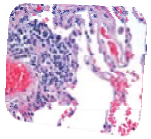
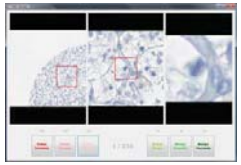
Structure learning and the generalization capacity of algorithms

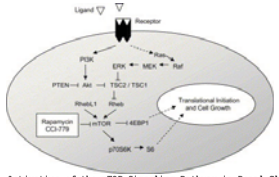
Joachim M. Buhmann

Computer Science Department, ETH Zurich

Modern information society and the Information Technology value chain

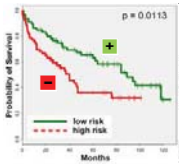






Activation of the mTOR Signaling Pathway in Renal Clear Cell Carcinoma. Robb et al., J Urology 177:346 (2007)

Data → **Information** → **Knowledge** → **Value**



Unsolved Problems in Computer Science:

1) What is useful information?

2) How can we design robust algorithms?

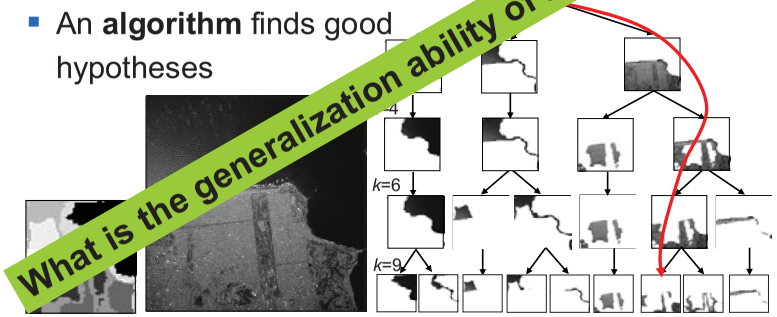
*my**Vision***: **Information theory** for **pattern analysis**

- **Approximation capacity** of algorithms
- Examples for cost function minimization
  - Cluster validation
  - Role based access control
  - (Robust SVD)
- Philosophical remarks

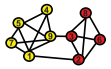
Pattern recognition and modeling

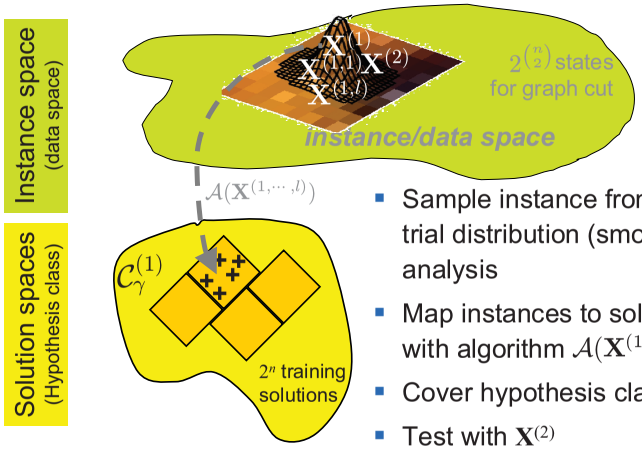
- Given are **data** $\mathbf{X}$ and **hypotheses**, i.e. interpretation of these data.
- An **algorithm** finds good hypotheses

What is the generalization ability of an algorithm?



Generalization of algorithms

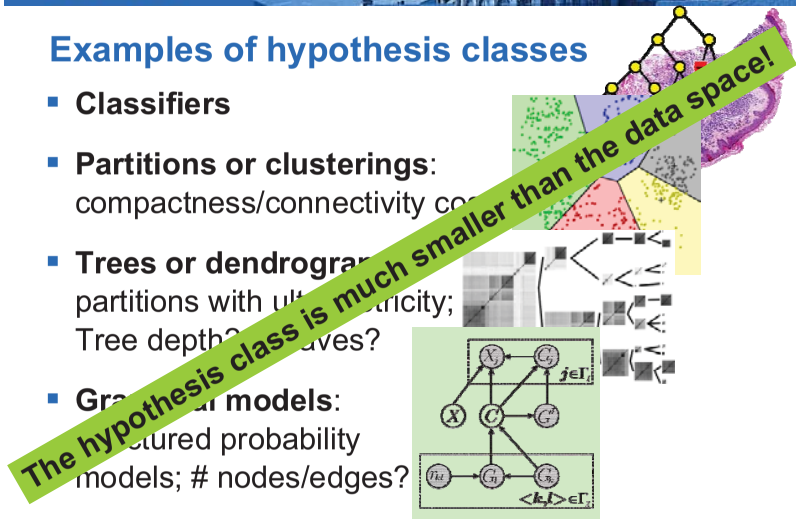




- Sample instance from a trial distribution (smoothed analysis
- Map instances to solutions with algorithm $\mathcal{A}(\mathbf{X}^{(1,\cdots,l)})$
- Cover hypothesis class
- Test with $\mathbf{X}^{(2)}$

Examples of hypothesis classes

- **Classifiers**
- **Partitions or clusterings**: compactness/connectivity co
- **Trees or dendrogra** partitions with ul tricity; Tree depth? aves?
- **Gra al models**: tured probability models; # nodes/edges?

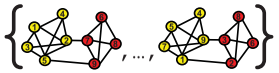


Information Theory & Pattern Recognition

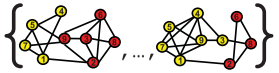
- IT-Components
  - **Code book** $\subsetneq$ **{strings}** = hypothesis class

0000000	0100101	1000011	1100110
0001111	0101010	1001100	1101001
0010110	0110011	1010101	1110000
0011001	0111100	1011010	1111111

  - Noisy **channel**
    1100110 -> 0110110
  - Decoder: minimize **Hamming distance**
    |0110110−1100110| <
    |0110110−1010101|
- Pattern Recognition elements
  - **Set of approximation sets** $\subsetneq$ **hypothesis class**



  - Noisy **optimization** problem



  - Decoding by **approximate optimization** of test instance

Symmetries of the Learning Problem

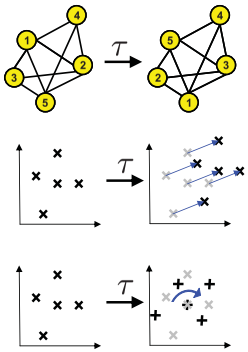
- Assume (!) that the algorithm $\mathcal{A}$ is equivariant under the transformations

$$\mathcal{T} = \{\tau : \tau \circ \mathcal{A}(\mathbf{X}) = \mathcal{A}(\tau_{\mathbf{X}} \circ \mathbf{X})\}$$

- **Idea**: define a code by transforming a given problem => codebook $\mathbb{T} = \{\tau_i \in \mathcal{T} : 1 \leq i \leq 2^{n\rho}\}$
- What is an appropriate set of transformations? That depends on the algorithm and the hypothesis class!

How to generate code problems?

1. Combinatorial optimization problems: **permutation** of e.g. vertices in graphs
2. Localization problems: **shifts** of data
3. Orientation problems (PCA, SVD): **rotations & projection**
4. Sparse linear regression: **Scaling & permutation**



Approximate Optimization

- Define weights for hypotheses with low costs

$$w_\beta(c, \mathbf{X}) = \exp(-\beta R(c, \mathbf{X})) \quad c \in \mathcal{C}(\mathbf{X})$$

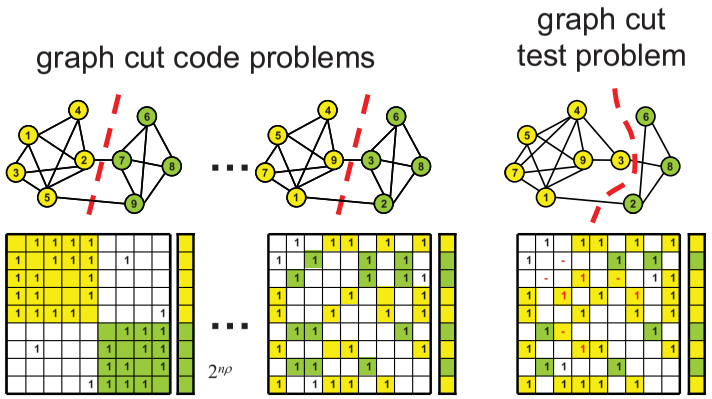
- Special case: **Approximation set**

$$w_\beta(c, \mathbf{X}) = \begin{cases} 1, & R(c, \mathbf{X}) \leq R\left(c^\perp, \mathbf{X}\right) + \gamma; \\ 0, & \text{otherwise.} \end{cases}$$

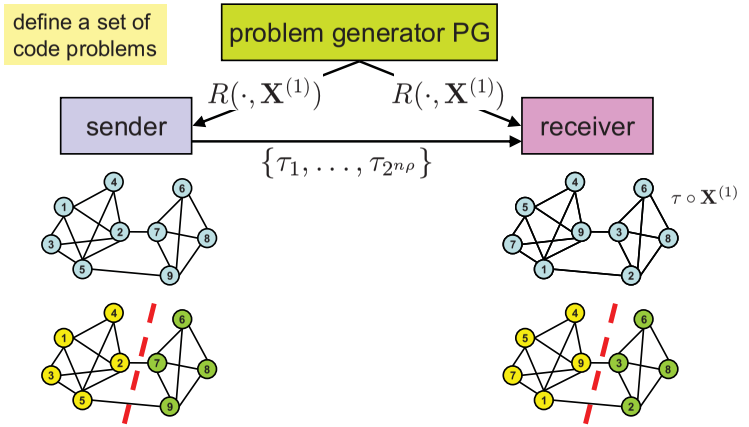
- Total weight of low cost hypotheses

$$W_\beta(\mathbf{X}) = \{w_\beta(c, \mathbf{X}) : c \in \mathcal{C}(\mathbf{X})\}$$

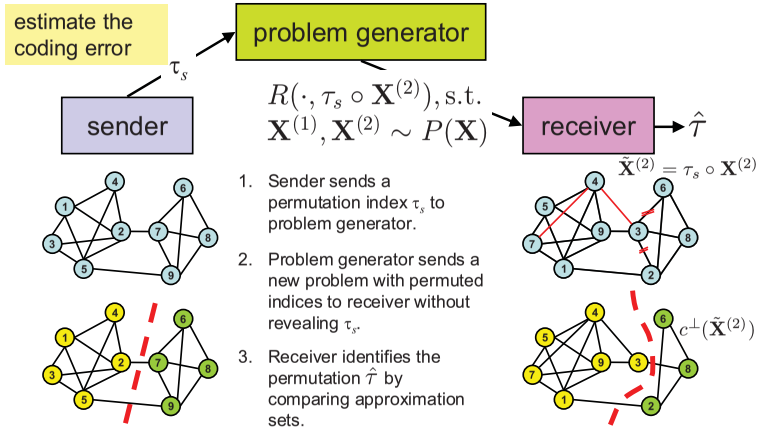
Code problem generation for Graph Cut



Coding with Graph Cut approximation sets



Communication by approximation sets



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

Communication Process

- Receiver **compares sets of hypothesis weights** $W_\beta(\mathbf{X}^{(1)})$ of training instance (code problem) with approximate clusterings $W_\beta(\mathbf{X}^{(2)})$ of the test data.
- Define a mapping $\psi : \mathbf{X}^{(1)} \to \mathbf{X}^{(2)}$
- **Decoding by maximizing weight overlap**

$$\hat{\tau} \in \arg\max_{\tau \in \mathbb{T}} \sum_{c \in \mathcal{C}(\mathbf{X}^{(1)})} w_\beta\left(c, \tau(\mathbf{X}^{(1)})\right) w_\beta\left(c, \tilde{\mathbf{X}}^{(2)}\right)$$

$$\text{with } \tilde{\mathbf{X}}^{(2)} := \psi \circ \tau_s \circ \psi^{-1}(\mathbf{X}^{(2)})$$

Condition of vanishing total error

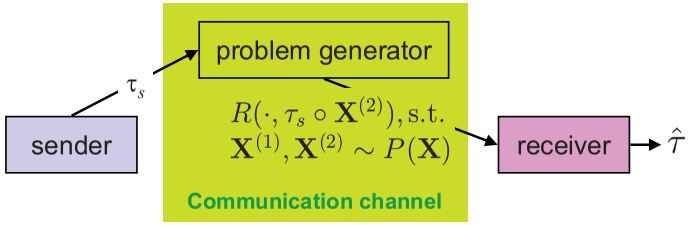
$\lim_{n\to\infty} P(\hat{\tau} \neq \tau_s | \tau_s) = 0$ yields

- Rate is bounded by mutual information

$$\begin{aligned} \rho \log 2 &< \frac{1}{n} \log \frac{|\{\tau_{\neq s}\}| Z_{\beta,s}^{(1\&2)}}{Z_\beta^{(1)} Z_\beta^{(2)}} \\ &= \frac{1}{n}\left(\log \frac{|\{\tau_{\neq s}\}|}{Z_\beta^{(1)}} + \log \frac{|\mathcal{C}^{(2)}|}{Z_\beta^{(2)}} - \log \frac{|\mathcal{C}^{(2)}|}{Z_{\beta,s}^{(1\&2)}}\right) \\ &\equiv \mathcal{I}_\gamma(\tau_s, \hat{\tau}) \end{aligned}$$

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

Model Selection by Maximization of Approximation Capacity

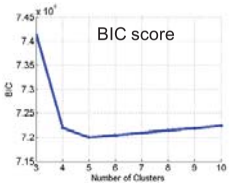


- Optimize the communication channel w.r.t. approximation quality γ (β), topology and metric of solution space, cost function $R(.,.)$, transfer function ψ

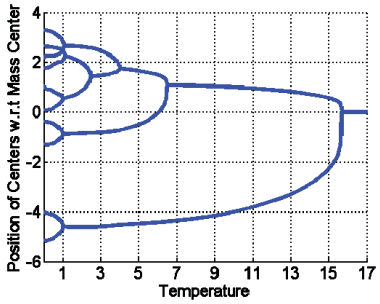
2d Mixture Model Estimation

Experimental Setting:

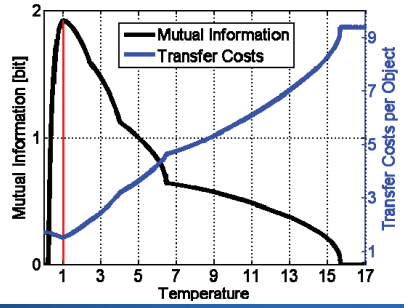
5 Gaussians, n=10000, d=2, k^{max}=10



Cluster splitting



Approximation Capacity



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.

⇒Bounds for robust optimization.

⇒**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.