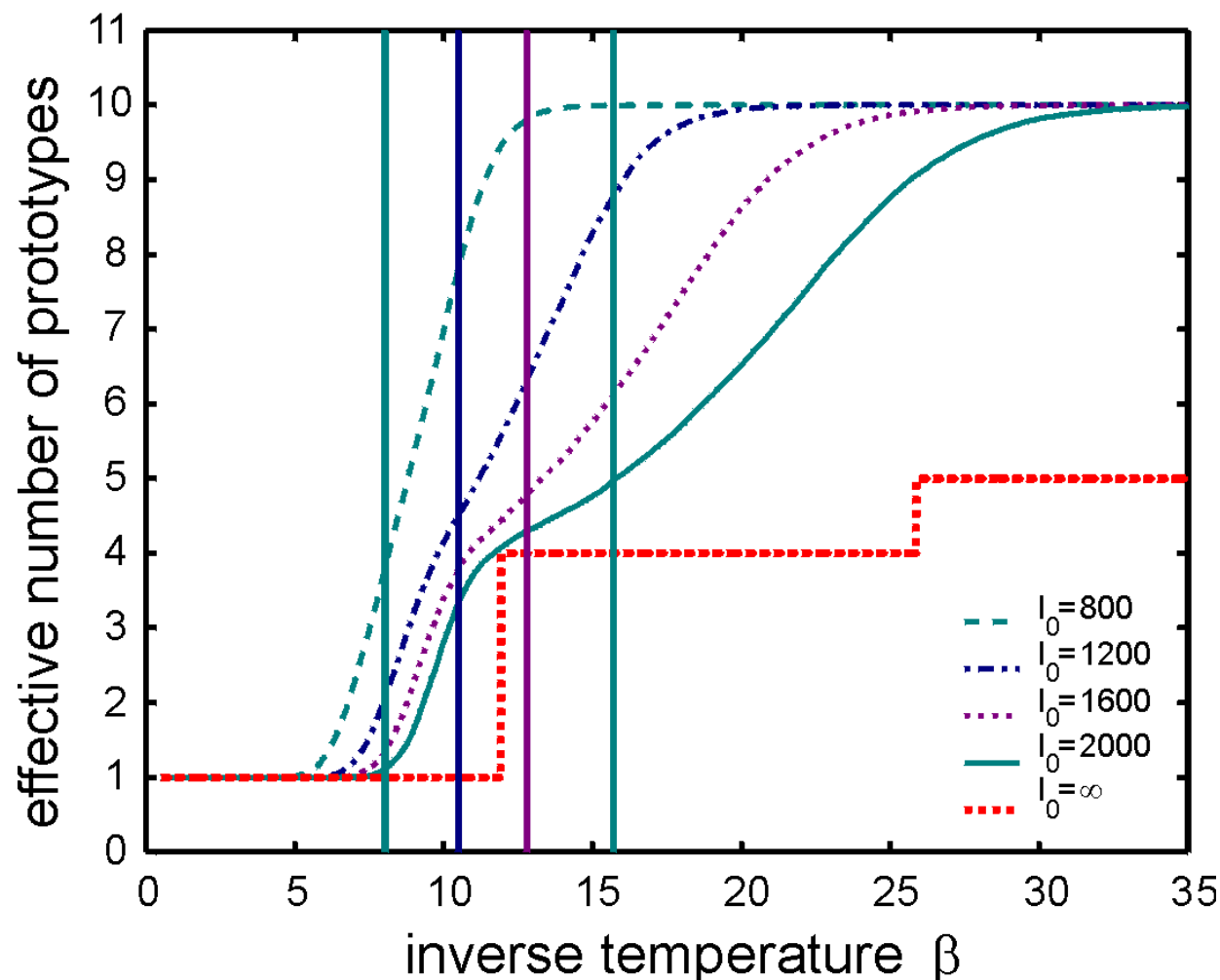
determine γ for minimum of bound $\left(\frac{d \text{ bound}}{d\gamma} = 0\right)$

$$\frac{d \text{ entropy}}{d \text{ energy}} = \frac{d \log |\mathcal{C}_\gamma|}{d\gamma} = T^{-1} \Rightarrow$$

$$\frac{1}{T^{\text{stop}}} \approx c \sqrt{\log\left(1 + \frac{|\mathcal{C}|}{|\mathcal{C}_\gamma|}\right) + \log \frac{2}{\delta}}$$

Gibbs Sampling with temperature $T \geq T^{\text{stop}}$

Estimate of Stopping Temperature



Experiment:

Data are drawn from a model with $k=5$ groups.

Inference algorithm assumes $k^{\max}=10$ groups.

Important: we do not infer more than 5 groups!

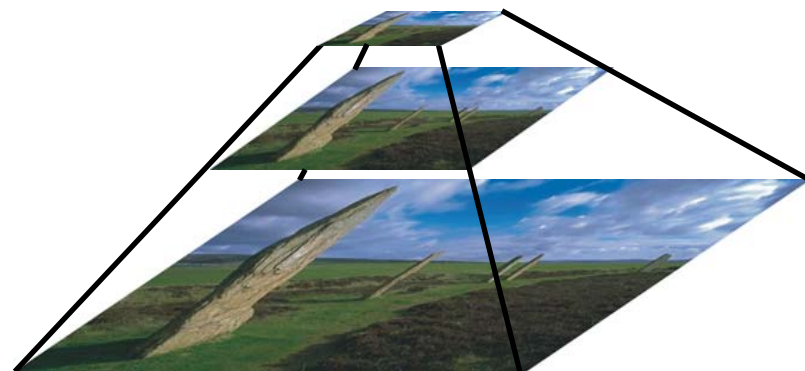
Inferred parameters are similar to the true parameter values.

Scales in Data Analysis and Vision

Coarsening of Variable Space



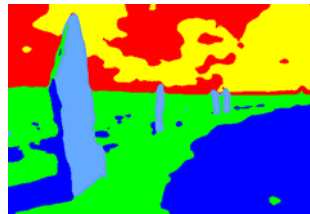
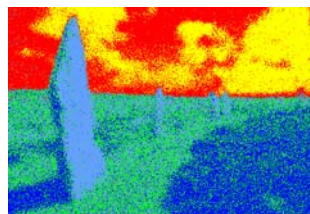
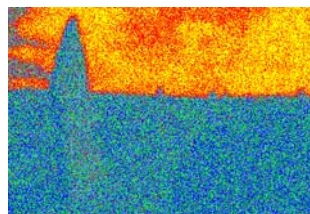
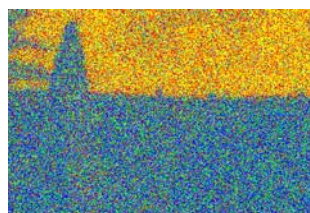
coarse



fine

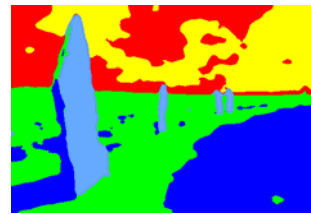
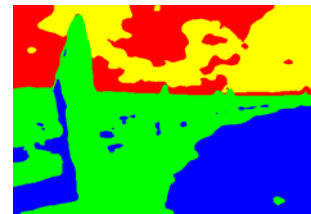
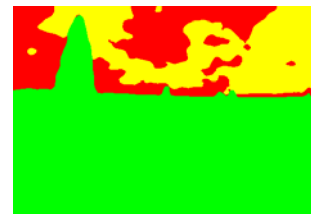
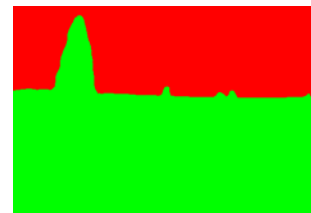
Increment Level of Resolution Pyramid

Coarsening of Optimization Criterion



Increase Regularization

Coarsening of Model Order



Reduce # of Segments

Conclusion & Open Issues

- Stability provides a convincing framework to adjust model complexity!
- What are the components of a theory which optimally trades stability against informativity?
- Empirical Risk Approximation requires a thorough Information Theory basis!
- What can we learn from clustering for other combinatorial optimization problems?