ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

SIMBAD

inf Informatik Computer Science

Information Theoretic Model Validation by Approximate Optimization

Joachim M. Buhmann

Computer Science Department, ETH Zurich

Overview

- Motivation of information theory for optimization
- Approximation capacity of a cost function
- Examples
  - Binary symmetric channel
  - Cluster validation
  - Role based access control
  - Robust SVD
- Conclusion and outlook

What is the central challenge of pattern recognition?

- **I)** Finding the "right" model? **II)** Validating a model?
- **Hypothesis**: **Validation of pattern recognition models is the fundamental challenge!**

⇒Algorithmic search for PR models should prefer ***noise tolerant*** and ***expressive*** models over brittle, simplistic ones! (***stability vs informativeness***)

⇒**Information theory** enables us to measure the **context sensitive information** content of models!

Information Theory & Pattern Recognition

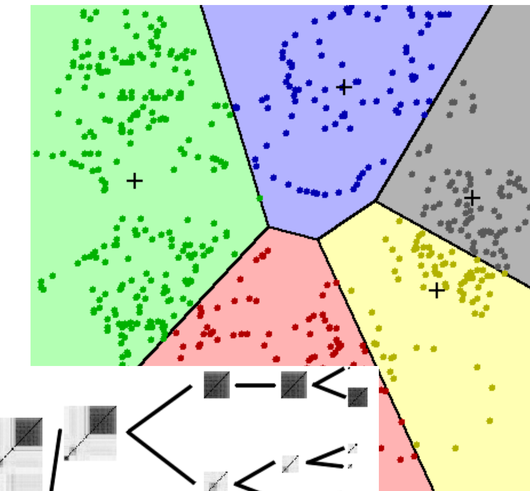
- IT-Components
  - **Code vectors $\subsetneq$ {strings}** = hypothesis class
  - Noisy **channel**
  - Decoder: minimize **Hamming distance**
- Criterion for error free communication
  => **mutual information**

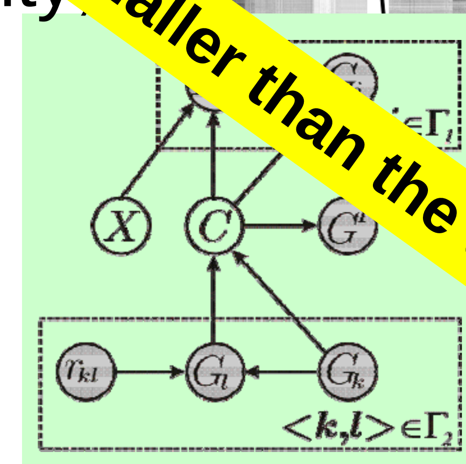
- Pattern Recognition elements
  - **Approximation sets $\subsetneq$ hypothesis class**
  - Noisy **optimization** problem
  - Decoding by **approximate optimization** of test instance
- model validation based on guaranteed approximations
  => **mutual information**

Examples of hypothesis classes

The hypothesis class is much smaller than the data space!

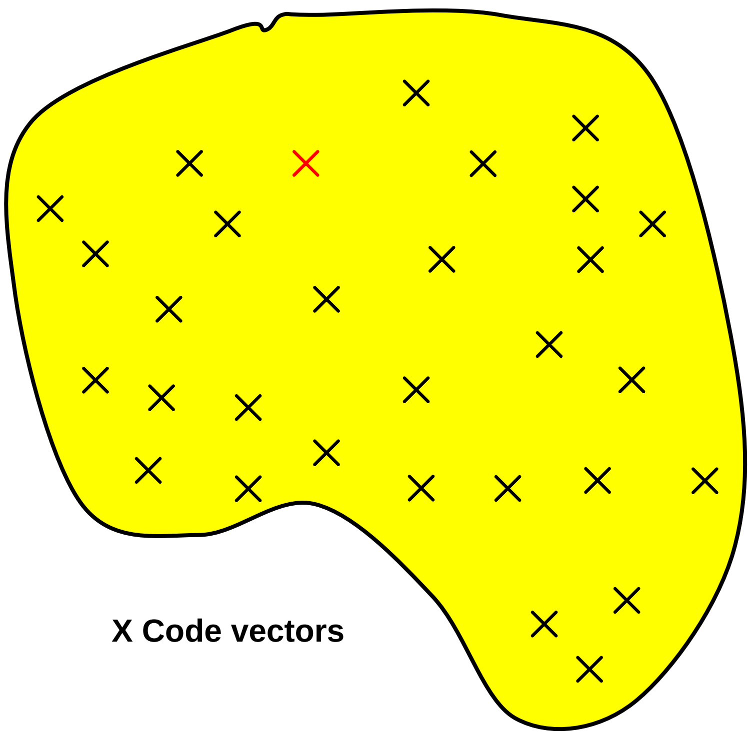
- **Partitions or clusterings**: compactness/connectivity costs
- **Trees or dendrograms**: partitions with ultrametricity; Tree depth? # leaves?
- **Graphical models**: structured probability models; # nodes/edges?

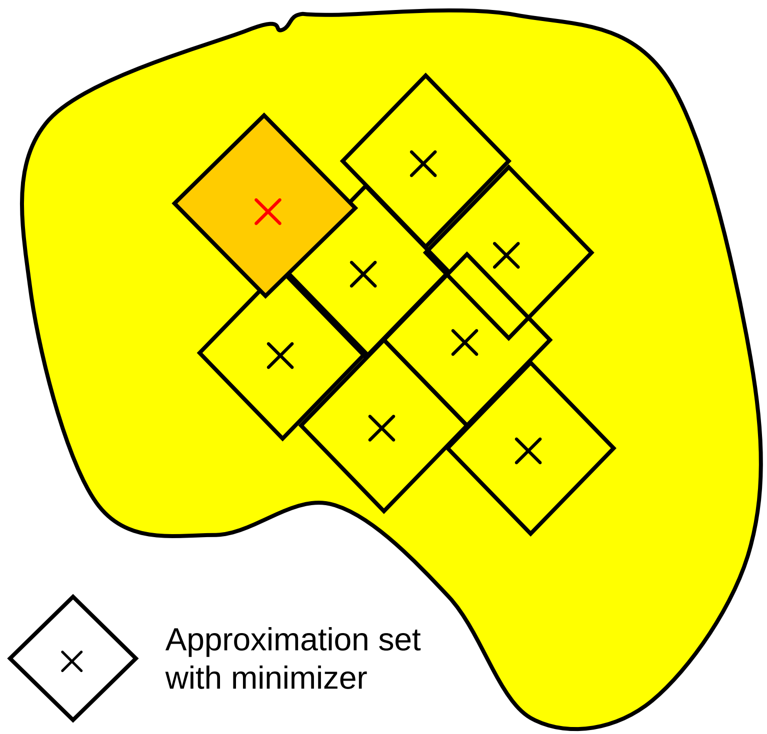




Code problems define approximation sets

- IT: **Space of strings** is partitioned by code vectors
- PR: **Hypothesis class** is partitioned by code problems





Pattern recognition as optimization

- Given: **data** $\mathbf{X} \in \mathcal{X}$ **in data (input) space** $\mathcal{X}$
- **Goal**: **Learn structure from data**, i.e., interpret data relative to a hypothesis class
- **Hypothesis class** $\mathcal{C}$ with hypotheses (solutions)

$$\begin{aligned} c \; : \; \mathcal{X} &\rightarrow \mathbb{K} \qquad \text{(e.g., } \mathbb{B}^n \text{ or } \{1,\dots,k\}^n\text{)} \\ \mathbf{X} &\mapsto c(\mathbf{X}) \end{aligned}$$

- **Cost function** to define a partial order on $\mathcal{C}$

$$\begin{aligned} R \; : \; \mathcal{C} \times \mathcal{X} &\rightarrow \mathbb{R}_{\geq 0} \\ (c, \mathbf{X}) &\mapsto R(c, \mathbf{X}) \end{aligned}$$

Symmetries of the Learning Problem

- Assume (!) that the cost function R is equivariant under the transformations

$$\Sigma = \{\sigma : R(c, \mathbf{X}) = R(\sigma \circ c, \sigma \circ \mathbf{X})\}$$

- Minimizer: $c^{\perp}(\mathbf{X}) = \arg\min_{c \in \mathcal{C}} R(c, \mathbf{X})$

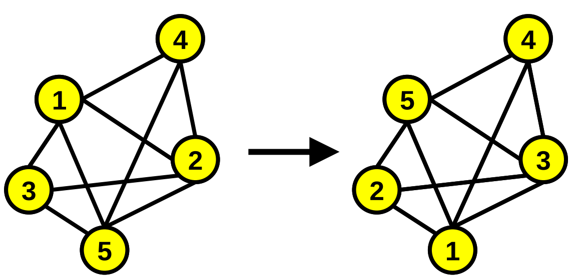
$$c^{\perp}(\sigma \circ \mathbf{X}) = \sigma \circ c^{\perp}(\mathbf{X})$$

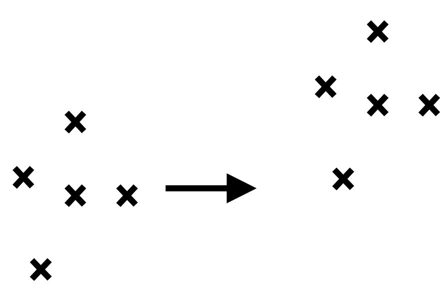
- **Approximation set:**

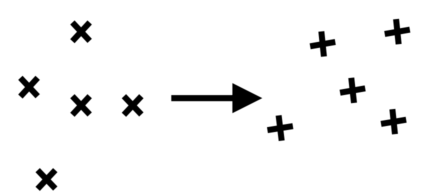
$$c \in \mathcal{C}_{\gamma}(\mathbf{X}) \equiv \left\{c : R(c, \mathbf{X}) \leq R\left(c^{\perp}, \mathbf{X}\right) + \gamma\right\}$$

How to generate code problems?

1. Combinatorial optimization problems: **permutation** of combinatorial components, e.g., vertices in graphs
2. Localization problems: **shifts** of data
3. Orientation problems (PCA, SVD): **rotations**

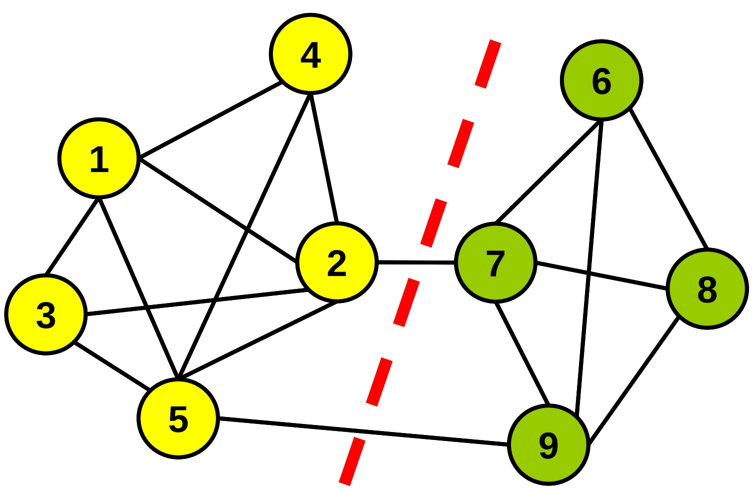






Ex.: Graph Cut - Clustering in two groups

- **Graph** representation: **vertices** denote **objects**
  edges express (dis)**similarities**
- **Hypothesis class:** all **cuts** of a graph

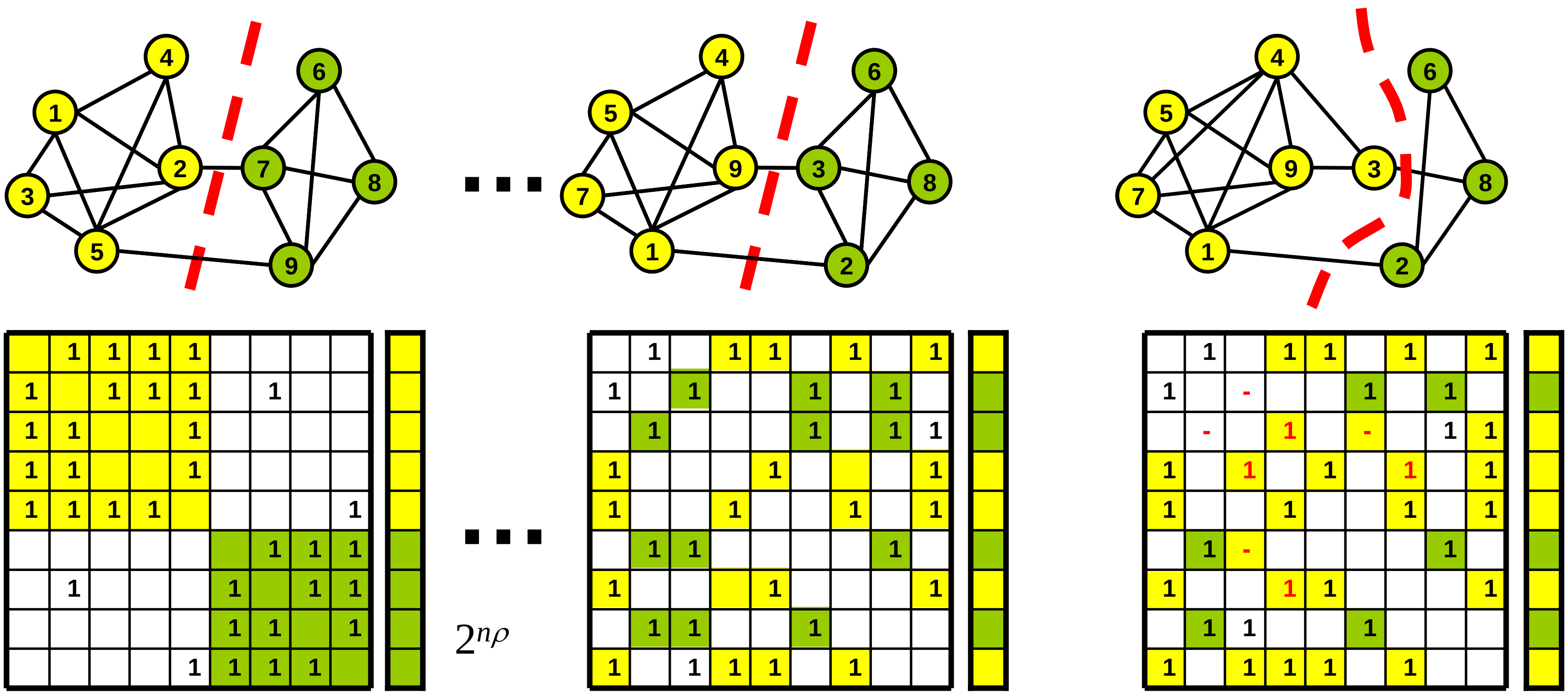


	1	1	1	1				
1		1	1	1		1		
1	1			1				
1	1			1				
1	1	1	1					1
						1	1	1
	1				1		1	1
					1	1		1
				1	1	1	1	

Code problem generation for Graph Cut

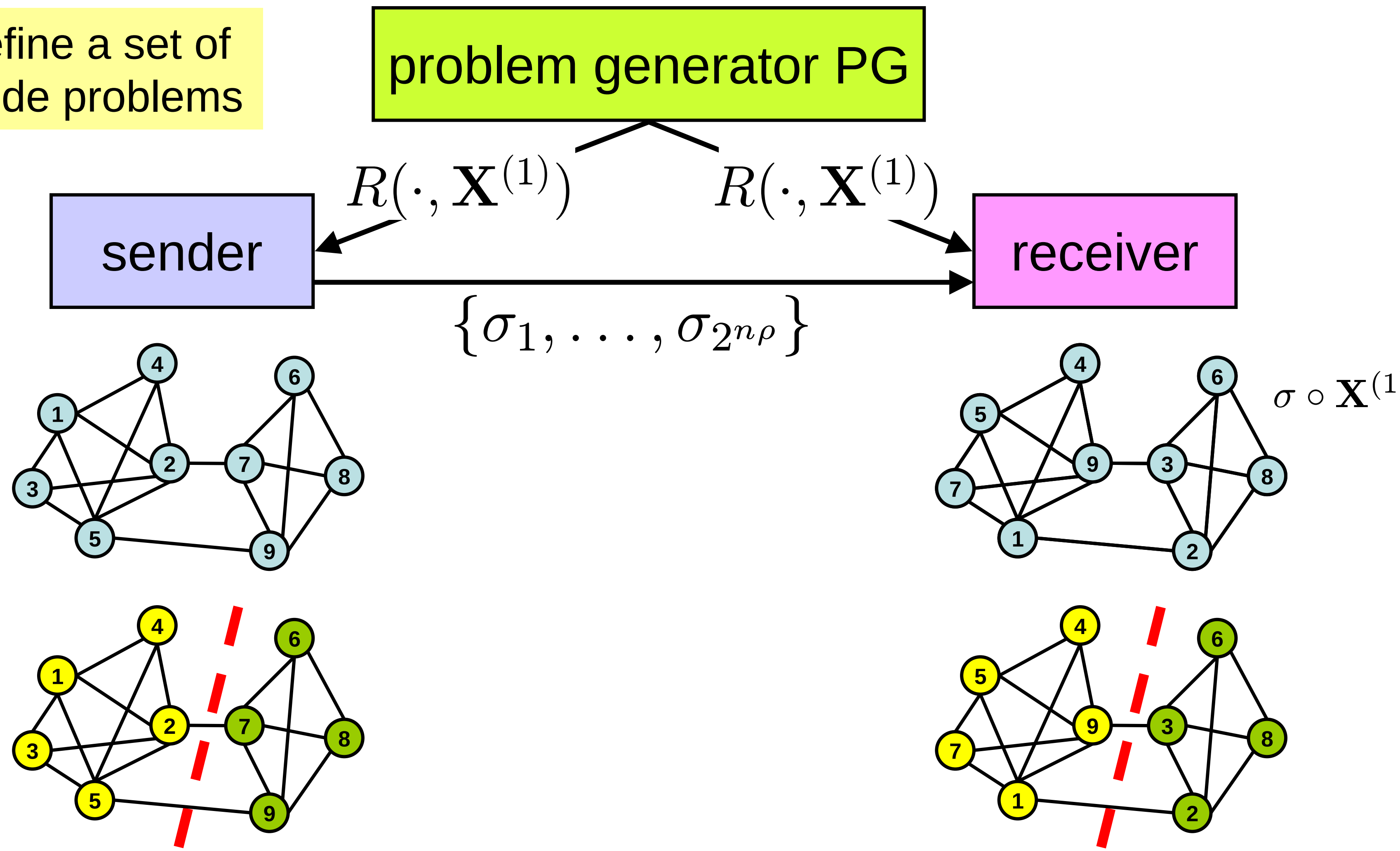
graph cut code problems

graph cut test problem



Coding with Graph Cut approximation sets

define a set of
code problems



Communication by approximation sets

estimate the coding error

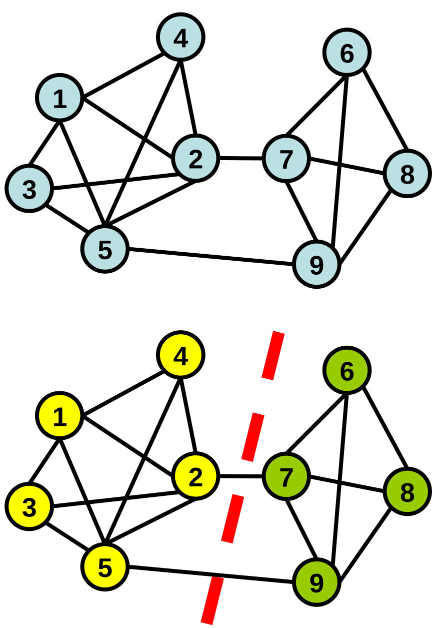
sender

σ_s

problem generator

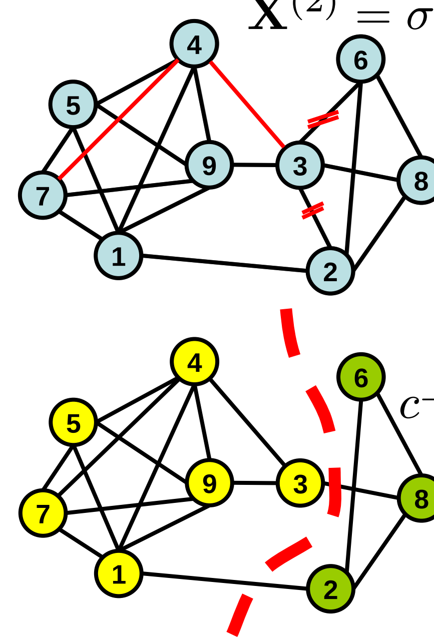
$$R(\cdot, \sigma_s \circ \mathbf{X}^{(2)}), \text{s.t.}$$
$$\mathbf{X}^{(1)}, \mathbf{X}^{(2)} \sim P(\mathbf{X})$$

receiver $\rightarrow \hat{\sigma}$



1. Sender sends a permutation index σ_s to problem generator.
2. Problem generator sends a new problem with permuted indices to receiver without revealing σ_s.
3. Receiver identifies the permutation $\hat{\sigma}$ by comparing approximation sets.

$\tilde{\mathbf{X}}^{(2)} = \sigma_s \circ \mathbf{X}^{(2)}$



$c^{\perp}(\tilde{\mathbf{X}}^{(2)})$

Communication Process

- Receiver has to **compare sets of hypotheses** $\mathcal{C}_\gamma(\mathbf{X}^{(1)})$ of training instance (code problem) with approximate clusterings $\mathcal{C}_\gamma(\mathbf{X}^{(2)})$ of the test data.
- Define a mapping $\psi : \mathcal{C}(\mathbf{X}^{(1)}) \to \mathcal{C}(\mathbf{X}^{(2)})$
- **Decoding by overlap maximization** $(\tilde{\mathbf{X}}^{(2)} := \sigma_s \circ \mathbf{X}^{(2)})$

$$\hat{\sigma} = \arg\max_{\sigma} \left| \psi \circ \mathcal{C}_\gamma(\sigma \circ \mathbf{X}^{(1)}) \cap \mathcal{C}_\gamma(\tilde{\mathbf{X}}^{(2)}) \right|$$

Error Events and Approximation Capacity

- Sender selects transformation $\sigma_s \;\Rightarrow\; \mathcal{C}_\gamma(\sigma_s \circ \mathbf{X}^{(1)})$
- **Joint approximation sets**

$$\Delta\mathcal{C}_j \quad = \quad \psi \circ \mathcal{C}_\gamma(\sigma_j \circ \mathbf{X}^{(1)}) \cap \mathcal{C}_\gamma(\sigma_s \circ \mathbf{X}^{(2)})$$
$$\text{for } 1 \leq j \leq 2^{n\rho},\ j \neq s$$

- **Error events** :

$$|\Delta\mathcal{C}_j| \geq |\Delta\mathcal{C}_s| \text{ for } 1 \leq j \leq 2^{n\rho},\ j \neq s$$

Error Probability

- Conditional error

$$
\begin{aligned}
P(\hat{\sigma} \neq \sigma_s | \sigma_s) &= P(\max_{j \neq s} |\Delta\mathcal{C}_j| > |\Delta\mathcal{C}_s| | \sigma_s) \\
&\leq \sum_{j \neq s} P\left(|\Delta\mathcal{C}_j| > |\Delta\mathcal{C}_s| \middle| \sigma_s\right)
\end{aligned}
$$

Union bound

Assume random transformations $\sigma \in \Sigma$

$$
\begin{aligned}
&\leq 2^{n\rho} P\left(|\Delta\mathcal{C}_{\neq s}| > |\Delta\mathcal{C}_s| \middle| \sigma_s\right) \\
&= 2^{n\rho} \mathbb{E}_{\mathbf{X}^{(1,2)}} \mathbb{E}_{\sigma_{\neq s}} \left[\mathbb{I}_{\{|\Delta\mathcal{C}_{\neq s}| \geq |\Delta\mathcal{C}_s|\}} \middle| \sigma_s \right]
\end{aligned}
$$

The random transformation statistically decouples the two approximation sets in $\Delta\mathcal{C}_{\neq s}$

Bounding of error

$$\mathbb{E}_{\sigma_{\neq s}}\left[\mathbb{I}_{\{|\Delta\mathcal{C}_{\neq s}|\geq|\Delta\mathcal{C}_s|\}}\right] \overset{(a)}{\leq} \frac{1}{|\{\sigma_{\neq s}\}|}\sum_{\{\sigma_{\neq s}\}}\frac{|\Delta\mathcal{C}_{\neq s}|}{|\Delta\mathcal{C}_s|}$$

$$\overset{(b)}{\leq} \frac{|\mathcal{C}_\gamma(\mathbf{X}^{(1)})||\mathcal{C}_\gamma(\mathbf{X}^{(2)})|}{|\{\sigma_{\neq s}\}||\Delta\mathcal{C}_s|}$$

$$\overset{(c)}{=} \exp\left(-n\mathcal{I}_\gamma(\sigma_{\neq s},\hat{\sigma})\right)$$

a) Bound on indicator function $\mathbb{I}_{\{x\geq a\}} \leq x/a$

b) Averaging over random transformation

c) Definition of mutual information

Condition of vanishing total error

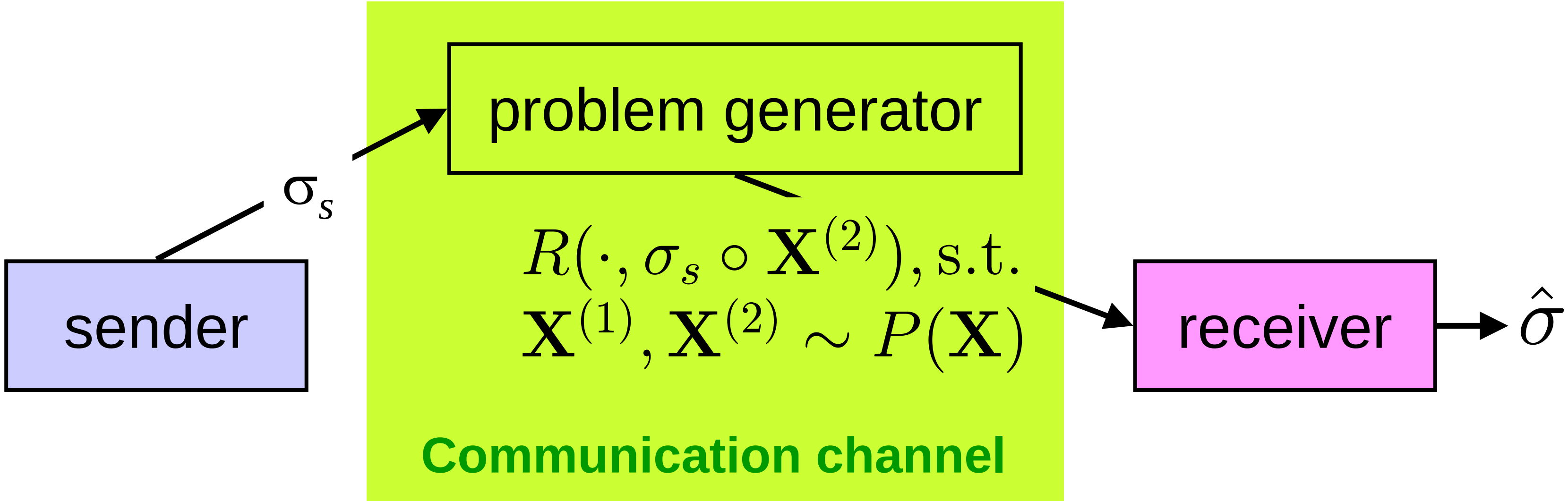
$\lim_{n\to\infty} P(\hat{\sigma} \neq \sigma_s | \sigma_s) = 0$ yields

- Rate is bounded by mutual information

$$
\begin{aligned}
\rho \log 2 \quad &< \quad \frac{1}{n} \log \frac{|\{\sigma_{\neq s}\}||\Delta\mathcal{C}_s|}{|\mathcal{C}_\gamma^{(1)}||\mathcal{C}_\gamma^{(2)}|} \\
&= \quad \frac{1}{n} \left(\log \frac{|\{\sigma_{\neq s}\}|}{|\mathcal{C}_\gamma^{(1)}|} + \log \frac{|\mathcal{C}^{(2)}|}{|\mathcal{C}_\gamma^{(2)}|} - \log \frac{|\mathcal{C}^{(2)}|}{|\Delta\mathcal{C}_s|} \right) \\
&\equiv \quad \mathcal{I}_\gamma(\sigma_{\neq s}, \hat{\sigma})
\end{aligned}
$$

- Lower bound: generalize Fano's inequality to ASC (work in progress)

Model Selection by Maximization of Approximation Capacity



- Optimize the communication channel w.r.t. approximation quality γ (β), topology and metric of solution space, cost function $R(.,.)$, transfer function ψ

Ex.: Binary Coding

- Hypothesis class: set of binary strings

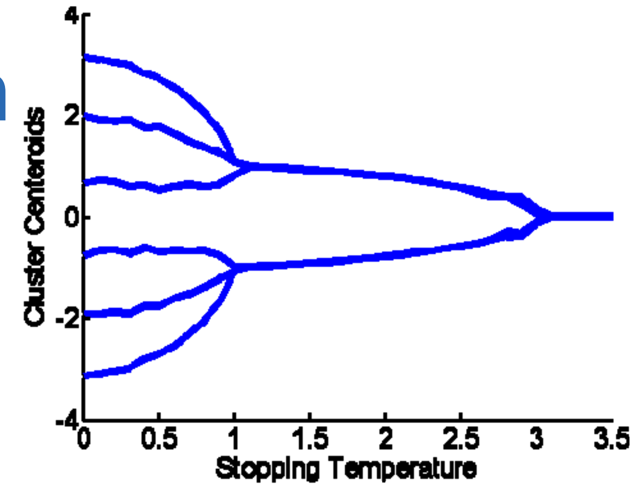
$$\xi^{(1)} = (\xi_1^{(1)}, \xi_2^{(1)}, \ldots, \xi_n^{(1)}), \xi^{(2)} \in \{-1, 1\}^n$$

- Costs: $R(s, \xi^{(1)}) = \sum_{i=1}^{n} \mathbb{I}_{\{s_i \neq \xi_i^{(1)}\}}$
- Mutual information: $\left(\delta = \frac{1}{n}|\{i : \xi_i^{(1)} \neq \xi_i^{(2)}\}|\right)$

$$\begin{aligned} \mathcal{I}_\beta &= \ln 2 + (1-\delta) \ln \cosh \beta - \ln(\cosh \beta + 1) \\ &\overset{(*)}{=} \ln 2 + (1-\delta)\ln(1-\delta) + \delta \ln \delta \\ \text{if} \quad & (*)\ \frac{d\mathcal{I}_\beta}{d\beta} = 0 \ \text{holds} \end{aligned}$$

- ASC for Hamming distance yield capacity of binary symmetric channel!

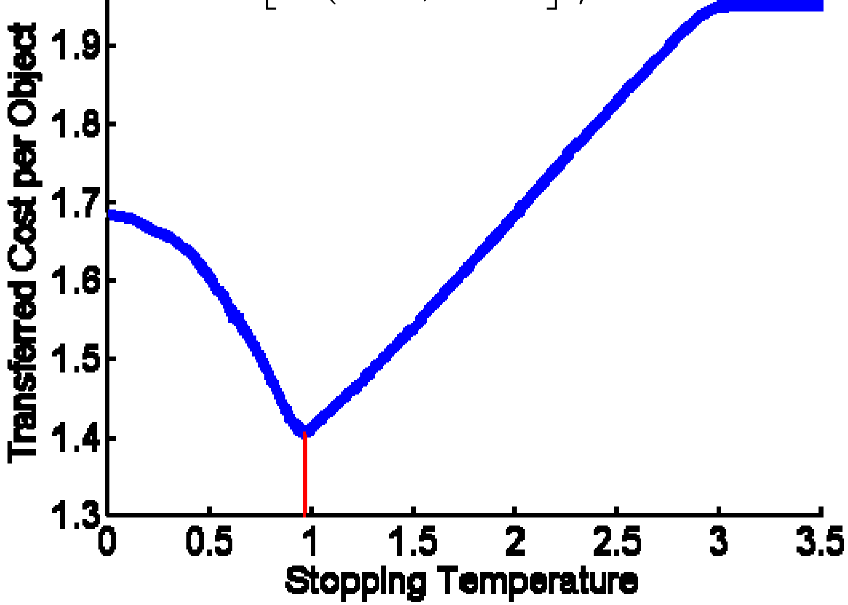
2d Mixture Model Estimation



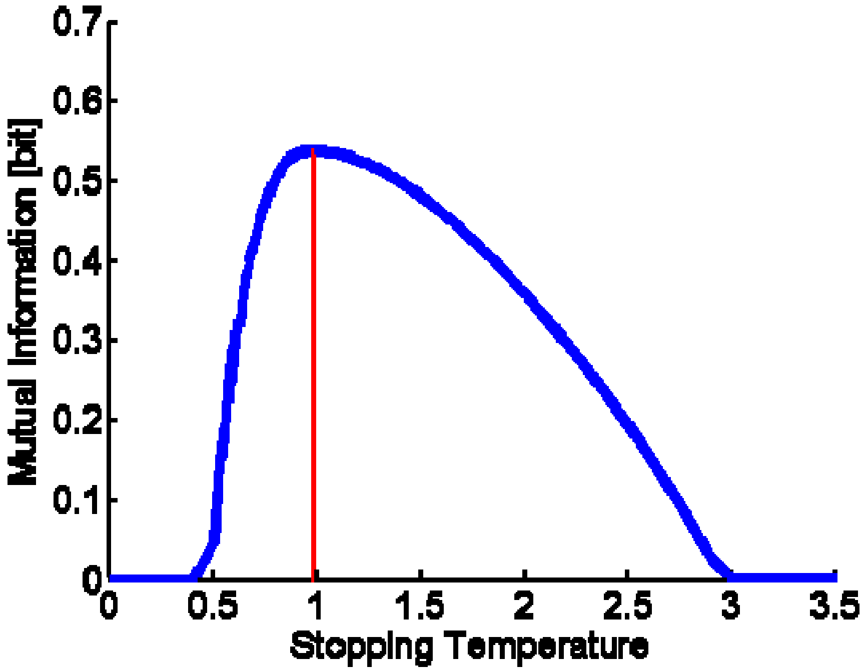
Experimental Setting:
2 source Gaussians,
n=10000, d=2, $\Delta\mu$=2

Generalization error

$$\mathbb{E}\left[R(c^{(1)}, \mathbf{X}^{(2)})\right]/n$$



Approximation Capacity



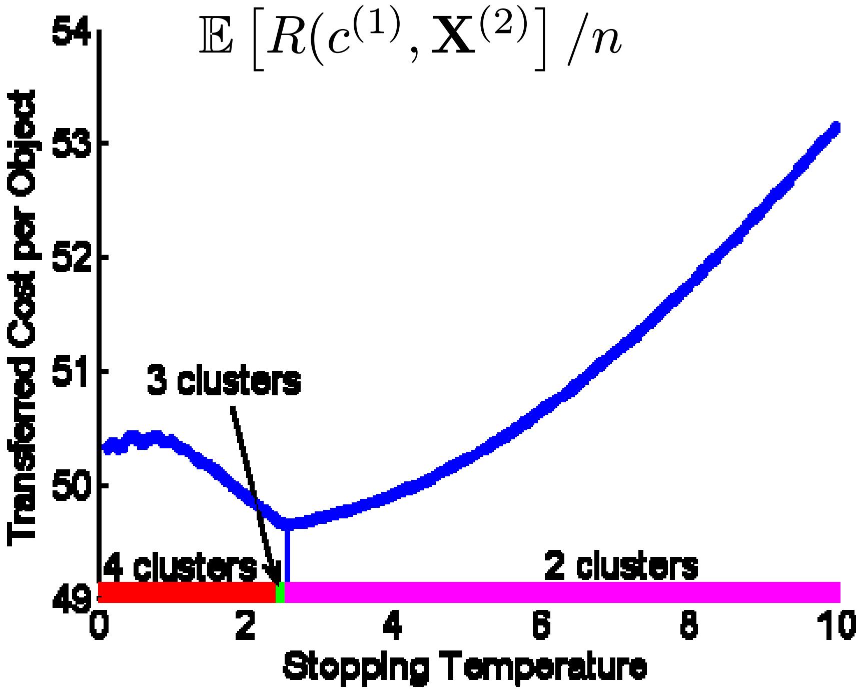
Gibbs sampling with 4 clusters

Experimental Setting:

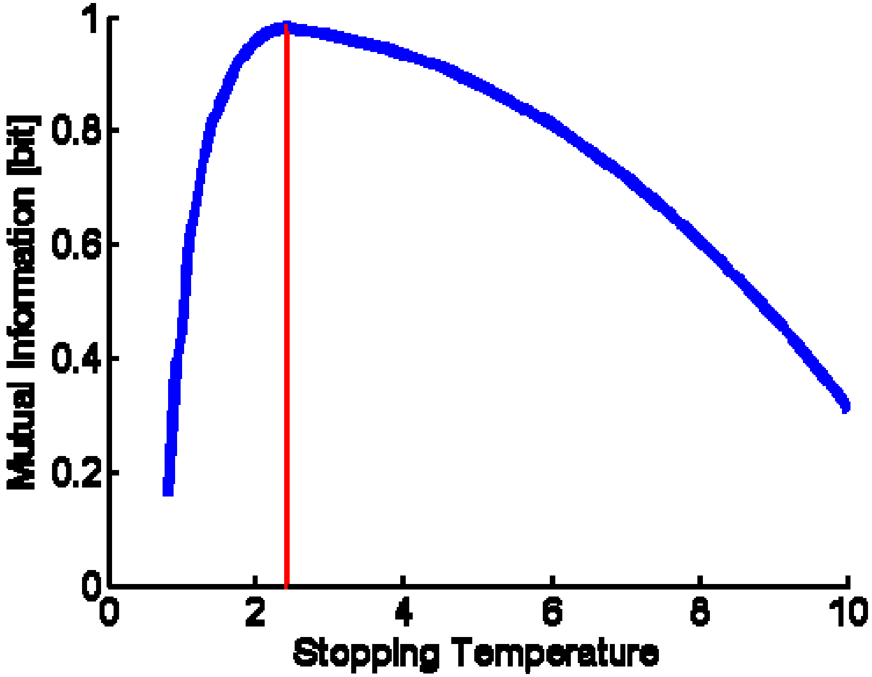
n=500, d=100, 2 source Gaussians

ordered phase, up to 4 estimated Gaussians

Empirical Generalization error



Approximation Capacity



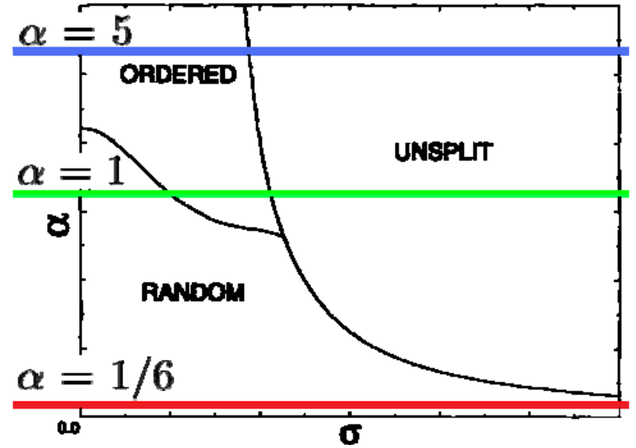
High Dimensional Density Estimation

Barkai, Sompolinsky, Phys Rev E 50:1766, 1994.

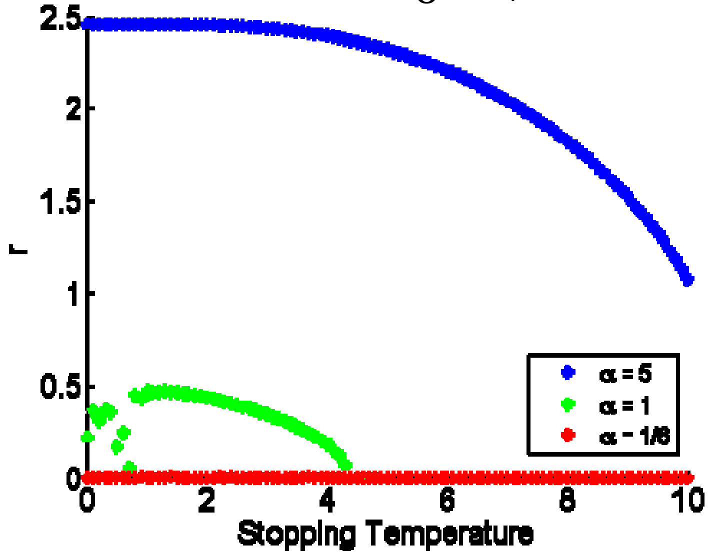
Phase Diagram:

mixture of 2 Gaussians

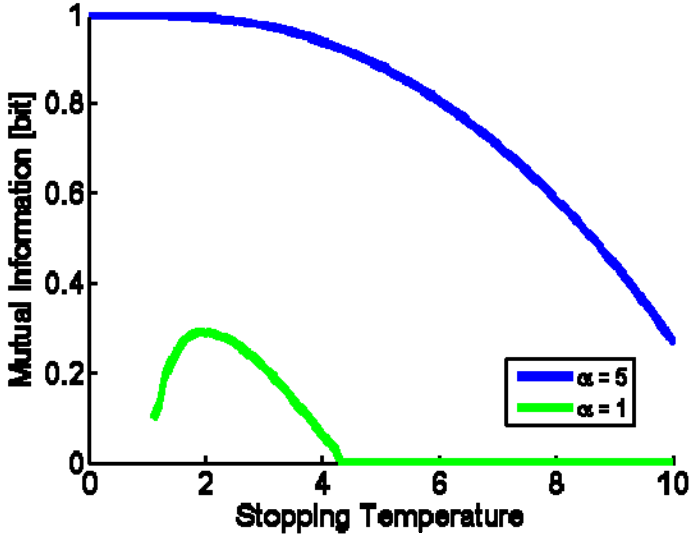
n=500; d=100, 500, 3000; α := n/d



Overlap: $r = u_0^{-1}\langle \Delta\mu, \Delta\mu_0 \rangle$

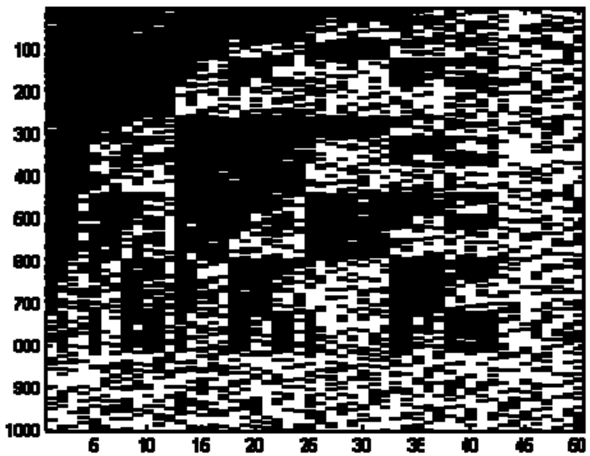


Approximation Capacity



Denoising Binary Matrices by rank-k approximation

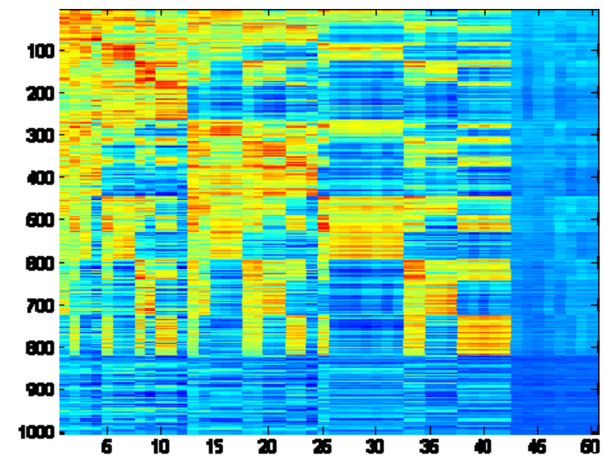
Boolean matrix with 40% random entries



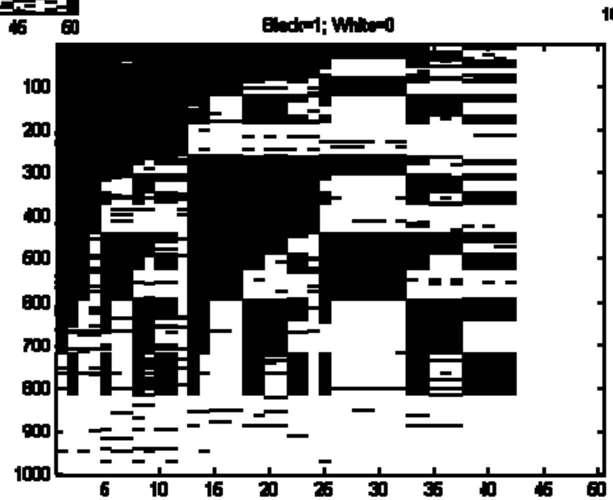
$X = USV$

continuous rank-k approximation

$X_5 = U_5 S_5 V_5$



Rounding as approximation $g(X_k) = \text{round}\ (X_k)$

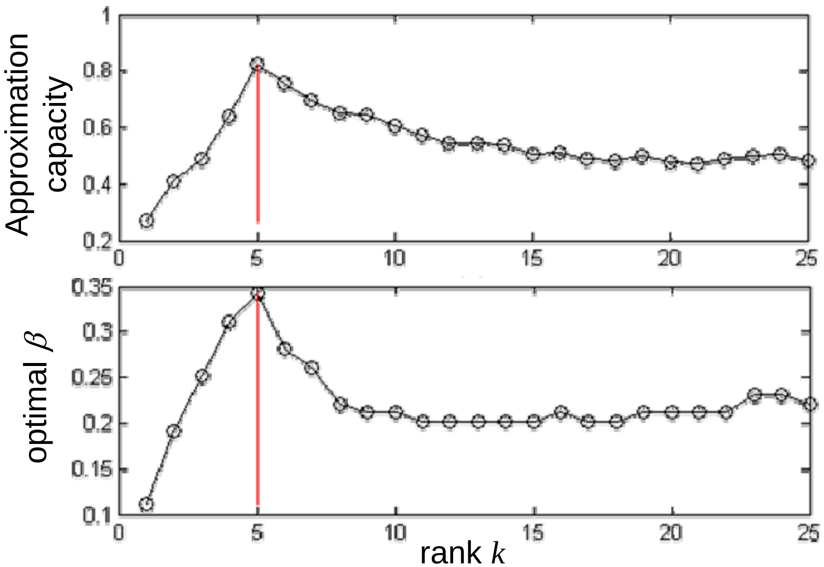


Maximum of approximation capacity selects optimal rank *k*

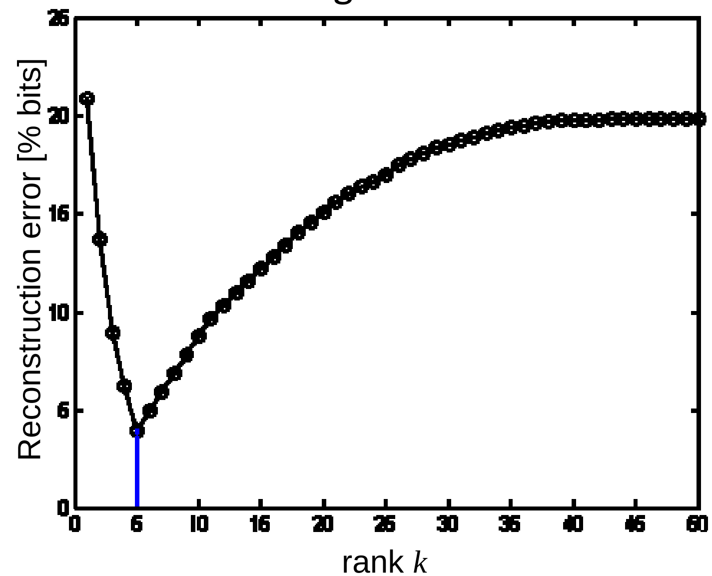
- Integrate over variations of the signal matrix U.

$$\mathcal{I}_\beta(\sigma_j, \hat{\sigma}) \quad = \quad \frac{1}{n} \log \frac{|\{\sigma_j\}||\Delta \mathcal{C}_\beta(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})\ |}{|\mathcal{C}_\beta(\mathbf{X}^{(1)})||\mathcal{C}_\beta(\mathbf{X}^{(2)})|}$$

Approximation capacity vs. k

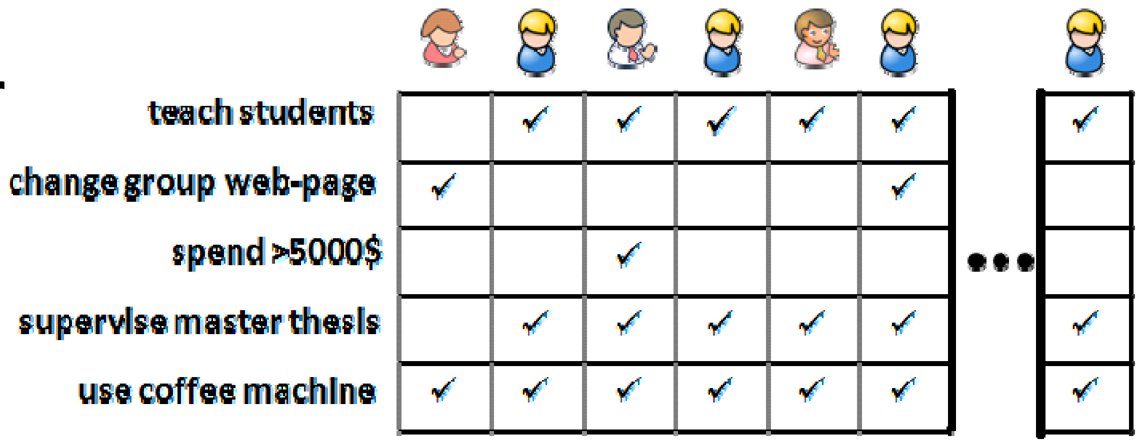


Difference to ground-truth vs. k

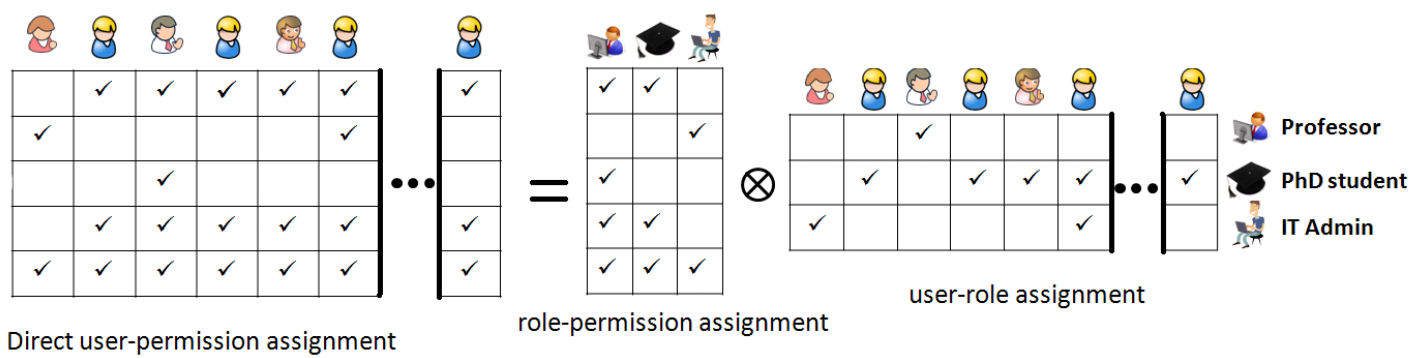


Role-Based Access Control

- **Given: Binary user permission matrix**



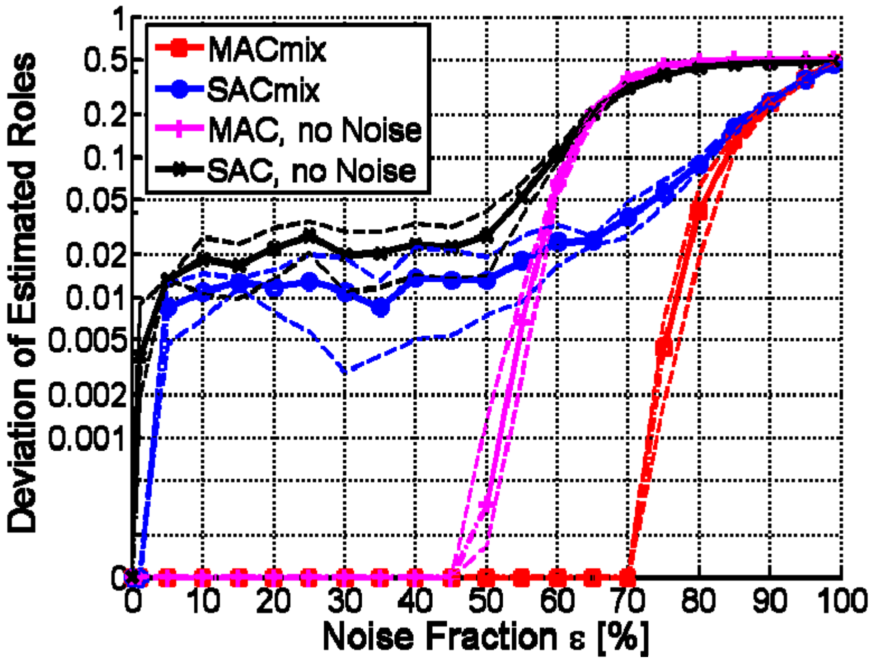
- **Discretional Access-Control**: Direct Assignments of users to permissions
- **Role-Based Access Control (RBAC)**: Permissions are granted via roles

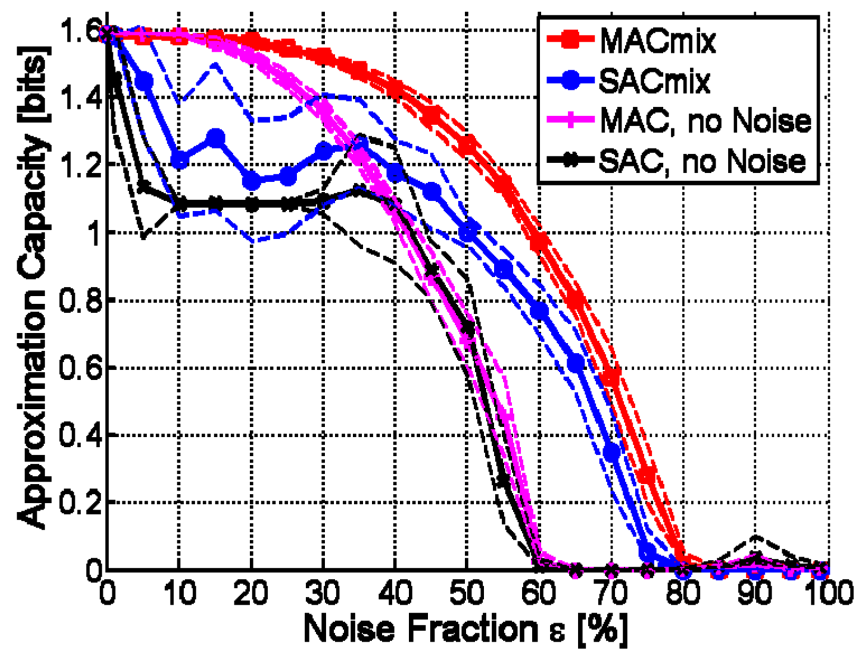


Role-Mining for RBAC

- **Role-Mining**: Given a user-permission assignment matrix **X**, find a set of roles **U** and assignments **Z** such that $\mathbf{X} \approx \mathbf{U} \otimes \mathbf{Z}$
- **Multi Assignment Clustering**: generative approach including noise model, inference with DA

Synthetic Data: Parameter Accuracy vs. Approximation Capacity

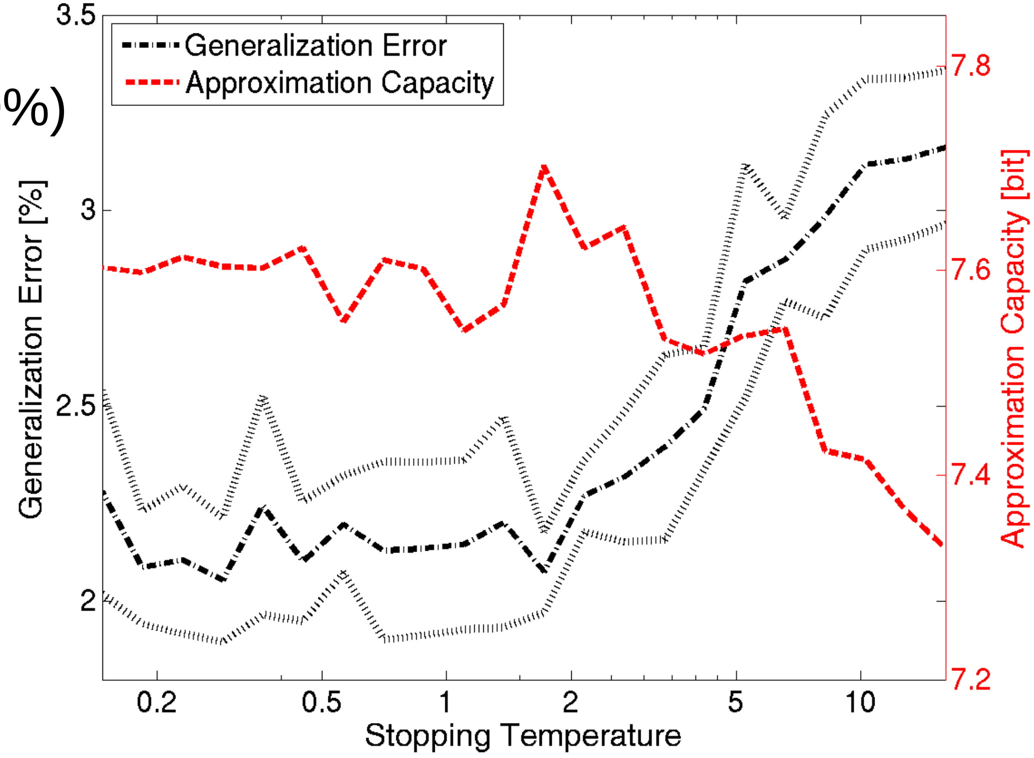




MAC: More accurate estimators for centroids, it yields higher approximation capacity than SAC.

Real-World Data: Prediction Error vs. Approximation Capacity

- **Generalization**: Can roles predict permissions of **new** users?
  1. Use few permissions (20%) to determine role set
  2. Predict hidden/missing permissions (80%).
- Centroids with maximal capacity yield minimal generalization error



Conclusion

- **Quantization**: Noise quantizes mathematical structures (hypothesis classes) **=>** symbols
- These symbols can be used for **coding**!
- Optimal error free coding scheme determines **approximation capacity** of a model class.

$\Rightarrow$Bounds for robust optimization.

$\Rightarrow$**Quantization** of hypothesis class measures **structure specific information** in data.

Future Work

- **Generalization**: replace approximation sets based on cost functions by smoothed outputs of **algorithms** ("smoothed generalization")
- **Model reduction** in dynamical systems: quantize sets of ODEs or PDEs (systems biology)
- Relate **statistical complexity**, i.e. the approximation capacity, to algorithmic or **computational complexity**.

Philosophical speculations

- We experience a **paradigm shift from model driven reasoning to algorithm dominated reasoning** (Bernard Chazelle "The Algorithm: Idiom of Modern Science")

=> **model validation** more essential than modeling since modeling can be algorithmically formulated as exploration of model space.

- *Ceterum censeo*: The coupling of **statistical complexity** and **algorithmic complexity** should be reconsidered in the light of **statistical learning theory** and information theory.