Context Sensitive Information: Model Validation by Information Theory

Joachim M. Buhmann

ETH Zürich, Departement Informatik, CAB G 69.2
Universitätsstraße 6, CH-8092 Zürich, Switzerland
jbuhmann@inf.ethz.ch
http://www.ml.inf.ethz.ch

Abstract. A theory of patterns analysis has to provide a criterion to filter out the relevant information to identify patterns. The set of potential patterns, also called hypothesis class of the problem, defines admissible explanations of the available data and it specifies the context for a patterns analysis task. Fluctuations in the measurements limit the precision which we can achieve to identify such patterns. Effectively, the distinguishible patterns define a code in a fictitious communication scenario where the selected cost function together with a stochastic data source plays the role of a noisy "channel". Maximizing the capacity of this channel determines the penalized costs of the pattern analysis problem with a data dependent regularization strength. The tradeoff between *informativeness* and *robustness* in statistical inference is mirrored in the balance between high information rate and zero communication error, thereby giving rise to a new notion of context sensitive information.

1 Introduction

We are drowning in data, but starving for knowledge!

This often deplored dilemma of the information age characterizes our rudimentary understanding of the concept "information": We are much more efficient in collecting and storing data than in analyzing them to extract the relevant bits for intelligent decision making. The information society is flooded with digital data but only a tiny fraction seems to enter the value chain of the information society. Raw data are generated at a much higher rate than we can convert these exascale data volumes into information and further refine them to knowledge. Only this transformation of digital data guarantees that digital information is a precious resource and that value is generated by information processing. To filter out the relevant information for solving a data analysis problem from the vast amount of superfluous signals and noise, we need a new concept of information – *context sensitive information*. This paper builds upon a theory of cluster analysis (see [3]) which enables the modeler to measure the informativeness of statistical models or, equivalently, to choose suitable cost functions for solving a pattern recognition problem.

J.-F. Martínez-Trinidad et al. (Eds.): MCPR 2011, LNCS 6718, pp. 12–21, 2011.

Pattern recognition in statistics and machine learning requires to infer patterns or structures in data which are stable in the presence of noise perturbations. Typical examples of such structures are data partitionings / clusterings, low dimensional embeddings of high dimensional or non-metric data or inference of discrete structures like trees from relational data. In almost all of these cases, the hypothesis class of the potential patterns is much smaller than the data space. Consequently, only a fraction of the noise fluctuations in the measurements will interfere with our search for optimal structures or interpretations of the data. Cost functions and algorithms, which preserve the relevant signal in the data to identify patterns but still filter out measurement noise as much as possible, should be prefered over brittle, noise sensitive methods.

In this paper, we formalize pattern analysis problems as optimization problems of appropriately chosen cost functions. The cost function assigns low costs to prefered patterns and high costs to unfavorable ones. How reliably we can distinguish patterns in the hypothesis class, is measured by a fictitious communication process between a sender and a receiver. The condition of error free communication determines the smallest sets of statistically indistinguishible patterns. The theoretical framework, which generalizes the model validation method for clustering by approximation set coding [3], is founded on maximum entropy inference.

2 Statistical Learning for Pattern Recognition

Given are a **set of objects** $\mathbf{O} = \{o_1, \ldots, o_n\} \in \mathcal{O}$ and **measurements** $\mathbf{X} \in \mathcal{X}$ to characterize these objects. $\mathcal{O}, \mathcal{X}$ denotes the object or measurement space, respectively. Such measurements might be d-dimensional vectors $\mathbf{X} = \{X_i \in \mathbb{R}^d, 1 \leq i \leq n\}$ or relations $\mathbf{D} = (D_{ij}) \in \mathbb{R}^{n \cdot n}$ which describe the (dis)-similarity between object o_i and o_j. More complicated data structures than vectors or relations, e.g., three-way data or graphs, are used in various applications. In the following, we use the generic notation $\mathbf{X}$ for measurements. Data denote object-measurement relations $\mathcal{O} \times \mathcal{X}$, e.g., vectorial data $\{X_i \ : \ 1 \leq i \leq n\}$ describe surjective relations between objects o_i and measurements $X_i := X(o_i)$.

The **hypotheses** of a pattern recognition problem are functions assigning data to patterns out of a pattern space $\mathcal{P}$, i.e.,

$$c \ : \ \mathcal{O} \times \mathcal{X} \to \mathcal{P}, \qquad (\mathbf{O}, \mathbf{X}) \mapsto c(\mathbf{O}, \mathbf{X}) \tag{1}$$

The pattern space for clustering problems is the set of possible assignments of data to k groups, i.e., $\mathcal{P} = \{1, \ldots, k\}^n$, $n = |\mathbf{O}|$ denoting the number of objects. For parameter estimation problems like PCA or SVD, the patterns are possible values of the orthogonal matrices and the pattern space is a subset of the d-dimensional Euclidean rotations. To simplify the notation, we omit the first argument of c in cases where $\mathbf{X}$ uniquely identifies the object set $\mathbf{O}$. A clustering is then defined as $c : \mathcal{X} \to \{1, \ldots, k\}^n$.

The **hypothesis class** for a pattern recognition problem is defined as the set of functions assigning data to elements of the pattern space, i.e.,

$$\mathcal{C}(\mathbf{X}) = \{c(\mathbf{O}, \mathbf{X}) : \mathbf{O} \in \mathcal{O}\} \tag{2}$$

For the clustering problem with n objects and given measurements we can distinguish $\mathcal{O}(k^n)$ such functions. In parameter estimation of parametric probability models we have to coarsen the continuous underlying space with regular grids or in a random fashion which yields $\mathcal{O}\big((\Omega/\epsilon)^p\big), \Omega \subset \mathbb{R}$ different functions, p being the number of parameters.

3 Empirical Risk Approximation

Exploratory pattern analysis in combination with model selection requires to assess the quality of hypotheses $c \in \mathcal{C}$, that are assignments of data to patterns. We adopt a cost function (risk) viewpoint in this paper which attributes a non-negative cost value

$$R : \mathcal{C} \times \mathcal{X} \to \mathbb{R}_+ , \qquad (c, \mathbf{X}) \mapsto R(c, \mathbf{X}) \tag{3}$$

to each hypothesis ($\mathbb{R}_+ := [0, \infty)$).

The classical theory of statistical learning [9,8] advocates to use the empirical minimizer as the solution of the inference problem. The best empirical pattern $c^\perp(\mathbf{X})$ minimizes the empirical risk (ERM) of the pattern analysis problem given the measurements $\mathbf{X}$, i.e.,

$$c^\perp(\mathbf{X}) \in \arg\min_c R(c, \mathbf{X}). \tag{4}$$

The ERM theory requires for learnability of classifications that the hypothesis class is not too complex (i.e., finite VC-dimension) and, as a consequence, the ERM solution $c^\perp(\mathbf{X})$ converges to the optimal solution which minimizes the expected risk. A corresponding criterion has been derived for regression [1].

This classical learning theory is not applicable when the size of the hypothesis class grows with the number of objects like in clustering or other optimization problems with a combinatorial flavor. Without strong regularization we cannot hope to identify a single solution which globally minimizes the expected risk in the asymptotic limit. Therefore, we replace the concept of a unique function as the solution for the learning problem with a weighted set of functions. The weights are defined as

$$w : \mathcal{C} \times \mathcal{X} \times \mathbb{R}_+ \to [0, 1] , \qquad (c, \mathbf{X}, \beta) \mapsto w_\beta(c, \mathbf{X}) . \tag{5}$$

The set of weights is denoted as $\mathcal{W}_\beta(\mathbf{X}) = \{w_\beta(c, \mathbf{X}) : c \in \mathcal{C}\}$.

How should we choose the weights $w_\beta(c, \mathbf{X})$ that large weights are only assigned to functions with low costs ? The partial ordering constraint

$$R(c, \mathbf{X}) \leq R(\tilde{c}, \mathbf{X}) \Leftrightarrow w_\beta(c, \mathbf{X}) \geq w_\beta(\tilde{c}, \mathbf{X}) , \tag{6}$$

ensures that functions with minimal costs $R(c^{\perp}, \mathbf{X})$ assume the maximal weight value which is normalized to one w.l.o.g., i.e., $0 \leq w_{\beta}(c, \mathbf{X}) \leq 1$. The non-negativity constraint of weights allows us to write the weights as $w_{\beta}(c, \mathbf{X}) = \exp\left(-\beta f(R(c, \mathbf{X}))\right)$ with the monotonic function $f(x)$. Since $f(x)$ amounts to a monotone rescaling of the costs $R(c, \mathbf{X})$ we resort w.l.o.g to the common choice of Boltzmann weights with the inverse computational temperature β, i.e.,

$$w_{\beta}(c, \mathbf{X}) = \exp\left(-\beta R(c, \mathbf{X})\right) \ . \tag{7}$$

4 Generalization and the Two Instance Scenario

To determine the optimal regularization of a pattern recognition method we have to define and estimate the generalization performance of hypotheses. We adopt the two instance scenario with training and test data described by respective object sets $\mathbf{O}^{(1)}, \mathbf{O}^{(2)}$ and measurements $\mathbf{X}^{(1)}, \mathbf{X}^{(2)} \sim \mathbb{P}(\mathbf{X})$. Both sets of measurements are drawn i.i.d. from the same probability distribution $\mathbb{P}(\mathbf{X})$. Furthermore, $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$ uniquely identify the training and test object sets $\mathbf{O}^{(1)}, \mathbf{O}^{(2)}$ so that it is sufficient to list $\mathbf{X}^{(j)}$ as references to object sets $\mathbf{O}^{(j)}, j = 1, 2$. The training and test data $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$ define two optimization problems $R(., \mathbf{X}^{(1)}), R(., \mathbf{X}^{(2)})$. The two instance scenario or two sample set scenario is widely used in statistics and statistical learning theory [8], i.e., to bound the deviation of empirical risk from expected risk, but also for two-terminal systems in information theory [5].

Statistical pattern analysis requires that inferred patterns have to generalize from training data to test data since noise in the data might render the ERM solution $c^{\perp}(\mathbf{X}^{(1)}) \neq c^{\perp}(\mathbf{X}^{(2)})$ unstable. How can we evaluate the generalization properties of solutions to a pattern recognition problem? Before we can compute the costs $R(., \mathbf{X}^{(2)})$ on test data of approximate solutions $c(\mathbf{X}^{(1)}) \in \mathcal{C}_{\gamma}(\mathbf{X}^{(1)})$ on training data we have to identify a pattern $\tilde{c}(\mathbf{X}^{(2)}) \in \mathcal{C}(\mathbf{X}^{(2)})$ which corresponds to $c(\mathbf{X}^{(1)})$. A priori, it is not clear how to compare patterns $c(\mathbf{X}^{(1)})$ for measurements $\mathbf{X}^{(1)}$ with patterns $c(\mathbf{X}^{(2)})$ for measurements $\mathbf{X}^{(2)}$. Therefore, we define a bijective mapping

$$\psi : \mathcal{X}^{(1)} \to \mathcal{X}^{(2)} , \qquad \mathbf{X}^{(1)} \mapsto \psi(\mathbf{X}^{(1)}) . \tag{8}$$

The mapping ψ allows us to identify a pattern hypothesis for training data $c \in \mathcal{C}(\mathbf{X}^{(1)})$ with a pattern hypothesis for test data $c \in \mathcal{C}(\psi(\mathbf{X}^{(2)}))$. The reader should note that such a mapping ψ might change the object indices. In cases when the measurements are elements of an underlying metric space, then a natural choice for ψ is the nearest neighbor mapping.

The mapping ψ enables us to evaluate pattern costs on test data $\mathbf{X}^{(2)}$ for patterns $c(\mathbf{X}^{(1)})$ selected on the basis of training data $\mathbf{X}^{(1)}$. Consequently, we can determine how many training patterns with large weights share also large weights on test data, i.e.,

$$\Delta Z_{\beta}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) := \sum_{c \in \mathcal{C}(\mathbf{X}^{(2)})} w_{\beta}(c, \psi(\mathbf{X}^{(1)}))\ w_{\beta}(c, \mathbf{X}^{(2)}) . \tag{9}$$

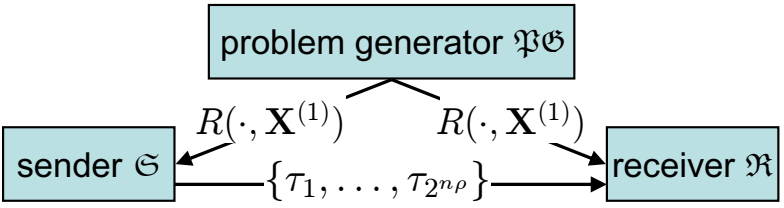

Fig. 1. Generation of a set of $2^{n\rho}$ code problems by e.g. permuting the object indices

A large subset of hypotheses with jointly large weights indicates that low cost hypotheses on training data $\mathbf{X}^{(1)}$ also perform with low costs on test data. The tradeoff between stability and informativeness for Boltzmann weights (7) is controlled by maximizing β under the constraint of large weight overlap $\Delta Z_\beta(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) / \sum_c w_\beta(c, \mathbf{X}^{(2)}) \approx 1$ for given risk function $R(., \mathbf{X})$.

5 Coding by Approximation

In the following, we describe a communication scenario with a sender $\mathfrak{S}$, a receiver $\mathfrak{R}$ and a problem generator $\mathfrak{PG}$. The problem generator serves as a noisy channel between sender and receiver. Communication takes place by approximately optimizing cost functions, i.e., by calculating weight sets $Z_\beta(\mathbf{X}^{(1)})$, $Z_\beta(\mathbf{X}^{(2)})$. This coding concept will be refered to as weighted approximation set coding (ASC) since the weights are concentrated on approximate minimizers of the optimization problem. The noisy channel is characterized by a pattern cost function $R(c, \mathbf{X})$ which determines the channel capacity of the ASC scenario. Validation and selection of pattern recognition models is then achieved by maximizing the channel capacity over a set of cost functions $R_\theta(., \mathbf{X}), \theta \in \Theta$ where θ indexes the various pattern recognition objectives.

Before we describe the communication protocol we have to define the code for communication. The objective $R(c, \mathbf{X}^{(1)})$ with the training data $\mathbf{X}^{(1)}$ define the noisy channel. We interpret the set of weights $\mathcal{W}_\beta(\mathbf{X}^{(1)})$ as the message to be communicated between sender and receiver. Since the peak of the weight distribution identifies patterns with low costs, we have to generate a set of weight sets in an analoguous way to Shannon's codebook. The two instance scenario, however, provides only one set of measurements $\mathbf{X}^{(1)}$, and consequently only one weight set. Therefore, we have to introduce a set of (random) equivariant transformations τ applied to the training data $\mathbf{X}^{(1)}$ such that the minimizer $c^\perp$ can be located at a respective (random) position in the hypothesis space. Formally, the equivariance condition states that

$$c(\tau \circ \mathbf{X}) = \tau \circ c(\mathbf{X}) \,, \tag{10}$$

i.e., a hypothesis on transformed data is equivalent to a transformation of the hypothesis on the original data. Special cases of such transformations τ are random permutations when optimizing combinatorial optimization cost functions

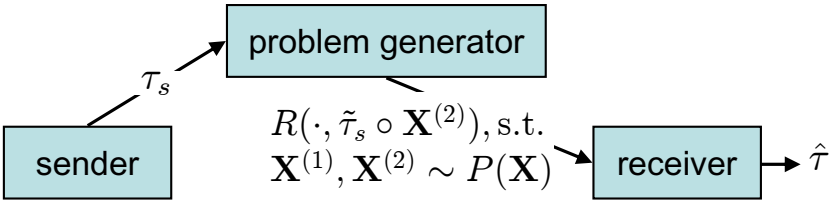

Fig. 2. Communication process: (1) the sender selects transformation τ_s, (2) the problem generator draws $\mathbf{X}^{(2)} \sim \mathbb{P}(\mathbf{X})$ and applies $\tilde{\tau}_s = \psi \circ \tau_s \circ \psi^{-1}$ to it, and the receiver estimates $\hat{\tau}$ based on $\tilde{\mathbf{X}} = \tilde{\tau}_s(\mathbf{X}^{(2)})$

like clustering models or graph cut problems. In parametric statistics, the transformations are parameter grids of e.g. rotations when estimating the orthogonal transformations of PCA or SVD. A possibly exponentially large set of transformations $\mathcal{T} = \{\tau_j : 1 \leq j \leq 2^{n\rho}\}$ then serves as the code in this communication process with a rate ρ. The set of transformations gives also rise to a set of equivalent optimization cost functions $R(c, \tau_1 \circ \mathbf{X}^{(1)}), \ldots, R(c, \tau_{2^{n\rho}} \circ \mathbf{X}^{(1)})$. It is important to note that we do not change the measurement values in this construction but their relation to the hypothesis class.

Sender $\mathfrak{S}$ and receiver $\mathfrak{R}$ agree on a cost function for pattern recognition $R(c, \mathbf{X}^{(1)})$ and on a mapping function ψ. The following procedure is then employed to generate the code for the communication process:

1. Sender $\mathfrak{S}$ and receiver $\mathfrak{R}$ obtain data $\mathbf{X}^{(1)}$ from the problem generator $\mathfrak{PG}$.
2. $\mathfrak{S}$ and $\mathfrak{R}$ calculate the weight set $\mathcal{W}_\beta(\mathbf{X}^{(1)})$.
3. $\mathfrak{S}$ generates a set of (random) transformations $\mathcal{T} := \{\tau_1, \ldots, \tau_{2^{n\rho}}\}$. The transformations define a set of optimization problems $R(c, \tau_j(\mathbf{X}^{(1)})), 1 \leq j \leq 2^{n\rho}$ to determine weight sets $\mathcal{W}_\beta(\tau_j(\mathbf{X}^{(1)})), 1 \leq j \leq 2^{n\rho}$.
4. $\mathfrak{S}$ sends the set of transformations $\mathcal{T}$ to $\mathfrak{R}$ who determines the set of weight sets $\{\mathcal{W}_\beta(\tau_j(\mathbf{X}^{(1)}))\}_{j=1}^{2^{n\rho}}$.

The rationale behind this procedure is the following: Given the measurements $\mathbf{X}^{(1)}$ the sender has randomly covered the hypothesis class $\mathcal{C}(\mathbf{X}^{(1)})$ by respective weight sets $\{\mathcal{W}_\beta(\tau_j(\mathbf{X}^{(1)})) : 1 \leq j \leq 2^{n\rho}\}$. Communication can take place if the weight sets are stable under the stochastic fluctuations of the measurements. The criterion for reliable communication is defined by the ability of the receiver to identify the transformation which has been selected by the sender. After this setup procedure, both sender and receiver have a list of weight sets available.

How is the communication between sender and receiver organized? During communication, the following steps take place as depicted in fig. 2:

1. The sender $\mathfrak{S}$ selects a transformation τ_s as message and send it to the problem generator $\mathfrak{PG}$.
2. $\mathfrak{PG}$ generates a new data set $\mathbf{X}^{(2)}$, establishes correspondence ψ between $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. $\mathfrak{PG}$ then applies the selected transformation τ_s, yielding $\tilde{\mathbf{X}} = \psi \circ \tau_s \circ \psi^{-1}(\mathbf{X}^{(2)})$.

3. $\mathfrak{PG}$ send $\tilde{\mathbf{X}}$ to the receiver $\mathfrak{R}$ without revealing τ_s.
4. $\mathfrak{R}$ calculates the weight set $\mathcal{W}_\beta(\tilde{\mathbf{X}})$.
5. $\mathfrak{R}$ estimates the selected transformation τ_s by using the decoding rule

$$\hat{\tau} = \arg\max_{\tau \in \mathcal{T}} \sum_{c \in \mathcal{C}(\mathbf{X}^{(2)})} w_\beta(c, \psi \circ \tau(\mathbf{X}^{(1)}))\, w_\beta(c, \tilde{\mathbf{X}})\,. \tag{11}$$

In the case of discrete hypothesis classes, then the communication channel is bounded from above by the cardinality of $\mathcal{C}(\mathbf{X})$ if two conditions hold: (i) the channel is noise free $\mathbf{X}^{(1)} \equiv \mathbf{X}^{(2)}$; (ii) the transformation set is sufficiently rich that every hypothesis can be selected as a global minimizer of the cost function.

6 Error Analysis of Approximation Set Coding

To determine the optimal approximation precision for an optimization problem $R(., \mathbf{X})$ we have to derive necessary and sufficient conditions which have to hold in order to reliably identify the transformations $\tau_s \in \mathcal{T}$. The parameter β, which controls the concentration of weights and thereby the resolution of the hypothesis class, has to be adapted to the size of the transformation set $|\mathcal{T}|$. Therefore, we analyse the error probability of the decoding rule (11) which is associated with a particular cost function $R(., \mathbf{X})$ and a rate ρ. The maximal value of β under the condition of zero error communication is defined as *approximation capacity* since it determines the approximation precision of the coding scheme.

A communication error occurs if the sender selects τ_s and the receiver decodes $\hat{\tau} = \tau_j, j \neq s$. To estimate the probability of this event, we introduce the weight overlaps

$$\Delta Z_\beta^j := \sum_{c \in \mathcal{C}(\mathbf{X}^{(2)})} w_\beta(c, \psi \circ \tau_j(\mathbf{X}^{(1)}))\, w_\beta(c, \tilde{\mathbf{X}})\,,\ \tau_j \in \mathcal{T}. \tag{12}$$

The number ΔZ_β^j measures the number of hypotheses which have low costs $R(c, \psi \circ \tau_j(\mathbf{X}^{(1)}))$ and $R(c, \tilde{\mathbf{X}})$.

The probability of a communication error is given by a substantial overlap ΔZ_β^j induced by $\tau_j \in \mathcal{T} \setminus \{\tau_s\}, 1 \leq j \leq 2^{n\rho}$, i.e.,

$$\begin{aligned} \mathbb{P}(\hat{\tau} \neq \tau_s | \tau_s) &= \mathbb{P}\left(\max_{1 \leq j \leq 2^{n\rho}} \Delta Z_\beta^j \geq \Delta Z_\beta^s \,\middle|\, \tau_s \right) \\ &= \mathbb{E}_{\mathbf{X}^{(1,2)}} \mathbb{E}_{\mathcal{T} \setminus \{\tau_s\}} \mathbb{I}\{\max_{j \neq s} \Delta Z_\beta^j \geq \Delta Z_\beta^s\} \end{aligned} \tag{13}$$

with the indicator function $\mathbb{I}\{expr\} = \begin{cases} 1 & expr \text{ is true,} \\ 0 & \text{otherwise.} \end{cases}$. The notation $\mathbf{X}^{(1,2)} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ denotes the expectation w.r.t. both training data $\mathbf{X}^{(1)}$ and test data $\mathbf{X}^{(2)}$. The expectation $\mathbb{E}_{\mathcal{T} \setminus \{\tau_s\}}$ is calculated w.r.t. the set of random transformations $\tau_j, 1 \leq j \leq 2^{n\rho}, j \neq s$ where we have conditioned on the sender

selected transformation τ_s. The joint probability distribution of all transformations $\mathbb{P}(\mathcal{T}) = \prod_{j=1}^{2^{n\rho}} \mathbb{P}(\tau_j)$ decomposes into product form since all transformations are randomly drawn from the set of all possible transformations $\{\tau_j\}$. It corresponds to the Shannon's random codebook design in information theory.

The error probability (13) can now be approximated by deriving an upper bound for the indicator function and, in a second step, by using the union bound for the maximum. The indicator function is bounded by $\mathbb{I}\{x \geq a\} \leq \frac{x}{a}$ for all $x \geq 0$.

The confusion probability with any other message $\tau_j, j \neq s$ for given training data $\mathbf{X}^{(1)}$ and test data $\mathbf{X}^{(2)}$ conditioned on τ_s is defined by

$$\begin{aligned}
\mathbb{E}_{\mathcal{T}\setminus\{\tau_s\}} \mathbb{I}\{\max_{j\neq s} \Delta Z_\beta^j \geq \Delta Z_\beta^s\} &\overset{(a)}{\leq} \mathbb{E}_{\mathcal{T}\setminus\{\tau_s\}} \frac{\max_{j\neq s} \Delta Z_\beta^j}{\Delta Z_\beta^s} \overset{(b)}{\leq} \mathbb{E}_{\mathcal{T}\setminus\{\tau_s\}} \sum_{\tau_j \neq \tau_s} \frac{\Delta Z_\beta^j}{\Delta Z_\beta^s} \\
&= \sum_{\tau_j \neq \tau_s} \frac{1}{|\{\tau_j\}|} \sum_{\{\tau_j\}} \frac{\Delta Z_\beta^j}{\Delta Z_\beta^s} \overset{(c)}{=} (2^{n\rho} - 1) \frac{Z_\beta^{(1)}\, Z_\beta^{(2)}}{|\{\tau_s\}| \Delta Z_\beta^s} \\
&\overset{(d)}{\leq} 2^{n\rho} \exp\left(-n \mathcal{I}_\beta(\tau_s, \hat{\tau})\right) \qquad (14)
\end{aligned}$$

with $Z_\beta^{(1,2)} := Z_\beta(\mathbf{X}^{(1,2)}) = \sum_{c \in \mathcal{C}(\mathbf{X}^{(1,2)})} w_\beta(c, \mathbf{X}^{(1,2)})$. In derivation (14) the expectation $\mathbb{E}_{\mathcal{T}\setminus\{\tau_s\}}\left[\mathbb{I}\{\Delta Z_\beta^j \geq \Delta Z_\beta^s\}\right]$ is conditioned on τ_s which has been omitted to increase the readability of the formulas. The summation $\sum_{\{\tau_j\}}$ is indexed by all possible realizations of the transformation τ_j that are uniformly selected. The first inequality (a) bounds the indicator function from above, the second inequality is due to the union bound of the maximum operator. Averaging over a random transformation τ_j (c) breaks any statistical dependency between sender and receiver weight sets which corresponds to the error case in jointly typical coding [4]; the number of possible transformations $|\{\tau_j\}|$ is identified with $|\{\tau_s\}|$ since all transformations have been chosen i.i.d. (d) We have introduced the mutual information between the sender message τ_s and the receiver message $\hat{\tau}$

$$\mathcal{I}_\beta(\tau_s, \hat{\tau}) = \frac{1}{n} \log\left(\frac{|\{\tau_s\}| \Delta Z_\beta^s}{Z_\beta^{(1)}\, Z_\beta^{(2)}}\right) = \frac{1}{n}\left(\log \frac{|\{\tau_s\}|}{Z_\beta^{(1)}} + \log \frac{|\mathcal{C}^{(2)}|}{Z_\beta^{(2)}} - \log \frac{|\mathcal{C}^{(2)}|}{\Delta Z_\beta^s}\right). \quad (15)$$

Inequality in step (d) results from dropping the correction -1.

The interpretation of eq. (15) is straightforward: The first logarithm measures the entropy of the number of transformations which can be resolved up to a minimal uncertainty encoded by the sum of the approximation weights $Z_\beta^{(1)}$ in the space of clusterings on the sender side. The logarithm $\log(|\mathcal{C}^{(2)}|/Z_\beta^{(2)})$ calculates the entropy of the receiver patterns which are quantized by $Z_\beta^{(2)}$. The third logarithm measures the joint entropy of $(\tau_s, \hat{\tau})$ which depends on the integrated weight product $\Delta Z_\beta^s = \sum_{c \in \mathcal{C}(\mathbf{X}^{(2)})} w_\beta(c, \psi \circ \tau_s \circ \mathbf{X}^{(1)})\, w_\beta(c, \tilde{\mathbf{X}})$. Inserting (14) into (13) yields the upper bound for the error probability

$$\mathbb{P}(\hat{\tau} \neq \tau_s | \tau_s) \leq \mathbb{E}_{\mathbf{X}^{(1,2)}} \exp(-n(\mathcal{I}_\beta(\tau_s, \hat{\tau}) - \rho \log 2)) \qquad (16)$$

The communication rate $n\rho \log 2$ is limited by the mutual information $\mathcal{I}_\beta(\tau_s, \hat{\tau})$ for asymptotically error-free communication.

7 Information Theoretical Model Selection

The analysis of the error probability suggests the following inference principle for controling the appropriate regularization strengths which implements a form of model selection: the approximation precision is controlled by β which has to be maximized to derive more precise solutions or patterns. For small β the rate ρ will be low since we resolve the space of solutions only in a coarse grained fashion. For too large β the error probability does not vanish which indicates confusions between $\tau_j, j \neq s$ and τ_s. The optimal β-value is given by the largest β or, equivalently the highest approximation precision

$$\beta^\star = \arg \max_{\beta \in [0,\infty)} \mathcal{I}_\beta(\tau_s, \hat{\tau}). \tag{17}$$

Another choice to be made in modeling is to select a suitable cost function $R(., \mathbf{X})$ for the pattern recognition problems at hand. Let us assume that a number of cost functions $\{R_\theta(., \mathbf{X}), \theta \in \Theta\}$ are considered as candidates. The approximation capacity $\mathcal{I}_\beta(\tau_s, \hat{\tau})$ depends on the cost function through the Gibbs weights. Therefore, we can rank the different models according to their $\mathcal{I}_\beta(\tau_s, \hat{\tau})$ values. Robust and informative cost functions yield a higher approximation capacity than simpler or more brittle models. A rational choice is to select the cost function

$$R^\star(c, \mathbf{X}) = \arg \max_{\theta \in \Theta} \mathcal{I}_\beta(\tau_s, \hat{\tau} | R_\theta) \tag{18}$$

where both the random variables τ_s and $\hat{\tau}$ depend on $R_\theta(c, \mathbf{X}), \theta \in \Theta$. The selection rule (18) prefers the model which is "expressive" enough to exhibit high information content (e.g., many clusters in clustering) and, at the same time robustly resists to noise in the data set. The bits or nats which are measured in the ASC communication setting are context sensitive since they refer to a hypothesis class $\mathcal{C}(\mathbf{X})$, i.e., how finely or coarsely functions can be resolved in $\mathcal{C}$.

8 Conclusion

Model selection and validation requires to estimate the generalization ability of models from training to test data. "Good" models show a high expressiveness and they are robust w.r.t. noise in the data. This tradeoff between *informativeness* and *robustness* ranks different models when they are tested on new data and it quantitatively describes the underfitting/overfitting dilemma. In this paper we have explored the idea to use weighted approximation sets of clustering solutions as a communication code. The *approximation capacity* of a cost function provides a selection criterion which renders various models comparable in terms of their respective bit rates. The number of reliably extractable bits of a pattern analysis cost function $R(., \mathbf{X})$ defines a "task sensitive information measure" since it only

accounts for the fluctuations in the data $\mathbf{X}$ which actually have an influence on identifying an individual pattern or a set of patterns.

The maximum entropy inference principle suggests that we should average over the statistically indistinguishible solutions in the optimal weighted approximation set. Such a model averaging strategy replaces the original cost function with the free energy and, thereby, it defines a continuation methods with maximal robustness. Algorithmically, maximum entropy inference can be implemented by annealing methods [7,2,6]. The urgent question in many data analysis applications, which regularization term should be used without introducing an unwanted bias, is naturally answered by the entropy. The second question, how the regularization parameter should be selected, in answered by ASC: Choose the parameter value which maximizes the approximation capacity!

ASC for model selection can be applied to all combinatorial or continuous optimization problems which depend on noisy data. The noise level is characterized by two sample sets $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$. Two samples provide far too little information to estimate the probability density of the measurements but two large samples contain sufficient information to determine the uncertainty in the solution space. The equivalence of ensemble averages and time averages of ergodic systems is heavily exploited in statistical mechanics and it also enables us in this paper to derive a model selection strategy based on two samples.

Acknowledgment. This work has been partially supported by the DFG-SNF research cluster FOR916 and by the FP7 EU project SIMBAD.

References

1. Alon, N., Ben-David, S., Cesa-Bianchi, N., Haussler, D.: Scale-sensitive dimensions, uniform convergence, and learnability. J. ACM 44(4), 615–631 (1997)
2. Buhmann, J.M., Kühnel, H.: Vector quantization with complexity costs. IEEE Tr. Information Theory 39(4), 1133–1145 (1993)
3. Buhmann, J.M.: Information theoretic model validation for clustering. In: IEEE International Symposium on Information Theory, Austin Texas. IEEE, New York (2010), http://arxiv.org/abs/1006.0375
4. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley & Sons, New York (1991)
5. Csiczár, I., Körner, J.: Information Theory: Coding theorems for discrete memoryless systems. Academic Press, New York (1981)
6. Hofmann, T., Buhmann, J.M.: Pairwise data clustering by deterministic annealing. IEEE Tr. Pattern Analysis and Machine Intelligence 19(1), 1–14 (1997)
7. Rose, K., Gurewitz, E., Fox, G.: Vector quantization by deterministic annealing. IEEE Transactions on Information Theory 38(4), 1249–1257 (1992)
8. Vapnik, V.N.: Estimation of Dependences Based on Empirical Data. Springer, Heidelberg (1982)
9. Vapnik, V.N., Chervonenkis, A.Y.: On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications 16, 264–280 (1971)