ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Cluster Stability and Robust Optimization
- An Idea

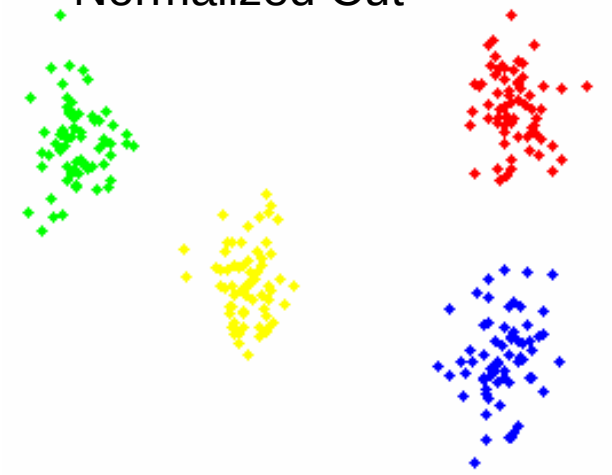
Joachim M. Buhmann

Institute for Computational Science, ETH Zurich

Grouping/Segmentation Principles

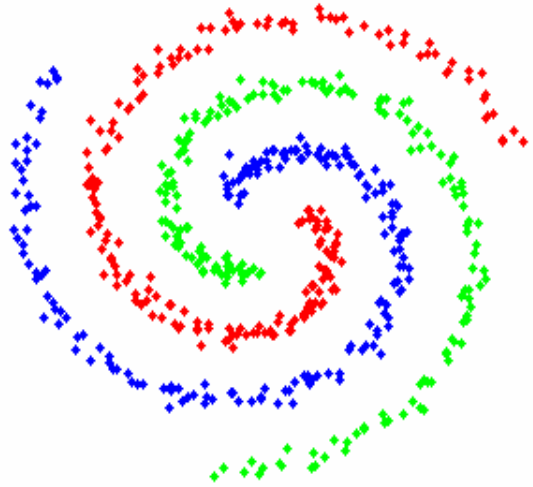
Compactness criterion

- 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 and an analogy to source channel coding.
- **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.,

$$c \;:\; \mathcal{O} \to \{1, \ldots, k\}$$
$$o \mapsto c(o) \in \mathcal{C} \quad \text{hypothesis class}$$

- **Clustering quality**: cost function

$$R \;:\; \mathcal{X} \times \mathcal{C} \;\to \mathbb{R}_+$$
$$(c, \mathbf{X}) \mapsto R(c, \mathbf{X}) = \textstyle\sum_{o \in \mathcal{O}} R_o(c, \mathbf{X})$$

Example: k-means clustering

- **Cost** per object:

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

$$y_\alpha : \text{centroids}$$

- **Optimal** clustering solution

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

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

- **Hypothesis** class
  - Vector quantization

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

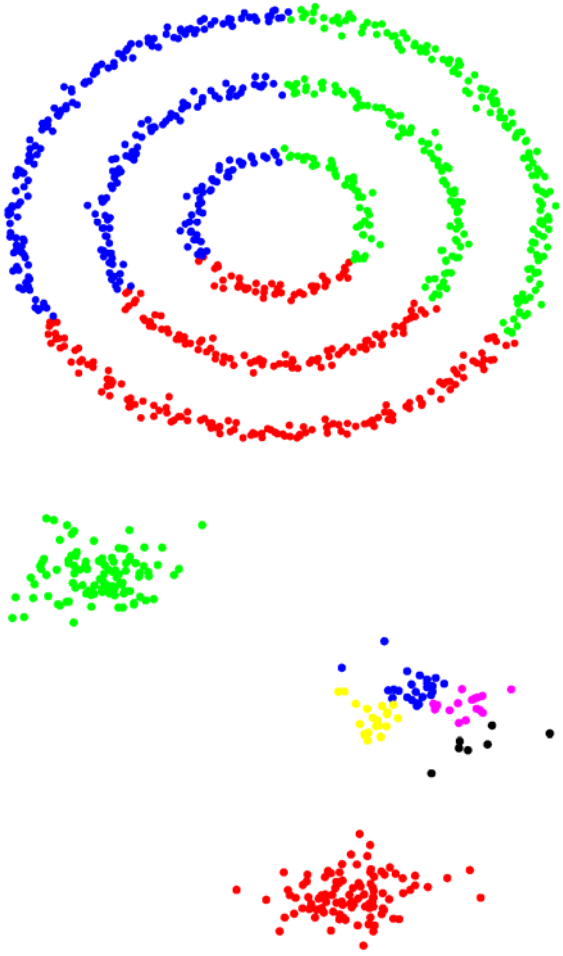
  - Mixture models

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

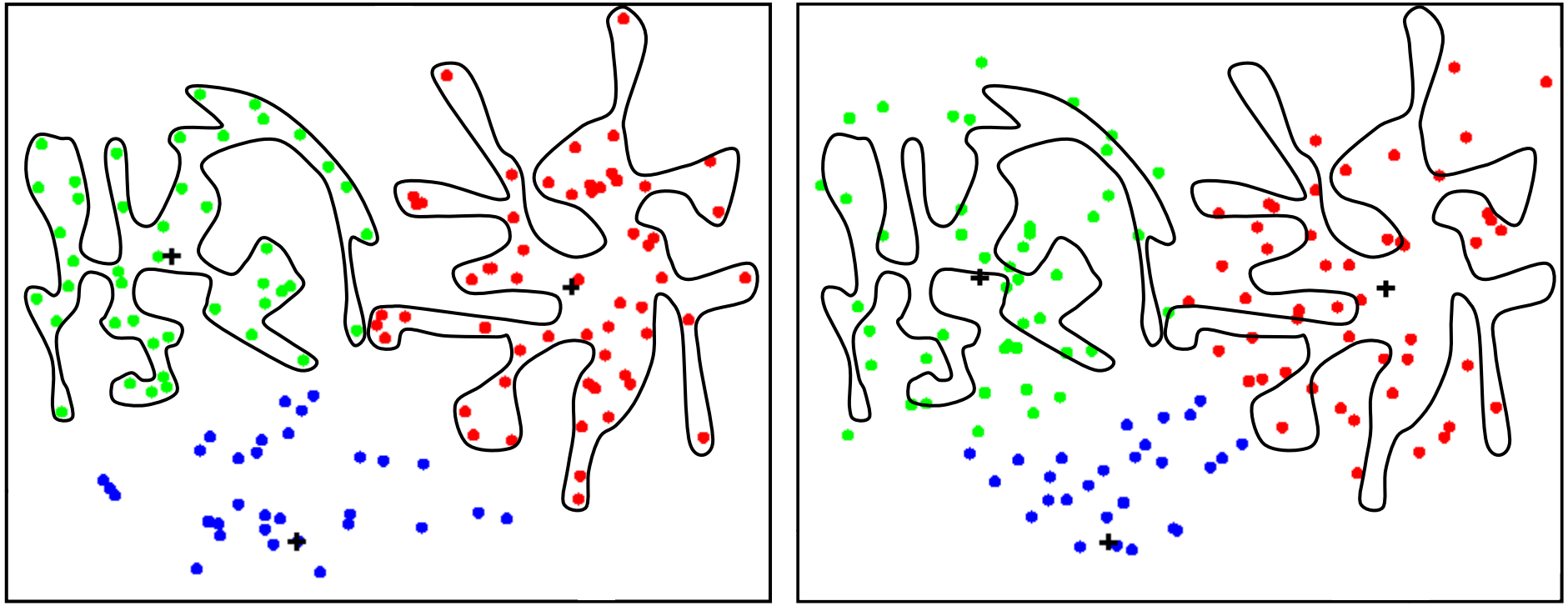
$$\dim_{\mathcal{VC}}(\mathcal{C}^{\text{MM}}) = \infty$$

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?

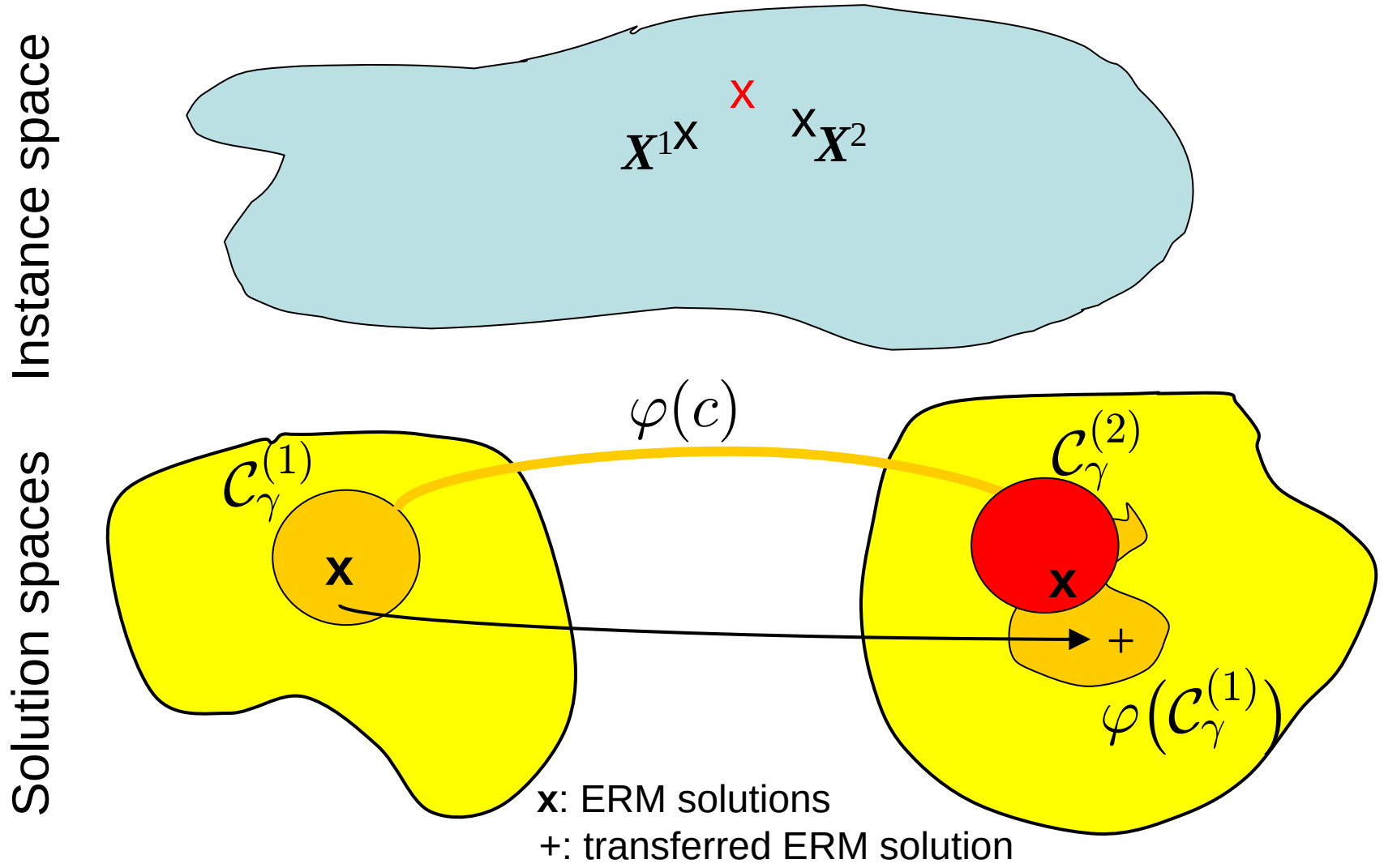


ng in Clustering



Requirement: Structures in two different data sets of the same data source should have approximately the same quality (costs)!

Two Instance Scenario



Information Theoretic Idea to Control Approximation

- Use **data partition** as a k-ary code
- **Communication** is achieved via **instances** since test instances are perceived as perturbed training instances
- Determine how well a partition can be approximated when you see a test instance.
- Space filling argument yields

  k^(entropy of partition type) / Card(approx. set)

Size of the Approximation Set?

- Optimality condition:
  - **Too small** => intersection empty $\varphi\left(\mathcal{C}_\gamma^{(1)}\right) \cap \mathcal{C}_\gamma^{(2)} = \emptyset$ 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

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

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

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

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

Vapnik-Chervonenkis Inequality

Bounding test performance of training solution

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

$$R\left(\varphi(c_\gamma), \mathbf{X}^{(2)}\right) - R\left(c_\gamma, \mathbf{X}^{(1)}\right) +$$

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

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

Bound on Expected Performance

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

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

$$2\max\left\{\mathbb{E}R\left(\varphi(c_\gamma), \mathbf{X}^{(2)}\right) - R\left(c_\gamma, \mathbf{X}^{(1)}\right),\right.$$

$$\left.\mathbb{E}R\left(\varphi^{-1}(c^{\perp}), \mathbf{X}^{(1)}\right) - \mathbb{E}R\left(c^{\perp}, \mathbf{X}^{(2)}\right)\right\}$$

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

Probability of Large Deviation

Estimate probability of large deviations

$$\mathbb{P}\left\{\mathbb{E}\mathcal{R}^{(2)}\left(\varphi(c_\gamma)\right) - \mathbb{E}\mathcal{R}^{(2)}\left(c^\perp\right) > \epsilon + \gamma\right\} \leq$$

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

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

1st term can be bounded in simple cases by Hoeffding or Bernstein inequality since $c^\#$ 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

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

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

Bound on expected risk

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

Relation to Gibbs Sampling

Relation to statistical mechanics of learning:

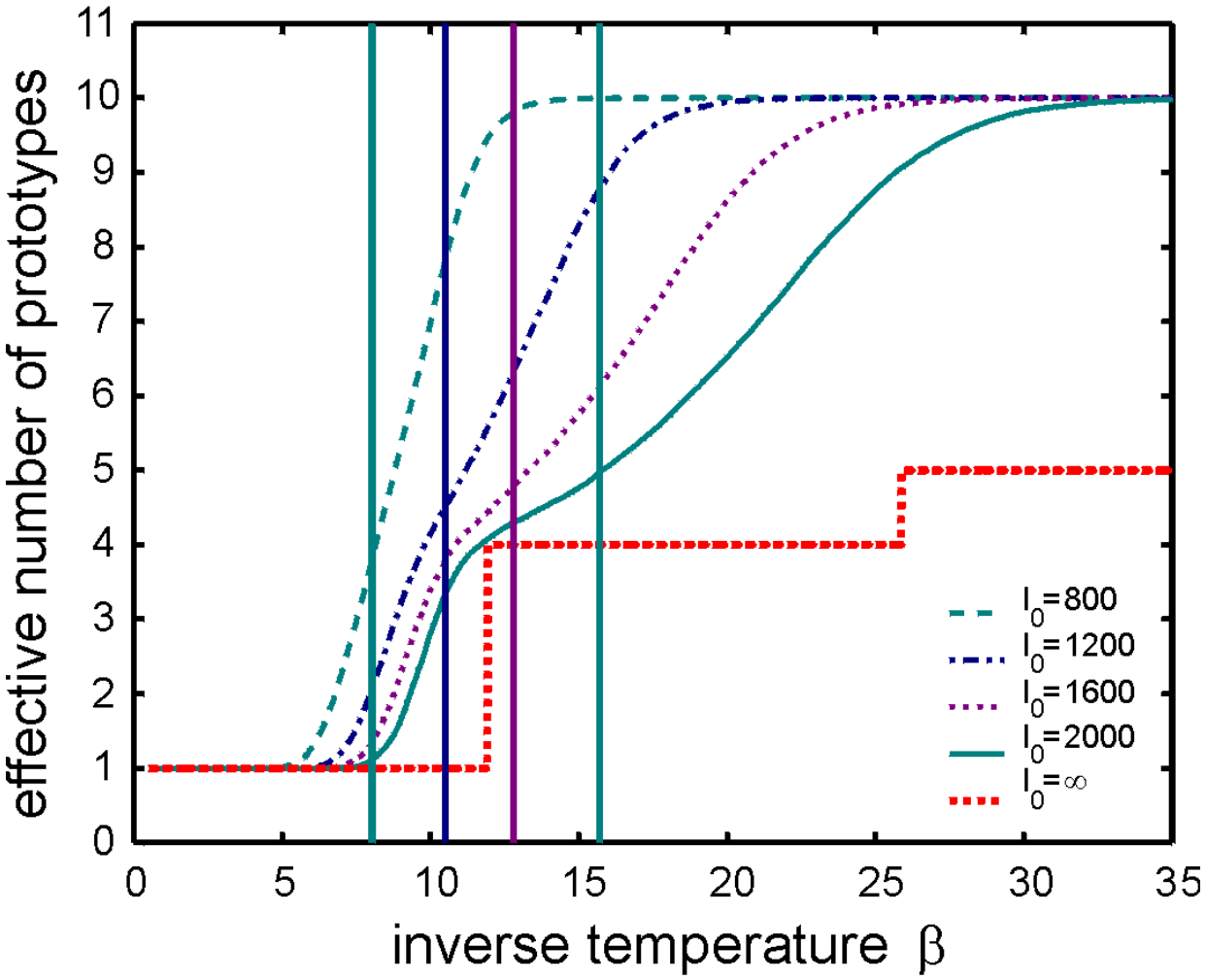
determine γ for minimum of bound $\left(\frac{d\,\textbf{bound}}{d\gamma} = 0\right)$

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

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

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

Estimate of Stopping Temperature



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

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

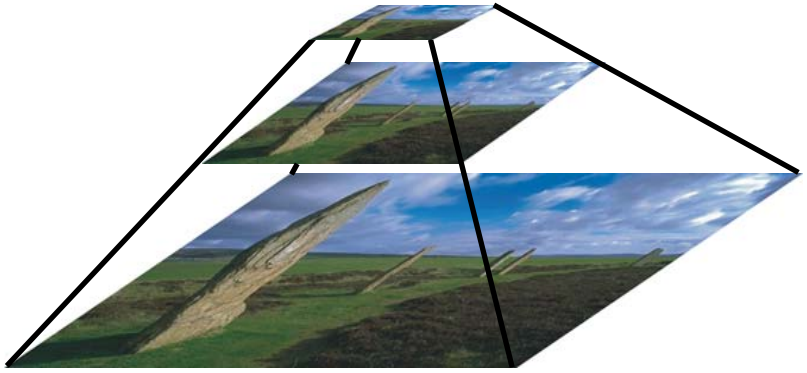
<u>Important</u>: we do not infer more than 5 groups!

Infered parameters are similar to the true parameter values.

Scales in Data Analysis and Vision

Coarsening of Variable Space

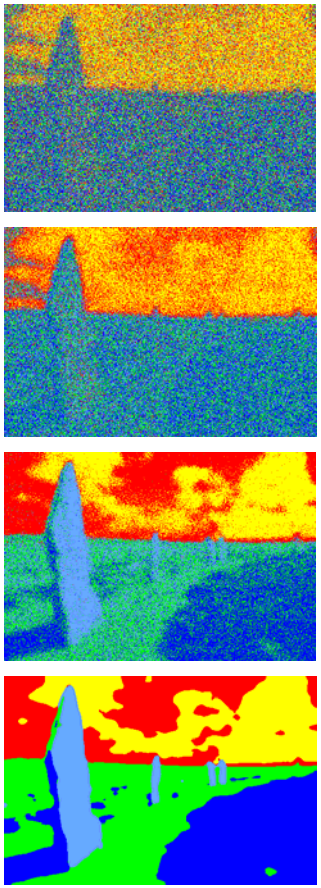
coarse



fine

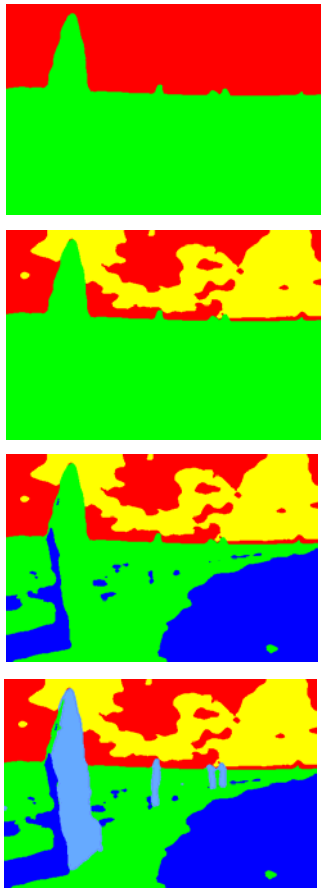
Increment Level of Resolution Pyramid

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?