Information Theoretic Model Validation by Approximate Optimization

Authors: *Joachim M. Buhmann, Morteza H. Chehreghani, Mario Frank, Andreas P. Streich*

Presenter: *Cheng Soon Ong*

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 mining for role-based access control (RBAC)
  - Robust SVD

Optimization approach to pattern recognition

- 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)

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

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

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

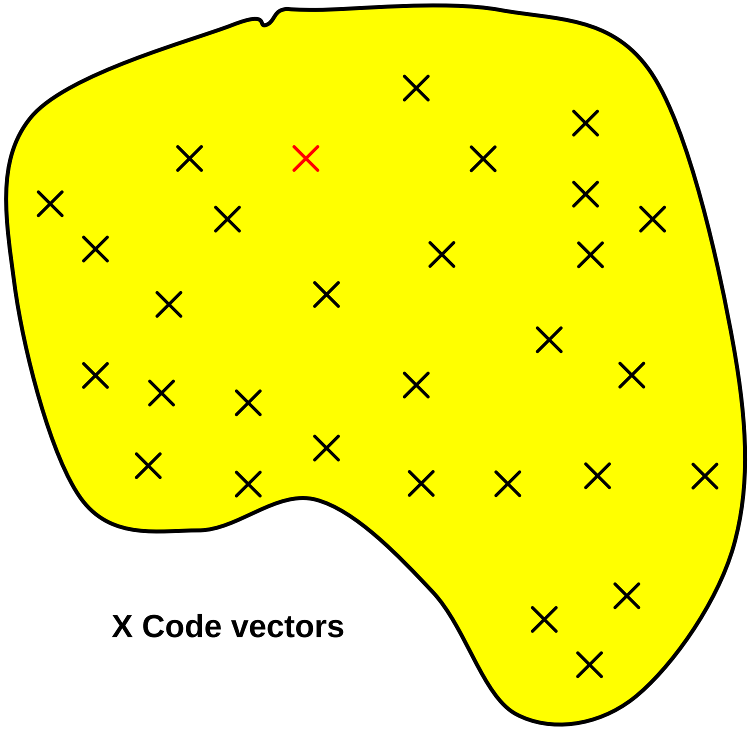
Pattern recognition and modeling

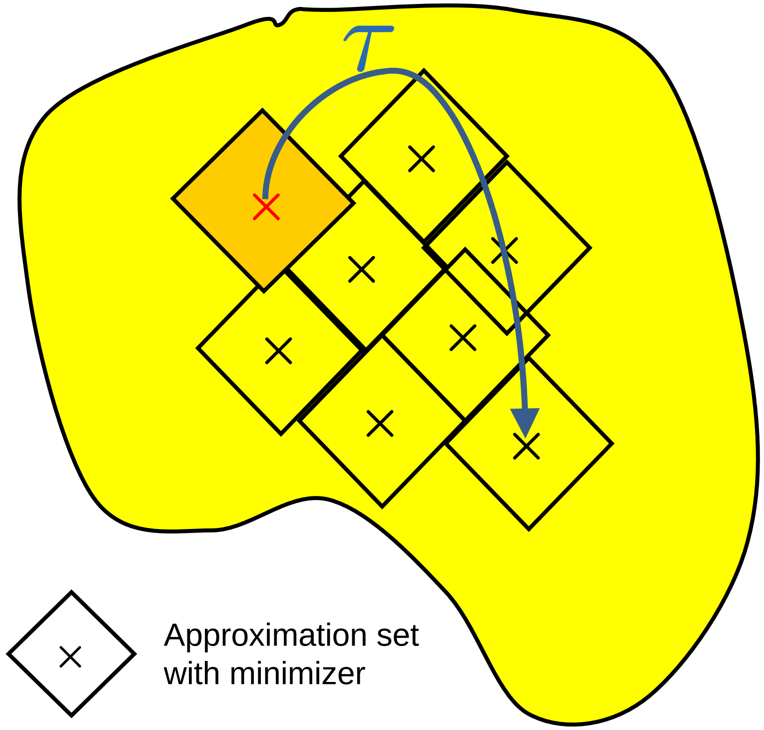
- Given are **data** and interpretations of these data, i.e., **hypotheses**.
- **Modeling** is (partial) ranking of the hypotheses encoded as data dependent costs.
  - Good/poor hypotheses have low/high costs
  - Optimal hypotheses minimize costs and are random variables.

$\Rightarrow$Search for hypotheses that have low costs on future data, i.e. **generalize** well!

Coding & pattern recognition with noisy data

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





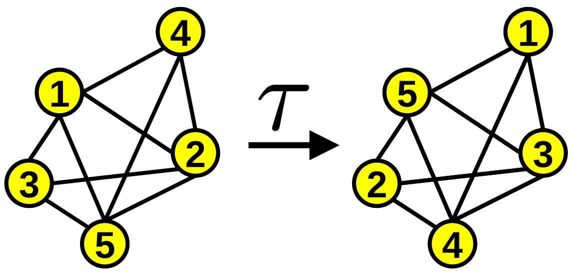
Coding by Code Problems

- **Idea**: define a code by transforming a given optimization problem

$\Rightarrow$codebook $\mathbb{T} = \{\tau_i \in \mathcal{T} : 1 \leq i \leq 2^{n\rho}\}$ with transformation set

$$\mathcal{T} = \{\tau : R(c, \mathbf{X}) = R(\tau \circ c, \tau_{\mathbf{X}} \circ \mathbf{X})\}$$

- Combinatorial optimization: *permutation* of vertices in a graph



- **Identifiable transformations $\mathcal{T}$ are <u>messages</u>!**

Asymptotical error-free communication $\lim_{n\to\infty} P(\hat{\tau} \neq \tau_s | \tau_s) = 0$ is possible if …

- ... mutual information $\mathcal{I}_\beta(\tau_s, \hat{\tau})$ is bounded by

$$\rho \quad < \quad \mathcal{I}_\beta(\tau_s, \hat{\tau}) \;\equiv\; \frac{1}{n} \log_2 \frac{|\mathcal{T}|\; Z_\beta^{(1\&2)}}{Z_\beta^{(1)}\; Z_\beta^{(2)}}$$

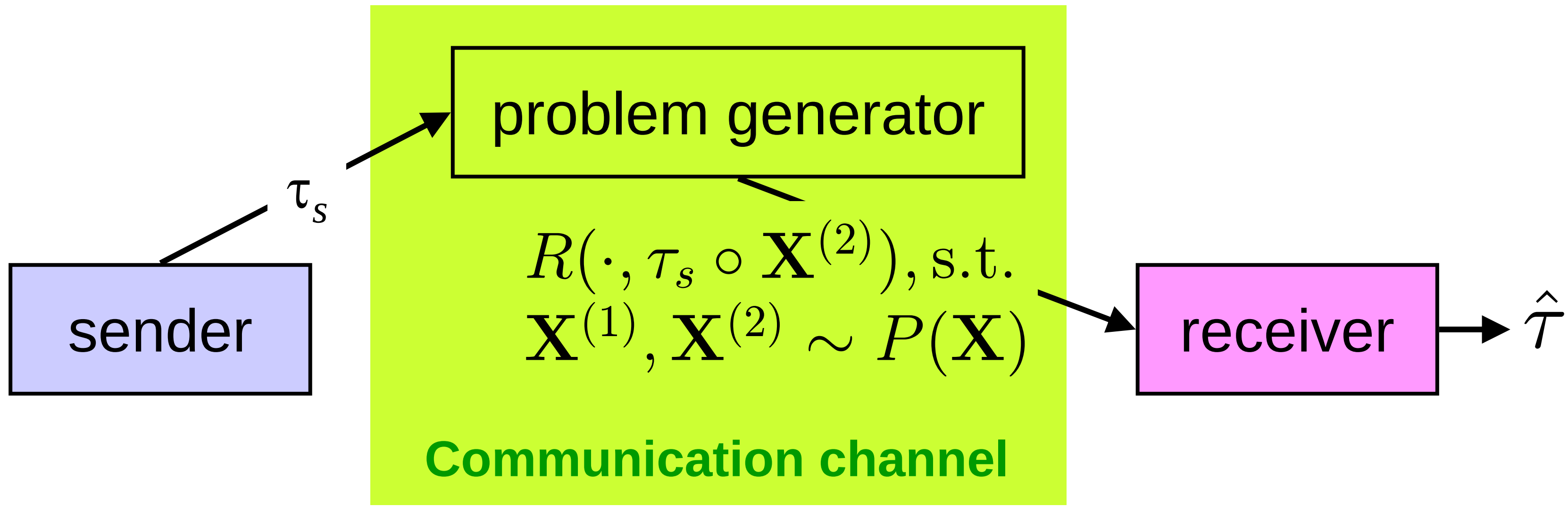
$$= \quad \frac{1}{n} \left(\log_2 \frac{|\mathcal{T}|}{Z_\beta^{(1)}} + \log_2 \frac{|\mathcal{C}^{(2)}|}{Z_\beta^{(2)}} - \log_2 \frac{|\mathcal{C}^{(2)}|}{Z_\beta^{(1\&2)}} \right)$$

Bound calculation involves partition functions for individual and joint costs

$$Z_\beta^{(\nu)} = \sum_{c \in \mathcal{C}(\mathbf{X}^{(\nu)}))} \exp(-\beta R(c, \mathbf{X}^{(\nu)})), \quad \nu = 1, 2$$

$$Z_\beta^{(1\&2)} = \sum_{c \in \mathcal{C}(\mathbf{X}^{(1)})} \exp\left(-\beta (R(c, \mathbf{X}^{(1)}) + R(c, \mathbf{X}^{(2)})\right)$$

Model Selection by Maximization of Approximation Capacity



- Maximize channel capacity w.r.t. approximation quality β, topology and metric of solution space, cost function $R(.,.)$

ASC for binary channel consistent with Shannon information theory

- Hypothesis class: set of binary strings $\xi^{(1)}, \xi^{(2)} \in \{-1, 1\}^n$

- Communication: $\xi^{(1)} \Rightarrow$

$\Rightarrow \xi^{(2)}$

$$\left(\delta = \tfrac{1}{n}|\{i : \xi_i^{(1)} \neq \xi_i^{(2)}\}|\right)$$

- Costs of string s: Hamming distance

$$R(s, \xi^{(1)}) = \sum_{i=1}^{n} \mathbb{I}_{\{s_i \neq \xi_i^{(1)}\}}$$

Channel capacity of BSC

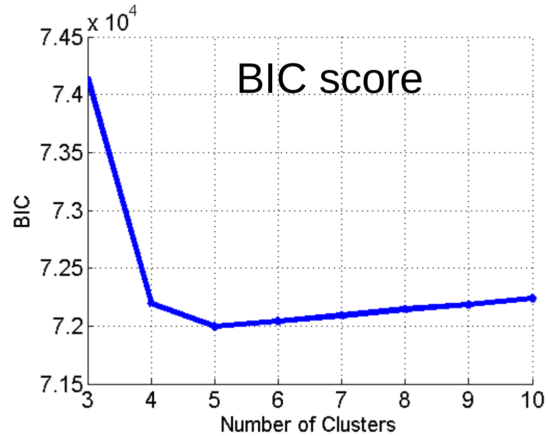
- Mutual information: $\mathcal{I}_\beta = \ln 2 + (1-\delta)\ln\cosh\beta - \ln(\cosh\beta + 1)$

for $(*)\ \frac{d\mathcal{I}_\beta}{d\beta} = 0 \quad \stackrel{(*)}{=} \ln 2 + (1-\delta)\ln(1-\delta) + \delta\ln\delta$

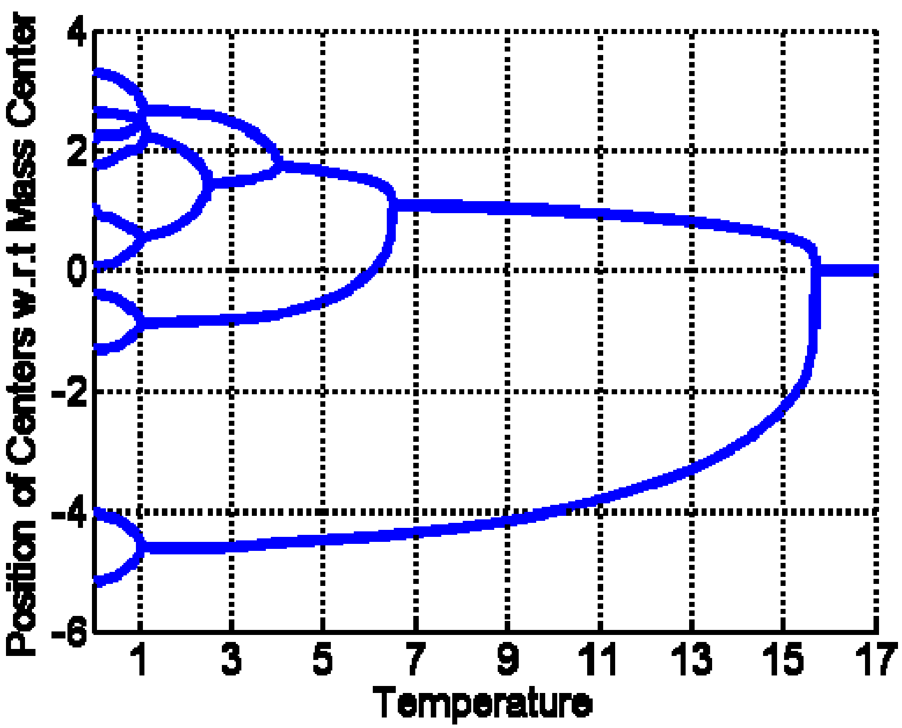
ASC selects optimal (true) number of clusters

Experimental Setting:

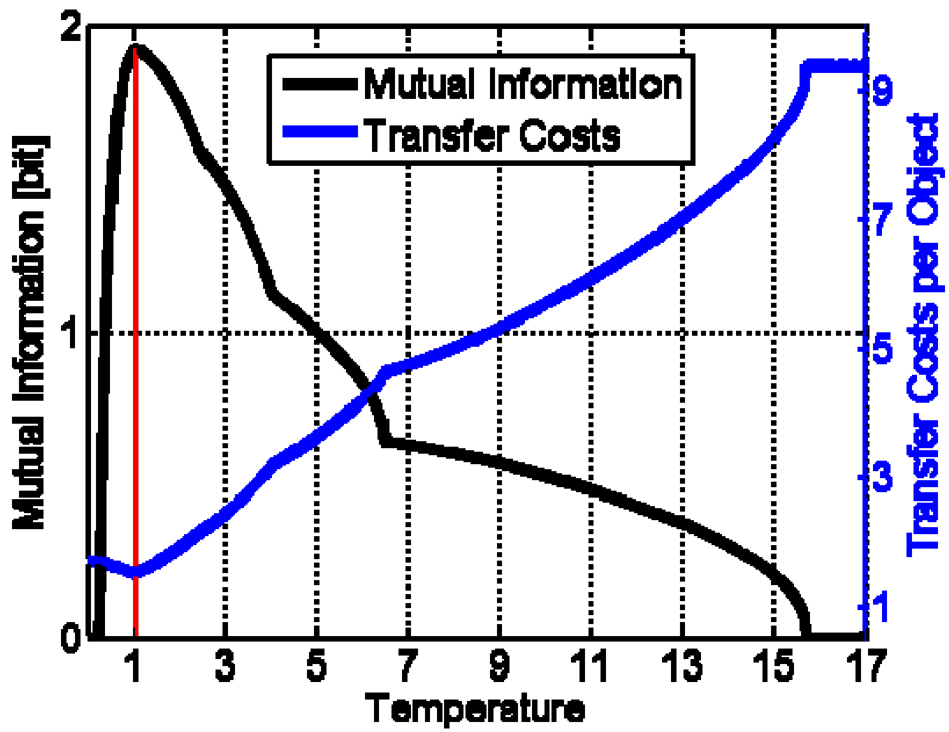
5 Gaussians, n=10000, d=2, k^{max}=10



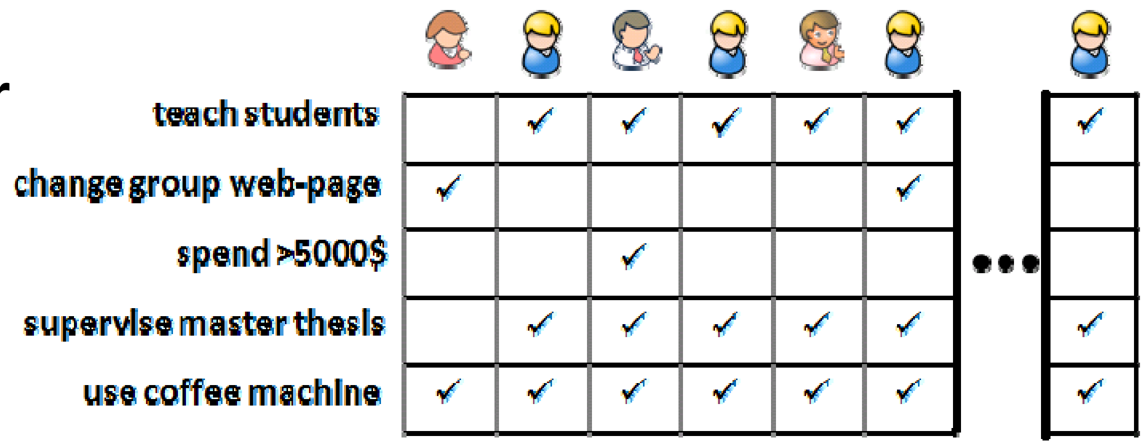
Cluster splitting



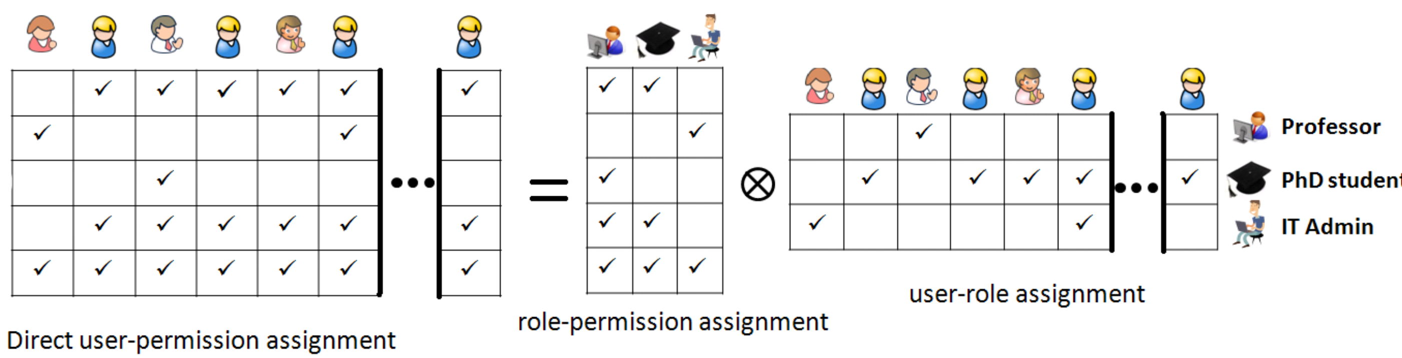
Approximation Capacity



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

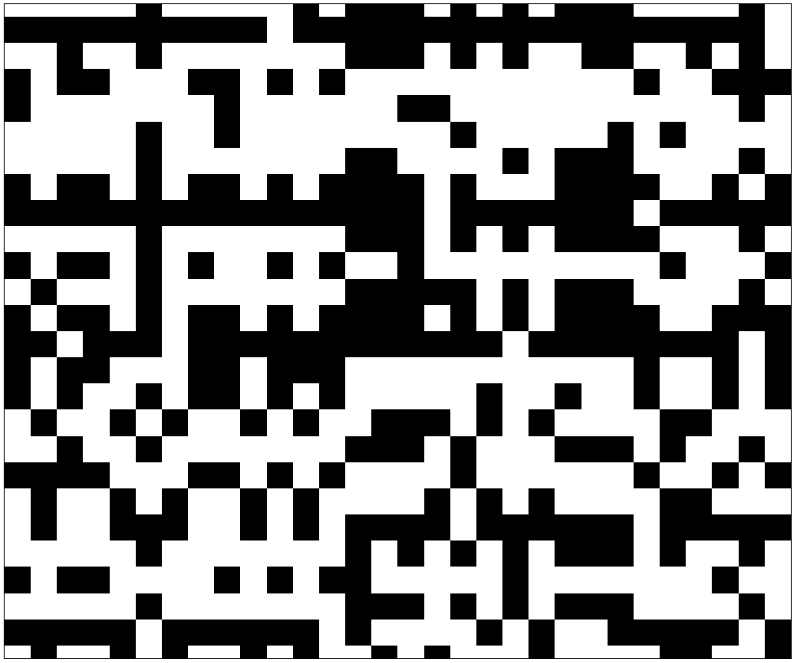


Role-Mining for RBAC

- **Role-Mining**: Given a user-permission assignment matrix **X**, find a set of roles **U** and assignments **Z** such that

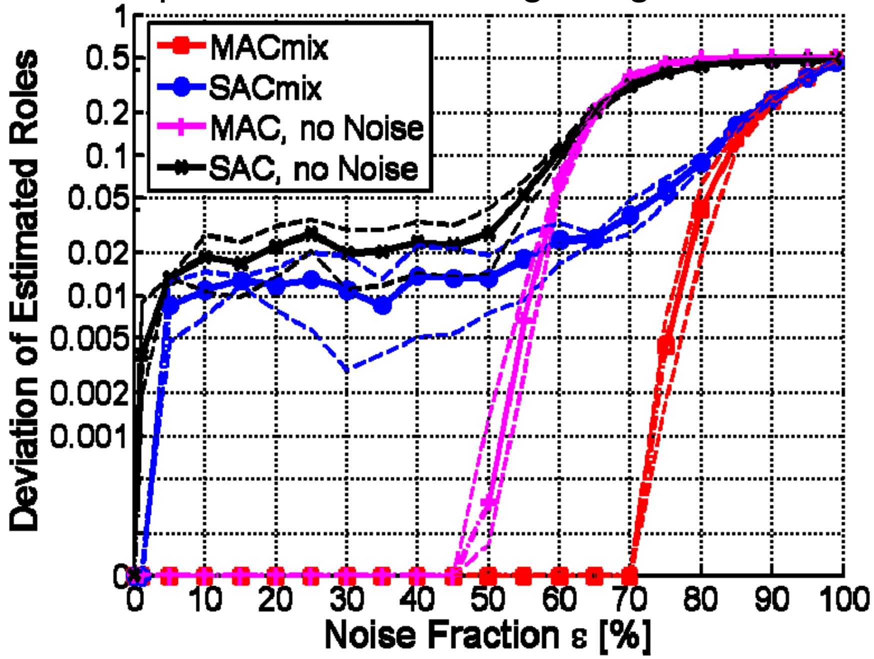
$$\mathbf{X} \approx \mathbf{U} \otimes \mathbf{Z}$$

- **Multi Assignment Clustering**: generative approach including noise model, inference with DA

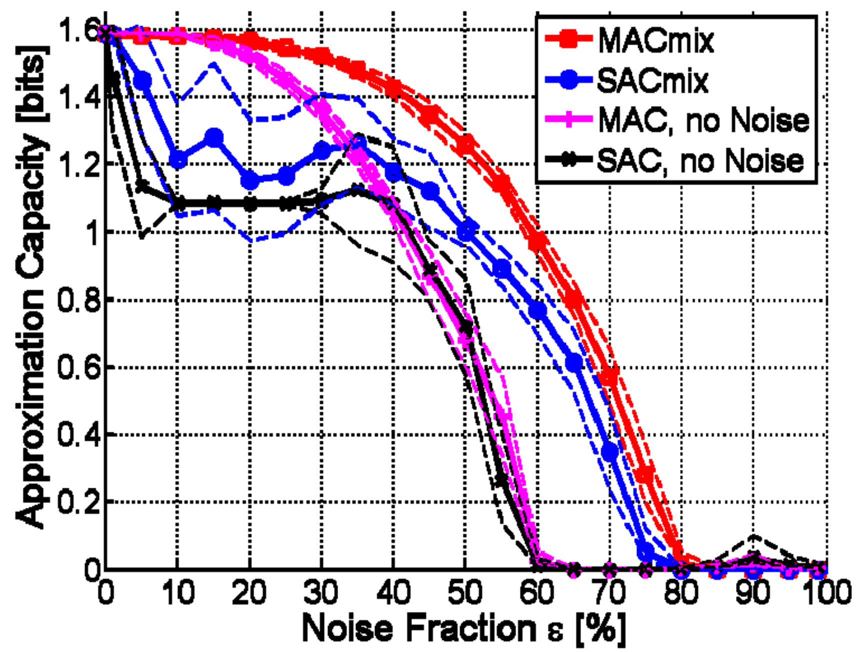


Synthetic Data: Parameter Accuracy vs. Approximation Capacity

computed with knowledge of ground truth



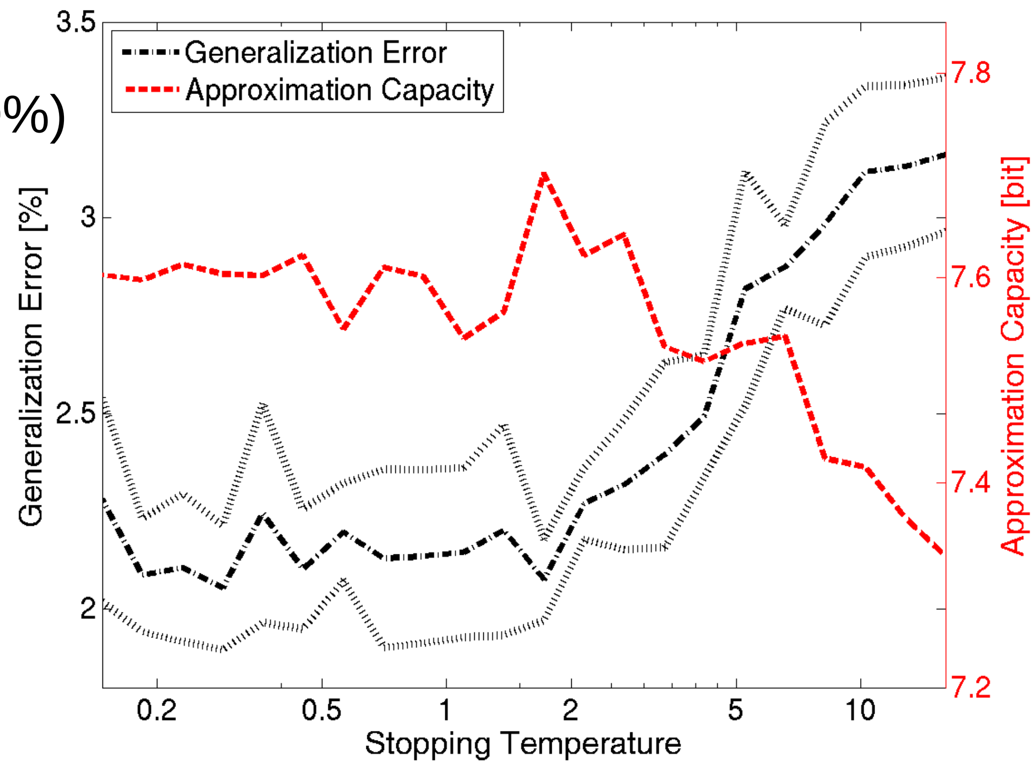
computed without knowledge of ground truth



ASC ranking of model variants complies with ranking according to ground truth.

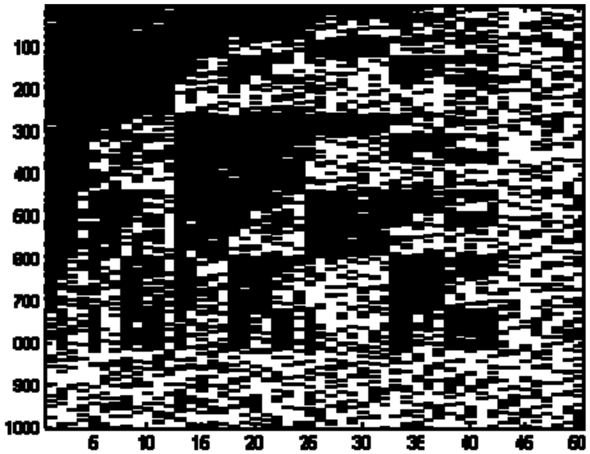
Real-World Data: Prediction Error complies with Approximation Capacity

- **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



Denoising Binary Matrices by truncated SVD

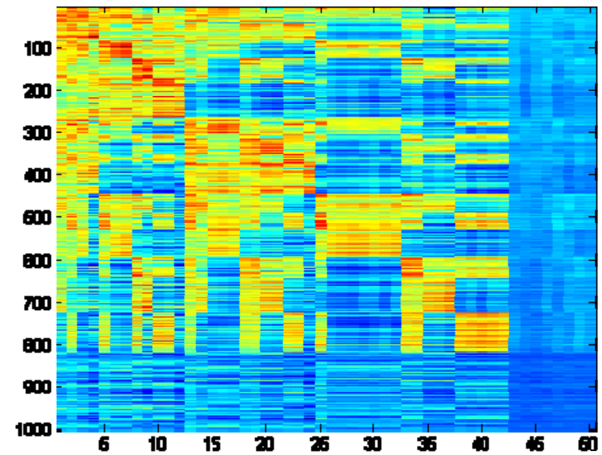
Boolean matrix with 40% random entries



$X = USV$

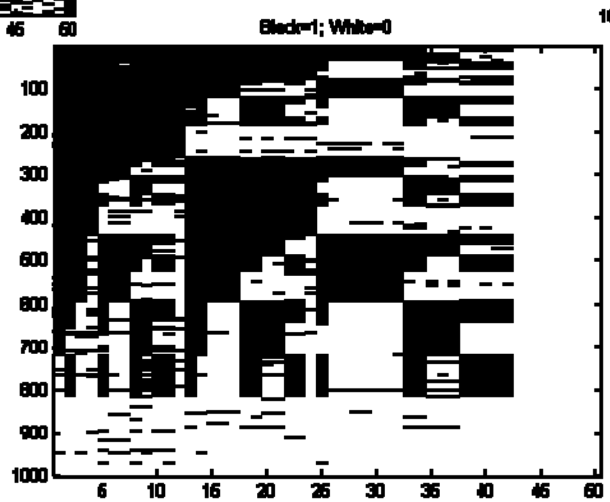
continuous rank-k approximation

$X_5 = U_5 S_5 V_5$



Rounding as approximation

$g(X_k) = \text{round}\ (X_k)$

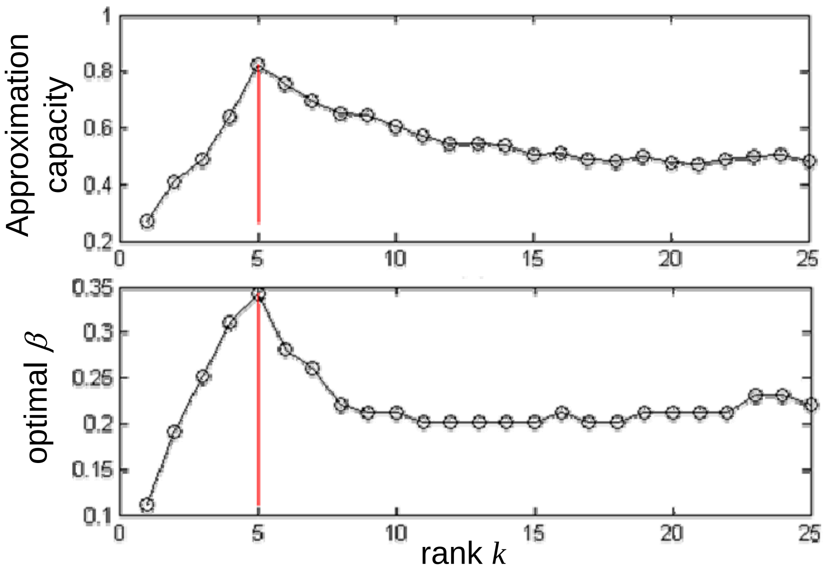


Maximum of approximation capacity selects optimal rank *k*

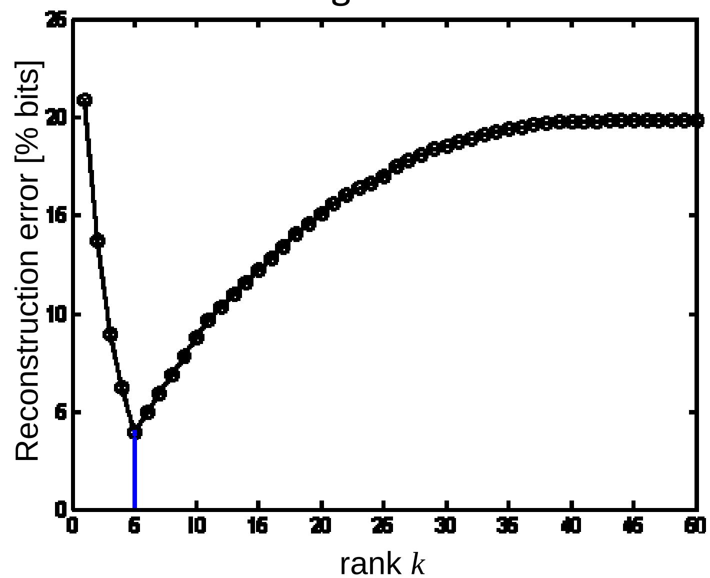
- Integrate over variations of the signal matrix U.

$$\mathcal{I}_\beta(\tau_s, \hat{\tau}) \quad = \quad \frac{1}{n} \log_2 \frac{|\mathcal{T}| \; Z_\beta^{(1\&2)}}{Z_\beta^{(1)} \; Z_\beta^{(2)}}$$

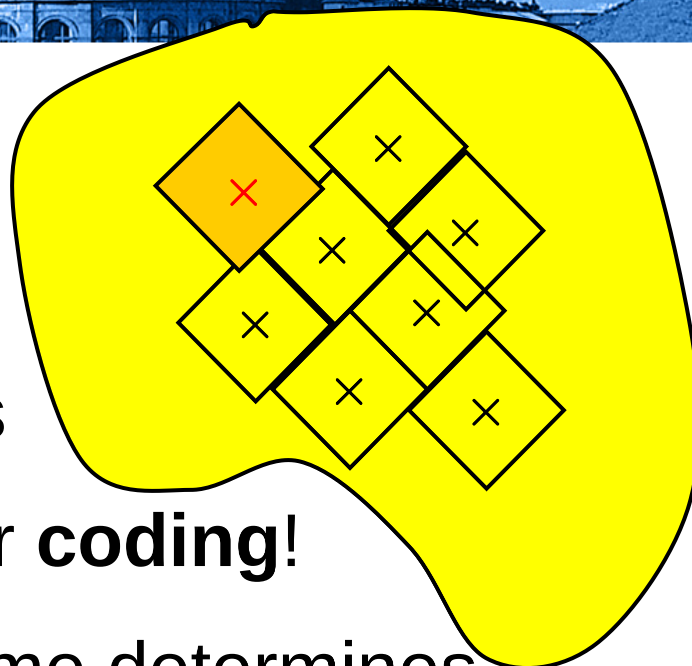
Approximation capacity vs. k



Difference to ground-truth vs. k



Conclusion



- **Quantization**: Noise quantizes hypothesis classes **=>** symbols
- These symbols can be used for **coding**!
- Optimal error free coding scheme determines **approximation capacity** of a cost function.

$\Rightarrow$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**.