Proceedings of the
FOURTH WORKSHOP ON
INFORMATION THEORETIC
METHODS IN SCIENCE AND
ENGINEERING

Edited by
Jorma Rissanen, Petri Myllynäki, Teemu Roos,
Ioan Tabus & Kenji Yamanishi
PREFACE

The Fourth Workshop on Information Theoretic Methods in Science and Engineering (WITMSE 2011) took place on August 7–10, 2011, in Helsinki, Finland. The workshop was hosted by the University of Helsinki and the Helsinki Institute for Information Technology HIIT. This was the fourth workshop in the series which started in 2008. The first one, as well as the following two organized in 2009 and 2010, respectively, were hosted by the Technical University of Tampere.

As the title of the workshop suggests, WITMSE seeks speakers from a variety of disciplines with emphasis on both theory and applications of information and coding theory with special interest in modeling. Since the beginning our plan has been, and still is, to keep the number of the participants small and to ensure the highest possible quality, which has been accomplished by inviting distinguished scholars as speakers.

The invitees include both plenary and regular invited speakers—although in this context, the word “regular” should be written in quotation marks as it is clear that both kinds of speakers have been outstanding. This year, the plenary talks were given by Mati Wax (Wavion Wireless Networks), Neri Merhav (Technion), and Veronica Gonzalez-Lopez and Jesus Garcia (State University of Campinas).

Outside the technical sessions the program included a reception offered by the Rector of the University of Helsinki, hosted by Vice-Rector Kimmo Kontula, and a Banquet including a chance to enjoy a Finnish sauna and take a dip in the Baltic Sea.

We would like to thank all the participants to our workshop. Many of the speakers kindly submitted written contributions to these proceedings, for which we are particularly grateful. We also want to thank the Federation of Finnish Learned Societies and the European Union Network-of-Excellence Pascal for sponsoring the workshop.

November 22, 2011
Helsinki, San Jose, Tampere, and Tokyo
Workshop Co-Chairs
Jorma Rissanen,
Petri Myllymäki,
Teemu Roos,
Ioan Tabus,
& Kenji Yamanishi
Contents

Preface ........................................................................................................... 3

Dror Baron: Information Complexity and Estimation ................................. 7
Fabio Benatti: Gacs Quantum Complexity and Quantum Entropy ............ 11
David R. Bickel: A Game-Theoretic Framework for Blending Bayesian and
Frequentist Methods of Statistical Inference ............................................. 15
Tim van Erven, Mark D. Reid, and Robert C. Williamson: Mixability is Bayes
Curvature Relative to Log Loss ................................................................. 19
Søren Forchhammer, Huynh Van Luong, and Xin Huang: Multiple LDPC Decoding
for Distributed Source and Video Coding .................................................. 21
Jesús E. García and Verónica A. González-López: Minimal Markov Models .... 25
Ciprian Doru Giurcăneanu, Seyed Alireza Razavi, and Antti Liski: Variable Selection
in Linear Regression: Several Approaches Based on Normalized Maximum Likelihood 29
Esfandiar Haghverdi: Aspects of Category Theory ....................................... 31
Peter Harremoës: Functional Dependences and Bayesian Networks ........... 35
Jenni Hukkanen, Edmond Sabo, and Ioan Tabus: MDL Based Interpretation for Overlapping Cell Nuclei in Histological Images ........................................... 39
Vladimir Katkovnik, Aram Danielyan, and Karen Egiazarian: Generalized Nash
Equilibrium Image Deblurring ................................................................. 41
Enes Makalic and Daniel Schmidt: MDL Multiple Hypothesis Testing ......... 45
Steven de Rooij and Wouter Koolen: Time Independent Switching between Two Source
Codes ........................................................................................................... 49
Narayana Santhanam and Venkat Anantharam: Universal Prediction over Large Alphabets ................................................................. 51
Ionuţ Şchiopu and Ioan Tăbuş: MDL Segmentation and Lossless Compression of Depth
Images ......................................................................................................... 55
Daniel F. Schmidt: A New Message Length Approximation for Parametric Estimation
and Model Selection .................................................................................. 59
Flemming Topsøe: Cognition and Inference in an Abstract Setting ................................. 67
Paul M.B. Vitányi: Information Distance: New Developments ................................. 71
Kazuho Watanabe: Minimum Variational Stochastic Complexity and Average General-
ization Error in Latent Variable Models ................................................................. 75
Hannes Wettig, Suvi Hiltunen, and Roman Yangarber: MDL-Based Modeling of Etymo-
logical Sounds Change in the Uralic Language Family ........................................... 79

Author Index .................................................................................................................. 83
INFORMATION COMPLEXITY AND ESTIMATION

Dror Baron

1Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695, USA, barondror@ncsu.edu

ABSTRACT

We consider an input $x$ generated by an unknown stationary ergodic source $X$ that enters a signal processing system $J$, resulting in $w = J(x)$. We observe $w$ through a noisy channel, $y = z(w)$; our goal is to estimate $x$ from $y$, $J$, and knowledge of $f_{Y | W}$. This is universal estimation, because $f_X$ is unknown. We provide a formulation that describes a trade-off between information complexity and noise. Initial theoretical, algorithmic, and experimental evidence is presented in support of our approach.

1. INTRODUCTION

Universal algorithms [1–5] achieve the best possible performance asymptotically—without knowing the input statistics. These algorithms have had tremendous impact in lossless compression, which is crucial for data backups and transmissions. In sharp contrast, universal algorithms have made much less impact in other areas.

Estimation algorithms attempt to recover an input from noisy measurements (Figure 1). Numerous estimation problems have received great attention including the additive noise scalar channel, $y = x + z$ [6]; linear matrix multiplication with additive noise, $y = Jx + z$ with applications including compressed sensing [7–9], finance, medical and seismic imaging; universal lossy compression [4, 5, 10], where the goal is to find compressible $x$ that is sufficiently close to $y$; nonlinear regression, where $J(x)$ is nonlinear; and distributed signal processing.

In these estimation problems, the common goal is to estimate the input $x$ from knowledge of the noisy measurements $y$ and measurement system $J$. To do so, we must exploit all statistical structure in $x$. A particularly challenging type of statistical structure is the appearance of spatial or temporal dependencies in data. In images, such dependencies can be captured by dictionary learning or employing energy compacting transforms. In other problems, the statistical dependencies might be more subtle. Following the lead of universal lossless compression, we assume that the input $x$ was generated by an unknown stationary ergodic source $X$. It is well known that stationary ergodic models capture the statistics of text files well, and hence the success of universal lossless compressors. Stationary ergodic models have also been incorporated in speech denoising and enhancement, and appear prominently in hidden Markov models.

One approach to universal estimation relies on Kolmogorov complexity [11]. For a prospective $\hat{x}$, the Kolmogorov complexity $K(\hat{x})$ is the length of the shortest computer program that can compute $\hat{x}$. Donoho [12] proposed a Kolmogorov-based estimator for the white scalar channel, $y = x + z$. Despite related extensions to compressed sensing [8, 9], what is missing in the literature is a universal approach in arbitrary measurement systems that would support noise and unknown stationary ergodic input distributions.

We propose to perform universal estimation in (potentially nonlinear) signal processing systems from noisy measurements. The algorithmic component of our work features a harmonious marriage of scalar quantization, universal lossless compression, and Markov chain Monte Carlo. We evaluate the estimated input $\hat{x}$ over a quantized grid and optimize for the trade-off between information complexity (lossless coding length) of $\hat{x}$ and how well $\hat{x}$ explains the measurements $y$. We report promising preliminary theoretical and numerical results.

2. INFORMATION COMPLEXITY FORMULATION

We focus on the setting where the lengths $M$ of the output $y$ and $N$ of the input $x$ both grow to infinity, $M, N \to \infty$. We further assume that their ratio is finite and positive, \( \lim_{N \to \infty} \frac{M}{N} = \delta > 0 \). Similar settings have been discussed in the literature, e.g., [13]. Since $x$ was generated by an unknown source, we must search for an estimation mechanism that is agnostic to the specific distribution $f_X$.

Kolmogorov complexity: For $x \in \mathbb{R}^N$, the Kolmogorov complexity [11] of $x$, denoted by $K(x)$, is the length of the shortest computer program that can compute $x$. To be more precise, $K(x)$ is the length of the shortest input to a Turing machine [14] that generates $x$ and then halts.

We limit our discussion to Turing machines whose “input tapes” consist of bits. Consider the shortest program $P(x)$ that generates $x$. From the perspective of a source encoder [6], we say that $P(x)$ is a code for $x$.

Having linked Turing machines [14] and data compression [6], let us temporarily limit the discussion to discrete valued $x$ generated by a stationary ergodic source $X$. Each such $x$ is generated with probability $p_X(x)$, and it is easily shown that the per-symbol Kolmogorov coding length $K(x)$ converges to the entropy rate $H$ almost surely, $\lim_{N \to \infty} \frac{1}{N} K(x) = H$ [6]. Noting that universal lossless compressors [1, 2] achieve $H$ asymptotically.
for discrete valued stationary ergodic sources [6], we see that these algorithms achieve the per-symbol Kolmogorov complexity almost surely.

**Kolmogorov sampler**: For additive white Gaussian noise, \( y = x + z \), Donoho [12] proposed the Kolmogorov sampler,

\[
\hat{x}_{KS} = \arg\min_{x} \left\{ K(x) - \log(f_Z(z = y - \hat{x})) \right\}.
\]

For stationary ergodic \( X \), \( \hat{x}_{KS} \) is sampled from the posterior \( f_{Y|X}(y|x) \), where the mean square error, \( E[(\hat{x}_{KS} - x)^2] \), is twice larger than the Bayesian minimum mean square error (MMSE) [12].

In a later paper, Donoho et al. discussed a Kolmogorov estimator for compressed sensing \( y = Jx \) [8]; their estimator ignores noise, and is of limited practical utility. For the noisy version of this problem, \( y = Jx + z \), Haupt and Nowak [9] described a complexity measure that, when optimized, produces the LASSO algorithm [15]. To the best of our knowledge, Haupt and Nowak did not pursue complexity based regularization beyond iid signals and additive white Gaussian noise (AWGN).

**Quantization and estimation**: The overwhelming majority of real numbers have infinite Kolmogorov complexity. Nonetheless, some scalars \( x \in \mathbb{R}^N \) can be represented by a finite length \( \mathcal{P}(x) \). In practice, it is impossible to compute \( K(x) \) even for discrete alphabets. At the same time, we have seen that universal lossless source codes [1, 2] achieve per-symbol Kolmogorov coding length almost surely [6]. To represent continuous valued \( \hat{x} \), we apply a scalar quantizer, \( Q : \hat{x} \in \mathbb{R}^N \rightarrow x' \in Q^N \), and then compress \( x' = Q(\hat{x}) \) with a universal lossless compressor \( U \) with coding length \( U(x') \), where quantization levels \( Q \subset \mathbb{R} \) consist of a finite subset of \( \mathbb{R} \), and performing an optimization over \( \hat{x} \in Q^N \) reduces the complexity of the estimation problem from infinite to combinatorial. Note that we generate \( x' \) by independently quantizing each entry of \( x \) with \( Q \). This *encoder* first describes the quantizer \( Q, \) and then comports \( Q(x) \). The coding length, which we desire to minimize, is denoted by \( U(Q(x)) \) or \( U(x) \).

It would seem that we must search for a good scalar quantizer \( Q \) (Section 3), but *data-independent* reproduction levels are of theoretical interest,

\[
\mathcal{R} \triangleq \left\{ \frac{\gamma^2}{2}, \frac{\gamma^2 - 1}{2}, \ldots, \frac{\gamma^2}{2} \right\}, \quad \gamma = \lceil \log(N) \rceil.
\]

As \( N \) increases, \( \mathcal{R} \) will quantize a broader range of values of \( x \) to a greater resolution. An encoder based on \( \mathcal{R} \) need not describe the structure of the data-independent quantizer, because \( N \) is known. That is, \( U(\mathcal{R}(x)) \) only accounts for the length of the universal code \( U \).

**Universal MAP estimation**: We perform maximum *a posteriori* (MAP) estimation over possible sequences \( \hat{x} \in \mathcal{R}^N \), where the prior \( p_X(x) = 2^{-U(x)} \) utilizes the coding length \( U(\hat{x}) \) of some universal lossless compressor [1, 2],

\[
\hat{x}_{MAP} = \arg\min_{\hat{x} \in \mathcal{R}^N} \{ U(\hat{x}) - \log(f_{Y|X}(y|w = J(\hat{x}))) \},
\]

where we note that \( \mathcal{R}(\hat{x}) = \hat{x} \) for \( \hat{x} \in \mathcal{R}^N \). Our MAP estimator is applicable to any signal processing system \( J \) and supports any probabilistic noise operators, it is closely related to universal prediction [2, 3].

**Estimation performance**: We have promising preliminary theoretical results using the data-independent quantizer \( \mathcal{R} \). In universal lossy source coding of analog (continuous valued) sources [4], we have shown with Weissman that \( \hat{x}_{MAP} \) achieves the rate distortion function for finite variance stationary ergodic sources in an approximate asymptotic sense. That is, \( U(\mathcal{R}(\hat{x})) \) offers a sufficiently good approximation to \( K(\hat{x}) \) in universal lossy compression, where we chose \( U(\hat{x}) \) to be empirical entropy of blocks of \( q = O(\log(N)) \) symbols in \( \hat{x} \). In universal compressed sensing [16], we have shown with Duarte that under minor technical conditions on \( f_x \), performing MAP estimation over the discrete alphabet \( \mathcal{R} \) converges to the MAP estimate over the continuous distribution \( f_x \) asymptotically, where we used i.i.d. zero-mean Gaussian noise \( z \in \mathbb{R}^M \) with known variance. It remains to be seen whether \( \mathcal{R} \) or other data-independent quantizers are useful for arbitrary nonlinear measurement systems.

In terms of the mean square error, we would expect \( \hat{x}_{MAP} \) to perform well in Donoho’s scalar channel setting, \( y = x + z \). With Duarte [16], we have promising results for the compressed sensing (linear matrix multiplication) channel, \( y = Jx + z \), where we approximated \( \hat{x}_{MAP} \) (1) by a Markov chain Monte Carlo (MCMC) [17] algorithm (Section 3). Figure 2 illustrates recovery results from Gaussian measurement matrices for a source with i.i.d. Bernoulli entries with nonzero probability of 3%. Our MCMC algorithm outperforms \( l_1 \)-norm minimization, which is a well-known compressed sensing reconstruction (estimation) algorithm [7], except when the number of measurements \( M \) is low. Comparing MCMC to the minimum mean square error (MMSE) achievable in the Bayesian regime with known statistics [13], the square error achieved by MCMC is three times larger. One is left to wonder whether the mean square error performance of our algorithm might also be double the MMSE, particularly in the limit of infinite computation (Section 3).
Taking Kolmogorov beyond MAP: The Kolmogorov sampler \(\hat{x}_{KS}\) samples from the posterior \([12]\); it throws away all the statistical information it has on signals \(\hat{x}\) that differ from \(\hat{x}_{KS}\). Seeing that the mean square error obtained by \(\hat{x}_{KS}\) is double the MMSE, there is great potential to reduce estimation error over our Kolmogorov-based MAP estimator \(\hat{x}_{MAP}\) \((1)\). We therefore propose Kolmogorov-based conditional expectation,

\[
\hat{x}_{MSE} = \frac{\sum_{x \in \mathbb{R}^N} \cdot 2^{-U(\hat{x})} f_{Y|W}(y|w = J(\hat{x}))}{\sum_{x \in \mathbb{R}^N} 2^{-U(\hat{x})} f_{Y|W}(y|w = J(\hat{x}))},
\]

where we employ the universal prior, \(p_X(\hat{x}) = 2^{-U(R(\hat{x}))}\). It is well known that conditional expectation achieves the MMSE of the Bayesian regime, and this estimator should perform well. Interestingly, when the signal to noise ratio (SNR) is low, the Bayesian MMSE is sizable, and achieving double the MMSE is unimpressive. In these low SNR settings, \(\hat{x}_{MSE}\) should estimate much better than \(\hat{x}_{MAP}\).

In some signal processing systems, one wants to minimize some other (not necessarily quadratic) distortion metric \(D(x, \hat{x})\). The universal prior is readily invoked by defining the Kolmogorov conditional probability,

\[
p_{X|Y}(x|y) = \frac{p_{Y|X}p_X}{p_Y} \propto p_{Y|X}p_X,
\]

and taking the minimizing expression gives the Kolmogorov-based estimator for \(D(\cdot)\),

\[
\hat{x}_D = \arg \min_{w} \left\{ \sum_{x \in \mathbb{R}^N} D(\hat{x}, w)f_{Y|W}(y|w = J(\hat{x}))2^{-U(\hat{x})} \right\}.
\]

For scalar channels and iid noise, Sivaramakrishnan and Weissman \([18]\) described a universal denoising algorithm that estimates \(x\) by \(\hat{x}_{SW}\), its expected error \(E[D(x, \hat{x}_{SW})]\) converges to the Bayesian risk asymptotically in an appropriate stochastic setting. For scalar channels and iid noise, our expected estimation error \(E[D(x, \hat{x}_D)]\) should also be asymptotically optimal. The performance in arbitrary signal processing systems \(J\) is an open question.

### 3. Algorithms

In principle, \(\hat{x}_{MAP}\) can be computed by evaluating the Kolmogorov-based posteriors of \(|R|\) possible sequences \(R(x)\). This is better than continuous estimation, but still computationally intractable. Instead, we perform this optimization using Markov chain Monte Carlo (MCMC) \([5, 17]\), where \(U(\hat{x}) = H_q(\hat{x})\) is the empirical entropy of blocks of \(q = O(\log(N))\) symbols of \(\hat{x}\).

**Markov chain Monte Carlo:** We use MCMC \([17]\) to approximate \(\hat{x}_{MAP}\), which is the globally optimal MAP minimizer. To keep things simple, assume that \(\hat{x} \in \mathbb{R}^N\) is a candidate estimate. Define the Boltzmann PDF,

\[
f_s(\hat{x}) \triangleq \frac{1}{\zeta_s} \exp(-s[H_q(\hat{x}) - \log(f_{Y|W}(y|w = J(\hat{x})))]),
\]

where \(H_q(\hat{x})\) is the empirical entropy of blocks of \(q\) symbols in \(x\) \([2, 4, 5, 16]\), \(q = O(\log(N))\) to ensure convergence of the empirical entropy to the entropy rate \([6]\), \(s > 0\) is inversely related to temperature in an analogous statistical physics heat-bath setting \([17]\), and \(\zeta_s\) is a normalization constant. To sample from the Boltzmann PDF \((2)\), we use a Gibbs sampler: in each iteration, a single element \(\hat{x}_n\) is generated by resampling from the PDF, while the rest of \(\hat{x}\) remains unchanged. The key idea is to reduce temperatures slowly enough for the randomness of Gibbs sampling to eventually drive MCMC out of any local minimum toward the globally optimal \(\hat{x}_{MAP}\).

**Adaptive quantizer:** Jalali and Weissman \([5]\) have used MCMC to approach the fundamental rate distortion (RD) limits \([6]\) in lossy compression of binary inputs. For continuous valued (analog) sources \([4]\), using the data-independent quantizer \(R\) in MCMC asymptotically achieves the RD function universally for stationary ergodic continuous amplitude sources. However, \(R\) grows with the input length, slowing down the convergence to the RD function, and is thus an impediment in practice.

To address this issue, we next propose an MCMC-based algorithm that uses an adaptive quantizer \(Q\). The ground-breaking work by Rose on the discrete nature of the Shannon codebook for iid sources when the Shannon lower bound is not tight \([19]\) suggests that, for most sources of practical interest, restriction of the quantizer \(Q\) to a smaller number of levels does not stand in the way of attaining the fundamental compression limits. When employed on such sources, our latter algorithm zeroes in on the finite quantizer, and thus enjoys rates of convergence commensurate with the small-quantizer setting.

**Numerical Results:** In universal lossy compression of analog sources \([4]\), we have developed an algorithm that optimizes the quantizer for square error, and have promising preliminary results. Figure 3 compares results for an
Figure 3. Universal lossy compression: Rate $R$ vs. distortion $D$ of entropy coding [20], results by Yang and Zhang [10], average rate and distortion of our universal lossy compression algorithm [4], and the RD function [6] for length-15000 iid Laplace inputs, $f_X(x) = \frac{1}{2}e^{-|x|}$.

iid Laplace input, $f_X(x) = \frac{1}{2}e^{-|x|}$, achieved by entropy coding [20], a deterministic approach by Yang and Zhang [10], and our universal MCMC algorithm [4].

In our universal compressed sensing work with Duarte [16], we focused on development of a fast routine for optimizing the quantizer; this routine greatly accelerates the algorithm. We have seen in Figure 2 for a source with i.i.d. Bernoulli entries with nonzero probability of 3% that MCMC outperforms $\ell_1$-norm minimization, except when the number of measurements $M$ is low. We have additional results, but omit these for brevity; MCMC generally estimates the input signal $x$ well, but much work remains to be done.

4. ACKNOWLEDGMENTS

I wish to thank Marco Duarte and Tsachy Weissman for permission to discuss our joint work [4, 16]. Thanks also to Neri Merhav, Deanna Needell, and Phil Schniter for enlightening discussions.

5. REFERENCES


GACS QUANTUM COMPLEXITY AND QUANTUM ENTROPY

Fabio Benatti\textsuperscript{1,2}

\textsuperscript{1}Department of Physics, University of Trieste and
\textsuperscript{2}Istituto Nazionale di Fisica Nucleare
Strada Costiera 11, I-34151, Trieste, ITALY, benatti@ts.infn.it

ABSTRACT

The development of quantum information theory motivated the extension to the quantum realm of notions from algorithmic complexity theory. Because of the structure of quantum mechanics, several inequivalent generalizations are possible. The same phenomenon characterizes the extension of the dynamical entropy of Kolmogorov and Sinai. In the following, we shall examine the relations between the quantum complexity introduced by Gacs and the quantum dynamical entropy proposed by Alicki and Fannes.

1. INTRODUCTION

Quantum information theory deals with what the transmission and manipulation of information when is the information carriers are quantum systems. Apart from its potential revolutionary applications to computational and cryptographical tasks [1], from a more abstract point of view, in view of the framework offered by classical algorithmic complexity theory [2], the use of quantum states raised the question of how complex they are and how complex is their dynamics [3,4,5,6,7]. From a statistical point of view, the von Neumann entropy rate of stationary quantum sources replaces the Shannon entropy rate of stationary classical sources. A stationary classical source can be viewed as a particular dynamical system where the dynamics is the shift along a classical spin chain and its entropy rate is the Kolmogorov-Sinai entropy of the shift. The Kolmogorov-Sinai entropy has several inequivalent extensions to quantum dynamical systems [8,9,10,11,12]; for quantum chains, namely one-dimensional lattices with a $d$-level quantum system at each site, one of these generalized dynamical entropies, the Alicki-Fannes entropy [9], differs from the von Neumann entropy rate by $\log d$. In the following we will relate this extra term to the Gacs quantum complexity [5] which is one among several proposals of how algorithmic complexity theory may be generalized to the quantum realm [3,4,5,6].

2. QUANTUM SOURCES

In quantum information theory, the central notion is that of a qubit, the most elementary quantum system described by a Hilbert space $\mathbb{H} = \mathbb{C}^2$ of dimension 2. The states of a qubit are of the form

$$|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad |1\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

and density matrices $\rho$. These are normalized ($\text{Tr} \rho = 1$), positive $2 \times 2$ matrices:

$$\rho = \begin{pmatrix} r & \sigma \\ \sigma^* & 1 - r \end{pmatrix}, \quad 0 \leq r \leq 1, \quad r(1-r) \leq |\sigma|^2,$$

such that, when spectralized, $\rho = r_i |r_i\rangle \langle r_i| + r_2 |r_2\rangle \langle r_2|$, their eigenvalues $r_{1,2}$ are probabilities.

To a qubit there corresponds the algebra $M = M_2$ of complex $2 \times 2$ matrices among which there are the quantum observables $A^\dagger = A$ (hermitian matrices) whose mean values with respect to $\rho$ are given by

$$\langle A \rangle_\rho = \text{Tr}(\rho A) = \sum_{i=1}^{2} r_i \langle r_i| A |r_i\rangle.$$  \hspace{1cm} (4)

The degree of mixedness of qubit states is measured by the von Neumann entropy

$$S(\rho) = -\text{Tr} \rho \log \rho = -\sum_{i=1}^{2} r_i \log r_i,$$ \hspace{1cm} (5)

which corresponds to the Shannon entropy of the probability distribution given by the eigenvalues of $\rho$.

2.1. Quantum chains as quantum sources

In the following, we shall consider systems consisting of infinitely many qubits; for $n$ of them the Hilbert space is $\mathbb{H} = \mathbb{C}^{2^n}$, the tensor product of the Hilbert spaces of the single qubits, their algebra is the tensor products of the single qubit matrix algebras,

$$M_{2^n} = M_2 \otimes M_2 \otimes \cdots M_2 =: M_{[0,n-1]}.$$ \hspace{1cm} (6)
while their state is a density matrix $\rho^{(n)}$ that is a positive matrix in $M_{[0,n−1]}$ of trace 1.

The notation $M_{[0,n−1]}$ indicates that the quantum system at hands can be thought as consisting of $n$ qubits located at the integer sites of an initial segment of length $n$ of a one-dimensional lattice. If the family of density matrices $\rho^{(n)}$ satisfies suitable compatibility conditions, letting $n → ∞$ yields a quantum chain $\mathcal{M} = \lim_n M_{[0,n−1]}$ equipped with a translation-invariant state $ω = \lim_n ω^{(n)}$ characterized by an entropy rate

$$s(ω) = \lim_{n → ∞} \frac{1}{n} S(\rho^{(n)}) .$$

(7)

This quantum spin chain is a model of a stationary quantum information source: if, instead of the matrix algebras $M_2$, we considered diagonal algebras and equip the resulting chain with a compatible sequence of diagonal density matrices $\rho^{(n)}$, then we would get a classical stationary binary source.

2.2. Alicki-Fannes quantum dynamical entropy

For a classical source the (Shannon) entropy rate corresponds to the Kolmogorov-Sinai dynamical entropy of the shift along the strings emitted by the source; we shall now focus upon the Alicki-Fannes entropy which is one of its possible quantum generalizations. In the case of a quantum source, the dynamics, that is the shift along the chain, is represented by an automorphism $\Theta$ of the algebra $\mathcal{M}$ that moves single site algebras $M_j$ one step to the right:

$$\Theta[M_{[0,n−1]}] = M_{[1,n]} .$$

(8)

In quantum mechanics, measurement processes alter the system state on which they are performed and their effects are generically described by partitions of unit:

$$\mathcal{X} = \{ X_i \}_{i=1}^p , \quad \sum_{i=1}^p X_i^\dagger X_i = 1 , \quad X_i ∈ \mathcal{M} .$$

(9)

These operators provide a tomography of a density matrix $ρ ∈ M_2$ by means of an auxiliary density matrix $ρ[\mathcal{X}] ∈ M_p$ whose entries are:

$$ρ[\mathcal{X}]_{ij} = \text{Tr}(ρX_j^∗ X_i) .$$

(10)

For quantum spin chains, a useful partition is provided by the so-called matrix units, $u_{ij} ∈ M_2 = M_0$:

$$\mathcal{U} = \left\{ u_{ij} = \frac{1}{\sqrt{d}} |i⟩⟨j| \right\} , \quad u_{ij} u_{kl} = δ_{jk} u_{il} ,$$

(11)

where $\{|i⟩\}_{i=1}^d$ in the orthonormal basis in (1). Under the dynamics, the matrix units evolve according to

$$M_0 ∋ u_{ij} → Θ(u_{ij}) = 1 ⊗ u_{ij} ∈ M_{[0,1]} .$$

(12)

After $n$ dynamical steps, one constructs refined partitions of unit: $\mathcal{U}^{(n)} = \{ u_{i(n),j(n)} \}$ whose $2^{2n}$ elements read

$$u_{i(n),j(n)} = u_{i_0,j_0} ⊗ u_{i_1,j_1} ⊗ \cdots u_{i_{n−1},j_{n−1}} .$$

(13)

This set of $2^{2n}$ matrices provides a tomography of the state $ρ^{(n)}$, namely a $2^{2n} × 2^{2n}$ partition dependent density matrix

$$ρ[\mathcal{U}^{(n)}] = \left[ \text{Tr}(ρ^{(n)} u_{i(n),j(n)}^∗ u_{k(n),ℓ(n)}) \right] ,$$

(14)

which describes how the dynamics influences the reconstruction of the system state in relation to a chosen partition of unit. Its matrix elements can be explicitly calculated using (11) and (13):

$$\text{Tr}(ρ^{(n)} u_{j(n),j(n)} u_{k(n),ℓ(n)}) = δ_{k,j} δ_{k,j} 1 \cdot δ_{k,j} δ_{k,j} (ℓ_0 \cdots ℓ_{n−1}) ρ^{(n)} |j_0 \cdots j_{n−1}) ,$$

so that the auxiliary density matrix reads

$$ρ[\mathcal{U}^{(n)}] = \frac{1}{d^n} ⊗ ρ^{(n)} .$$

(15)

The latter has thus von Neumann entropy

$$S(ρ[\mathcal{U}^{(n)}]) = S(ρ^{(n)}) + n \log d ,$$

(16)

and entropy rate

$$h_ω(Θ, \mathcal{U}) = \lim_{n → ∞} \frac{1}{n} S(ρ[\mathcal{U}^{(n)}]) = s(ω) + \log d .$$

(17)

This is the Alicki-Fannes entropy of the quantum chain as it is the maximum one can obtain by varying over all possible partitions of unit.

How can the extra term $\log d$ be interpreted? For this, in the following, we rely on quantum algorithmic complexity tools.

3. CLASSICAL ALGORITHMIC COMPLEXITY

While the entropy rate characterize the complexity of a classical binary source from a statistical point of view, the algorithmic complexity refers to individual strings $i^{(n)} ∈ Ω^n_2 = \{0, 1\}^n$; it is the length $ℓ(p)$ of the shortest binary program $p$ which run by a prefix universal Turing machine (UTM) outputs $i^{(n)}$:

$$K(i^{(n)}) = \min_{p : T(p) = i^{(n)}} \{ ℓ(p) : T(p) = i^{(n)} \} .$$

(18)

The algorithmic complexity of individual strings is intimately connected with the notion of universal probability of a string:

$$\mathbb{P}(i^{(n)}) := \sum_{p : T(p) = i^{(n)}} 2^{−ℓ(p)} .$$

(19)

It turns out that, apart from an additive constant, independent of the string,

$$K(i^{(n)}) = − \log \mathbb{P}(i^{(n)}) + C .$$

(20)

The universality of $\mathbb{P}$ relies upon the fact that, for any semi-computable semi-measure $μ$ there exists a constant $C_μ > 0$ dependent on $μ$ only such that

$$μ(i^{(n)}) ≤ C_μ \mathbb{P}(i^{(n)}) .$$

(21)
Individual and statistical complexity are related by the following two results: for any semi-computable probability \( \pi^{(n)} = \{p(i^{(n)})\} \), the Shannon entropy is closed to the average Kolmogorov complexity:

\[
H(\pi^{(n)}) = -\sum_{i^{(n)} \in \Omega_{i}^{n}} p(i^{(n)}) \log p(i^{(n)}) \\
\approx \sum_{i^{(n)} \in \Omega_{i}^{n}} p(i^{(n)}) K(i^{(n)}). \tag{22}
\]

Furthermore, if the source is ergodic, in the limit of binary strings going to sequences the complexity per symbol equals the source entropy rate, for almost all of them:

\[
k(i) := \lim_{n \to +\infty} \frac{1}{n} K(i^{(n)}) = h(\pi) = \lim_{n \to +\infty} \frac{1}{n} H(\pi^{(n)}). \tag{23}
\]

4. GACS QUANTUM COMPLEXITY

Differently from other approaches that rely upon classical or quantum Turing machines, in [5] a quantum complexity is constructed by generalizing the notion of universal probability to that of universal semi-density matrix.

In order to do so, given a quantum system with Hilbert space \( \mathbb{H} \), one relies on elementary vectors \( |\Psi\rangle \) which are those with computable coefficients when expanded with respect to the basis obtained by tensor product of the vectors in (1). It is thus possible to describe them by means of a binary string of length \( n \): \( |\Psi\rangle \leftrightarrow i_{\Psi} \in \Omega_{i}^{n} \) with associated universal probability \( \mathbb{P}(i_{\Psi}) \). The universal semi-density matrix is then defined by a non-normalized convex combination of projectors onto all elementary vector states:

\[
\mathbb{D} = \sum_{\Psi} \mathbb{P}(i_{\Psi}) |\Psi\rangle\langle \Psi|. \tag{24}
\]

It is universal since, for any non-normalized density matrix \( \rho \in \mathbb{M} \) whose entries are semi-computable, there exists a constant \( C_{\rho} \geq 0 \) depending only on \( \rho \) such that

\[
\rho \leq C_{\rho} \mathbb{D}. \tag{25}
\]

Then, one defines the Gacs operator algorithmic complexity, Gacs complexity for short, as

\[
\kappa_{q} = -\log \mathbb{D}. \tag{26}
\]

It turns out that the von Neumann entropy of any semi-computable density matrix is close to the average Gacs complexity [5]:

\[
S(\rho) \simeq \text{Tr}(\rho \kappa_{q}^{(n)}). \tag{27}
\]

4.1. Gacs Complexity and Alicki-Fannes Entropy

In order to establish a relation between the Alicki-Fannes entropy and the Gacs complexity, it is necessary to examine closely the auxiliary matrix \( \rho^{(n)} \) introduced in (15).

Let us first associate to the state on the quantum chain segment of length \( n \), \( M_{d^{n}} \ni \rho^{(n)} = \sum_{i} r_{i}^{(n)} \langle r_{i}^{(n)} | \rangle | r_{i}^{(n)} \rangle \langle r_{i}^{(n)} | \)

the vector state

\[
\sqrt{\rho^{(n)}} = \sum_{i} \sqrt{r_{i}^{(n)}} | r_{i}^{(n)} \rangle \otimes | r_{i}^{(n)} \rangle \tag{28}
\]
in the doubled Hilbert space \( \mathbb{C}^{2n} \otimes \mathbb{C}^{2n} \). Then, we associate to the auxiliary matrix \( \xi^{(n)} \) in \( M_{2^{2n}} \) the vector state

\[
|\xi^{(n)}\rangle = \sum_{i} \sum_{\ell} \sqrt{r_{i}^{(n)}} u_{\ell}^{(n)} \langle r_{i}^{(n)} | \ell \rangle \otimes | r_{i}^{(n)} \rangle \otimes | (k^{(n)})^{(n)} \rangle \tag{29}
\]
in the Hilbert space \( \mathbb{C}^{2n} \otimes \mathbb{C}^{2n} \otimes \mathbb{C}^{2^{2n}} \), where \( |(k^{(n)})^{(n)}\rangle \) enumerate the standard basis vectors in \( \mathbb{C}^{2^{2n}} \).

Via this double purification, by partial tracing out, the first two Hilbert spaces first, and then the third one, one gets the marginal states

\[
\text{Tr}_{I,II} \left( |\xi^{(n)}\rangle\langle \xi^{(n)} | \right) = \rho^{(n)} \tag{30}
\]

\[
\text{Tr}_{III} \left( |\xi^{(n)}\rangle\langle \xi^{(n)} | \right) = \sum_{i,j} \sqrt{r_{i}^{(n)} r_{j}^{(n)}} \times
\]

\[
\times \sum_{(k^{(n)})^{(n)}} u_{\ell}^{(n)} | r_{i}^{(n)} \rangle \langle r_{j}^{(n)} | \ell \rangle \otimes | r_{i}^{(n)} \rangle \langle r_{j}^{(n)} | \right) \tag{31}
\]

The previous two steps have first purified the density matrix for the initial length \( n \) segment of the quantum chain and then coupled it to an ancilla system described by a Hilbert space of dimension given by the number of elements in the partition of unit \( U^{(n)} \). As the marginal density matrices of a pure state have the same entropy, it turns out that

\[
S(\rho^{(n)}) = S(\xi^{(n)}) = S(\rho^{(n)}) + n \log d \tag{32}
\]

If we now consider the third system and take into account that the matrix units provide a computable partition and thus a computable density matrix \( R[U^{(n)}] \), relation (27) then states that the average Gacs complexity is close to the von Neumann entropy. Thus,

\[
\text{Tr} \left( R[U^{(n)}] \kappa_{q}^{(n)} \right) \simeq S(\rho^{(n)}) = S(\rho^{(n)}) + n \log d. \tag{33}
\]

5. CONCLUSION

While for classical dynamical systems there are essentially one dynamical entropy and one algorithmic complexity closely connected to each other, quantum mechanics offers different inequivalent options. We have considered the Alicki-Fannes entropy for a stationary quantum chain as an instance of quantum dynamical entropy; it is based on a tomographic reconstruction of the density matrix on longer and longer segments of the chain and gives an extra term with respect to the entropy rate of the chain. We have showed that this extra term can be related to the Gacs quantum complexity of the auxiliary system which provides the quantum tomography.

6. REFERENCES


1. MOTIVATION

Papers compiled in Good (1983) made first attempts at combining attractive aspects of Bayesian and frequentist approaches to statistical inference. While the hybrid inference approach of Yuan (2009) succeeded in leveraging Bayesian point estimators with maximum likelihood estimates, hybrid inference does not yet cover the case of a parameter of interest that has a partially known prior. Since such partial knowledge of a prior occurs in many scientific inference situations, it calls for a theoretical framework for method development that appropriately blends Bayesian and frequentist methods by meeting these criteria:

1. **Complete knowledge of the prior.** If the prior is known, the corresponding posterior is used for inference. Among statisticians, this principle is almost universally acknowledged. However, it is rarely the case of the prior is essentially known.

2. **Negligible knowledge of the prior.** If there is no reliable knowledge of a prior, inference is based on methods that do not require such knowledge. This principle motivates not only the development of confidence intervals and p-values but also Bayesian posteriors derived from improper and data-dependent priors. Accordingly, blended inference must allow the use of such methods when applicable.

3. **Continuum between extremes.** Inference relies on the prior to the extent that it is known while relying on the other methods to the extent that it is not known. Thus, there is a gradation of methodology between the above two extremes. This intermediate scenario calls for a careful balance between pure Bayesian methods on one hand and impure Bayesian or non-Bayesian methods on the other hand.

Instead of framing the knowledge of a prior in terms of confidence intervals, as in pure empirical Bayes approaches, the full version of this extended abstract (Bickel, 2011a) frames it more generally in terms of a set of plausible priors, as in interval probability (Weichselberger, 2000; Augustin, 2002, 2004) and robust Bayesian (Berger, 1984) approaches. Whereas the concept of an unknown prior cannot arise in strict Bayesian statistics, it does arise in robust Bayesian statistics when the levels of belief of an intelligent agent have not been fully assessed (Berger, 1984). Unknown priors also occur in many more objective contexts involving purely frequentist interpretations of probability in terms of variability in the observable world rather than the uncertainty.
in the mind of an agent. For example, frequency-based priors are routinely estimated under random effects and empirical Bayes models; see, e.g., Efron (2010). (Bickel (2011a) comments further on interpretations of probability and relaxes the assumption of a true prior.)

With respect to the problem at hand, the most relevant robust Bayesian approaches are the \textit{minimax Bayes risk} (“\textit{Γ}-minimax”) practice of minimizing the maximum Bayes risk (Robbins, 1951; Berger, 1985; Vidakovic, 2000) and the \textit{maxmin expected utility} (“conditional \textit{Γ}-minimax”) practice of maximizing the minimum posterior expected payoff or, equivalently, minimizing the maximum posterior expected loss (Gilboa and Schmeidler, 1989; DasGupta and Studden, 1989; Vidakovic, 2000; Augustin, 2002, 2004). Augustin (2004) reviews both methods in terms of interval probabilities that need not be subjective. With typical loss functions, the former method meets the above criteria for classical minimax alternatives to Bayesian methods but does not apply to other attractive alternatives. For example, several confidence intervals, \textit{p}-values, and objective-Bayes posteriors routinely used in biostatistics are not minimax optimal. (Fraser and Reid (1990) and Fraser (2004) argued that requiring the optimality of frequentist procedures can lead to trade-offs between hypothetical samples that potentially mislead scientists or yield pathological procedures.) Optimality in the classical sense is not required of the alternative procedures under the framework outlined below, which can be understood in terms of maxmin expected utility with a payoff function that incorporates the alternative procedures to be used as a benchmark for the Bayesian posteriors.

\section{2. HEURISTIC OVERVIEW}

To define a general theory of blended inference that meets a formal statement of the three criteria, Bickel (2011a) introduced a variation of a zero-sum game of Topsøe (1979), Harremoës and Topsøe (2001), and Topsøe (2007). (The discrete version of the game also appeared in Pfaffelhuber (1977); Grünwald and Philip Dawid (2004) interpreted it as the codelength special case of the maxmin expected utility problem. Alternative versions of minimax optimal codelength, including recent generalizations of normalized maximum likelihood (NML) (e.g., Rissanen and Roos, 2007; Bickel, 2011b), have also been applied to statistical inference.) The “nature” opponent selects a prior consistent with the available knowledge as the “statistician” player selects a posterior distribution with the aim of maximizing the minimum information gained relative to one or more alternative methods. Such benchmark methods may be confidence interval procedures, frequentist hypothesis tests, or other techniques that are not necessarily Bayesian.

From that game, Bickel (2011a) derived a widely applicable framework for testing hypotheses. For concreteness, the motivating results are heuristically summarized here. Consider the problem of testing \( H_0 : \theta = 0 \), the hypothesis that a real-valued parameter \( \theta \) of interest is equal to the point 0 on the real line \( \mathbb{R} \). The observed data vector \( x \) is modeled as a realization of a random variable denoted by \( X \). Let \( p(x) \) denote the \textit{p}-value resulting from a statistical test.

The \textit{p}-value for a simple (point) null hypothesis is often smaller than Bayesian posterior probabilities of the hypothesis (Lindley, 1957; Berger and Sellke, 1987). Suppose \( \theta^\ast \) has an unknown prior distribution according to which the prior probability of \( H_0 \) is \( \pi_0 \). While \( \pi_0 \) is unknown, it is assumed to be no less than some known lower bound denoted by \( \pi_{0,\ast} \).

Following the methodology of Berger et al. (1994), Sellke et al. (2001) found a generally applicable lower bound on the Bayes factor. As Bickel (2011a) explains, that bound immediately leads to

\[
\Pr (H_0|p(X) = p(x)) = \left(1 - \left(\frac{1 - \pi_{0,\ast}}{\pi_{0,\ast} ep(x) \log p(x)}\right)^{-1}\right)^{-1}
\]

as a lower bound on the posterior probability of the null hypothesis for \( p(x) < 1/e \) and to \( \pi_{0,\ast} \) as a lower bound on the probability if \( p(x) \geq 1/e \).

In addition to \( \Pr (H_0|p(X) = p(x)) \), the unknown Bayesian posterior probability of \( H_0 \), there is a frequentist posterior probability of \( H_0 \) that will guide selection of a posterior probability for
inference based on $\pi_0 \geq \pi_0$ and other constraints summarized by
\[ \Pr(H_0|p(X) = p(x)) \geq \Pr(H_0|p(X) = p(x)). \]
While it is incorrect to interpret the p-value $p(x)$ as a Bayesian posterior probability, it is seen in Bickel (2011a) that $p(x)$ is a confidence posterior probability that $H_0$ is true.

With the confidence posterior as the benchmark, the solution to the optimization problem described above gives the blended posterior probability that the null hypothesis is true. It is simply the maximum of the p-value and the lower bound on the Bayesian posterior probability:
\[ \Pr(H_0; p(x)) = p(x) \lor \Pr(H_0|p(X) = p(x)). \]

By plotting $\Pr(H_0; p(x))$ as a function of $p(x)$ and $\pi_0$, figures in Bickel (2011a) illustrate each of the above criteria for blended inference:

1. **Complete knowledge of the prior.** In this example, the prior is only known when $\pi_0 = 1$, in which case
\[ \Pr(H_0; p(x)) = \Pr(H_0|p(X) = p(x)) = 1 \]
for all $p(x)$. Thus, the p-value is ignored in the presence of a known prior.

2. **Negligible knowledge of the prior.** There is no knowledge of the prior when $\pi_0 = 0$ and negligible knowledge when $\pi_0$ is so low that $\Pr(H_0|p(X) = p(x)) \leq p(x)$. In such cases, $\Pr(H_0; p(x)) = p(x)$, and the Bayesian posteriors are ignored.

3. **Continuum between extremes.** When $\pi_0$ is of intermediate value in the sense that $\Pr(H_0|p(X) = p(x))$ is exclusively between $p(x)$ and 1,
\[ \Pr(H_0; p(x)) = \Pr(H_0|p(X) = p(x)) < 1. \]
Consequently, $\Pr(H_0; p(x))$ increases gradually from $p(x)$ to 1 as $\pi_0$ increases (Bickel, 2011a). In this case, the blended posterior lies in the set of allowed Bayesian posteriors but is on the boundary of that set that is closest to the p-value. Thus, both the p-value and the Bayesian posteriors influence the blended posterior and thus the inferences made on its basis.

The plotted parameter distribution is presented in Bickel (2011a) as a widely applicable blended posterior.

Bickel (2011a) offers additional details and generalizations.

3. REFERENCES


MIXABILITY IS BAYES RISK CURVATURE RELATIVE TO LOG LOSS

Tim van Erven1, Mark D. Reid2 and Robert C. Williamson2

1Centrum Wiskunde & Informatica (CWI)
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands, E-mail: tim@timvanerven.nl
2ANU and NICTA
Canberra ACT 0200, Australia, E-mail: {Mark.Reid, Bob.Williamson}@anu.edu.au

ABSTRACT

Given $K$ codes, a standard result from source coding tells us how to design a single universal code with codelengths within $\log(K)$ bits of the best code, on any data sequence. Translated to the online learning setting of prediction with expert advice, this result implies that for logarithmic loss one can guarantee constant regret, which does not grow with the number of outcomes that need to be predicted. In this setting, it is known for which other losses the same guarantee can be given: these are the losses that are mixable.

We show that among the mixable losses, log loss is special: in fact, one may understand the class of mixable losses as those that behave like log loss in an essential way. More specifically, a loss is mixable if and only if the curvature of its Bayes risk is at least as large as the curvature of the Bayes risk for log loss (for which the Bayes risk equals the entropy).

1. INTRODUCTION

For $n \in \mathbb{N}$, let $\mathcal{Y} = \{1, \ldots, n\}$ be the outcome space. We will consider a prediction game where the loss of the learner making predictions $v_1, v_2, \ldots \in \mathcal{Y}$ is measured by a loss function $\ell: \mathcal{Y} \times \mathcal{Y} \to [0, \infty]$ cumulatively: for $T \in \mathbb{N}$,

$$\text{Loss}(T) := \sum_{t=1}^{T} \ell(y_t, v_t),$$

where $y_1, y_2, \ldots \in \mathcal{Y}$ are outcomes. A loss $\ell$ is called $\eta$-mixable if for every distribution $P$ on actions $\mathcal{Y}$ there exists a single action $v_P$ such that

$$\ell(v_P, y) \leq \frac{1}{\eta} \log \mathbb{E}_{v \sim P} \left[ e^{-\eta \ell(v, y)} \right]$$

for all $y \in \mathcal{Y}$.

A loss is called mixable if there exists any $\eta > 0$ such that it is $\eta$-mixable.

The learner has access to predictions $v^t_i, t = 1, 2, \ldots, i \in \{1, \ldots, N\}$ generated by $N$ experts $\mathcal{E}_1, \ldots, \mathcal{E}_N$ that attempt to predict the same sequence. The goal of the learner is to predict nearly as well as the best expert. A strategy for the learner, called a merging strategy, is a function $\mathcal{M}: \bigcup_{t=1}^{\infty} (\mathcal{Y}^{t-1} \times (\mathcal{Y}^N)^t) \to \mathcal{Y}$, which takes the outcomes $y_1, \ldots, y_{t-1}$ and predictions $v^s_i$, $i = 1, \ldots, N$ for times $s = 1, \ldots, t$ and outputs an aggregated prediction $v^M_i$, incurring loss $\ell(y_t, v^M_i)$ when $y_t$ is revealed. After $T$ rounds, the loss of $\mathcal{M}$ is $\text{Loss}_{\mathcal{M}}(T) = \sum_{t=1}^{T} \ell(y_t, v^M_i)$ and the loss of expert $\mathcal{E}_i$ is $\text{Loss}_{\mathcal{E}_i}(T) = \sum_{t=1}^{T} \ell(y_t, v^i_t)$. When $\mathcal{M}$ is the aggregating algorithm (which can be used for all losses considered in this paper) $[2]$, $\eta$-mixability implies for all $t \in \mathbb{N}$, all $i \in \{1, \ldots, N\}$,

$$\text{Loss}_{\mathcal{M}}(t) \leq \text{Loss}_{\mathcal{E}_i}(t) + \frac{\ln N}{\eta}. \quad (1)$$

Conversely, if the loss function $\ell$ is not mixable, then it is not possible to predict as well as the best expert up to an additive constant using any merging strategy.

Thus determining $\eta_i$ (the largest $\eta$ such that $\ell$ is $\eta$-mixable) is equivalent to precisely bounding the prediction error of the aggregating algorithm. The mixability of several binary losses and the Brier score in the multiclass case [3] is known. However a general characterisation of $\eta_i$ in terms of other key properties of the loss has been missing. We show how $\eta_i$ depends upon the curvature of the conditional Bayes risk for $\ell$ when $\ell$ is a strictly proper multiclass loss.

2. PROPER MULTICLASS LOSSES

Let $\Delta^n := \{(x_1, \ldots, x_n) : x_i \geq 0, \sum_{i=1}^{n} x_i = 1\}$ denote the $n$-simplex, which is the set of all probability vectors on $n$ outcomes. We consider multiclass losses for class probability estimation, where $\mathcal{Y} = \{1, \ldots, n\}$ is the set of possible classes. A loss function $\ell: \Delta^n \to [0, \infty]^n$ assigns a loss vector $\ell(q) = (\ell_1(q), \ldots, \ell_n(q))$ to each distribution $q \in \Delta^n$ where $\ell_i(q) = (\ell(i, q)$ traditionally) is the penalty for predicting $q$ when outcome $i \in \mathcal{Y}$ occurs. If the outcomes are distributed with probability $p \in \Delta^n$ then the risk for predicting $q$ is just the expected loss

$$L(p, q) := \sum_{i=1}^{n} p_i \ell_i(q).$$

---

These results have previously appeared in the COLT 2011 proceedings [1]. More details can be found there.
The Bayes risk for $p$ is the minimal achievable risk for that outcome distribution,

$$L(p) := \inf_{q \in \Delta^n} L(p, q).$$

We say that a loss is proper whenever the minimal risk is always achieved by predicting the true outcome distribution, that is, $L(p) = L(p, p)$ for all $p \in \Delta^n$. We say a proper loss is strictly proper if there exists no $q \neq p$ such that $L(p, q) = L(p)$. The log loss $\ell_{\log}(p) := (-\ln(p_1), \ldots, -\ln(p_n))^T$ is strictly proper. Its corresponding Bayes risk is $L_{\log}(p) = -\sum_{i=1}^n p_i \ln(p_i)$, which is the entropy of $p$.

3. THE BINARY CASE

Consider first the binary case, where $n = 2$. Then for continuous, twice differentiable losses $\ell$ it is known [4] that

$$\eta_{\ell} = \min_{p \in (0, 1)} \frac{\ell'_{\ell}(p)\ell''_{\ell}(p) - \ell'_1(\ell)\ell'_2(\ell)}{p(1-p)w(p)},$$

(2)

When a binary loss $\ell$ is differentiable, properness implies the stationarity condition [5]

$$p\ell'(p) + (1-p)\ell'_2(p) = 0,$$

from which it follows that

$$\frac{\ell'_1(p)}{p} = \frac{\ell'_2(p)}{p} = w(p) =: w_{\ell}(p),$$

where $w$ or $w_{\ell}$ is called the weight function [5]. By differentiating twice, one also finds that $L''(p) = -w(p)$. Substituting these expressions into (2) and simplifying, one finds that many factors cancel, leading to

$$\eta_{\ell} = \min_{p \in (0, 1)} \frac{1}{p(1-p)w(p)}.$$

Observing further that $L''_{\log}(p) = \frac{1}{p(1-p)}$ and so $w_{\log}(p) = \frac{1}{p(1-p)}$, we obtain the simple expression

$$\eta_{\ell} = \min_{p \in (0, 1)} \frac{w_{\log}(p)}{w_{\ell}(p)} = \min_{p \in (0, 1)} \frac{L''_{\log}(p)}{L''(p)}.$$   

(3)

That is, the mixability constant of binary proper losses is the minimal ratio of the weight functions for log loss and the loss in question. In the next section we will show how (3) generalises to the multiclass case ($n > 2$). That there is a relationship between Bayes risk and mixability was also pointed out (in a less explicit form) by Kalnishkan, Vovk and V’yugin [6].

4. THE MULTICLASS CASE

Because probabilities sum up to one, any $p \in \Delta^n$ is fully determined by its first $n-1$ components $\hat{p} = (p_1, \ldots, p_{n-1})$. Let $\hat{\Delta}^n = \{\hat{p} : \hat{p} \in \Delta^n\}$ be the set of such $(n-1)$-dimensional vectors. We have been implicit about this in the previous section, but for the derivatives of $L$ to make sense in the multiclass case, we need to define it as a function of $\hat{p}$ rather than the full vector $p$:

$$L(\hat{p}) = \sum_{i=1}^{n-1} p_i \ell_i(\hat{p}) + \left(1 - \sum_{i=1}^{n-1} p_i\right) \ell_n(\hat{p}).$$

Let $HL(\hat{p})$ denote the Hessian of $L(\hat{p})$ and for any matrix $A$ let $\lambda_{\max} A$ denotes its maximum eigenvalue. Then in the multiclass case we obtain the following generalisation of (3):

**Theorem 1.** Suppose a loss $\ell$ satisfies Condition 1. Then its mixability constant is

$$\eta_{\ell} = \min_{\hat{p} \in \hat{\Delta}^n} \lambda_{\max} \left(\left(HL(\hat{p})\right)^{-1} \cdot H_{\log}(\hat{p})\right).$$

(4)

The condition we require is as follows:

**Condition 1.** The loss $\ell$ is strictly proper, continuous on $\Delta^n$, and continuously differentiable on the relative interior $\text{relint}(\Delta^n)$ of its domain.

5. CONCLUSION

Under Condition 1, we have shown that mixability of a loss is determined by whether the curvature of its Bayes risk is as least as large as the curvature of the Bayes risk for log loss.

6. REFERENCES


MULTIPLE LDPC DECODING FOR DISTRIBUTED SOURCE AND VIDEO CODING

Søren Forchhammer, Huynh Van Luong and Xin Huang

DTU Fotonik, Technical University of Denmark, Building 343, Lyngby 2800, Denmark
Email: {sofo, hulu, xhua}@fotonik.dtu.dk

ABSTRACT
Distributed source coding (DSC) is a coding paradigm for systems which fully or partly exploit the source statistics at the decoder to reduce the computational burden at the encoder. Distributed video coding (DVC) is one example. This paper considers the use of Low Density Parity Check Accumulate (LDPCA) codes in a DSC scheme with feed-back. To improve the LDPC coding performance in the context of DSC and DVC, while retaining short encoder blocks, this paper proposes multiple parallel LDPC decoding. The proposed scheme passes soft information between decoders to enhance performance. Experimental results on DVC show that the LDPCA performance implies a loss compared to the conditional entropy, but also that the proposed scheme reduces the DVC bit rate up to 3.9% and improves the rate-distortion (RD) performance of a Transform Domain Wyner-Ziv (TDWZ) video codec.

1. INTRODUCTION
Distributed source coding as e.g. distributed video coding [1] proposes to fully or partly exploit the redundancy at the decoder, rather than at the encoder. The Slepian-Wolf theorem [2] states, it is possible to achieve the same rate by independently encoding but jointly decoding two statistically dependent signals as for typical joint encoding and decoding (with a vanishing error probability). The Wyner-Ziv theorem [3] extends the Slepian-Wolf theorem to the lossy case, becoming the theoretical basis for DSC, where source data X are lossy coded and decoded based on a correlated source Y at the decoder.

The source data X may be predicted using the side information Y at the decoder and thereafter the prediction errors may be corrected using an error-correcting code. The coding efficiency of the error correcting code, an LDPC Accumulate (LDPCA) codec [8] in this paper, plays a key role in distributed source coding. The scheme we consider utilizes feed-back from the decoder to the encoder. To improve the performance, a Wyner-Ziv decoder with multiple LDPCA decoders is proposed in this work. The proposed scheme is inspired by the work in [9] using joint bitplane LDPC decoding. Different from [9], the proposed Wyner-Ziv codec utilizes multiple LDPCA decoders in parallel and passes soft information between decoders. The modifications only involve the buffer part and the decoder, while the LDPCA encoder is not changed. The objective is to increase performance by modifying the decoder, while using the same (short) encoding blocks for low-complexity and to allow for fairly fine granularity for adaptive updating of the decoder estimates.

2. DISTRIBUTED SOURCE CODING
Based on work on distributed video coding, we shall outline one approach to distributed source coding, which codes the data X given the side information Y. The distributed video codec (TDWZ codec) will be described in Section 4. Here we note that the problem is lossy coding of coefficients of the source based on side information (key frames in DVC). The coefficients are quantized and thereafter they are decomposed into bitplanes, which are fed to a rate-compatible LDPCA encoder [8] starting from the most significant bitplane (MSB) to least significant bitplane (LSB). For each encoded bitplane, the corresponding accumulated syndrome is stored in a buffer together with an 8-bit Cyclic Redundancy Check (CRC).

The decoder requests bits through a feedback channel as shown in Figure 1. We shall use the terms frames, bands, coefficients and bit-planes from DVC, where coefficients just refer to the (possibly transformed) values we want to code and bitplanes refer to any collection of source bits of the same significance as, e.g. MSB and LSB from a given set of coefficients. A band is a set of coefficients, e.g. a frequency band and finally a frame is a set of bands forming an instance of X, e.g. an image frame.

At the decoder, the side information Y is used to predict the value of X and a corresponding noise residue, which expresses the conditional probabilities (Pr) fed to the LDPC decoder for each bitplane. In our DSC scheme, we predict the coefficient values and the residual error may be modelled by a Laplacian distribution. Thereafter the LDPCA decoder starts to decode the various bitplanes, ordered from MSB to LSB, to correct the bit errors [4]. After all the bitplanes are successfully decoded, the Wyner-Ziv frame can be decoded.

For the LDPCA decoding, a Belief-Propagation (BP) algorithm is used to retrieve each transmitted bitplane. The BP algorithm is a soft-decoding approach, which passes a Log-Likelihood Ratio (LLR) of Pr back and forth between source nodes and the syndrome nodes. Let \(X=(b_{m-1},...,b_t,b_0)\) denote a quantized coefficient of a Wyner-Ziv frame, where \(b_{m-1}\) is an MSB bit and \(b_0\) is an LSB bit and \(Y\) denotes a quantized coefficient of the side...
information. The LLR of a bit \( b_i \) \((0 \leq i \leq m-1)\) of the \( i \)th significant bitplane is described as:

\[
L(b_i) = \log \left( \frac{Pr(b_{i}=0|Y,b_{i+1},...,b_{m-1})}{Pr(b_{i}=1|Y,b_{i+1},...,b_{m-1})} \right)
\]  

(1)

where \( b_{i+1},...,b_{m-1} \) represent bits from previous successfully decoded bits of the transformed coefficient. The decoder utilizes information from previous successfully decoded bitplanes to calculate soft information for the future bitplanes.

As one approach to improve the performance of the LDPCA codec, a decoder may iteratively exchange information between the decoding processes of the bitplanes and refine the soft-input for each bitplane during the decoding process. Thus, a Wyner-Ziv codec with multiple LDPCA decoders is proposed. The multiple LDPCA decoders are running in parallel to keep refining the soft-input in each iteration. Each LDPCA decoder operates on the syndromes for one bitplane, but the correlation between bitplanes is exploited by passing beliefs from one bitplane to another. Once a bitplane is successfully decoded, the corresponding LDPCA decoder no longer requests syndrome bits from the buffer and the rest of the LDPCA decoders are reinitialized.

The proposed Wyner-Ziv codec using multiple LDPCA decoders is depicted in Figures 1 and 2. Soft information is exchanged between the LDPCA decoders using the so-called bitplane correlation model to reform soft-input based on feedback from the LDPCA decoders and the estimated noise distribution from the noise model. The new soft-input information of the source \( X \) is estimated and updated, expressing \( X-Y \) using the Laplacian parameter calculated by the noise model.

The main difference between this approach and [4] is that the LLR of a bit \( b_i \) \((0 \leq i \leq m-1)\) of the \( i \)th significant bitplane is computed conditional on the binary distributions \( \beta_k \) of bits of the other bitplanes, \( b_k \) \((k \neq i)\). This means that the LLR is calculated by using soft information from the other bit-planes. Let \( \beta_k \) denote a probability of bitplane \( k \). The LLR described in (1) is here generalized for a bit \( b_i \) of bitplane \( i \) as:

\[
L(b_i) = \log \left( \frac{Pr(b_i=0|Y,b_{i+1},...,b_{m-1},\beta_1,...,\beta_{i-1},\beta_{i+1},...,\beta_{m-1})}{Pr(b_i=1|Y,b_{i+1},...,b_{m-1},\beta_1,...,\beta_{i-1},\beta_{i+1},...,\beta_{m-1})} \right)
\]  

(2)

where \( \beta_k \) are soft values for the same coefficient as \( b_k \).

The method involves both bitplane (bit) and coefficient (symbol) levels to update soft side information via one BP algorithm. Similar to [9], the key idea is to use the BP mechanism during the decoding of a frame and to convert the LLR back and forth between symbol level and bit level. Distinctly, in the proposed method, the soft-input is only updated after the multiple LDPCA decoders of one coefficient band are completely processed (using a certain number of iterations) at bit level based on the given syndrome bits. Let \( \Pr^{t-1}(b_i) \) denote the probability of bit \( b_i \) at the iteration \( t-1 \) at bit level. The LLR of bit \( b_i \) is updated for iteration \( t \) as an approximation of (2):

\[
L^{\text{t}}(b_i) = \log \left( \frac{\sum_{X \in \mathcal{X}} \Pr(X|Y) \prod_{k \neq i} \Pr^{t-1}(b_k)}{\sum_{X \in \mathcal{X}} \Pr(X|Y) \prod_{k \neq i} \Pr^{t-1}(b_k)} \right)
\]  

(3)

where \( X=(b_0,...,b_{i-1},b_i,b_{i+1},...,b_{m-1}) \) and \( S_t=[X=S;b_i=0], S_f=[X=S;b_i=1] \). \( \Pr(X|Y) \) is calculated at coefficient level by using the updated noise distribution between the side information coefficient and the original Wyner-Ziv coefficient via the noise model as shown in Figure 1.

The LLRs at iteration \( t \) noted by \( L^{\text{t}}(b_i) \), are in turn input to multiple LDPCA decoders. After one LDPCA is processed, \( L^{\text{t}}(b_i) \) is temporarily achieved as output. The updated \( \Pr^{t}(b_i) \) values are obtained based on the LLR definition:

\[
\Pr^{t}(b_i=0) = \frac{\Pr^{t-1}(b_i=0)}{1-\Pr^{t-1}(b_i=0)}
\]  

(4)

i.e. for the next iteration, we have:

\[
\Pr^{t}(b_i=0) = \frac{1}{2} \left( 1 + \tanh \left( \frac{L^{\text{t}}(b_i)}{2} \right) \right)
\]  

(5)

This \( \Pr^{t}(b_i) \) is used as a new probability of bit \( b_i \) to compute new LLRs, \( L^{\text{t+1}}(b_i) \), for the next iteration of multiple LDPCA decoding based on (3). Since all LDPCA decoders are running in parallel, once a bitplane is successfully decoded, the re-initialization procedure is performed. The new soft-inputs for the rest of the bit-planes are assigned conditional on the successfully decoded bitplane. Once a LDPCA decoder has successfully decoded a bitplane, it will no longer request syndromes from the buffer. Assume \( b_i \) is successfully decoded with value \( 0 \), then \( \Pr^{t}(b_i=0)=1 \) and the iteration count is reset as \( t=0 \). In addition, the remaining unfinished bitplanes are re-initialized by \( \Pr^{t}(b_i=0)=1/2 \). The updated \( L^{\text{t+1}}(b_i) \) and \( \Pr^{t}(b_i) \) are iteratively operated up to a maximum numbers of iterations \( (T_{\text{max}}) \) with the given syndrome bits. If they are not successful after this number of iterations, the LDPCA decoders request more syndrome bits from the buffer via
the feedback channel. Then a new process is started until all the bitplanes of the current set of coefficients are successfully decoded. Let $N_{\text{max}}$ denote a maximum numbers of syndromes.

Overall, the multiple LDPCA decoding is handled as follows:

1. **Initiate parameters.** Iteration count $t=0$; Number of syndrome bits $n=0$; For all bits $h_i$, $Pr^{(0)}(b_i=0)=1/2$.
2. **Increase and check conditions.**
   a. **Syndrome bit condition:** Increase $n=n+1$. If $n>N_{\text{max}}$, then end, else go to Step 2.b.
   b. **Iteration count condition:** Increase $t=t+1$. If $t<T_{\text{max}}$, go to Step 3, else return to 2.a.
3. **Compute the LLRs.** At bit level, (3) is computed to get the LLRs, $L^{(t)}(b_i)$.
4. **Check if any LDPCA is successfully decoded?**
   a. **No:** Compute probabilities of bitplanes. $L^{(t)}(b_i)$ are forwarded to multiple LDPCA decoders where $L^{(t)}(b_i)$ are received from LDPCA outputs. New probabilities of bitplanes, $Pr^{(t)}(b_i)$, by (5).
   b. **Yes:** Re-initialize the process. Assume LDPCA $(b_i)$ is successfully decoded with value $b_i=0$, assign $Pr^{(0)}(b_i=0)=1$. Reset iteration count $t=0$ and the remaining unfinished LDPCA decoders by $Pr^{(0)}(b_i=0)=1/2$.
5. **Check all LDPCA decoders.** The process is ended if all bitplanes are successfully decoded, otherwise, go to Step 2.b.

The procedure above is repeated for all bands of coefficients for which Wyner-Ziv bits are transmitted. In some cases, the length of the required syndromes consumed for the LSB is (close to) 1 bit per symbol, even though there is still some correlation. This is due to a (relative) loss in the LDPCA decoder. This may be reduced by first coding the LSB independently, as the entropy of the LSB often is close to 1 bit/symbol, and then apply multiple decoding to the remaining bitplanes after decoding the LSB and updating the soft information for the remaining bit-planes.

![Figure 2: Architecture of feedback channel based Transform Domain Wyner-Ziv video codec](image)

**4. STATE-OF-THE-ART TRANSFORM DOMAIN WYNER-ZIV VIDEO CODING**

Transform Domain Wyner-Ziv (TDWZ) video coding is a popular approach to DVC [1]. It has been improved by e.g. advanced side information generation schemes [5], finer noise models [4][5] and refinement schemes [7]. Despite the advances in practical TDWZ video coding, the RD performance still trails the performance of conventional video coding, such as H.264/AVC. The architecture of a TDWZ video codec is depicted in Figure 2. It basically follows the same architecture as the DISCOVER one [4]. However, a better side information generation scheme [5] and an improved noise model [6] are adopted. At the encoder, periodically one frame out of $N$ in the video sequence is named as key frame and intermediate frames are WZ frames. The key frames are intra coded by using low complexity video coding as H.264/AVC Intra, while the WZ frames in between are coded with a Wyner-Ziv approach. WZ frames are transformed using a 4x4 block size and the transformed coefficients within the same frequency band are grouped together and then quantized. At the decoder, a side information frame is interpolated and the corresponding noise residue is generated by using previously decoded frames. The noise residue is modeled assuming a Laplacian distribution of $|x-y|$. The data are decoded using single or multiple LDPCA decoding as outlined. After all the bit-planes are successfully decoded, the Wyner-Ziv frame can be decoded through combined de-quantization and reconstruction followed by an inverse transform.

**5. PERFORMANCE EVALUATION**

In this section, the RD performance of the proposed approach [10] is presented and compared with the state-of-the-art TDWZ video codec described in Section 4 as well as relevant benchmarks. The test sequences are 149 frames of Foreman, Hall Monitor, Soccer, and Coast-guard (15Hz, QCIF). GOP (group of pictures) size is 2, where odd frames are coded as a key frame using H.246/AVC Intra and even frames are coded using Wyner-Ziv coding. Eight RD points ($Q$) are considered corresponding to eight 4x4 quantization matrices [4], which also determine the number of bitplanes, $m$, of each DCT coefficient band. The proposed model uses $m$ regular LDPC accumulate decoders [8], with a length of 1584 bits each, for the 1584 transform coefficients. The $m$ LDPCA each decodes one bitplane.

Table 1 shows rate and PSNR values of the proposed TDWZ codec with multiple LDPCA decoders (WZMD) as well as the savings in total rate, $\Delta R$ (in %), and WZ rate, $\Delta R_{WZ}$ (in %), compared with the state-of-the-art TDWZ codec [6]. The WZMD achieves a reduction of bit-rate for WZ frames up to 1.8% for Foreman; 2.59% for Hall Monitor; 2.26% for Soccer; 1.82% for Coast-guard.

In some cases, the length of the required syndromes for the LSB is (close to) 1 bit per symbol, even though there is still some correlation. This LDPCA decoder loss, which may be reduced by first coding the LSB independently and thereafter apply WZMD to the remaining bitplanes. This is called WZMD(LSB). Up to three LSB bitplanes may be sent first. Deviating from distributed encoding, the Ideal Code Length ICL may be interpreted as the number of bits required by a backward adaptive prediction video coding scheme applying ideal arithmetic coding to the calculated soft-input values, $Pr$, which the encoder can also calculate if it duplicates the processing of the decoder. The decision is based on thresholding the
ICL for the LSB. For 1-5 bitplanes the LSB is evaluated. For 6 and 7 bitplanes, 2 and 3 LSB bitplanes are evaluated, respectively. The thresholds applied are 0.89, 0.95 and 0.98.

As a result, the coding efficiency in terms of bit-rate is improved. Table 2 depicts the WZ bit rate savings for WZMD and WZMD(LSB) compared with TDWZ [6]. The results shows that WZ rate savings up to 3.9% for Foreman and 3.77% for Soccer. In a follow-up work [11] we have included inter bitplane correlation refinement in the loop of WZMD coding for additional performance.

The experimental results in Fig. 3 demonstrate that the proposed approach significantly improves overall RD performance compared with the DISCOVER codec, with PSNR gains up to about 0.7 dB for Foreman and 0.9 dB for Soccer. The performance of H.264/AVC (Intra), the H.264/AVC (No Motion), and ICL codecs are also included. The WZMD is more efficient than H.264/AVC (Intra) for Foreman.

6. CONCLUSION

This paper considers LDPCA for DSC and DVC and proposes a Wyner-Ziv codec using multiple parallel LDPC decoding by passing soft information between the bitplanes during the decoding process. Experimental results show that the proposed multiple LDPC decoding can improve the coding efficiency of DVC (TDWZ) in terms of WZ rate savings up to 3.9% compared with the corresponding single LDPC TDWZ [6] coding.

7. REFERENCES


Table 1. RD performance comparison

<table>
<thead>
<tr>
<th>Q</th>
<th>Foreman</th>
<th>Hall</th>
<th>Soccer</th>
<th>Coast-guard</th>
</tr>
</thead>
<tbody>
<tr>
<td>ΔR</td>
<td>WZ</td>
<td>WZMD(LSB)</td>
<td>WZ</td>
<td>WZMD(LSB)</td>
</tr>
<tr>
<td>1</td>
<td>20.85</td>
<td>25.74</td>
<td>28.65</td>
<td>1.32</td>
</tr>
<tr>
<td>2</td>
<td>0.78</td>
<td>1.51</td>
<td>3.60</td>
<td>1.11</td>
</tr>
<tr>
<td>3</td>
<td>0.79</td>
<td>0.99</td>
<td>2.41</td>
<td>1.38</td>
</tr>
<tr>
<td>4</td>
<td>0.68</td>
<td>0.88</td>
<td>1.88</td>
<td>1.88</td>
</tr>
<tr>
<td>5</td>
<td>0.78</td>
<td>0.80</td>
<td>1.80</td>
<td>1.80</td>
</tr>
<tr>
<td>6</td>
<td>0.82</td>
<td>0.95</td>
<td>2.26</td>
<td>1.35</td>
</tr>
<tr>
<td>7</td>
<td>0.71</td>
<td>0.86</td>
<td>1.99</td>
<td>1.14</td>
</tr>
<tr>
<td>8</td>
<td>0.66</td>
<td>0.79</td>
<td>1.62</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Table 2. Bit rate savings (in %) of WZMD and WZMD (LSB)

<table>
<thead>
<tr>
<th>Q</th>
<th>Foreman</th>
<th>Soccer</th>
</tr>
</thead>
<tbody>
<tr>
<td>WZMD</td>
<td>WZMD(LSB)</td>
<td>WZMD</td>
</tr>
<tr>
<td>ΔR</td>
<td>WZ</td>
<td>ΔR</td>
</tr>
<tr>
<td>1</td>
<td>0.49</td>
<td>1.32</td>
</tr>
<tr>
<td>2</td>
<td>0.62</td>
<td>1.51</td>
</tr>
<tr>
<td>3</td>
<td>0.53</td>
<td>1.30</td>
</tr>
<tr>
<td>4</td>
<td>0.68</td>
<td>0.88</td>
</tr>
<tr>
<td>5</td>
<td>0.78</td>
<td>0.80</td>
</tr>
<tr>
<td>6</td>
<td>0.82</td>
<td>0.95</td>
</tr>
<tr>
<td>7</td>
<td>0.71</td>
<td>0.86</td>
</tr>
<tr>
<td>8</td>
<td>0.66</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Figure 3. RD performance comparison

MINIMAL MARKOV MODELS

Jesús E. García and Verónica A. González-López

Department of statistic, State University of Campinas, Rua Sergio Buarque de Holanda,651 Cidade Universitária-Barão Geraldo CEP 13083-859 Campinas, SP, Brazil.

ABSTRACT

In this work we introduce a new and richer class of finite order Markov chain models and address the following model selection problem, find the Markov model with the minimal set of parameters (called here minimal Markov model) which is enough to represent a source as a Markov chain of finite order. Let us call $M$ the order of the chain and $A$ the finite alphabet, to determine the minimal Markov model, we define an equivalence relation on the state space $A^M$, such that all the sequences of size $M$ with the same transition probabilities are put in the same part. In this way we have one set of $(|A| - 1)$ transition probabilities for each part, obtaining a model with a minimal number of parameters. We show that the model can be selected consistently using the Bayesian information criterion.

1. INTRODUCTION

In this work we consider discrete stationary processes over a finite alphabet $A$ of size $|A|$. Markov chains of finite order are widely used to model stationary processes with finite memory. A limitation, in terms of point estimation, with full Markov chains models of finite order $M$ is that the number of parameters $(|A|^M(|A| - 1))$ grows exponentially with the order $M$. Another limitation is that the class of full Markov chains is not very rich, fixed the alphabet $A$ there is just one model for each order $M$ and in practical situations could be necessary a more flexible structure in terms of number of parameters. For an extensive discussion of those two limitations see [1]. A richer class of finite order Markov models introduced by [2] and [1] are the variable length Markov chain models (VLMC). In the VLMC class, each model is identified by a prefix tree $T$ called context tree. For a given model with a context tree $T$, the final number of parameters for the model is $|T|(|A| - 1)$ and depending on the tree, this produce a parsimonious model. In [3] is proved that the bayesian information criterion (BIC) can be used to consistently choose the VLMC model in an efficient way using the context tree maximization (CTM) algorithm.

In data sources is very common to have temporal dependence and also a great degree of redundancy. Here for redundancy we mean different sequences of symbols having the same effect for the future law of the process. For example in linguistic we have words that are synonymous, in some cases exchanging a word for a synonymous, does not change the law for the future of the sequence, we can think that they are equivalent, there are also sequences of multiple words which are equivalent. Any meaningful model for this kind of datasets should retrieve this kind of equivalence or redundancy. For this to be done we need to get rid of the exigence of the suffix tree structure and admit any partition of the state space as the structure for the Markov model.

In this paper we introduce a larger class of finite order Markov models. In our class, each model is determined by choosing a partition of the state space, our class of models includes the full Markov chain models and the VLMC models because a context tree can be seen as a particular partition of the state space (see example 3.1).

For this larger class of models we address the problem of model selection, showing that the model can be selected consistently using the BIC criterion. We show that to apply the BIC criterion, it is not necessary to find a global maximum inside the set of partitions, which will be impossible even for a moderate size of the state space. Instead, it is possible to start with an initial partition, for example the state space, and then refine this partition step by step. For an extended version of this abstract see [4].

2. MARKOV CHAIN WITH PARTITION $\mathcal{L}$

Let $(X_t)$ be a discrete time order $M$ Markov chain on a finite alphabet $A$. Let us call $\mathcal{S} = A^M$ the state space. Denote the string $a_ma_{m+1}\ldots a_n$ by $a_{m:n}$, where $a_i \in A$, $m \leq i \leq n$.

For each $a \in A$ and $s \in \mathcal{S}$,

$$P(a|s) = \text{Prob}(X_t = a|X_{t-M}^{t-1} = s);$$

Let $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ be a partition of $\mathcal{S}$,

$$P(L, a) = \sum_{s \in L} \text{Prob}(X_{t-M}^{t-1} = s, X_t = a), \ a \in A, \ L \in \mathcal{L}; \quad (1)$$

$$P(L) = \sum_{s \in L} \text{Prob}(X_{t-M}^{t-1} = s), \ L \in \mathcal{L}; \quad (2)$$

if $P(L) > 0, \forall a \in A$, we define

$$P(a|L) = \frac{P(L, a)}{P(L)}. \quad (3)$$
In order to define our model we need to introduce the following equivalence relation.

**Definition 2.1** Let \((X_t)\) be a discrete time order \(M\) Markov chain on a finite alphabet \(A\). We will say that \(s, r \in S\) are equivalent (denoted by \(s \sim_p r\)) if \(P(a|s) = P(a|r)\) \(\forall a \in A\).

For any \(s \in S\), the equivalence class of \(s\) is given by \([s] = \{r \in S | s \sim_p r\}\).

**Remark 2.1** The equivalence relationship defines a partition of \(S\). The parts of this partition are the equivalence class. The class are the subsets of \(S\) with the same transition probabilities i.e. \(s, r \in S\) belongs to different classes if and only if they have different transition probabilities.

**Remark 2.2** We can think that each element of \(S\) on the same equivalence class activates the same random mechanism to choose the next element in the Markov chain.

We can define now the Markov chain with partition \(\mathcal{L}\).

**Definition 2.2** Let \((X_t)\) be a discrete time, order \(M\) Markov chain on \(A\) and let \(\mathcal{L} = \{L_1, L_2, \ldots, L_K\}\) be a partition of \(S\). We will say that \((X_t)\) is a Markov chain with partition \(\mathcal{L}\) if this partition is the one defined by the equivalence relationship \(\sim_p\), introduced by definition 2.1.

**Remark 2.3** The set of parameters for a Markov chain over the alphabet \(A = \{a_1, a_2, \ldots, a_{|A|}\}\) with partition \(\mathcal{L} = \{L_1, L_2, \ldots, L_K\}\) can be denoted by:

\[
\{P(a_i|L_j) : 1 \leq i < |A|, 1 \leq j \leq K\}.
\]

If we know the equivalence relationship for a given Markov chain, then we need \((|A| - 1)\) transition probabilities for each class to specify the model. Then the number of parameters for the model is \(|\mathcal{L}|(|A| - 1)\).

### 3. Minimal Markov Models Selection

Let \(x^n_t\) be a sample of the process \((X_t)\), \(s \in S\), \(a \in A\) and \(n > M\). We denote by \(N_n(s, a)\) the number of occurrences of the string \(s\) followed by \(a\) in the sample \(x^n_t\),

\[
N_n(s, a) = \left| \{ t : M < t \leq n, x^{t-1}_{t-M} = s, x_t = a \} \right|, \tag{4}
\]

the number of occurrences of \(s\) in the sample \(x^n_t\) is denoted by \(N_n(s)\) and

\[
N_n(s) = \left| \{ t : M < t \leq n, x^{t-1}_{t-M} = s \} \right|. \tag{5}
\]

The number of occurrences of elements into \(L\) followed by \(a\) is given by,

\[
N_n^L(a) = \sum_{s \in L} N_n(s, a), \quad L \in \mathcal{L}; \tag{6}
\]

the accumulated number of \(N_n(s)\) for \(s\) in \(L\) is denoted by,

\[
N_n^L(L) = \sum_{s \in L} N_n(s), \quad L \in \mathcal{L}. \tag{7}
\]

As a consequence, if we write \(P(x^n_t) = \text{Prob}(X^n_t = x^n_t)\), we obtain under the assumption of a hypothetical partition \(\mathcal{L}\) of \(S\),

\[
P(x^n_t) = P(x^n_1) \prod_{L \in \mathcal{L}, a \in A} P(a|L) N_n^L(a). \tag{8}
\]

In the same way that [3] we will define our BIC criterion using a modified maximum likelihood. We will call maximum likelihood to the maximization of the second term in the equation (8) for the given observation. For the sequence \(x^n_t\), will be

\[
\text{ML}(\mathcal{L}, x^n_t) = \prod_{L \in \mathcal{L}, a \in A} \left( \frac{N_n^L(a)}{N_n^L(L)} \right)^{N_n^L(a)}. \tag{9}
\]

The BIC is given by the next definition.

**Definition 3.1** Given a sample \(x^n_t\) of the process \((X_t)\), a discrete time order \(M\) Markov chain on a finite alphabet \(A\) with \(S = A^M\) the state space and \(\mathcal{L}\) a partition of \(S\). The BIC of the model given by definition 2.2 and according to the modified likelihood (9) is given by

\[
\text{BIC}(\mathcal{L}, x^n_t) = \ln (\text{ML}(\mathcal{L}, x^n_t)) - \frac{(|A| - 1)|L|}{2} \ln(n). \tag{10}
\]

#### 3.1. Good partitions of \(S\)

Let \((X_t)\) be a discrete time order \(M\) Markov chain on a finite alphabet \(A\) and \(S = A^M\) the state space.

**Definition 3.2** Let \(\mathcal{L} = \{L_1, L_2, \ldots, L_K\}\) be a partition of \(S\). \(L \in \mathcal{L}\) is a good partition if \(\mathcal{L}\) and \(S\) are a good partition of \(S\) if for each \(i \in \{1, \ldots, K\}\), \(L_i\) verify definition 3.2.

**Remark 3.1** \(\mathcal{L} = S\) is a good partition of \(S\).

If \(\mathcal{L}\) is a good partition of \(S\), we define for each part \(L \in \mathcal{L}\)

\[
P(a|L) = \text{Prob}(X_t = a|X^{t-1}_{t-M} = s) \quad \forall a \in A, \quad (10)
\]

where \(s\) is some element into \(L\).

#### 3.2. Minimal Markov models and VLMC models

The family of minimal Markov models includes the VLMC models.

Let \((X_t)\) be a finite order Markov chain taking values on \(A\) and \(T\) a set of sequences of symbols from \(A\) such that no string in \(T\) is a suffix of another string in \(T\), for each \(s \in T\), \(d(T) = \max \{l(s), s \in T\}\) where \(l(s)\) denote the length of the string \(s\), with \(l(\emptyset) = 0\) if the string is the empty string.
Definition 3.4 \( T \) is a context tree for the process \((X_t)\) if for any sequence of symbols in \( A \), \( x_1^n \) sample of the process with \( n \geq d(T) \), there exist \( s \in T \) such that

\[
P(X_{n+1} = a | X^n_1 = x^n_1) = P(X_{n+1} = a | X^n_{n-1(s)+1} = s)
\]

where \( d(T) \) is the depth of the tree.

Each model in the family of variable length Markov chain (VLMC) models, is identified by its context tree. For more details see [2] and [1].

The context tree for a VLMC with finite depth \( M \) defines a good partition on the space \( S = A^M \) as illustrated by the next example.

Example 3.1 Let be a VLMC over the alphabet \( A = \{0, 1\} \) with depth \( M = 3 \) and contexts,

\[
\{0\}, \{01\}, \{011\}, \{111\}
\]

This context tree correspond to the good partition

\[
\{L_1, L_2, L_3, L_4\}
\]

where

\[
L_1 = \{000\}, \{100\}, \{010\}, \{110\}, \\
L_2 = \{001\}, \{101\}, \\
L_3 = \{011\} \text{ and } L_4 = \{111\}.
\]

3.3. Smaller good partitions

Definition 3.5 Let \( L^{ij} \) denote the partition

\[
L^{ij} = \{L_1, \ldots, L_{i-1}, L_{ij}, L_{i+1}, \ldots, L_{j-1}, L_{j+1}, \ldots, L_K\},
\]

where \( L = \{L_1, \ldots, L_K\} \) is a partition of \( S \), and for \( 1 \leq i < j \leq K \) with \( L_{ij} = L_i \cup L_j \).

Notation 3.1 For a \( a \in A \) we write,

\[
P(L_{ij}, a) = P(L_i, a) + P(L_j, a); \\
P(L_{ij}) = P(L_i) + P(L_j).
\]

\[
N_n^{L^{ij}}(L_{ij}, a) = N_n^{L_i}(L_i, a) + N_n^{L_j}(L_j, a); \\
N_n^{L^{ij}}(L_{ij}) = N_n^{L_i}(L_i) + N_n^{L_j}(L_j).
\]

Remark 3.2 If \( L \) is a good partition and \( P(\cdot|L_i) = P(\cdot|L_j) \) then \( L^{ij} \) is a good partition.

The next lemma allows to prove the main result of this section, nevertheless this lemma could be applied to partitions (it is not necessary good partitions) with at least two good parts.

Theorem 3.1 Let \((X_t)\) be a Markov chain of order \( M \) over a finite alphabet \( A \), \( S = A^M \) the state space and \( x_1^n \) a sample of the Markov process. Let \( L = \{L_1, L_2, \ldots, L_K\} \) be a partition of \( S \) and suppose that exist \( i \) and \( j \), \( i \neq j \) such that \( L_i \) and \( L_j \) verified the definition 3.2. Then,

\[
P(a|L_i) = P(a|L_j) \quad \forall a \in A
\]

if, and only if, eventually almost surely as \( n \to \infty \),

\[
BIC(L^{ij}, x_1^n) > BIC(L, x_1^n).
\]

Where \( L^{ij} \) partition is defined under \( L \) by definition 3.5.

In the following corollary, we show that the BIC criterion provides a consistent way of detecting smaller good partition.

Corollary 3.1 Under the assumptions of theorem 3.1, if \( L = \{L_1, L_2, \ldots, L_K\} \) is a good partition of \( S \) and \( 1 \leq i < j \leq K \), then \( P(a|L_i) = P(a|L_j) \) \( \forall a \in A \) if, and only if, eventually almost surely as \( n \to \infty \),

\[
BIC(L, x_1^n) < BIC(L^{ij}, x_1^n).
\]

Definition 3.6 Let be \((X_t)\) a Markov chain of order \( M \), with finite alphabet \( A \) and state space \( S = A^M \), \( x_1^n \) a sample of the process and let \( L = \{L_1, L_2, \ldots, L_K\} \) be a good partition of \( S \),

\[
d_L(i, j) = \frac{1}{\ln(n)} \sum_{a \in A} \left( N_n^L(L_i, a) \ln \left( \frac{N_n^L(L_i, a)}{N_n^L(L_j, a)} \right) + N_n^L(L_j, a) \ln \left( \frac{N_n^L(L_j, a)}{N_n^L(L_i, a)} \right) - N_n^{L^{ij}}(L_{ij}, a) \ln \left( \frac{N_n^{L^{ij}}(L_{ij}, a)}{N_n^L(L_{ij})} \right) \right) \tag{13}
\]

Corollary 3.2 Under the assumptions of theorem 3.1,

\[
BIC(L, x_1^n) - BIC(L^{ij}, x_1^n) < 0
\]

\[\iff d_L(i, j) < \frac{(|A| - 1)}{2}.\]

Remark 3.3 The results will remain valid if we replace the constant \(\frac{|A| - 1}{2}\) by an arbitrary constant, positive and finite value \( v \), into the definition 3.1.

Remark 3.4 Under the assumptions of theorem 3.1, if \( P(a|L_i) \neq P(a|L_j) \) for some \( a \in A \), then eventually almost surely as \( n \to \infty \), \( BIC(L, x_1^n) > BIC(L^{ij}, x_1^n) \) where \( L^{ij} \) verified the definition 3.5.

3.4. Minimal good partition

The smaller good partition in the universe of all possible good partitions of \( S \) is the partition defined by the equivalence relation 2.1. We will call this partition of minimal good partition. Our objective is to find this minimal good partition.

Definition 3.7 Let \((X_t)\) be a discrete time order \( M \) Markov chain on a finite alphabet \( A \), \( S = A^M \) the state space. A partition \( L = \{L_1, L_2, \ldots, L_K\} \) of \( S \) is the minimal good partition of \( S \) if it is the partition corresponding to the equivalence relation established in definition 2.1.

Remark 3.5 For a discrete time order \( M \) Markov chain on a finite alphabet \( A \) with \( S = A^M \) the state space, \( \exists \) minimal good partition of \( S \).
The next theorem shows that for \( n \) large enough we achieve the partition \( L^* \) which is the minimal good partition.

**Theorem 3.2** Let \( (X_t) \) be a Markov chain of order \( M \) over a finite alphabet \( A, S = A^M \) the state space and \( x^n_1 \) a sample of the Markov process. Let \( \mathcal{P} \) be the set of all the partitions of \( S \). Define,

\[
L^*_n = \arg\max_{L \in \mathcal{P}} \{ \text{BIC}(\mathcal{L}, x^n_1) \}
\]

then, eventually almost surely as \( n \to \infty \),

\[
L^* = L^*_n
\]

where \( L^* \) is the minimal good partition of \( S \), following definition 3.7.

3.5. Minimal good partition estimation algorithm

We use the corollary 3.2 to build our algorithm.

Consider \( x^n_1 \) a sample of the Markov process \( (X_t) \), of order \( M \) over a finite alphabet \( A, S = A^M \) the state space. Let \( L = \{L_1, L_2, \ldots, L_K\} \) be a good partition of \( S \).

**Algorithm 3.1** (MMM algorithm for good partitions)

**Input:** \( L \).

**Output:** \( \hat{L}_n \).

\[ i \leftarrow 0, \ j \leftarrow 1 \]

\[ \text{while } i < K - 1 \]

\[ \quad i \leftarrow i + 1 \]

\[ \text{while } j < K \]

\[ \quad j \leftarrow j + 1 \]

\[ \quad d \leftarrow d_L(i, j) \]

\[ \text{while } d < 1 \]

\[ \quad L_i \leftarrow L_i \cup L_j \]

\[ \text{for each } i \in \{j, \ldots, K - 1\} \text{ do } \]

\[ \quad L_i \leftarrow L_i \cup L_{i+1} \]

\[ \quad K \leftarrow K - 1 \]

\[ \quad \mathcal{L} \leftarrow \{L_1, L_2, \ldots, L_K\} \]

\[ \quad d \leftarrow d_L(i, j) \]

**Return:** \( \hat{L}_n = \{L_1, L_2, \ldots, L_K\} \)

4. CONCLUSIONS

Our main motivation to define the minimal Markov models is, in the first place, the concept of partitioning the state space in parts in which the states are equivalent, this allow us to model the redundancy that appears in many processes in practice such as in genetics, linguistics, etc. Each part in the state space has a very specific, clear and practical meaning, any sequence of symbol in the same part has the same effect on the future distribution of the process. In other words, they activate the same random mechanism to choose the next symbol on the process. We can think about each part of the resulting minimal good partition as being a set of synonymous in the state space \( S \).

In the second place our motivation for developing this methodology is to prove that for a stationary, finite memory process it is theoretically possible to find consistently a minimal Markov model to represent this process and that this can be accomplished in practice. The utilitarian implication of the fact that the model selection process can be started from a context tree partition, is that minimal Markov models can be easily fitted to stationary sources where VLMC models already work.

5. REFERENCES


VARIABLE SELECTION IN LINEAR REGRESSION: SEVERAL APPROACHES BASED ON NORMALIZED MAXIMUM LIKELIHOOD

Ciprian Doru Giurcăneanu, Seyed Alireza Razavi, Antti Liski

Department of Signal Processing
Tampere University of Technology, Finland
ciprian.giurcaneanu@tut.fi, alireza.razavi@tut.fi, antti.liski@tut.fi

ABSTRACT

In this talk, we discuss the application of the normalized maximum likelihood (NML) for model selection in Gaussian linear regression. All the results which will be presented have been recently published in [1].

1. INTRODUCTION

The use of the NML for model selection in Gaussian linear regression poses troubles because the normalization coefficient is not finite. The most elegant solution has been proposed by Rissanen and consists in applying a particular constrain for the data space [2]. The resulting criterion is independent of arbitrarily selected hyper-parameters. Surprisingly, for about one decade, it was totally ignored the important fact that the closed-form expression of the criterion depends on the particular constraint which has been involved in its derivation. Only recently, it was shown in [3] that two other criteria can be obtained by employing constraints which are different of the one used in [2]. Hence, novel NML-based criteria can be devised by enforcing various constraints.

2. MAIN RESULTS

Our talk is focused on the following aspects:

- We demonstrate that the methodology introduced by Rissanen can be applied in a more general framework, and not only for the ellipsoidal constraints which have been considered in [2, 3]. We also discuss the particular case of rhomboidal constraint.
- We analyze the relationship between Rissanen criterion and the two criteria that have been introduced in [3].
- We compare experimentally the capabilities of the NML-based selection rules against BIC [4], $\text{AIC}_{c3}$ [5], CME [6], $\text{MML}_g$ and $\text{MML}_u$ [7].
- We also provide some guidance on the use of various criteria in model selection.

3. REFERENCES


ASPECTS OF CATEGORY THEORY

Esfandiar Haghverdi

School of Informatics and Computing
Indiana University
150 S. Woodlawn Ave.,
Bloomington, IN 47405 USA
ehaghver@indiana.edu

ABSTRACT

I give a very brief introduction to category theory. The paper covers the basic definitions, some important basic constructions and spends some time discussing the important notion of categorical trace that has recently found many applications in computer science and the proof theory of Linear Logic among other things. The goal is to introduce categorical ideas and notions to Information Theory community with the hopes that it might find some applications.

1. INTRODUCTION

Category theory was invented by S. Eilenberg and S. Mac Lane in 1945 in their study of algebraic topology. The first invented notion was that of a natural transformation which led to the formulation of a category and that of a functor. Since then category theory has found many applications in diverse areas from proof theory and logic to the foundations of quantum mechanics, semantics of programming languages, to the seminal work of Grothendieck on redefining the foundations of algebraic geometry. The aim of this paper is to introduce the basic notions of category theory to Information Theory community. I have emphasized the important notion of the categorical trace invented by Joyal, Street, and Verity [1] as I believe that this notion in particular, can find many applications in the field of Information Theory.

2. BASIC DEFINITIONS

A category \( \mathcal{C} \) is a collection \( ob(\mathcal{C}) \) of objects and for each pair \( A, B \) of objects a set \( \mathcal{C}(A, B) \) called the homset of morphisms from \( A \) to \( B \). Such categories are called locally small because \( \mathcal{C}(A, B) \) is a set. For a general category it can be a class. In this paper, we shall only consider locally small categories. For each triple of objects \( A, B, C \) there is a composition operation

\[
\circ_{A,B,C} : \mathcal{C}(A, B) \times \mathcal{C}(B, C) \to \mathcal{C}(A, C)
\]

and for each object \( A \) there is an identity morphism \( 1_A : A \to A \) subject to the following conditions:

- For any morphisms \( f : A \to B, g : B \to C, \) and \( h : C \to D, \)

\[
h \circ (g \circ f) = (h \circ g) \circ f
\]

- For any morphism \( f : A \to B, 1_B \circ f = f = f \circ 1_A. \)

For readability we shall write \( gf \) instead of \( g \circ f \). Here are some examples of categories:

1. The category \( \text{Sets} \) of sets and mappings, where objects are sets and morphisms are mappings. Composition is the usual functional composition which is associative and has the identity function as its neutral element.

2. The category \( \text{Rel} \) of sets and relations, where objects are sets and a morphism \( R : A \to B \) is a binary relation from \( A \) to \( B \). Given relations \( R : A \to B \) and \( S : B \to C, SR : A \to C \) is defined as

\[
(a, c) \in SR \iff \exists b \in B, (a, b) \in R \text{ and } (b, c) \in S.
\]

It can be easily checked that this composition operation is associative with the identity relation as its neutral element.

3. The category \( \text{Pfn} \) of sets and partial functions, where objects are sets and a morphism \( f : A \to B \) is a partial function from the set \( A \) to \( B \).

4. The category \( \text{SRel} \), where objects are measurable spaces and a morphism \( f : (X, \mathcal{F}_X) \to (Y, \mathcal{F}_Y) \) is map \( f : X \times \mathcal{F}_Y \to [0,1] \) such that for a fixed \( x \in X, f(x,.) : \mathcal{F}_Y \to [0,1] \) is a probability measure, and for a fixed \( B \in \mathcal{F}_Y, f(.,B) : X \to [0,1] \) is a bounded measurable function. Given \( f : X \to Y, g : Y \to Z, \)

\[
gf(x,C) = \int_Y g(y,C)f(x,dy).
\]

The identity morphism \( 1_X : X \to X \) is given by \( 1_X(x, A) = 1 \) if \( x \in A, \) and \( 0 \) otherwise.
5. The category $\text{FDVec}_k$ has infinite-dimensional vector spaces over a field $k$ as objects and a morphism $f : V \to W$ is a linear transformation between the vector space $V$ to $W$.

6. The category $\text{Graph}$ of graphs and graph homomorphisms where objects are graphs and a morphism $f : G \to H$ is a graph homomorphism, that is an edge preserving map from the vertex set of $G$ to that of $H$.

The basic idea and the main message of category theory is the emphasis on relationships, over that of objects themselves. For example, from a category theory perspective, set theory is about functions, linear algebra is about vector spaces and linear transformations, where the monoidal product is the familiar tensor product of vector spaces and the tensor unit is the ground field.

A traced symmetric monoidal category is a symmetric monoidal category $(\mathcal{C}, \otimes, I, s)$ with a family of functions $Tr^U_{X,Y} : \mathcal{C}(X \otimes U, Y \otimes U) \to \mathcal{C}(X, Y)$ called a trace, subject to the following axioms:

- **Natural in $X$**, 
  $$Tr^U_{X,Y}(f) g = Tr^U_{X,Y}(f(g \otimes 1_U))$$
  where $f : X \otimes U \to Y \otimes U$, $g : X' \to X$,

- **Natural in $Y$**, 
  $$g Tr^U_{X,Y}(f) = Tr^U_{X,Y}((g \otimes 1_U)f)$$
  where $f : X \otimes U \to Y \otimes U$, $g : Y \to Y'$,

- **Dinatural in $U$**, 
  $$Tr^U_{X,Y}((1_Y \otimes g)f) = Tr^{U'}_{X,Y}(f(1_X \otimes g))$$
  where $f : X \otimes U \to Y \otimes U'$, $g : U' \to U$,

- **Vanishing (I,II)**, 
  $$Tr^U_{X,Y}(f) = f$$
  and
  $$Tr^{U \otimes V}_{X,Y}(g) = Tr^U_{X,Y}(Tr^V_{X,Y}(Tr^{U \otimes U \otimes Y \otimes U}(g)))$$
  for $f : X \otimes I \to Y \otimes I$ and $g : X \otimes V \otimes Y \otimes U \otimes V$, 

- **Superposing**, 
  $$g \otimes Tr^U_{X,Y}(f) = Tr^V_{W \otimes X \otimes Z \otimes Y}(g \otimes f)$$
  for $f : X \otimes U \to Y \otimes U$ and $g : W \to Z$,

- **Yanking**, 
  $$Tr^U_{I,U}(s_{U,U}) = 1_U$$

3. CATEGORICAL TRACE

In recent years the notion of categorical trace, introduced by Joyal, Street, and Verity [1] has gained a lot of attention ranging in applications from lambda calculus [2], to semantics of linear logic [3]. Before we introduce this notion we shall briefly mention the notion of a monoidal (tensor) category. A monoidal category $\mathcal{C}$ is a category together with a bifunctor $\otimes : \mathcal{C} \times \mathcal{C} \to \mathcal{C}$ (monoidal or tensor product), an object $I$ in $\mathcal{C}$ called the tensor unit and structure isomorphisms $\alpha_{A,B,C} : A \otimes (B \otimes C) \to (A \otimes B) \otimes C$ (associativity), $\rho_A : A \otimes I \to A$ (right unit), and $\lambda_A : I \otimes A \to A$ (left unit) such that certain diagrams commute, for details see the canonical reference [4]. A symmetric monoidal category is a monoidal category with a natural isomorphism $s_{A,B} : A \otimes B \to B \otimes A$ such that $s_{B,A}s_{A,B} = 1_{A\otimes B}$ plus some more conditions. The canonical example to keep in mind is that of vector spaces and linear transformations, where the monoidal product is the familiar tensor product of vector spaces and the tensor unit is the ground field.

A traced symmetric monoidal category is a symmetric monoidal category $(\mathcal{C}, \otimes, I, s)$ with a family of functions $Tr^U_{X,Y} : \mathcal{C}(X \otimes U, Y \otimes U) \to \mathcal{C}(X, Y)$ called a trace, subject to the following axioms:
3.2. Examples

Consider the category $\text{FDVec}_k$ of finite dimensional vector spaces and linear transformations. Given $f : V \otimes U \rightarrow W \otimes U$, and $\{v_i\}, \{u_j\}, \{w_k\}$ bases for $V, U, W$ respectively, write $f$ as $f(v_i \otimes u_j) = \sum_{k,m} a_{ij}^{km} w_k \otimes u_m$, then

$$\text{Tr}_{V,W}^U(f)(v_i) = \sum_{j,k} a_{ij}^{kj} w_k.$$ 

Note that this is a simple generalization of the notion of trace for a matrix: sum of diagonal entries. Here the diagonal entries are block matrices of size $\text{dim}(W) \times \text{dim}(V)$ and one is adding $\text{dim}(U)$ many such diagonal blocks. Clearly if we choose $\text{dim}(V) = \text{dim}(W) = 1$, that is when $V \cong W \cong k$, we get back the usual trace.

As a second example consider the category $\text{Rel}$ of sets and binary relations with $X \times Y = X \times Y$. Note that this is a tensor product with unit $I = \{\ast\}$. Given $R : X \otimes U \rightarrow Y \otimes Y$, $\text{Tr}_X^Y(R) : X \rightarrow Y$ is defined by

$$(x,y) \in \text{Tr}(R) \text{ iff } \exists u.(x,u,y,u) \in R.$$ 

3.3. Traced Unique Decomposition Categories

Finally we shall consider an important class of traced categories called traced Unique Decomposition Categories (traced UDCs) without getting into too much detail. These are categories with special additive structures on their hom-sets called $\Sigma$-monoids.

A $\Sigma$-monoid is a pair $(M, \Sigma)$, where $M$ is a nonempty set and $\Sigma$ is a partial operation on countable families in $M$ where $\{x_i\}_{i \in I}$ is summable if $\Sigma_{i \in I} x_i$ is defined subject to:

- **Partition-Associativity:** $\{x_i\}_{i \in I}$ and $\{f_j\}_{j \in J}$ is a countable partition of $I$
  $$\Sigma_{i \in I} x_i = \Sigma_{j \in J} (\Sigma_{i \in I_j} x_i).$$
- **Unary sum:** $\Sigma_{i \in \{j\}} x_i = x_j$.

Here are some quick facts about $\Sigma$-monoids:

- $\Sigma_{i \in \emptyset} x_i$ exists and is denoted by $0$. It is a countable additive identity.
- Sum is commutative and associative whenever defined.
- $\Sigma_{i \in I} x_{\phi(i)}$ is defined for any permutation $\phi$ of $I$, whenever $\Sigma_{i \in I} x_i$ exists.
- There are no additive inverses: $x + y = 0$ implies $x = y = 0$.

Let us look at some examples. The set $M = \text{PInj}(X, Y)$ of partial injective functions where a family $\{f_i\}$ is summable if $f_i$ and $f_j$ have disjoint domains and codomains for all $i \neq j$. The sum is then defined as $\langle \Sigma_{i \in I} f_i \rangle(x) = \{ f_j(x), \text{ if } x \in \text{Dom}(f_j) \text{ for some } j \in I; $}

$$\text{undefined, otherwise.}$$

The set $M = \text{PInj}(X, Y)$ of partial functions from a set $X$ to a set $Y$ where a family $\{f_i\}$ is summable if $f_i$ and $f_j$ have disjoint domains for all $i \neq j$, and $\langle \Sigma_{i \in I} f_i \rangle(x)$ is defined as above.

The set $M = \text{Rel}(X, Y)$ of binary relations from $X$ to $Y$. Here all families are summable, and $\Sigma_{I,R} = \bigcup_{R_{i}} R_{i}$.

Finally here is a non-example, let $M$ be an $\omega$-complete poset, that is a partially ordered set where all countable chains have suprema. Define a family $\{x_i\}$ to be summable if it is a countable chain, and let the sum be $\Sigma_{i \in I} x_i = \sup_{X \in I} x_i$. Suppose $x, y, z$ are in this family, with $x \leq z, y \leq z$ and $x, y$ incomparable, then $x + (y + z)$ is defined but $(x + y) + z$ is not defined.

A unique decomposition category $C$ is a symmetric monoidal category where (1) Every homset is a $\Sigma$-monoid, (2) Composition distributes over sum whenever defined, satisfying the axiom:

$$(A) \text{ For all } j \in I (f \text{ finite) there are morphisms}$$

- quasi injection: $t_j : X_j \rightarrow \otimes_{i \in I} X_j$, and
- quasi projection: $p_j : \otimes_{i \in I} X_i \rightarrow X_j$, such that

  $$\rho_k t_j = 1_{X_j}, \text{ if } j = k \text{ and } 0_{X_j X_k}, \text{ otherwise},$$

  $$\sum_{i \in I} t_i \rho_i = 1_{\otimes_{i \in I} X_i}.$$ 

**Proposition 1** [Matricial Representation]

For $f : \otimes_{j \in J} X_j \rightarrow \otimes_{j \in J} Y_j$, there exists a unique family $\{f_{ij}\}_{i \in J, j \in I} : X_j \rightarrow Y_i$ with $f = \sum_{i \in J, j \in I} f_{ij} p_j$, namely, $f_{ij} = p_j f_{ij}$. In particular, for $|I| = m, |J| = n$

$$f = \begin{bmatrix} f_{11} & \cdots & f_{1n} \\ \vdots & \ddots & \vdots \\ f_{m1} & \cdots & f_{mn} \end{bmatrix}$$

Here are a few examples, the category $\text{PInj}$ of sets and partial injective functions with disjoint union as the monoidal product. Here, $p_j : \otimes_{i \in I} X_i \rightarrow X_j, p_j(x, i)$ is
undefined for $i \neq j$ and $\rho_j(x, j) = x$, and $\iota_j : X_j \to \otimes_{i \in I} X_i$ by $\iota_j(x) = (x, j)$.

The category Rel with disjoint union as the monoidal product, $\rho_j : \otimes_{i \in I} X_i \to X_j, \rho_j = \{(x, j) \mid x \in X_j\}$. And $\iota_j : X_j \to \otimes_{i \in I} X_i, \iota_j = \{(x, (x, j)) \mid x \in X_j\} = \rho_j^{op}$.

**Proposition 2** [Standard Trace Formula]

Let $C$ be a unique decomposition category such that for every $X, Y, U$ and $f : X \otimes U \to Y \otimes U$, the sum $f_{11} + \sum_{n=0}^{\infty} f_{12} f_{21}^{n} f_{21}$ exists, where $f_{ij}$ are the components of $f$. Then, $C$ is traced and

$$Tr_{X,Y}^{U}(f) = f_{11} + \sum_{n=0}^{\infty} f_{12} f_{21}^{n} f_{21}.$$ 

Let us calculate some traces. Let $C$ be a traced UDC. Then given any $f : X \otimes U \to Y \otimes U$, $Tr_{X,Y}^{U}(f)$ exists.

- Let $f : X \otimes U \to Y \otimes U$ be given by $[\begin{matrix} g & 0 \\ h & 0 \end{matrix}]$.

Then

$$Tr_{X,Y}^{U}(f) = Tr_{X,Y}^{U} \left([\begin{matrix} g & 0 \\ h & 0 \end{matrix}]\right) = g + \sum_{n=0}^{\infty} h^{n}.$$ 

- $g + 0 h = g + 0 = g$.

- Let $f : X \otimes U \to Y \otimes U$ be given by $[\begin{matrix} g & 0 \\ 0 & h \end{matrix}]$.

Then

$$Tr_{X,Y}^{U}(f) = Tr_{X,Y}^{U} \left([\begin{matrix} g & 0 \\ 0 & h \end{matrix}]\right) = g + \sum_{n=0}^{\infty} 0 h^{n}$$

- $g + 0 = g$.

4. CONCLUSION

In this paper, I have tried to give a glimpse of category theory with special attention to the notion of trace hoping that such ideas might find some applications in Information Theory and/or Statistical Inference. Even though the treatment here is very cursory, I hope I was able to convey some of the general philosophy and the spirit of thinking in terms of categories. At this point the reader will naturally wonder what she or he might be able to do with categories. Let me try to point out some possibilities below.

- **Unification**

Category theory, above all offers a unifying language whereof one can define once and for all a single notion or construction which upon instantiation yields the extant concrete definitions. We saw one such example here, namely that of a categorical product which yields the Cartesian product in Sets and graph product in Graphs and direct product of vector spaces in FDVec$_k$, etc.

- **Technology transfer**

Once you have organized your objects of study into a category you can start trying different constructions on them. For example you might wonder what the coproduct for two graphs could be or what kind of tensor (monoidal) products you could define on graphs, etc. Thus general categorical constructions (for example limits and colimits) might yield new, unknown constructions on the objects that you study. On the other hand, there might be some constructions that are possible in categories with additional structure. Once you prove that your category has this additional structure, you can transfer the existing technology and apply it to objects of your study.

- **Mathematical analysis**

Quite often it is much more productive to study certain mathematical objects via their representations in terms of more familiar objects which are easier to understand or study, or it is the case that there are a lot of techniques developed for their study. For example think of group representation theory where one represents groups as linear transformations on vector spaces, or think of homology theory where one studies topological spaces using groups (homology groups). Such relations are often functorial, for example the functor that associates an abelian group to a topological space, etc. The lesson here, thus is to find functorial relations (representations) from the category of objects of study into one which is much better understood, in order to obtain new results about the former category.

5. REFERENCES


FUNCTIONAL DEPENDENCES AND BAYESIAN NETWORKS

Peter Harremoës
Copenhagen Business College
Copenhagen, Denmark
harremoes@ieee.org

ABSTRACT

Abstract Bayesian networks and conditional independence is studied via functional dependences. Armstrong’s axioms known from the theory of relational databases is used to reformulate the concept of Bayesian networks into the theory of meet-semidistributive lattices. Lattice theory provides us with a richer language to discuss causation than graph theory does.

1. INTRODUCTION

The theory of Bayesian networks is normally considered as a theory of conditional independence. In this paper we shall take a different point of view and consider the theory of Bayesian networks as a theory of functional dependence. Bayesian networks has become a popular model of causation among computer scientists. Functional dependencies has definitely played a role in philosophers attempts to understand the concepts cause and effect, and from a philosophical point of view the ideas presented here are not all that new and discussion and references can be found in [1, Section 1.4]. Nevertheless this point of view has not been explored in sufficient detail in relation to Bayesian networks.

This work use ideas from three areas. The theory of functional dependencies in relational databases was pioneered by [2]. Now the basic results can be found in any textbook like [3] on relational databases, but in the database community this branch of research is not really active anymore.

Lattice theory is now a mature research area with unified notation, plenty of results and applications in all branches of mathematics. The theory of lattices was developed by Birkhoff and his reprinted textbook [4] on this subject is still worth reading. A comprehensive modern expositions can be found in [5]. For the more specialized theory of semimodular lattices we refer to [6], that we shall follow with regard to notation and terminology.

The study of time and causation has an enormous literature that started about 2500 years ago. Here we will only point at a few books that present the most important scientific positions on this subject [7, 8, 9, 10, 11, 12]. The study of causation via Bayesian networks is much younger and a good overview of this approach to the study of causation can be found in [1, 13]. Many of the definitions that we use may be found in these textbooks.

In this short paper proofs will be omitted.

2. FUNCTIONAL DEPENDENCE LATTICES

There are various reasons to study functional dependences in databases. The two most important are:

Compression One can make a more compact representation of data when functional dependences are known. How to make such compact representations is the subject of the theory of relational databases.

Control If one has direct control over some of the attributes/variables in a database then the functional dependences will tell what the effect will be on other attributes. If for instance one has a database of all members of a club and want to send a letter to all members in a specific area, one may do this by sending to all members for which part of their zip code has a certain value. The relation between zip code, town and areas will tell how this can be done most efficiently.

We note that these two reasons are also the main reasons to study causality. In this section we shall describe functional dependences in databases. The relation between functional dependence and lattices has previously been studied by M. Levene[14].

First we shall consider a set of variables and subsets of this set of variables. The set of subsets is also called the power set and we can define an ordering by \( A \subseteq B \) if \( B \subseteq A \). With this ordering the power set is a lattice with \( A \cap B = A \cup B \) and \( A \cup B = A \cap B \). With the ordering defined in this way specifying the values of the variables in \( A \) and specifying the values of the variables in \( B \) is the same as specifying the values of the variables in \( A \cap B \).

We note that the smallest element in the lattice \( \bot \) is the whole set and the largest element in the lattice \( \top \) equals the empty set \( \emptyset \).

Inspired by Armstrong’s theory of relational databases we say that a relation \( \leq \) in a lattice \( L \) satisfies Armstrong’s axioms if it satisfies the following properties

Reflexivity If \( X \leq Y \), then \( X \leq Y \).

Augmentation If \( X \leq Y \), then \( X \leq Z \leq Y \). \( \wedge \).

Transitivity If \( X \leq Y \) and \( Y \leq Z \), then \( X \leq Y \).

In the database literature the lattice is a Boolean algebra of subsets of a set of variables and \( Y \leq X \) would mean that any variable in \( Y \) can be is a function of the variables.
in X. For functional dependence ≤ in a database obviously satisfies Armstrong’s axioms if A ≤ B means B ⊆ A. The following results have been proved by Armstrong.

**Theorem 1** If a relation ≤ in a lattice L satisfies Armstrong’s axioms then it also satisfies the following properties

- **Decomposition** If Z ≤ X ∧ Y, then Z ≤ X and Z ≤ Y.
- **Psuedotransitivity** If Y ≤ X and X ∧ Z ≤ W, then Y ∧ Z ≤ W.
- **Union** If Z ≤ X and Z ≤ Y, then Z ≤ X ∧ Y.

Let L denote a lattice with a relation ≤ such that Armstrong’s axioms are satisfied. For simplicity we will assume that L is finite. For X ∈ L we define cl_L(X) as X \ Y, where the meet is taken over all Y_i such that Y_i ≤ X. In a set of random variables where ≤ denotes functional dependence cl_L(X) is the set of all variables determined by the variables X. We say that X is closed if cl_L(X) = X.

**Theorem 2** Assume that (L, ∨, ∧) is a lattice with a pre-ordering ≤ satisfying Armstrong’s axioms. Let L≤ denote the set of closed elements in L. Then L≤ is a lattice with ordering ≤. The lattice operations are given by X ∨≤ Y = X ∨ Y and X ∧≤ Y = cl_L(X ∨ Y).

This theorem essentially dates back to Armstrong when L was a power set with inclusion as ordering and the proof is the same, so the proof is omitted here. The theorem as it is formulated here probably in the literature on lattices although the author has not been able to locate a good reference. An element X in a lattice is meet-irreducible if X = Y ∧ Z implies X = Y or X = Z.

**Theorem 3** If an element X in a lattice of functional dependencies is irreducible then there exists a variable x such that X = cl_L({x}).

We consider three variables a, b and c that denote real numbers. Assume that c = (a + b)^2. Then the associated lattice is the lattice that is normally called S_c.

Even simple examples of functional dependence lattices may be complicated to describe if they are not based on simple causal relations between the variables.

This example concern fruit from a supermarket. Variable X tells whether the supermarket will sell it at normal price, or at a reduced price because it is close to the expiration date, or whether it is thrown out because the expiration date has been exceeded. Variable Z describes whether the fruit tastes very fresh, is eatable, or looks disgusting. The variable Y tells whether the fruit will make you sick or not. The functional dependencies are given by Z ≤ Y and X ∨ Y = X ∨ Z. The lattice is N_5. This is the standard example of a lattice that is not modular.

**Theorem 4** Any lattice is equivalent with a functional dependence lattice.

Since any lattice is equivalent to a lattice of functional dependence so all what can be said about functional dependence can be expressed in the language of lattices. In the rest of this paper we will primarily use the language of lattices although we primarily think of these lattices as lattices of functional dependence.

### 3. ORDERING OF IRREDUCIBLE ELEMENTS

We now recall the definition of modularity. An lattice element A is said to be right modular if for all Y. A lattice element B is said to be left modular if for all Y. Note that a lattice element A is often called modular if it is right modular.

An element A in a lattice L is said to be distributive if

\[ A ∨ (Y ∧ Z) = (A ∨ Y) ∧ (A ∨ Z) \]

for all Y, Z ∈ L. It is dual distributive if

\[ A ∧ (Y ∨ Z) = (A ∧ Y) ∨ (A ∧ Z) \]

for all Y, Z ∈ L.

An element A is standard if

\[ X ∧ (A ∨ Z) = (X ∧ A) ∨ (X ∧ Z) \]

for all Y, Z ∈ L. Similarly, A is dual standard if

\[ X ∨ (A ∧ Z) = (X ∨ A) ∧ (X ∨ Z) \]

for all Y, Z ∈ L.

An element X is separating if X ∨ Y = X ∨ Z and X ∧ Y = X ∧ Z implies Y = Z for all Y, Z ∈ L.

**Theorem 5** An element is (dually) standard if and only if it is separating and (dually) distributive. If an element is distributive then it is left modular. If an element is dually distributive then it is right modular.

Let A and B denote elements of a lattice. We write A ≤ B if for any dual standard element D ≤ A we have D ≤ B.

**Theorem 6** The relation ≤ is a pre-ordering.
elements of the standard sub lattice will in general not be injective.

On a set of variables with functional dependences the preordering can be used to define a sublattice where $c$ is injective. The sublattice is defined by merging variables together if they are equivalent in the preordering $\leq$. Thus the sublattice is generated by elements of the form $\bigwedge_{A \leq B} A$ where $A$ and $B$ denote irreducible elements. From now on we will assume that $c$ is injective on irreducible elements.

**Theorem 7** Assume that $B$ is meet irreducible and that $A \not\leq B$ and $A \wedge C \leq B$. Then $C \leq B$.

Let $A$ be meet irreducible. Then a backward chain from $A$ is a maximal totally ordered set of irreducible elements with the last element dominated by $A$.

**Theorem 8** Let $D$ be a set of meet irreducible elements. If there exists a backward chain $C$ from $A$ that does not intersect $D$ then $A \not\leq \bigwedge_{B \leq D} B$.

4. THE LOCALLY DETERMINISTIC LATTICES AND THEIR PROPERTIES

A lattice $L$ is said to be a locally deterministic lattice if for any set $\mathcal{D}$ of meet irreducible elements of a lattice and any $B \geq \bigwedge_{\mathcal{D}} A$ there exists a backward chain $C$ from $B$ that intersects $\mathcal{D}$.

As we have seen for any lattice $L$ one get a preordering of the irreducible elements that we will denote $\leq$. We will now study some of the properties of locally deterministic lattices. First we note that if $A \leq B$ and $B \leq A$ then any maximal chain through $A$ intersects the set $\{B\}$ so $A \leq B$. Similarly $B \leq A$ implying that $A = B$. Hence the preordering is an ordering and the mapping $c$ is injective.

Next we shall see that the ordering defines a lattice.

In a poset a set of variables $A$ graphically determines a set of variables $B$ if any maximal chain through a variable $v \in B$ intersects $A$ in a variable prior to $v$. When $A$ graphically determines $B$ we write $A \leq_g B$.

The graphical closure $cl_g(A)$ of a set of variables $A$ is largest set $B$ such that $A \leq_g B$.

**Theorem 9** The ordering $\leq_g$ defines a locally deterministic lattice on the set of graphically closed subsets of a poset.

A closure operator $cl$ is called a convex-hull operator if it satisfies the so-called antiexchange property, i.e. If $x, y \not\in cl(A)$ and $x \in cl(A \cup \{y\})$ and $y \in cl(A \cup \{x\})$ then $x = y$.

**Theorem 10** A convex-hull operator gives a lower locally distributive lattice with $\cup$ and $\cap$ as lattice operations. The graphical closure operator is a convex-hull operator. The graphical closure lattice is upper locally distributive with $\wedge$ and $\vee$.

A functional dependence structure may be quite complicated. For instance we may think of four variables $a, b, c$ and $d$ such that $c$ is a function of $a$ and $b$ and $b$ is a function $c$ and $d$. Let us introduce $X = \{a, d\}$ and $Y = \{b\}$ and $Z = \{c\}$.

**Theorem 11** A locally deterministic lattice can be modelled by a Bayesian network.

In the previous example $X \geq Y \wedge Z$ and $Y \geq X \wedge Z$. Hence $X \wedge Z = Y \wedge Z$. To handle this kind of complication we make the following definition.

A lattice said to be semi-convex if $X \wedge Y = X \wedge Z$ and $X \vee Z = Y \vee Z$ implies $X \leq Z$. A lattice is said to be meet-semidistributive if $X \wedge Z = Y \wedge Z$ implies that $(X \vee Y) \vee Z = X \vee Z$ for all $X, Y$ and $Z$.

Note that a lattice is semi-convex if $X \wedge Y = X \wedge Z$ and $X \vee Z = Y \vee Z$ is equivalent to $X \leq Y \leq Z$.

If a lattice is meet-semidistributive then it is semi-convex.

We note that the lattice $M_5$ is not semi-convex, and semi-convexity efficiently rule out the possibility of anything like a causal loop.

A lattice is join-distributive (also called Upper Locally Distributive, ULD), if any element in the lattice has a unique decomposition as a meet of meet irreducible elements.

A locally deterministic lattice is join-distributive (upper locally distributive, ULD). In particular the a graphical closure lattice is semi-convex, upper semidistributive and $M$-symmetric.

The lattice $N_5$ is not a locally deterministic lattice. To see this we note that $N_5$ is not semimodular.

**Theorem 12** Let $X$ and $Y$ denote meet irreducible elements in a locally deterministic lattice. Then $X \geq Y$ or $Y \geq X$ or $X \vee Y = \top$.

To get a better understanding of the relation between the upsets in a poset and the ordering $\leq$ of the irreducible elements we need the following results.

**Theorem 13** For an element $A$ in a locally deterministic lattice the following five conditions are equivalent:

1. The element $A$ is an upset in the poset of irreducible elements.
2. The element $A$ is dually standard.
3. The element $A$ is dually distributive.
4. The element $A$ is right modular.

We see that in a locally deterministic lattice there are several equivalent ways of defining the ordering of the irreducible elements. If the lattice is not locally deterministic these different definitions may not be equivalent. If the ordering $\preceq$ is translated into the notion of causation we see that notion of causation splits up in different concepts when we leave the simplest situations where any variable is determined by its parents.
5. INDEPENDENCE AND FUNCTIONAL DEPENDENCE

In the theory of Bayesian networks one studies the relation \((A /\bot B \mid C)\), where \(A, B\), and \(C\) are disjoint subsets of a set \(M\) of random variables. The power set of \(M\) is a Boolean lattice with inclusion as ordering, \(\cup\) as join, and \(\cap\) as meet. We shall say that a relation \(\langle \cdot \not\parallel \cdot \mid \cdot \rangle\), on a lattice \((L, \vee, \wedge)\), is a semi-graphoid relation, if it satisfies the following axioms:

**Symmetry**  \((X \not\parallel Y \mid W)\), if and only if \((Y \not\parallel X \mid W)\).

**\(I\)-Decomposition** If \(Y \subseteq Z\) then \((X \not\parallel Y \mid W)\) implies \((X \not\parallel Z \mid W)\).

**Weak Union** \((X \not\parallel Y \wedge Z \mid W)\) implies \((X \not\parallel Z \mid Y \wedge W)\).

**Contraction** \((X \not\parallel Y \mid W)\), and \((X \not\parallel Z \mid Y \wedge W)\) implies \((X \not\parallel Y \wedge Z \mid W)\).

These propositions should hold for all \(X, Y, Z, W \in L\). If \(L\) is the power set of a set and the relation \(\langle \cdot \not\parallel \cdot \mid \cdot \rangle\) is only defined for disjoint sets then the definition coincides with the definition given in [1].

In this paper we are also interested in the case where the subsets are not disjoint. In a power set of random variables we note that if \(A\) is independent of \(A\) given \(C\) then \(A\) is a function of \(C\) almost surely. Hence we introduce the following additional axioms that are fulfilled for random variables.

**Auto-independence** For all subsets \(A\) we have \((A \not\parallel A \mid A)\).

**Forced independence** For all subsets \(A, B\), and \(C\) we have that \((A \not\parallel A \mid C)\), implies \((A \not\parallel B \mid C)\).

A semi-graphoid relation is said to be super-graphoid if it satisfies auto-independence and forced independence.

If \((A \not\parallel A \mid C)\) we write \(C \subseteq A\) and say that \(A\) depends functionally on \(C\). We can use the basic properties of semi-graphoid relations to prove properties of functional dependence.

**Theorem 14** If \(\langle \cdot \not\parallel \cdot \mid \cdot \rangle\) is a super-graphoid relation on a lattice then \(\forall A\), satisfies Armstrong’s axioms.

**Theorem 15** If \((L, \vee, \wedge)\) is a lattice with a super-graphoid relation \(\langle \cdot \not\parallel \cdot \mid \cdot \rangle\), then this relation restricted to the lattice \(L_{\leq}^\ast\), is also super-graphoid.

The significance of this theorem is that if we start with a super-graphoid relation on a power set of random variables then this super-graphoid relation is also super-graphoid when restricted to the set of random variables that are closed under functional dependence.

6. DISCUSSION

Bayesian networks may be considered as deterministic Bayesian networks where some of the variables are hidden. Conditional entropy and conditional mutual information between variables can be interpreted as measures of the minimal entropy of such hidden variables. In this way the quantities of information theory become the right tools to study how much observed variables deviate from being modelled by a deterministic Bayesian network. This also leads to the study of Shannon and non-Shannon inequalities for entropy between random variables with a functional dependence structure. This is a new area because the entropy inequalities have so far only been studied on power sets of variables.

7. REFERENCES


MDL BASED INTERPRETATION FOR OVERLAPPING CELL NUCLEI IN HISTOLOGICAL IMAGES

Jenni Hukkanen\textsuperscript{1}, Edmond Sabo\textsuperscript{2}, and Ioan Tabus\textsuperscript{1}

\textsuperscript{1} Department of Signal Processing, Tampere University of Technology, P.O.Box 553, FIN-33101 Tampere, FINLAND, Jenni.Hukkanen@tut.fi
\textsuperscript{2} Department of Pathology, Rappaport Faculty of Medicine, Technion, Haifa, Israel, e.sabo@rambam.health.gov.il

ABSTRACT

In this extended abstract we present our current work concerning the development of minimum description length (MDL) principle based solutions for medical image segmentation problems. Our application is the interpretation of overlapping nuclei in histological images in terms of individual nuclei within a parametric shape family. MDL is a well suited approach for this application as MDL provides an efficient tool for comparing various competing geometrical structures composed of different numbers of elliptical shapes, each shape representing one nucleus. Our MDL criterion, developed for solving the problem of overlapping objects, involves different cost terms than the existing MDL criteria for image segmentation and in addition our resulting description is fully implementable, since it does not use asymptotic expressions for the involved codelengths. In experiments we have compared the results of automatic segmentation and human subject segmentations including the segmentation given by an expert pathologist.

1. INTRODUCTION

Histological images are 2D images taken from thin slices of tissue samples. They provide important information to expert pathologists for medical diagnosis and evaluation of the grade of the disease. The staining used in histological images is typically hematoxylin and eosin (H&E) staining, which highlights nuclei of the cells in the image. Unfortunately, the thickness of the tissue slice is in practice higher than a single nuclei layer, which results into clumps of overlapping nuclei in the two-dimensional image. Hence, the nuclei segmentation algorithm should not only segment well separated individual nuclei, but also should separate overlapping and occluding nuclei into individual ones.

Traditional segmentation algorithms are intended only for producing binary segmentation results, where the problem of overlapping nuclei is not resolved. There exist algorithms for splitting clumps of nuclei from binary segmentation. Unfortunately, binary segmentation results for H&E stained histological images can be noisy and unreliable.

The shape of the nuclei is in many cases almost elliptical and hence in our approach we interpret clumps of nuclei as unions of elliptical shapes such that each ellipse represents one nucleus. Our recently proposed SNEF algorithm \cite{1} is an ellipse fitting based algorithm for cell nuclei segmentation. It has a number of options, which may lead to different ellipse proposals for the interpretation. In \cite{2, 3} we proposed a minimum description length (MDL) principle \cite{4} based criterion for comparison between different interpretations involving different numbers and arrangements of ellipses.

2. METHODS

MDL provides a tool for comparison between different statistical models representing geometrical structures, since by using the MDL criterion one can choose that description, which best explains the data, for a given data set and class of models, providing a natural trade-off between the complexity of the model and the fitting of the data.

The MDL for image segmentation was first introduced in Leclerc \cite{5}. The main idea in \cite{5} and our approach \cite{2, 3} is the lossless description of the image to be segmented using a total codelength $L(Y; \Omega, \beta)$, involving the following terms:

$$L(Y; \Omega, \beta) = L(\Omega) + L(\beta|\Omega) + L(Y; \Omega, \beta),$$

where $L(\Omega)$ is the codelength for describing the contour, which splits the image into foreground and background, $L(\beta|\Omega)$ is the cost of describing the coding parameters in each of the regions, and $L(Y; \Omega, \beta)$ is the codelength for encoding the image given the contour and using the coding distributions. Since our problem is not the segmentation, but interpreting a region of possibly overlapping nuclei by ellipses, the resulting costs for the contour have a different form than in \cite{5}. We use Golomb-Rice codes for encoding the residuals, which is known to be efficient in the field of lossless image coding and provides a fully implementable coding algorithm.

In order to minimize locally the MDL criterion we have introduced an iterative algorithm for updating the parameters of ellipses \cite{2}. 

3. RESULTS AND CONCLUSIONS

We run the SNEF algorithm for each image four times, using different thresholds for binarization of gradients and intensity values, resulting in four alternative segmentations and then we compute for each segmentation the value of MDL criterion. For each image the best SNEF segmentation is taken as the one with the lowest value of MDL criterion. We compared the obtained best SNEF segmentation results to the segmentations provided by human subjects. The subjects were allowed to give multiple interpretations for each image and also to specify their belief towards the interpretation. The ground truth was computed as a weighted average of the subjects interpretations.

We found in [2] that our local iterative algorithm, which optimizes the parameters of ellipses in order to minimize the MDL, decreases the variability in the MDL value of the provided human interpretations. We also noticed that after the iterative algorithm the deviations of the MDL values obtained by the best SNEF with respect to the MDL of the ground truth are in general lower than two times the standard deviation of MDL for the human segmentations.

One can use before the SNEF algorithm different pre-processing stages, which can effectively protect against the artifacts in the original images and in [3] we studied the effects of preprocessing using smoothing by Gaussian filtering and rescaling at various downsizing scales. We noticed a good correlation between the highest MDL and the highest (supervised) similarity index, which measures the overlapping areas between the ground truth and the provided segmentation.

As a conclusion, the proposed MDL based criterion for comparison between different interpretations of cell nuclei in histological images offers a good selection tool, matching closely supervised criteria, which will require segmentations provided by human subjects.

4. REFERENCES


GENERALIZED NASH EQUILIBRIUM IMAGE DEBLURRING

Vladimir Katkovnik, Aram Danielyan and Karen Egiazarian

Department of Signal Processing, Tampere University of Technology, Tampere, Finland, e-mail: firstname.lastname@tut.fi.

ABSTRACT

We compare two different formulations of the deblurring problem: one (variational) is defined by minimization of a single objective function and another one is based on a generalized Nash equilibrium balance of two objective functions. The latter results in the algorithm where the denoising and deblurring operations are decoupled. For image modeling we use the recent BM3D-frames. Simulation experiments show that the decoupled algorithm derived from the generalized Nash equilibrium formulation and using BM3D-frames demonstrates the best numerical and visual results and shows superiority with respect to the state of the art in the field.

1. INTRODUCTION

Image restoration from blurry and noisy observations is considered. Assuming a circular shift-invariant blur operator and additive zero-mean white Gaussian noise the observation model is expressed as

\[ z = Ay + \sigma e, \]

where \( z, y \in \mathbb{R}^N \) are vectors representing the observed and true images, respectively, \( A \) is an \( N \times N \) blur matrix, \( e \sim \mathcal{N}(0_{N \times 1}, I_{N \times N}) \) is a vector of i.i.d. Gaussian random variables, and \( \sigma \) is the standard deviation of the noise. The deblurring problem is to reconstruct \( y \) from \( z \).

Image modeling lies at the core of image reconstruction problems. Recent trends are concentrated on sparse representation techniques, where the image is assumed to be defined as a combination of few atomic functions taken from a certain dictionary. It follows that the image can be parameterized and approximated locally or nonlocally by these functions. To enable sparse approximations, the dictionary should be rich enough to grasp all variety of images. Clearly, the classical orthonormal bases are too limited for this task, and one needs to consider overcomplete systems with a number of elements essentially larger than the dimensionality of approximated images. Frames are generalization of the concept of basis to the case when atomic functions are linearly dependent and form overcomplete systems. There is a vast amount of literature devoted to the sparsity based models and methods. An excellent introduction and overview of this area can be found in the recent book [1].

The block-matching 3D (BM3D) image denoising, originated in [2], is formalized in [3] and [4] in terms of the overcomplete sparse frame representation. The analysis and synthesis developed in BM3D are interpreted as a general sparse image modeling applicable to various image processing problems.

In this paper we discuss two different variational formulations of the image deblurring proposed in our recent papers [3] and [4]: single objective function optimization vs. fixed point of two objective functions (generalized Nash equilibrium). The latter approach results in the algorithm where denoising and deblurring operations are decoupled. It is shown by simulation experiments that the best image reconstruction both visually and numerically is obtained by the algorithm based on this decoupling. To the best of our knowledge, this algorithm provides results which are the state-of-art in the field.

1.1. BM3D-frame image modeling

It has been shown in [3] that provided a fixed grouping the BM3D analysis/synthesis can be given in the matrix form linking the image \( y \in \mathbb{R}^N \) and its groupwise spectrum vector \( \omega \in \mathbb{R}^M, M \gg N \), by the forward and backward transforms

\[ \omega = \Phi \cdot y, y = \Psi \cdot \omega. \]

It is proved in [3] that the matrices \( \Phi^T \Phi \) and \( \Psi \Psi^T \) are diagonal with positive items; \( \Phi = I_{N \times N} \). The last formula enables perfect reconstruction of the image \( y \) from its groupwise spectrum \( \omega \). It is shown also that \( \Phi \) and \( \Psi^T \) are full column rank matrices. The rows of \( \Phi \) constitute a frame in \( \mathbb{R}^N \), and the columns of \( \Psi \) constitute a frame in \( \mathbb{R}^N \) dual to \( \Phi \). These frames are non-tight, \( \Phi^T \cdot \Phi \neq \lambda I_{N \times N} \) and \( \Psi \Psi^T \cdot \Psi \neq \lambda I_{N \times N}, \lambda > 0 \). If the group weights in BM3D synthesis [2] are equal to 1, then \( \Psi = (\Phi^T \Phi)^{-1} \Phi^T \), however in general, \( \Psi \neq (\Phi^T \Phi)^{-1} \Phi^T \).

Once BM3D groups are defined, the operators \( \Phi, \Phi^T, \Psi \) and \( \Psi^T \) can be implemented efficiently since all of them perform groupwise separable 3D transforms. To build the groups the block matching (grouping) procedure from [2] is used. The BM3D-frames are nonlocal and data adaptive, which make them quite different from the other popular frames used for image modeling.

1.2. Variational formulations for deblurring

The analysis (using the analysis matrix \( \Phi \)) and synthesis (using the synthesis matrix \( \Psi \)) variational image reconstructions are conventional for an overcomplete image modeling [1]. For a Gaussian noise these reconstructions can be given in the form of constrained optimization, respectively, for analysis

\[ \hat{y} = \arg \min_y \left\{ \frac{1}{2\mu} ||z - Ay ||_2^2 + \tau \cdot ||\omega ||_p \right\}, \quad \omega = \Phi y, \]
and synthesis
\[
\hat{\omega} = \arg\min_\omega \left\{ \frac{1}{2\mu} \| z - Ay \|^2_2 + \tau \cdot \| \omega \|_p \mid y = \Psi \omega \right\}.
\] (4)

These two formulations are studied thoroughly in literature assuming that the frames are tight, \( \Phi^T \Phi = I_{N \times N} \), \( \Psi = \Phi^T \). The last formula says that the analysis matrix \( \Phi \) defines completely the synthesis one and vice versa. For the non-tight BM3D-frames these matrices do not define each other, and for image reconstruction we need to use both the analysis and synthesis operators.

In this way we arrive to the combined analysis/synthesis formulation of image reconstruction [4]:
\[
(\hat{\omega}, \hat{y}) = \arg\min_{\omega, y} \left\{ \frac{1}{2\mu} \| z - Ay \|^2_2 + \tau \cdot \| \omega \|_p \mid y = \Phi y, y = \Psi \omega \right\},
\] (5)
where the analysis and synthesis links between the image and spectrum are considered as constraints.

For \( p = 1 \) and \( p = 0 \) (3)-(5) are defining \( l_2-l_1 \) and \( l_2-l_0 \) optimization problems, respectively.

Let us replace the constraints in (3)-(5) by the quadratic assumptions (e.g. Theorem 6 in [5]). We see that we can say that \( \hat{\omega} \) is a solution if \( \hat{\omega} \) is a fixed point of the point-to-set mapping \( S \). If the feasible sets \( X_v(x^{-v}) \) do not depend on the rival players’ strategies, so we have \( X_v(x^{-v}) = X_v \) for some set \( X_v \subseteq \mathbb{R}^{n_v} \) and all \( v = 1, \ldots, N \), the GNEP reduces to the standard Nash equilibrium problem (NEP for short).

For non-empty and convex \( X_v(x^{-v}) \) the existence of the solution can be guaranteed provided some reasonable assumptions (e.g. Theorem 6 in [5]).

The sets \( X_v(x^{-v}) \) can be given by equality or inequality constraints, for instance, as
\[
X_v(x^{-v}) = \{ x^v : g_v(x^v, x^{-v}) \leq 0 \}.
\] (10)

Recall, that a Pareto vector optimization is different from the Nash equilibrium by a simultaneous optimization on \( x \). A vector \( \hat{x} \), \( \hat{x}^v \in X_v(\hat{x}^{-v}) \), is Pareto optimal, if there exists no other vector \( y \) such that
\[
f_v(y) \leq f_v(\hat{x}), \text{ for all } v = 1, \ldots, N
\] and \( f_i(y) < f_i(\hat{x}) \) for at least one \( i \), provided that \( y^v \in X_v(x^{-v}) \). It is known that the GNEP and Pareto optimization, in general, give different solutions.

1.4. Deburring as generalized Nash equilibrium problem

Let us formulate the deburring for observations (1) as the following GNEP problem [3]:
\[
\begin{align*}
\hat{y} &= \arg\min_y \frac{1}{2\mu} \| z - Ay \|^2_2 \text{ subject to } \| y - \Psi \hat{\omega} \|_2^2 \leq \varepsilon_1, \\
\hat{\omega} &= \arg\min_\omega \tau \cdot \| \omega \|_p \text{ subject to } \| \omega - \Phi \hat{y} \|_2^2 \leq \varepsilon_2,
\end{align*}
\] (11)
where \( \varepsilon_1, \varepsilon_2 > 0 \), and the inequality constraints \( \| y - \Psi \hat{\omega} \|_2^2 \leq \varepsilon_1 \) and \( \| \omega - \Phi \hat{y} \|_2^2 \leq \varepsilon_2 \) relax the equalities (2).

Two groups of variables \( y \) and \( \omega \) define two players in the formulation (9) with the corresponding objective and restriction functions. For the algorithm development we replace (11) by unconstrained optimization problems:
\[
\begin{align*}
\hat{y} &= \arg\min_y L_1(y, \omega) \\
\hat{\omega} &= \arg\min_\omega L_2(\hat{y}, \omega),
\end{align*}
\] (12)
where
\[ \mathcal{L}_1(\mathbf{y}, \omega) = \frac{1}{2\sigma^2} \| z - \mathbf{A}\mathbf{y}\|_2^2 + \frac{1}{2\gamma} \| \mathbf{y} - \mathbf{P}\omega\|_2^2, \]
\[ \mathcal{L}_2(\mathbf{y}, \omega) = \tau \cdot \|\omega\|_p + \frac{1}{2\kappa} \| \omega - \Phi\mathbf{y}\|_2^2. \] (14)

A solution \((\hat{\mathbf{y}}, \hat{\omega})\) of (12) is a fixed point or Nash equilibrium of the two objective functions \(\mathcal{L}_1\) and \(\mathcal{L}_2\). Minimization of the quadratic \(\mathcal{L}_1\) on \(\mathbf{y}\) results in a linear solution which is a regularized inverse of the blur operator \(\mathbf{A}\). Minimization of the non-quadratic \(\mathcal{L}_2\) on \(\omega\) results in a nonlinear hard- or soft-thresholding solution filtering noise.

### 1.5. IDD-BM3D algorithm

The proposed iterative algorithm follows from the alternating solution of (12):

\[
\begin{cases}
    y_{t+1} = \arg \min_{\mathbf{y}} \mathcal{L}_1(\mathbf{y}, \omega_t) \\
    \omega_{t+1} = \arg \min_{\omega} \mathcal{L}_2(\mathbf{y}_{t+1}, \omega) 
\end{cases} , \quad t = 0, 1, \ldots . \] (15)

We call this algorithm Iterative Decoupled Deblurring BM3D (IDD-BM3D). Details of the fast implementation of the algorithm using FFT calculations as well as the convergence statement can be seen in [3].

### 1.6. Simulation experiments

Here some test-results from [3] are presented. We consider six deblurring scenarios used as the benchmarks in many publications. The blur point spread function (PSF) \(h(x_1, x_2)\) and the variance of the noise \(\sigma^2\) for each scenario are summarized in Table 1.

![Table 1. Blur PSF and noise variance used in each scenario.](image)

In the experiments with the combined analysis/synthesis algorithm we optimize the parameters of the objective function (6). Nevertheless, the results obtained by this algorithm are not better and only close to those obtained by the analysis algorithm. The comparison is definitely in favor of the IDD-BM3D algorithm based on GNEP. We wish to note also that optimization of the parameters in the objective function (6) for the combined algorithm actually gives the results close to those which can be obtained using the Pareto optimization. Thus, the simulation results demonstrate also the advantage of GNEP versus the Pareto optimization.

The experiments with the IDD-BM3D algorithm can be reproduced using the Matlab program available as a part of the BM3D package¹.

### 1.7. Acknowledgement

This work was supported by the Academy of Finland: project no. 213462, 2006-2011 (Finnish Programme for Centres of Excellence in Research) and project no. 138207, 2011-2014, and by Tampere Doctoral Programme in Information Science and Engineering (TISE).

### 2. REFERENCES


¹http://www.cs.tut.fi/~foi/GCF-BM3D
Table 2. Comparison of the output ISNR [dB] of the proposed deblurring algorithms. Row corresponding to “Input PSNR” contain PSNR [dB] of the input blurry images. Blurred signal-to-noise ratio (BSNR) is defined as $10 \log_{10} \left( \frac{\text{var}(A_y)}{N\sigma^2} \right)$, where \( \text{var}() \) is the variance.

<table>
<thead>
<tr>
<th>Method</th>
<th>Scenario</th>
<th>Thresh.</th>
<th>Weights $g_r$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cameraman (256x256)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BSNR</td>
<td></td>
<td></td>
<td></td>
<td>31.87</td>
<td>25.85</td>
<td>40.00</td>
<td>18.53</td>
<td>29.19</td>
<td>17.76</td>
</tr>
<tr>
<td>Input PSNR</td>
<td></td>
<td></td>
<td></td>
<td>22.23</td>
<td>22.16</td>
<td>20.76</td>
<td>24.62</td>
<td>23.36</td>
<td>29.82</td>
</tr>
<tr>
<td>Synthesis</td>
<td>soft</td>
<td>unit</td>
<td></td>
<td>6.30</td>
<td>4.60</td>
<td>7.88</td>
<td>2.06</td>
<td>2.98</td>
<td>2.84</td>
</tr>
<tr>
<td>Analysis</td>
<td>soft</td>
<td>unit</td>
<td></td>
<td>7.88</td>
<td>5.75</td>
<td>9.22</td>
<td>3.00</td>
<td>3.67</td>
<td>3.92</td>
</tr>
<tr>
<td>IDD-BM3D</td>
<td>soft</td>
<td>adaptive</td>
<td></td>
<td>8.41</td>
<td>6.41</td>
<td>9.59</td>
<td>3.38</td>
<td>3.98</td>
<td>4.14</td>
</tr>
<tr>
<td>IDD-BM3D</td>
<td>hard</td>
<td>adaptive</td>
<td></td>
<td>8.85</td>
<td>7.12</td>
<td>10.45</td>
<td>3.98</td>
<td>4.31</td>
<td>4.89</td>
</tr>
</tbody>
</table>

Figure 1. Deblurring of the Cameraman image, scenario 3. From left to right and from top to bottom are presented zoomed fragments of the following images: original, blurred noisy, reconstructed by CGMK [6] (ISNR 9.15), L0-AbS [7] (ISNR 9.10), DEB-BM3D [8] (ISNR 8.34) and by proposed IDD-BM3D method (ISNR 10.45).
MDL MULTIPLE HYPOTHESIS TESTING

Enes Makalic and Daniel Schmidt

Centre for MEGA Epidemiology, The University of Melbourne, Level 3, 207 Bouverie Street, Carlton VIC 3053, Australia, {emakalic,dschmidt}@unimelb.edu.au

ABSTRACT

This paper examines the problem of simultaneously testing many independent multiple hypotheses within the minimum encoding framework. We introduce an efficient coding scheme for nominating the accepted hypotheses in addition to compressing the data given these hypotheses. This formulation reveals an interesting connection between multiple hypothesis testing and mixture modelling with the class labels corresponding to the accepted hypotheses in each test. An advantage of the resulting method is that it provides a posterior distribution over the space of tested hypotheses which may be easily integrated into decision theoretic post-testing analysis.

1. INTRODUCTION

Consider the problem of performing \( m \) hypothesis tests from \( m \) samples of data \( Y^m = (y_1, \ldots, y_m) \), where each sample \( y_i \in \mathcal{Y} \subset \mathbb{R}^n \) \((1 \leq i \leq m)\). It is assumed that there exist \( K \geq 2 \) candidate hypotheses under consideration for each test. In a standard frequentist approach to hypothesis testing, one has two candidate hypotheses \((K = 2)\) deemed the ‘null’ and alternative hypothesis respectively, and generally proceeds by performing \( m \) independent hypothesis tests. In order to determine whether a null hypothesis is rejected, one must also specify a significance level, \( \alpha > 0 \), which is often taken to be about \( \alpha = 0.05 \). The \( m \) tests yield \( p \)-values \( p = (p_1, \ldots, p_m) \) which may be ‘corrected’ for multiple testing using, say, Bonferroni-type procedures.

In contrast, minimum encoding methods, such as Minimum Message Length (MML) [1] and Minimum Description Length (MDL) [2, 3], treat all candidate hypotheses on an equal footing (that is, specification of the null and alternative hypotheses is not required) and can automatically determine a suitable significance level solely from the observed data. As an example, given \( m \) data sets of non-negative integers, it may be of interest to determine whether each data set was generated by a Poisson, geometric or a negative binomial distribution; that is, there are \( K = 3 \) candidate hypothesis for each test. To date, such methods have largely been applied to single hypothesis testing and nested model selection problems with great success [4]. It would be of interest if the minimum encoding approach could be extended to the problem of testing multiple hypotheses. This paper considers a minimum encoding approach to the the multiple hypothesis testing problem which can be framed as a special case of latent variable inference. In light of this, the evidence for each hypothesis is deemed to be a latent variable and is inferred from the data. An intriguing consequence of this approach is the close connection to the mixture modelling problem.

The minimum encoding approach to inference advocates choosing the hypothesis that most compresses the data as optimal. This has been formalised in the notion of universal models which are a practical approximation to Kolmogorov complexity [5]. A model \( \hat{p}(\cdot) \) is universal relative to a set of distributions \( p(\cdot|\theta) \) indexed by parameter vector \( \theta \in \Theta \subset \mathbb{R}^p \), if for all \( \epsilon > 0 \) there exists an \( n > 0 \) such that

\[
\max_{y^n \in \mathcal{Y}^n} \left\{ \frac{1}{n} \log \frac{p(y^n|\hat{\theta}_{\text{ML}}(y^n))}{\hat{p}(y^n)} \right\} \leq \epsilon \tag{1}
\]

where \( \hat{\theta}_{\text{ML}}(y^n) \) is the maximum likelihood estimator of \( \theta \) and \( \mathcal{Y}^n \subset \mathbb{R}^n \) is the data space. This implies that the relative difference between the optimal non-transmittable code, \( -\log p(y^n|\hat{\theta}_{\text{ML}}(y^n)) \), and the universal model approaches zero as \( n \to \infty \); this difference in code lengths is often referred to as \textit{regret} in the literature.

There exist a range of universal models in the literature, including the Normalized Maximum Likelihood [2, 6] code, the Minimum Message Length [1] codes and sequential codes [7]. Under suitable regularity conditions, the codelength for data \( y^n \) using a universal model \( \hat{p}(\cdot) \) satisfies

\[
-\log \hat{p}(y^n) = -\log p(y^n|\hat{\theta}_{\text{ML}}(y^n)) + \frac{k}{2} \log n + O(1) \tag{2}
\]

as \( n \to \infty \), where \( p > 0 \) is the number of free parameters. This is the well known Bayesian Information Criterion (BIC) [8]. For finite \( n \), the \( O(1) \) term can be arbitrarily large and can have significant effect on inference. Thus, the choice of universal model largely determines the \( O(1) \) term for the model class under consideration.

When applying minimum encoding procedures to a single test consisting of \( K \) competing hypotheses, one generally proceeds by determining the codelengths of the \( K \) candidate models, say \( I_k(y^n) \), and selecting the model with the shortest codelength as the best explanation of the data. Strictly, one also needs a preamble code stating
which of the $K$ models, say model $k$, is subsequently used to compress the data. Let $I(k)$ denote the length of the preamble code. The particular form of a preamble code induces a prior distribution over the support $\{1, \ldots, K\}$. When the number of competing hypotheses $K$ is small, or the models form a nested structure, using an uninformative uniform ‘prior’ distribution over the $K$ competing hypotheses (that is, $I(k) = \log K$) generally yields satisfactory results [4]. Model selection is then performed by finding the candidate model $k$ such that

$$
\hat{k} = \arg \min_k \{ I(k) + I_k(y^n) \} \quad (3)
$$

In words, the ‘accepted’ hypothesis is the model whose total codelength, which comprises the preamble code, $I(k)$, and the data code $I_k(y^n)$, is the shortest. Here, the codelengths $I_k(y^n)$ are assumed to be found by using any suitable universal model.

2. MULTIPLE HYPOTHESIS TESTING

A generalisation of the minimum encoding approach to testing $m$ independent hypotheses is now discussed. Let $k^m \in \{1, \ldots, K\}^m$ denote the set of accepted hypotheses

$$
k^m = \arg \min_{k^m} \left\{ \sum_{i=1}^m I(k_i) + I_{k_i}(y_i^n) \right\} \quad (4)
$$

Choosing a uniform preamble code for each $k_i$ (that is, $I(k_i) = \log K, 1 \leq i \leq m$), reduces to the single hypothesis testing procedure. While this choice of code expresses prior ignorance about which hypotheses are likely to be ‘true’, the resulting codelength is optimal only in the case that all $K$ hypotheses are equally likely to occur. In practice, we may expect that one hypothesis is more likely for example, the conventional ‘null’ hypothesis) rendering the uniform prior code inefficient in this setting.

We conjecture that a suitable preamble code for $k^m$ must: (1) attain shorter codelengths than the uniform prior for the majority of data $k^m \in \{1, \ldots, K\}^m$, (2) be invariant to relabelling of the candidate hypotheses, and (3) be invariant to permutations of the set $k$. Encoding the accepted hypotheses $k^m$ as data arising from a multinomial distribution with cell probabilities $\theta = (\theta_1, \ldots, \theta_K)$ satisfies all three requirements. In practice, the ideal cell probabilities are unknown and one may use a suitable universal model to compress the set $k^m$, obtaining a codelength $I(k^m)$ which is for almost all cases shorter than the naive codelength $m \log K$. Even in the worst case of maximum entropy (all $\theta$ being identical), for large $m$, by use of (2), the codelength $I(k^m)$ exceeds that of the naive codelength only by approximately $K/2 \log m$ nits.

To examine the behaviour of this multinomial prior we consider the case of two competing hypotheses (that is, $K = 2$). For $m$ sufficiently large, the codelength for $k$ is

$$
I_m(k) = -h_1 \log \left( \frac{h_1}{m} \right) - h_2 \log \left( \frac{h_2}{m} \right) + O(\log m) \quad (5)
$$

where $h_j = \sum_{i=1}^m \mathcal{I}(k_i = j)$ is the number of times hypothesis $j$ is chosen, and $\mathcal{I}(\cdot)$ is the indicator function. The codelength (5) is a symmetric concave function with two minima at $(h_1 = 0, h_2 = m)$ and $(h_1 = m, h_2 = 0)$, and a maximum at $h_1 = h_2$. Some new light may be shed on the behaviour of this prior by instead viewing the issue of variable selection in regression models as a problem of multiple hypothesis testing. Our requirement for invariance under relabelling implies that all hypotheses should be treated on the same footing in terms of a codelength for $k^m$; thus, there should be no difference between a small number of included regressors or a small number of omitted regressors. Of course, the inclusion of more regressors increases the complexity of the resulting model, but this is captured in the subsequent $\sum_{i=1}^m I_k(\cdot)$ codelengths.

A further argument for the merits of the multinomial coding can be drawn from the theory of algorithmic complexity. A so-called ‘universal’ prior, based on algorithmic complexity, can be defined over a set of strings as $\pi^*(k^m) \propto \exp(-K(k^m|M))$, where $K(k^m|M)$ is the algorithmic complexity of the data $k^m$ with respect to a computer $M$ (see, pp. 133–135). Such a universal prior is known to have strong theoretical properties; see for example, [9]. Enforcing the invariance restrictions results in a reduced set of possible programs executable by the machine $M$ with which to compress the data $k^m$, so that the algorithmic complexity is approximately proportional to the Shannon entropy of the data. The set of possible programs is restricted to include only those that assign the same codelength to all permutations of $k^m$. Thus, the chosen multinomial prior over the set of class allocations exhibits similar behaviour to the universal prior $\pi^*(\cdot)$.

3. MIXTURE MODELLING OF HYPOTHESES

The total codelength $I_k(Y^m, k)$ for all the datasets $Y^m = (y_1, \ldots, y_m)$ and the chosen hypotheses $k$ may be written in the following form:

$$
-\sum_{i=1}^m \sum_{j=1}^K \mathcal{I}(k_i = j) \log \theta_j + \sum_{j=1}^K \sum_{i=1}^m \mathcal{I}(k_i = j)I_j(y^n_i) = I_k(Y^m, k) \quad (6)
$$

where $\theta$ denotes the cell probabilities of a $K$-nomial distribution, and $I_j(y^n_i)$ is the codelength of dataset $y^n_i$ coded using hypothesis $j$. Strictly, the term $\log 1/\Gamma(K + 1)$ should be added to (6) to account for the fact that the labelling of the hypotheses is arbitrary. This encoding is known as the ‘hard assignment’ codelength in the literature as each dataset is encoded by only one of the $K$ candidate hypothesis.

Examining (6) reveals a close connection with the problem of mixture modelling where the hypothesis chosen to compress each data set is viewed as a latent variable. Thus, for a given a set of competing hypotheses, the multiple hypothesis testing problem is synonymous with estimating class labels of a mixture model in which the $K$ hypotheses take the role of the ‘classes’, the indicators $k^m$.
are the class labels and the $y^n_i$ are the ‘data points’. Importantly, treating the data set codelengths as ‘likelihoods’ and maximising (4) is identical to maximum likelihood estimation of class labels in a regular $K$-component mixture model which is known to suffer from problems of inconsistency [1]. In multiple hypothesis testing, marginalising the class labels is not an option because the primary reason for doing the testing disappears with the marginalisation.

### 3.1. Partial Assignment of Hypotheses

The problem of inferring the class labels in a mixture model is synonymous with the allocation of data sets to hypotheses. That is, we can view the class label as a latent variable which determines the optimal hypothesis for a given data set and must be inferred from the available data. The codelengths for $I(k^m)$ and $I(Y^m|k^m)$ are optimal in isolation; this amounts to making an assumption that the data set and must be inferred from the available data. The cardinal rule in the MML and MDL principles is that no parameter should be stated critically depends on the codelengths assigned to the data by the competing hypotheses, so that the assumption of independence results in a code-word longer than could be formulated if the dependence is taken into account.

The reason for the inefficiency is that the hard assignment code assumes the class labels to be known with certainty. However, the class labels are unknown and are themselves being estimated from the data, taking on the role of nuisance parameters. The problem was first identified by Wallace, who subsequently introduced an ingenious way of optimally coding discrete parameters for the problem of regular mixture modelling ([1], pp. 275–295). The new encoding leads to a scheme where both the nuisance class labels as well as the mixture components and their parameters are simultaneously estimated from the data in a consistent fashion.

As an example, consider the task of coding data $y^n$ with two candidate hypotheses, say $I_1(\cdot)$ and $I_2(\cdot)$. Suppose that the difference in codelength when using hypothesis $I_1(\cdot)$ over the alternative hypothesis is small; that is, $I_1(y^n) - I_2(y^n) \leq \epsilon$, where $\epsilon > 0$ is a small. The hard assignment approach accepts hypothesis $I_2(\cdot)$ and ignores the fact that the alternative hypothesis results in a compression that is almost as good. Wallace’s ‘partial assignment’ procedure takes advantage of the fact that the candidate hypotheses yield similar codelengths in encoding the class labels (see [1] for details). Using partial assignment, the total codelength is

$$I_s(Y^m, k) = \sum_{i=1}^{m} \sum_{j=1}^{K} r_{ij} \log \frac{r_{ij}}{\theta_j} + \sum_{i=1}^{m} \sum_{j=1}^{K} r_{ij} I_s(y^n_i)$$

where

$$r_{ij} = \frac{\exp (-I_j(y^n_i)) \theta_j}{\sum_{q=1}^{K} \exp (-I_q(y^n_i)) \theta_q}$$

are the posterior probabilities of data set $y^n_i$ belonging to hypothesis $j$ for test $i$; note, as in Section 2, we omit the term $\log 1/\Gamma(K + 1)$.

The first term, $I_s(k^m)$, encodes the class labels (that is, assigns data to hypotheses) optimally based on the posterior probability of the data belonging to the candidate hypotheses. As such, the codelength $I_s(k^m)$ is always shorter than stating the class labels with absolute certainty as in the hard assignment approach. Conversely, the second term $I_s(Y^m|k^m)$ denotes the codelength of data $Y^m$ under the $K$ different hypotheses using mixture proportions $r_{ij}$. In hard assignment, since the class label is stated precisely, the codelength of the data, $I_h(Y^m|k^m)$, must necessarily be shorter than the corresponding partial assignment code. However, the total codelength $I_s(Y^m, k^m)$ is generally significantly shorter than $I_h(Y^m, k^m)$, unless the posterior probability of the class labels for each data set is (approximately) one.

**Remark 2:** Minimising $I_s(Y^m, k^m)$ over the class labels $k^m$, yields the probability of accepting each of the $K$ hypotheses for all $m$ tests. The posterior probability of a hypothesis being accepted can readily be used in a decision theoretic analysis. For example, suppose the tests determine whether a particular drug is to be approved based on $m$ possible side-effects of the drug. The seriousness of each side-effect can be assigned a utility (often of a monetary nature), and based on the posterior probabilities a decision theoretic analysis undertaken to determine whether the drug should be accepted.

### 3.2. Algorithm

To minimise the partial assignment codelength (7) the following Expectation-Maximisation (EM) [10] type algorithm can be used.

1. Initialise $\theta$ using, for example

$$\theta_j \leftarrow \frac{1}{m} \sum_{i=1}^{m} I(k_i = \arg \min_{q} \{ I_q(y^n) \})$$

for all $j = 1, 2, \ldots, K$. This is the rate of acceptance of hypothesis $j$ when the effects of multiple testing are not taken into consideration and is equivalent to using a uniform code over the class labels.

2. Update the posterior probabilities using (8) given the mixing proportions $\theta$.

3. Re-estimate $\theta$ by

$$\theta_j \leftarrow \frac{1}{m} \sum_{i=1}^{m} r_{ij}, \ (1 \leq j \leq K).$$

4. Repeat steps (2)–(4) until convergence.
Preliminary empirical testing shows that the algorithm is not particularly sensitive to the initial choice of $\theta$; simulation suggestions that setting $\theta_i = 1/K \ (1 \leq i \leq K)$ converges to the same solution at roughly the same rate as the initialisation using (9).

3.3. Application Example

A Genome Wide Association Study (GWAS) involves determining whether there exists any genetic association with observable traits; for example, testing whether a particular genetic variant increases the risk of cancer. In the case considered here, one observes a binary data vector $x^n \in \{0,1\}^n$ that denotes the presence of the trait of interest (also known as the phenotype) in each of the $n$ people, and a matrix $G \in \{0,1\}^{(n \times m)}$ of measured genetic information (genotypes) for each person. Each of the $m$ columns of the matrix denotes the presence of a genetic mutation at a particular locus in the DNA (known as a single nucleotide polymorphism, or SNP). The aim is to determine whether there exists any association between the $m$ SNPs and the observed data (phenotype) vector $x^n$.

Under the usual assumption of independence between SNPs, the standard approach is to perform $m$ independent tests of association. For each test, one constructs a $2 \times 2$ contingency table from $x^n$ and $G$ with entries $y = \{y_{11}, y_{12}, y_{21}, y_{22}\}$, where the sum of all the entries in a contingency table is equal to $n$. Given a contingency table, one computes a test statistic, such as $\chi^2$ or Fisher’s exact test, and decides that the SNP is associated with the phenotype if this statistic is sufficiently large. An alternative approach based on minimum encoding is to apply the method of Section 3. In a slight shift from the usual way of viewing the problem, we look to test whether the genotype is dependent on the phenotype by compressing the data under two different hypotheses ($K = 2$): (1) the genotype is independent of the phenotype and the data may be compressed concisely by two binomial distributions, and (2) the genotype is dependent on the phenotype, and the data is best compressed using one quanodinal distribution.

Formally, let $\phi_{ij}$ denote the probability of each cell in a contingency table, such that $\sum_{ij} \phi_{ij} = 1$ for $i, j = 1, 2$, and consider a sampling scheme in which only the sample size $n$ is fixed. Under the independence assumption, $\phi_{ij} = p_i q_j$ where $\sum_{i=1}^{2} p_i = \sum_{j=1}^{2} q_j = 1$ and the contingency table can be compressed using the two parameter universal model for the (constrained) multinomial distribution

$$\binom{n}{y} p_{11}^{y_{11}+y_{12}} (1-p_1)^{y_{21}+y_{22}} q_1^{y_{11}+y_{21}} (1-q_1)^{y_{12}+y_{22}}.$$

If the genotype depends on the phenotype vector, the contingency table may instead be compressed using a universal model for the (unconstrained) quanodinal distribution. In both cases, one may use the the Normalised Maximum Likelihood (NML) universal model [2] for which codelengths can be computed in $O(n)$ time using the clever algorithm of [11]; an accurate approximation is given in [12].

The algorithm in Section 3.2 may then be used to minimise the mixture codelength (7) and determine which SNPs are associated with the phenotype. Given the resulting posterior probabilities, one could choose to accept that there is association for SNP $j$ if the corresponding posterior probability is greater than $0.5 \ (1 \leq j \leq m)$; otherwise, we accept the hypothesis that the genotype and phenotype for SNP $j$ are independent.

4. REFERENCES

TIME INDEPENDENT SWITCHING BETWEEN TWO SOURCE CODES

Steven de Rooij1 and Wouter Koolen1,2

1Centrum Wiskunde & Informatica (CWI)
Science Park 123, P.O. Box 94079, 1090 GB Amsterdam, Netherlands. Email: s.de.rooij@cwi.nl
2Royal Holloway, University of London, Department of Computer Science
Egham, Surrey, TW20 0EX, United Kingdom. Email: wouter@cs.rhul.ac.uk

ABSTRACT

Switching algorithms for combining sequential codes usually come with bounds, relating their performance to that of the best fixed strategy that switches $m$ times. The overhead is parameterized by the number of outcomes $n$ and the number of switches $m$; it may be difficult to judge whether this is acceptable. We present an alternative algorithm whose overhead does not depend on $n$; instead our bound expresses the overhead in terms of the benefit of switching in the first place.

1. INTRODUCTION

In 1998, Paul Volf and Frans Willems published an algorithm that runs two separate source coding algorithms, say, the Context Tree Weighting and LZ77 data compression schemes, in parallel, determining on-line how their predictive models should be combined so that any part of the data is compressed almost as well as if we used the best of the two algorithms for it [1].

Around the same time, many related results on “expert tracking” appeared in the online learning community. A general overview is provided in [2]; publications most relevant to this work are [3, 4, 5, 6].

For all these approaches, performance guarantees of the following form are given. Let $L_r(x^n)$ denote the number of bits a code $x$ uses to encode outcomes $x^n = x_1, \ldots, x_n$. Fix two codes $A$ and $B$ with length functions $L_A$ and $L_B$, and let $L_d$ denote the number of bits used by the algorithm that switches between $A$ and $B$ at fixed times $t = t_1, \ldots, t_m$. For all tracking algorithms in the listed references, bounds on the maximal performance difference are given. For example, for Volf and Willems’ algorithm, one may show that

$$L_{\text{run}}(x^n) - L_d(x^n) \leq nH(m/n) + \frac{1}{2} \log n + 3,$$  \hspace{1cm} (1)

where $H(p) = -p \log p - (1-p) \log(1-p)$ is the binary entropy function. This bound can be interpreted as the number of bits required to encode where the switches occur.

Being valid for any time sequence $t$ and data sequence $x^n$, performance guarantees such as (1) are very robust: no assumptions about the data, stochastic or otherwise, are required. However, it can be difficult to relate the overhead expressed by the bound to the benefit of switching in the first place. For example, suppose that $A$ starts out a little better than $B$ — there is an $n_0$ such that $L_A(x^{n_0}) < L_B(x^{n_0}) - C$; but $B$ is better than $A$ in the long run — $L_B(x^n) \ll L_A(x^n)$ from some $n$ onwards. Then we can improve our code length by $C$ bits by switching once from $A$ to $B$ at the appropriate time, but if we use Volf and Willems’ algorithm to do so, this improvement is dwarfed by the maximal overhead of $(3/2) \log n + O(1)$, so the bound does not tell us whether or not using the algorithm would be wise.

In our paper “Switching Investments” [7] we develop an alternative strategy for which, roughly speaking, the maximal impact of each pair of switches on the regret is measured in terms of the benefit of having those two switches in the first place. (The paper is presented in terms of investment strategies, but the setting is equivalent to coding, as explained in Section 4.) We will now look at an example to explain the rough idea, before discussing our algorithm and its regret bound in more detail in the last two sections.

2. EXAMPLE

We have created some fictitious code lengths for coding 4000 outcomes with $A$ and $B$. The solid line in Figure 1 shows the difference $\Lambda(i) = L_A(x^i) - L_B(x^i)$ as a function of the number $i$ of processed outcomes. On the whole code $B$ clearly outperforms code $A$, but there are substantial intervals where the situation is reversed. Ideally, we would switch every time $\Lambda$ changes direction, but that is...
clearly not feasible. But it looks as if it might be worthwhile to switch from one code to the other at the points indicated by the dashed black line, which can be thought of as a regularisation, or smoothing, of the actual behaviour of the code length difference.

With (1) we can bound the overhead in terms of the numbers on the horizontal axis, which may or may not have anything to do with the sizes of the fluctuations in code length. Our main idea is to parameterize the reference strategy not by its switching times \( t \), but instead by a sequence \( \delta = \delta_1, \ldots, \delta_m \) specifying the number of bits that \( \Lambda \) must increase or decrease between subsequent switches, as measured along the vertical axis. Note that we can identify any sequence of local extrema using such a sequence, and it provides the information we need to switch at the appropriate times as we process the data sequentially. The strategy that corresponds to the dashed line has \( \delta = 42, 6, 46, 14, 44, 55, 33, 13, 13 \).

### 3. REGRET BOUND

Our algorithm, with code length function denoted \( L_m \), uses a fixed prior density function \( \pi \) (that has to satisfy some mild restrictions). We prove the following bound for all \( \delta \), including the one highlighted in Figure 1:

\[
L_m(x^n) - L_\delta(x^n) \leq \sum_{i=1}^m \left( -\log \pi(\delta_i) \right) + (m-1)\alpha + \beta, \tag{2}
\]

where \( m \) is the length of \( \delta \) and \( \alpha \) and \( \beta \) are small constants that depend on the used prior density function (for example, for \( \pi(x) = 3(x + 3)^{-2} \) we have \( \alpha = 0.042 \) bits per switch and \( \beta = 6.13 \) bits). Note that the second term is constant per switch, and the last two terms are additive constants; the main contribution to the regret comes from the first term, which can now be interpreted as the number of bits required to encode \( \delta \) (to an appropriate precision).

Comparing (2) to (1), we find that all dependence on the sample size has disappeared; instead the overhead is expressed in terms of the sizes of the fluctuations of the code length difference. If we use a fat-tailed prior, the overhead is only logarithmic in the fluctuation size, which means that if the fluctuations are sufficiently large our switching approach is certain to yield an improvement over \( A \) and \( B \).

### 4. THE ALGORITHM

The algorithm is easiest to explain as the implementation of a Bayesian prediction strategy. By the Kraft inequality, each code length function has an associated probability distribution such that the \( \log \) of the probability of a sequence of outcomes is equal to the code length. Thus, our results can be stated either in terms of coding, or in terms of prediction with logarithmic loss. In the paper we use a third formalism, that of online investment — which turns out to be equivalent yet again.

Now that we think about the set of reference codes a set of random processes \( \mathcal{M} = \{ P_\delta \mid \delta \in \Delta \} \), where \( \Delta = [0, \infty)^\infty \) is the set of all infinite sequences of positive differences, we can define a universal model by putting a prior distribution \( \pi \) on the elements of \( \mathcal{M} \). The marginal probability of a sequence of outcomes becomes

\[
P_u(x^n) = \int_\mathcal{M} \pi(\delta) P_{\delta}(x^n) \, d\delta,
\]

and as usual we can predict the next outcome by conditioning

\[
P_u(X_{n+1} = x \mid x^n) = \frac{P_u(x^n, X_{n+1} = x)}{P_u(x^n)}.
\]

(From this probabilistic setting we can get back to coding using e.g. arithmetic coding.) It is not immediately clear that these predictions can be evaluated efficiently, but if the prior \( \pi \) is a product distribution \( \pi(\delta) = \prod_i \pi(\delta_i) \), then the height and direction of the last switch is a sufficient statistic and we need to maintain only a linear number of weights in the prediction process, reducing the running time to \( O(n) \) per outcome. In fact, if the density \( \pi \) is chosen to be exponential, even more weights can be lumped together and the running time is further reduced to amortized \( O(1) \) per outcome.

### 5. REFERENCES


ABSTRACT

Insurance transfers losses associated with risks to the insurer for a price, the premium. Considering a natural probabilistic framework for the insurance problem, we derive a necessary and sufficient condition on loss models such that the insurer remains solvent despite the losses taken on. In particular, there need not be any upper bound on the loss—rather it is the structure of the model space that decides insurability.

Insurance is a way of managing losses associated with risks—for example, floods, network outages, and earthquakes—primarily by transferring risk to another entity—the insurer, for a price, the premium. The insurer attempts to break even by balancing the possible loss that may be suffered by a few (risk) with the guaranteed payments of many (premium).

In 1903, Filip Lundberg [1] defined and formulated this scenario in its natural probabilistic setting as part of his thesis. In particular, Lundberg formulated a collective risk problem pooling together the risk of all the insured. There is an underlying risk model—a probability measure on loss sequences. Typically, the model itself is unknown, but can be imagined to belong to a known class of risk models. Suppose the insurance company sets a premium to be paid by the insured regularly—say, once at a time prior to entering the insurance game. However, we require the insurer enters the game with probability 1 no matter what loss models are in force, and the insurer can’t quit once entered.

A preliminary question is, then, what are necessary and sufficient conditions for a class of measures on infinite loss sequences to be insurable? In this paper, we provide a partial answer. If losses can be modelled as i.i.d. samples from a set of distributions we determine a necessary and sufficient condition on the set of distributions for insurability.

We adopt the collective risk approach, namely, we abstract the problem without loss of generality to include just two players in the insurance game—the insured and the insurer. We denote the sequence of losses by \( \{X_i\}_{i \geq 1} \), and we assume that \( X_i \in \mathbb{N} \) for all \( i \geq 1 \), where \( \mathbb{N} \) denotes the set of natural numbers, \( \{0, 1, 2, \ldots\} \). \( \mathcal{P}^\infty \) is a collection of measures on infinite length loss sequences. In this paper, we deal with only i.i.d. measures. Consequently, we denote by \( \mathcal{P} \) the set of distributions on \( \mathbb{N} \) obtained as single letter marginals of \( \mathcal{P}^\infty \).

Let \( \mathbb{N}^* \) be the collection of all finite length strings of natural numbers. The insurer’s scheme \( \Phi \) is a mapping from \( \mathbb{N}^* \rightarrow \mathbb{R}^+ \), and is interpreted as the premium demanded by the insurer from the insured after a loss sequence is observed. The insurer can observe the loss for a time prior to entering the insurance game. However, we require the insurer enters the game with probability 1 no matter what loss models are in force, and the insurer cannot quit once entered.

We adopt another abstraction without loss of generality: at any stage if the insurer is surprised by a loss bigger than the premium charged in that round, the insurer...
goes bankrupt. To see why this simplification does not involve any loss of generality, imagine the sequence of premiums set in the paper to represent the cumulative premium thus far.

To eliminate trivial schemes that do not enter the game at all, we require that for all \( p \in \mathcal{P} \), the insurer enters the game with probability 1.

A class \( \mathcal{P}^\infty \) of measures is insurable if \( \forall \eta > 0 \), there exists a premium scheme \( \Phi \) such that \( \forall p \in \mathcal{P}^\infty \), \( p(\Phi \text{ goes bankrupt}) < \eta \) and if, in addition, for all \( p \in \mathcal{P}^\infty \), \( \lim_{n \to \infty} p(\{X^n : \Phi(X^n) < \infty\}) = 1 \).

In Section 2, we consider an example of each of insurable and non-insurable classes.

1. RESULTS

We model the loss at each time by numbers in \( \mathbb{N} = \{0, 1, \ldots\} \). A loss distribution is a distribution over \( \mathbb{N} \), and let \( \mathcal{P} \) be a set of loss distributions. \( \mathcal{P}^\infty \) is the collection of i.i.d. measures over infinite sequences from \( \mathbb{N} \) such that the set of marginals over \( \mathbb{N} \) they induce is \( \mathcal{P} \). We call \( \mathcal{P} \) the set of single letter marginals of \( \mathcal{P}^\infty \). Each \( p \in \mathcal{P} \) is assumed to have finite support, and the span of \( p \in \mathcal{P} \) is the highest number which has probability \( > 0 \) under \( p \).

An insurer’s scheme \( \Phi \) is a mapping from \( \mathbb{N}^* \to \mathbb{R}^+ \), and is interpreted as the premium demanded by the insurer from the insured after a loss sequence is observed. For convenience, we assume \( \Phi(x^n) = \infty \) on every sequence \( x^n \) of losses on which \( \Phi \) has not entered.

Note however that the support over all distributions \( p \in \mathcal{P} \) of \( p \) need not be bounded. Thus, we do not assume an upper bound on the possible loss.

The crux of insurability is this: we would like close distributions to be similar in their span. We first define what distributions are close, followed by what distributions have “similar” span. We will then specify the necessary and sufficient conditions for insurability.

1.1. Close distributions

Insurability of \( \mathcal{P}^\infty \) depends on the neighborhoods of the probability distributions among its single letter marginals \( \mathcal{P} \). The relevant “distance” between distributions in \( \mathcal{P} \) that decides the neighborhood is

\[
J(p, q) = D\left(p\|\frac{p + q}{2}\right) + D\left(q\|\frac{p + q}{2}\right).
\]

1.2. Cumulative distribution functions

In this paper, we phrase the notion of similarity in span in terms of the cumulative distribution function. Note that we are dealing with distributions over a discrete (countable) support, so a few non-standard definitions related to the cumulative distribution functions need to be clarified.

For our purposes cumulative distribution function of any distribution \( p \) is a function from \( \mathbb{R} \to [0, 1] \), and will be denoted by \( F_p \). We obtain \( F_p \) by first defining \( F_p \) on points in the support of \( p \) and the point at infinity. We define \( F_p \) for all other points by linearly interpolating between the values in the support of \( p \).

Let \( F_p^{-1}(1) \) be the smallest number \( y \) such that \( F_p(y) = 1 \), and let \( F_p^{-1}(x) = 0 \) for all \( 0 \leq x < F_p(0) \). Note that for \( 0 \leq x \leq 1 \), \( F_p^{-1}(x) \) is now uniquely defined.

Two technical observations are in order since we are dealing discrete distributions. Consider a distribution \( p \) with support \( A \subset \mathbb{N} \). For \( \delta > 0 \), let \( (T \text{ for tail}) \)

\[
T_{\delta} = \{y \in A : y \geq F_p^{-1}(1 - \delta)\},
\]

and let \( (H \text{ for head}) \)

\[
H_{\delta} = \{y \in A : y \leq 2F_p^{-1}(1 - \delta/2)\}.
\]

It is easy to see that

\[
p(T_{\delta}) > \delta \text{ and } p(H_{\delta}) > 1 - \delta.
\]

Suppose, for some \( \delta \), \( F_p^{-1}(1 - \delta) > 0 \) and the premium is set to \( F_p^{-1}(1 - \delta) \), the probability under \( p \) of the loss exceeding the premium is \( \geq \delta \). If the premium is set to \( 2F_p^{-1}(1 - \delta/2) \), the probability that the loss exceeds the premium is \( \leq \delta \). We will use these observations in the proofs to follow.

1.3. Necessary and sufficient conditions for insurability

Existence of close distributions with very different spans is what kills insurability. A scheme could be “deceived” by some process \( p \in \mathcal{P}^\infty \) into setting low premiums, while a close enough distribution lurks with a high loss. The conditions for insurability of \( \mathcal{P}^\infty \) are phrased in terms of its single letter marginals \( \mathcal{P} \).

Formally, a distribution \( p \) in \( \mathcal{P} \) is deceptive if \( \forall \epsilon > 0, \exists \delta > 0 \) so that no matter what function \( f : \mathbb{R} \to \mathbb{R} \) is chosen, \( \exists \) a bad distribution \( q \in \mathcal{P} \) such that

\[
J(p, q) \leq \epsilon
\]

and

\[
F_q^{-1}(1 - \delta) > f(F_p^{-1}(1 - \delta)).
\]

2. EXAMPLES

The set \( \mathcal{N}^\infty \) is the class of i.i.d. processes whose single letter marginals have finite moment. Namely, \( \forall p \in \mathcal{N}^\infty \), \( E_pX_1 < \infty \).

Theorem 1. \( \mathcal{N}^\infty \) is not insurable.

Proof Note that the loss measure that puts probability 1 on the all-0 zero sequences exists in \( \mathcal{N}^\infty \). Since we consider only schemes that enter with probability 1 no matter what \( p \in \mathcal{N}^\infty \) is in force, every insurer must therefore enter after seeing a finite number of zeros.

Fix any scheme. Denote the premiums charged at time \( i \) by \( \Phi(X^i) \). Suppose the scheme enters the game after seeing \( N \) losses of size 0. To show that \( \mathcal{N}^\infty \) is not insurable, we show that \( \exists \eta > 0 \) such that for all schemes \( \Phi \), \( \exists p \in \mathcal{N}^\infty \) such that

\[
p(\Phi \text{ goes bankrupt}) \geq \eta.
\]
Fix some $\delta = 1 - \eta$. Let $\epsilon$ be small enough that
\[(1 - \epsilon)^N > 1 - \delta/2,\]
and let $M$ be a number large enough that
\[(1 - \epsilon)^M < \delta/2.\]
Note that since $1 - \delta/2 \geq \delta/2$, $N < M$.

Let $L$ be greater than any of premiums charged by $\Phi$ for the sequences $0^N, 0^{N+1}, \ldots 0^M$. Let $p \in N^\infty$ satisfy, for all $i$,
\[p(X_i) = \begin{cases} 1 - \epsilon & \text{if } X_i = 0 \\ \epsilon & \text{if } X_i = L. \end{cases}\]

For the process $p$, the insurer is bankrupted on all sequences that contain loss $L$ in between the $N$'th and $M$'th step. The sequences in question have probabilities (under $p$)
\[(1 - \epsilon)^N\epsilon, (1 - \epsilon)^{N+1}\epsilon, \ldots, (1 - \epsilon)^{N+M-1}\]
and they also form a prefix free set. Therefore, summing up the geometric series and using the assumptions on $\epsilon$ above,
\[p(\Phi \text{ is bankrupted}) \geq 1 - \delta/2 - \delta/2 = \eta.\]

One can verify that every distribution in $N^\infty$ is deceptive.

A monotone distribution on numbers satisfies for all $i$, probability of $i \geq$ probability of $i + 1$. Let $M^\infty$ be the set of all monotone i.i.d. loss processes with finite support. It will follow from Section 3 that

**Theorem 2.** $M^\infty$ is not insurable.

The above results mean that while insurability seems related to weak compressibility [8], it is not identical.

Consider $U$, the collection of all uniform distributions over a finite support of form $\{m, \ldots M\}$, with $m$ and $M$ being arbitrary. Let the losses be sampled i.i.d. from one of the distributions in $U$—call these processes $U^\infty$.

**Theorem 3.** $U^\infty$ is insurable.

**Proof** If the threshold probability of ruin is $\eta$, set the premiums $\Phi$ as follows. For all sequences $\pi$ with length $\leq \log \frac{1}{\eta} + 1$, $\Phi(\pi) = \infty$. For all sequences longer than $\log \frac{1}{\eta} + 1$, the premium is twice the largest loss observed thus far. It is easy to see this scheme is bankrupted with probability $\leq \eta$.

**3. NECESSARY AND SUFFICIENT CONDITION FOR INSURABILITY**

Note that according to the conventions adopted with defining cumulative distribution functions in Section 1.2, if for a sequence $x$, $F^{-1}_\eta(1 - \delta) > \Phi(x)$, the scheme $\Phi$ will be bankrupted with probability $\geq \delta$ in the next step.

$P^\infty$ is a set of i.i.d. measures over infinite sequences from $N$, and let $P$ denote the collection of their single letter marginals.

**Theorem 4.** $P^\infty$ is insurable iff no $p \in P$ is deceptive.
MDL SEGMENTATION AND LOSSLESS COMPRESSION OF DEPTH IMAGES

Ionut Şchiopu and Ioan Tăbus

Department of Signal Processing, Tampere University of Technology, P.O.Box 553, FIN-33101 Tampere, FINLAND
ionut.schiopu@tut.fi and ioan.tabus@tut.fi

ABSTRACT
This paper proposes methods for lossless compression of depth images, where the intermediate stage of image segmentation produces a minimum description length (MDL) segmentation, which is realized so that the overall description of the regions obtained and of the residuals obtained over each region is minimized. The existing methods for image segmentation based on minimum description length normally consider prediction over the regions by using planar or second order models. Differently, in here nonlinear prediction is utilized in each region, making it possible to achieve compression ratios much better than the current standards for lossless compression. The standard lossless compression methods, which were designed for natural images, prove to not be the most effective way to encode depth images, because depth images are more redundant and have different regularities than natural images. The newly proposed technique reduces the size of the compressed files in average to 55% of the standard JPEG-LS for a wide range of depth image material.

1. BACKGROUND TO MDL SEGMENTATION
Finding the segmentation of an image based on the minimum description length was first considered in [2], where the optimization of the MDL criterion was done through a costly continuation method. A much faster and more intuitive optimization process was introduced later by using region merging segmentation driven by MDL criteria in [1][3], where multichannel images were also considered. The lossless description of the image is realized by encoding four items: 1) the contours of the regions which make up the segmentation description; 2) the parameters of the polynomial models over each region; 3) the parameters of one multivariate (in the case of multichannel image) Gaussian model for the residuals over each region; and 4) the residuals over each region, encoded based on the statistical models specified at item 3). This representation of the image has the significance of decomposing the original image in a cartoon like image where each region has a given color, plus a residual image, which will include texture and random noise as well. The fact that the residual image is modeled over each region only as a multivariate Gaussian will leave unexploited a lot of redundancies present in regular textures or other local regularities inside regions, and as a result the overall compression of the above scheme is inferior to most lossless compression schemes. A similar approach was recently used for the lossy compression of depth images in [4]. We present in this paper a different scheme, where the modeling of the regions is done by more efficient predictive tools and as a consequence the compression obtained with our scheme is competitive and even exceeding by a large margin the performance of the best lossless coders. At a conceptual level, finding the segmentation which minimizes the description length in our scheme is really the minimum description length segmentation of the dept image. In the following we give a brief account of the algorithmic solutions involved and for a detailed description we refer to [5].

2. A NEW METHOD FOR MDL SEGMENTATION
We start by defining the costs for each region. A region is defined by the set Ω of pixel coordinates \((x,y)\) which belong to that region. The image graylevel at pixel \((x,y)\) is \(I(x,y) \in \{0,2^B - 1\}\), where the number of depth planes is usually \(B = 8\). We want to predict the depth \(I(x_t,y_t)\) at a current pixel \((x_t,y_t)\) by using the values \(I(x_i,y_i)\) at the pixels \((x_i,y_i)\) from a causal neighborhood \(N(x_t,y_t)\) of the pixel \((x_t,y_t)\). We note that this predictive principle is used in all competitive lossless image compression schemes. The shape and size of the neighborhood selected in different compression schemes varies quite much, the most simple shapes being those including only the west, north, north-west pixels. Here we adopt two scanning orders for the pixels in a neighborhood: horizontal, i.e. along the rows of the image, or vertical, i.e. along the columns of the image, and use over each region that scanning which gives the best results. The causal neighborhood for the horizontal scanning will have four pixels: W,NW,N, and NE, while the neighbors for the vertical scanning will be: W,NW,N, and SW.

2.1. Prediction
We want to define a suitable segmentation of the image, where any region \(Ω\) contains pixels with identical, or similar graylevel values, and the pixels outside \(Ω\) may have very different graylevel values. However a region may contain also pixels having the property that their values are well predictable using a given prediction method and a given prediction mask, while the pixels outside the re-
region are not anymore predictable. For this reason, the causal neighborhood $N(x_t, y_t)$ is restricted only to pixels belonging to $\Omega$, and all pixels from outside $\Omega$ are excluded. Close to the borders of the regions, the neighborhood $N(x_t, y_t)$ will not have exactly 4 pixels, it may have 3, 2, or even 1, zero even pixels. We need to define a suitable predictor function and we consider here $\hat{I}(x_t, y_t) = \text{median}(I(x_t, y_t)| (x_t, y_t) \in N(x_t, y_t) \cap \Omega)$, having the support with a varying number of pixels. The pixels having the intersection $N(x_t, y_t) \cap \Omega$ empty are collected in a set $\Delta$ and we use for them the prediction $\hat{I}(x_t, y_t) = \text{mean}\{I(x_t, y_t)| (x_t, y_t) \in \Delta\}$ which is encoded and transmitted for each region as a header.

2.2. Encoding of prediction residuals

We fix a scanning order of the pixels in the region $\Omega$ having $N$ pixels, and denote by $(x_t, y_t)$ the coordinates at location $t$ in the region, with $t = 1, \ldots, N$. We define the residuals $\epsilon(x_t, y_t) = I(x_t, y_t) - \hat{I}(x_t, y_t)$ for all pixels $(x_t, y_t) \in \Omega$ and denote the minimum and maximum residuals by $m$ and $M$. Both $m$ and $M$ will be encoded before encoding the region, with $m$ encoded uniformly in $\{-2^B + 1, 2^B - 1\}$, and $m_a = M - m$ encoded after that uniformly in $\{0, 2^B - m - 1\}$. Since $m$ and $M$ will be available also at decoder, we are going to encode $\epsilon(x_t, y_t) = \epsilon(x_t, y_t) - m \in \{0, M - m\}$. In case $m = M$ there is no need to encode the residual, for all pixels in the region we have $I(x_t, y_t) = m = M$.

In the case $m \neq M$ we need to encode the residuals $\epsilon(x_t, y_t)$, which will be done by using the adaptive distribution collected while encoding the residuals, in the agreed scanning order. Let denote the counts $n_i(i), i = 0, \ldots, M - m$, which tell how many times $\epsilon(x_t, y_t)$ was equal to $i$ for all $j \leq t$. The distribution of the residuals is tracked and used adaptively, so that at the current pixel $(x_t, y_t)$ both encoder and decoder have available the set of counts $n_{t-1}(i), i = 0, \ldots, M - m$, which tell the frequency of the symbols observed up to and including $t - 1$’th pixel. We can thus encode $\epsilon(x_t, y_t)$ using $L_t = - \log_2(n_{t-1}(\epsilon(x_t, y_t)) \sim \sum_{i} n_{t-1}(i))$. The overall codelength will add all such elementary codelengths over one region.

2.3. Encoding of region contours

The overall segmentation is formed of the contours separating the regions. There are a number of strategies for encoding these contours and their starting points. The order in which boundaries are transmitted will affect the number of starting points (which we also call anchors) and of ending points. We tested a number of heuristics and choose the one offering the lowest cost. We encode the vertex points and D. Steele, “A fast algorithm for MDL-based multi-band image segmentation,” in IEEE Conference on Computer Vi-

2.4. The overall segmentation method

With the costs as defined above, a MDL segmentation can be obtained by starting from an initial over-segmentation obtained e.g. as in [3] and then performing a very laborious sequential merging process, where two neighbor regions are merged if the overall code-length is better after merging. In order to accelerate the segmentation process we found a much faster procedure, probably suboptimal but still extremely efficient, summarized in the following: take as initial over-mentation of the image the split into regions depending on a variable called Threshold, defined as follows. A pixel will belong to a given region if the absolute value of the difference between the pixel depth and the depth of one of its 4-connectivity neighbors is smaller than Threshold. At the first step we find the regions with Threshold = 1 (they are constant regions) and keep the regions which are large enough as they are (no further merging is attempted). The smaller regions are collected together and a new split process takes place, allowing this time more variability inside regions, by increasing the Threshold to 2. Again, the regions which are large enough are kept unchanged. The process continues with Thresholds 3 and 4, and only the remaining small regions are checked if their merging produces improvements in the overall codelength. Finally, the encoding of remaining very small regions of up to 4 pixels is performed in a number of specific ways. The detailed processing is illustrated in Figure 5 and we send for details to [5]. The overall encoding strategy is summarized in Figure 1.

2.5. Experimental results

We illustrate the segmentation algorithm by the segmentation in Figure 4 for the depth image presented in Figure 3 (for completeness we also show the corresponding color image in Figure 2). The resulting MDL segmentation relies on more regions than a human will tend to associate to the image if he would intend to get only a sketchy cartoon of the image. However, the complete lossless representation of the image require such an over-segmented image in order to obtain really a minimum description of the whole depth image.

For illustrating the lossless compression performance we present results for a set of images from [8]. The comparison with the standard JPEG-LS [7] compressor (using the implementation provided in [9]) and with the PNG compressor (PNG being the format normally used for storing depth image in the public databases) is illustrated in Table 1, where compression factor (CF) is defined as compressed size over initial size, showing a very good performance of our encoder. All results are double checked for perfect reconstruction of the original file after decoding.

More comparisons in [5] for about 200 frames of two depth image sequences show that indeed the lossless compression performance of the presented scheme overpasses significantly that of commonly used standard lossless image compression methods.

3. REFERENCES

Table 1: Results and compression factors (CF) for the set of different depth images. With bold text are presented the best results.

<table>
<thead>
<tr>
<th>Image Name</th>
<th>Initial size (bits)</th>
<th>PNG size (bits)</th>
<th>CharLS size (bits)</th>
<th>Our new method result (bits)</th>
<th>CF PNG [%]</th>
<th>CF CharLS [%]</th>
<th>CF of our newmeth. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art</td>
<td>1370480</td>
<td>212704</td>
<td>12194736</td>
<td>107080</td>
<td>15.52</td>
<td>14.20</td>
<td>7.81</td>
</tr>
<tr>
<td>Books</td>
<td>1370480</td>
<td>152720</td>
<td>159560</td>
<td>104208</td>
<td>11.14</td>
<td>11.64</td>
<td>7.60</td>
</tr>
<tr>
<td>Dools</td>
<td>1370480</td>
<td>232496</td>
<td>215952</td>
<td>177840</td>
<td>16.96</td>
<td>16.11</td>
<td>12.97</td>
</tr>
<tr>
<td>Laundry</td>
<td>1323120</td>
<td>187112</td>
<td>174224</td>
<td>113768</td>
<td>13.16</td>
<td>12.35</td>
<td>8.60</td>
</tr>
<tr>
<td>Moebius</td>
<td>1370480</td>
<td>170272</td>
<td>193848</td>
<td>109424</td>
<td>12.42</td>
<td>11.74</td>
<td>7.98</td>
</tr>
<tr>
<td>Reindeer</td>
<td>1323120</td>
<td>174224</td>
<td>187112</td>
<td>113768</td>
<td>13.16</td>
<td>12.35</td>
<td>8.60</td>
</tr>
<tr>
<td>Lampshade</td>
<td>11544000</td>
<td>300561</td>
<td>611288</td>
<td>245768</td>
<td>2.60</td>
<td>2.59</td>
<td>2.13</td>
</tr>
</tbody>
</table>

Figure 1: The encoding strategy for depth image compression.


Figure 2: Reindeer: color image.

Figure 3: Reindeer: depth image.

Figure 4: Segmentation of image Reindeer

Figure 5: Image segmentation diagram.
A NEW MESSAGE LENGTH APPROXIMATION FOR PARAMETER ESTIMATION AND MODEL SELECTION

Daniel F. Schmidt
Centre for MEGA Epidemiology, The University of Melbourne, Melbourne, AUSTRALIA, dschmidt@unimelb.edu.au

ABSTRACT
This paper examines Bayesian two-part coding schemes as tools for parameter estimation and model selection. The Wallace–Freeman message length approximation to strict minimum message length can be used to obtain two-part message lengths. However, this approximation relies on some strong assumptions regarding the likelihood function and prior distribution which do not hold for a large range of models. We present a new two-part message length formula that is more widely applicable than the popular Wallace–Freeman message length approximation, while remaining significantly easier to compute than the exact strict minimum message length procedure.

1. MML TWO-PART CODES
Consider the problem of choosing a plausible explanation for some observed data \( y^n = (y_1, \ldots, y_n) \in \mathbb{Y}^n \subseteq \mathbb{R}^n \). The possible explanations are the distributions, or (fully specified) models, contained in a countable set of parametric model structures \( \gamma \in \Gamma \). Let \( p_\gamma(y^n|\theta) \) denote the model, in model structure \( \gamma \), indexed by \( \theta \in \Theta_\gamma \subseteq \mathbb{R}^{k_\gamma} \). The minimum encoding approach [1, 2] to inference suggests that the model that most compresses the data is the most plausible explanation. One way to compress the data is by two-part coding, in which the model and the data are compressed together as a two-part message. This idea is central to the minimum message length principle (MML). The MML principle is explicitly Bayesian in nature, so we further assume that a suitable prior distribution, \( \pi_\gamma(\theta) \), \( \theta \in \Theta_\gamma \), exists for all \( \gamma \in \Gamma \).

The first part of the message, or *assertion*, states which model, \( p_\gamma(\cdot|\theta) \), from the structure \( \gamma \), is to be used to compress the data. The second part, or *detail*, states the data \( y^n \) using the nominated model from \( \gamma \). Let the length of these two terms be denoted by \( I(\theta; \gamma) \) and \( I(y^n; \theta; \gamma) \), respectively. Further, let \( I(\gamma) \) denote the length of a preamble code stating which structure from \( \Gamma \) is being used. Estimation of both a model structure, as well as the model parameters, may be simultaneously performed by solving

\[
\left\{ \hat{\gamma}, \hat{\theta}_\gamma \right\} = \arg \min_{\gamma \in \Gamma, \theta \in \Theta_\gamma} \left\{ I(\gamma) + I(\theta; \gamma) + I(y^n; \theta; \gamma) \right\}.
\]

In the strict MML (SMML) procedure, the assertion and detail codes are constructed so that for a given structure \( \gamma \), the expected joint codelength is minimised, the expectation being taken with respect to the marginal distribution of the data. The optimisation problem implicit in this minimisation is in general NP-hard [3], and thus the procedure is impractical for all but the simplest of problems.

1.1. The Wallace–Freeman Codelength
Under suitable regularity conditions, Wallace and Freeman proposed an approximate codelength formula which we shall refer to as MML87 [4]. For a structure \( \gamma \) with \( k \) free parameters, the MML87 assertion and detail lengths for a model \( \theta \in \Theta_\gamma \), are

\[
\begin{align*}
I_{87}(\theta; \gamma) & = - \log \left( \frac{\pi_\gamma(\theta)}{|J_{\gamma}(\theta)|^{\frac{1}{2}}} \right) + \frac{k}{2} \log \kappa_k, \quad (1) \\
I_{87}(y^n; \theta; \gamma) & = - \log p_\gamma(y^n|\theta) + \frac{k}{2}, \quad (2)
\end{align*}
\]

where \( J_{\gamma}(\theta^*) = \mathbb{E}_\theta \left[ \left( \frac{\partial^2 \log p_\gamma(y^n|\theta)}{\partial \theta \partial \theta'} \right)_{\theta^* = \theta^*} \right] \) is the Fisher information matrix and \( \kappa_k \) is the normalised second moment of an optimal quantising lattice in \( k \) dimensions. Wallace has shown that if the local curvature of the prior distribution \( \pi_\gamma(\cdot) \) is “small” in comparison to the curvature of the negative log-likelihood, the MML87 codelength (1–2) is virtually indistinguishable (pp. 230–231, [1]) from the exact SMML codelengths.

The Wallace–Freeman approximation is computationally tractable. However, the accuracy of the approximation depends crucially on the behaviour of the likelihood function and the prior. If the Fisher information matrix is near singular, or the curvature of the prior is too great, the MML87 codelength can be a poor approximation to the SMML codelength. This paper introduces a new two-part codelength formula, named “MML08” after the year of its introduction [5], that is robust to these problems, while remaining significantly easier to compute than the exact SMML codes.

2. RANDOM CODING AND MMLD
We now discuss the MMLD approximation [6] that was specifically proposed to provide a more robust alternative.

---

1 We acknowledge that this use of the term “model” differs from much of the traditional statistical literature. This is done to keep the terminology in this paper consistent with the MML literature.
to the MML87 approximation, and which forms the basis for the new MML08 code length presented in Section 3. We provide a derivation of the MMLD code length that differs from the one in [1] (pp. 210–213), and discuss how data may be transmitted using a model structure \( \gamma \) without the need to perform a complete discretization of the parameter space \( \Theta \), by using Wallace’s ingenious procedure of random coding.

2.1. Random Coding

To transmit data \( y^n \) via random coding it is required that both the receiver and transmitter have access to a pseudo random number generator capable of sampling from the prior \( \pi_\gamma(\cdot) \), and that both generators are initialized with the same seed. The transmitter repeatedly samples models from the prior distribution until they generate one that lies inside a set \( S \subseteq \Theta_\gamma \). The transmitter then sends the number of draws required to arrive at the model, say \( d \), to the receiver using a universal code for the integers, with code length \( l^*(d) \). This is the message assertion. The receiver then makes \( d \) draws from their random number generator to arrive at the same parameter vector. The transmitter may then use this model, say \( \theta_d \), to send the data; this is the detail of the message. The total message length is then

\[
I(y^n, d, \theta_d; \gamma) = l^*(d) - \log p_{\gamma}(y^n | \theta_d).
\]

The length of the code required to transmit a string \( y^n \) using random coding is a random variable that depends crucially on the choice of \( S \). One wishes for the messages to be short on average and so \( S \) is chosen to minimise the average expected random coding message length, i.e.,

\[
\arg \min_{S \subseteq \Theta_\gamma} \mathbb{E}[l^*(d) - \log p_{\gamma}(y^n | \theta_d)] ,
\]

where the expectation is taken with respect to the random variables \( (d, \theta_d) \). The MMLD message length is found by approximating \( \mathbb{E}[l^*(d)] \) and then solving for the minimising set \( S \). This is detailed in the next section.

2.2. MMLD and Average Code lengths

Observe that the random variables \( d \) and \( \theta_d \) are independent; it thus suffices to find the expectations for both components of the random coding message length individually. Let

\[
q_\gamma(S) = \mathbb{P}(\theta \in S) = \int_S \pi_\gamma(\theta) d\theta
\]

be the probability that a model \( \theta \) sampled randomly from \( \pi_\gamma(\cdot) \) lies in \( S \). The number of draws, \( d \), required for a model to fall in \( S \) is a random variable following a geometric distribution with parameter \( q_\gamma(S) \). To transmit \( d \) to the receiver we use a universal code for integers, such as the log-star code [2], or Wallace tree code [1]. The log-star code length for integer \( d \) is

\[
l^*(d) = \log d + \log \log d + \ldots
\]

where the iterated logarithms continue until they become negative. Given that \( \mathbb{E}[d] = 1/q_\gamma(S) \) and \( \text{var}(d) = (1 - q_\gamma(S))/q_\gamma(S)^2 \), we can use the approximation \( \mathbb{E}[l^*(d)] = \log 1/q_\gamma(S) + O(\log \log 1/q_\gamma(S)) \). Using only the dominant term we arrive at the expression for the average length of the assertion of a random coding message based on the set \( S \),

\[
I(S; \gamma) = \log 1/q_\gamma(S).
\]

It remains to determine the average length of the detail. The distribution of \( \theta_d \), i.e., the first randomly generated model to lie in \( S \), is

\[
p(\theta_d) = \int_S \pi_\gamma(\theta) d\theta , \quad \theta_d \in S,
\]

so that the average detail length of a message based on the set \( S \) is given by

\[
I(y^n|S; \gamma) = -\frac{1}{q_\gamma(S)} \int_S \pi_\gamma(\theta) \log p_{\gamma}(y^n | \theta) d\theta . \tag{3}
\]

The average total random coding message based on \( S \) is

\[
I(y^n; S; \gamma) = I(S; \gamma) + I(y^n|S; \gamma). \tag{3}
\]

It is informative to define the “round-off” error as

\[
r_\gamma(y^n, S) = I(y^n|S; \gamma) + \log p_{\gamma}(y^n | \hat{\theta}_{ML}) , \tag{4}
\]

where \( \hat{\theta}_{ML} = \arg \max_{\theta \in \Theta_\gamma} \{ p_{\gamma}(y^n | \theta) \} \) is the maximum likelihood estimate. The quantity \( r_\gamma(y^n, S) \) can be interpreted as the increase in the length of the detail over the “maximum-likelihood” code incurred by using a quantised estimate, represented by \( S \), in place of \( \hat{\theta}_{ML} \), to transmit the data. Let \( \Omega_\gamma(y^n) \) denote the set that solves

\[
\min_{S \subseteq \Theta_\gamma} \left\{ I(S; \gamma) - \log p_{\gamma}(y^n | \hat{\theta}_{ML}) + r_\gamma(y^n, S) \right\} . \tag{5}
\]

In the MML literature, this set is called the uncertainty region, as it includes all models that are considered to be plausible explanations of the data. The MMLD code length is then given by

\[
I_D(y^n; \gamma) \equiv I(y^n, \Omega_\gamma(y^n); \gamma), \tag{3}
\]

Examining the minimisation problem (5) shows that the MMLD code length can be interpreted as balancing the accuracy to which maximum likelihood estimates are stated against the evidence in the data.

Unfortunately, direct replacement of MML87 by the MMLD approximation is not possible. The MMLD code length may be used to select a model structure \( \gamma \), but offers no guidance for selection of a suitable point estimate. This is because the MMLD messages are essentially (redundant) one-part codes. The random coding procedure on which they are based is in theory two-part; however, the MMLD procedure, by integrating out the random variables \( (d, \theta_d) \) to arrive at a sensible measure of message length, removes the ability to transmit the data using an arbitrary model from \( \Theta_\gamma \). The data alone determines the uncertainty region \( \Omega_\gamma(y^n) \), and in this sense the MMLD message length offers a codebook over \( \gamma^n \) only. The next section proposes a new message length formula that addresses this issue.

3. THE MML08 CODE LENGTH

The main contribution of this paper is to present a generalization of the MMLD message length equation that allows
one to derive point estimates explicitly by minimising the joint model and data codelength. In this way it acts as a replacement for MML87 when the Wallace–Freeman assumptions do not apply, and is significantly easier to compute than the exact SMML codelength.

3.1. Model Cost

Define the quantity

\[ D_D(y^n; \gamma) = \log 1/q_\gamma(\gamma) + r_\gamma(y^n, \Omega_\gamma(y^n)), \]

so that \( I_D(y^n; \gamma) = D_D(y^n; \gamma) - \log p_\gamma(y^n|\theta_{\text{ML}}). \) We call \( D_D(\cdot; \gamma) \) the model cost; it is the extra number of nits (nats) required to name the model used to transmit the data \( y^n \), the “model” being described by the uncertainty region \( \Omega_\gamma(y^n) \). In the case of MMLD the model cost is also the regret of the MMLD message length with respect to the “ideal” maximum likelihood codelength, though this is in general not the case for other MML approximations. In particular, comparing \( D_D(y^n; \gamma) \) to the MML87 model cost

\[ D_{87}(\theta; \gamma) = -\log p_\gamma(\gamma) + \frac{1}{2} \log |\mathcal{M}_\gamma(\theta)| + \frac{k}{2} (\log k + 1) \]

it is clear the fundamental difference between MML87 and MMLD is that \( D_{87}(\theta; \gamma) \) depends on the chosen model \( \theta \) used to encode the data, while \( D_D(y^n; \gamma) \) depends on the data, and only offers a measure for complexity of a model structure \( \gamma \). Thus, the MML87 codelength allows one to perform point estimation and model structure selection by minimising the sum of model cost for a particular model, say \( \theta \in \Theta_\gamma \), and the negative log-likelihood of the data using this model, i.e.,

\[
\{ \hat{\gamma}_{87}, \hat{\theta}_{87} \} = \arg \min_{\gamma \in \Gamma, \theta \in \Theta_\gamma} \{ I(\gamma) + D_{87}(\theta; \gamma) - \log p_\gamma(y^n|\theta) \}.
\]  

It would clearly be advantageous to have an analogue of (6) for a robust MMLD-like approximation.

3.2. MML08 Message Length

Examining (3) it is clear to see that \( y^n \) enters the round-off function (4) only through the negative log-likelihood function. Thus, following the arguments of Wallace and Freeman, we wish to find the expected increase in codelength due to quantisation of a model \( \theta^* \) to some region \( S \). Rewrite \( r_\gamma(y^n, S) \) as

\[
r_\gamma(y^n, S) = -\frac{1}{q_\gamma(S)} \int_S \pi_\gamma(\theta) \log \frac{p_\gamma(y^n|\theta)}{p_\gamma(y^n|\theta_{\text{ML}})} d\theta.
\]

As in the Wallace–Freeman approximation [4], we can replace the dependency on a particular string \( y^n \) by a dependency on a particular model \( \theta^* \) by assuming that the data \( y^n \sim p_\gamma(\cdot|\theta^*) \), and finding the expected inflation in codelength due to quantisation of \( \theta^* \). The average codelength for coding data \( y^n \sim p_\gamma(\cdot|\theta^*) \) using model \( \theta \) is simply \( \text{E}_{\theta^*}[\log 1/p_\gamma(y^n|\theta)] \), and this expression obtains a minimum when \( \theta = \theta^* \) (the entropy). Thus, the expected excess codelength for coding data \( y^n \) coming from \( p_\gamma(\cdot|\theta^*) \) using model \( \theta \) in place of the optimal model \( \theta^* \) is

\[
\Delta_\gamma(\theta^*||\theta) = \text{E}_{\theta^*}[\log \frac{p_\gamma(y^n|\theta^*)}{p_\gamma(y^n|\theta)}],
\]

which is the well known Kullback–Leibler (KL) divergence [7] between a generating model \( \theta^* \) and approximating model \( \theta \). Note that the KL divergence in (7) is defined for \( n \) data points, which in the i.i.d. case is simply \( n \) times the KL divergence for a single datapoint.

Now the overall expected increase in codelength due to quantisation of the model \( \theta^* \) to the region \( S \) is given by

\[
r_\gamma(\theta^*, S) = \frac{1}{q_\gamma(S)} \int_S \pi_\gamma(\theta) \Delta_\gamma(\theta^*||\theta) d\theta,
\]

and we can find the quantisation cell that minimises the sum of the assertion plus the expected round off by solving

\[
\Omega_\gamma(\theta^*) = \arg \min_{\theta \in \Theta_\gamma} \{ \log 1/q_\gamma(S) + r_\gamma(\theta^*, S) \}.
\]

We call the set \( \Omega_\gamma(\theta^*) \) the expected uncertainty region for the model \( \theta^* \); in contrast to \( \Omega_\gamma(y^n) \), it depends only on the expected behaviour of the model \( \theta^* \). We now define the MML08 model cost for a model \( \theta^* \) by

\[
D_{\text{os}}(\theta^*; \gamma) = -\log q_\gamma(\Omega_\gamma(\theta^*)) + r_\gamma(\theta^*, \Omega_\gamma(\theta^*)).
\]

Given that \( D_{\text{os}}(\theta^*; \gamma) \) depends only on the model, \( \theta^* \), and not on the data \( y^n \), we can compute a valid joint message length for any pair \( (\theta^*, y^n) \in \Theta_\gamma \times \mathcal{Y}^n \); this is

\[
I_{\text{os}}(y^n, \theta^*; \gamma) = D_{\text{os}}(\theta^*; \gamma) - \log p_\gamma(y^n|\theta^*).
\]

We call (9) the “MML08” message length approximation. Explicit point estimation, as well as model structure estimation, can be performed for a given \( y^n \) by comparing candidate models \( \theta^* \in \Theta_\gamma, \gamma \in \Gamma \), on their joint MML08 codelength, and choosing the model which yields the shortest message length, i.e.,

\[
\{ \hat{\gamma}_{\text{os}}, \hat{\theta}_{\text{os}} \} = \arg \min_{\gamma \in \Gamma, \theta \in \Theta_\gamma} \{ I(\gamma) + I_{\text{os}}(y^n, \theta^*; \gamma) \}.
\]

From (10) we see that the MML08 codelength balances the accuracy to which some particular model, \( \theta^* \in \Theta_\gamma \), is stated, against the evidence for that particular model that is present in the data. This is in contrast to the MML codelength, which implicitly quantises the maximum likelihood estimate. The MML08 message length may be split into assertion and detail components

\[
I_{\text{os}}(\theta^*; \gamma) = -\log q_\gamma(\Omega_\gamma(\theta^*)),
\]

\[
I_{\text{os}}(y^n; \theta^*; \gamma) = -\log p_\gamma(y^n|\theta^*) + r_\gamma(\theta^*, \Omega_\gamma(\theta^*)).
\]

The MML08 message length approximation generalizes the MMLD approximation, which can be recovered by setting \( \theta^* = \theta_{\text{ML}} \) and replacing the KL divergence with \( \log p_\gamma(y^n|\theta_{\text{ML}})/p_\gamma(y^n|\theta) \), i.e., the empirical KL divergence.

Finally, we note that the MML08 model cost is very robust in the sense that \( D_{\text{os}}(\theta^*; \gamma) \geq 0 \) for all \( \theta \in \Theta_\gamma \). In contrast, the MML87 model cost can be (nonsensically) negative if the conditions under which the MML87 approximation was derived are violated.
4. PROPERTIES OF MML08 CODELENGTHS

Assuming that the prior distribution and KL divergence are differentiable functions of $\theta^*$, we have the following properties. The proofs are given in [5].

Property 1. The MML08 model cost (8) is invariant under differentiable, one-to-one transformations of the parameters $\theta^*$, that is

$$I_{08}(y^n, \theta^*; \gamma) = I_{08}(y^n, \phi; \gamma),$$

where $\phi = g(\theta^*)$ are the transformed parameters, $g(\cdot)$ is a differentiable one-to-one function, and the prior distribution $\pi(\cdot)$ is appropriately transformed.

This property has the important implication that inferences made by minimising the MML08 message length will be invariant to the choice of model parameterisation. This property is shared by the MML87 and SMML estimators, but not (in general) by Bayes estimators.

Property 2. The model cost (8) satisfies the “Boundary Rule” [1]; that is

$$\Omega_\gamma(\theta^*) = \{ \theta \in \Theta_\gamma : \Delta_\gamma(\theta^*|\theta) \leq \delta_\gamma(\theta^*) \},$$

where $\delta_\gamma(\theta^*)$ is the Kullback–Leibler divergence of any model on the boundary of $\Omega_\gamma(\theta^*)$.

This property implies that the expected uncertainty region can be completely, and uniquely, defined by the value of the Kullback–Leibler divergence at the boundary of the region, $\delta_\gamma(\theta^*)$. This property also suggests intriguing links with the normalised maximum likelihood code, and the concept of distinguishable distributions [2]; these links are interesting topics for future research.

5. LARGE SAMPLE BEHAVIOUR

The large sample behaviour of the MML08 approximation is now examined. Under the regularity conditions used in the derivation of the MML87 approximation [1], we have

$$\text{vol} \left( \Omega_\gamma(\theta^*) \right) = q(\Omega_\gamma(\theta^*)) + o_n(1),$$

$$\Delta_\gamma(\theta^*|\theta) = \frac{1}{2} \| \theta - \theta^* \|^2_{\text{A}^{-1}(\theta^*)} + o_n(1),$$

where $\| x \|^2_{\text{A}^{-1}} = x^T \text{A}^{-1} x$. Ignoring terms of order $o_n(1)$, the KL divergence is a quadratic function of $\theta$, and the uncertainty region will be a $k$-dimensional ellipse. Following similar arguments to those in [1], coupled with the rules for integration of polynomials over balls [8], the following assertion and detail lengths can be derived

$$I_{08}(\theta; \gamma) = -\log \left( \frac{\pi_\gamma(\theta)}{\pi_\gamma(\theta^*)} \right) - \frac{k}{2} \log(\pi(k + 2)) + \log \Gamma \left( \frac{k}{2} + 1 \right) + o_n(1),$$

$$I_{08}(y^n|\theta; \gamma) = -\log p_\gamma(y^n|\theta) + \frac{k}{2} + o_n(1).$$

Comparing (11–12) to (1–2), it is clear that for large $n$ and sufficiently regular likelihood functions and prior distributions, the MML08 codeengths and MML87 codeengths differ only in their respective dimensionality constants. As (11–12) make use of elliptical uncertainty regions, which do not tessellate, the large sample MML08 codeength is actually slightly shorter than the MML87 codeength for $k > 1$. Interestingly, by assuming that the uncertainty region is congruent to an optimal quantising cell that tessellates the parameter space, the MML08 codeength can be used as a basis for a novel derivation of the MML87 approximation (as done in Chapter 2, [5]).

Under suitable regularity conditions, the large sample MML08 formulae (11–12) can be used to show that, asymptotically, as the sample size $n \to \infty$, with the number of parameters $k$ fixed, the MML08 codeength is equivalent to the Bayesian information criterion [9]. The usual consistency properties of maximum likelihood parameter estimation, and Bayesian information criterion model structure selection, follow as a consequence.

6. ACKNOWLEDGEMENTS

The author would like to thank E. Makalic for comments and suggestions that improved this paper considerably.

7. REFERENCES

ADAPTIVE SAMPLING FOR SPARSE RECOVERY

Ali Tajer and H. Vincent Poor

Department of Electrical Engineering, Princeton University, Princeton, NJ 08544
Email: {tajer, poor}@princeton.edu

ABSTRACT

Consider \( n \) data sequences, each consisting of independent and identically distributed elements drawn from one of the two possible zero-mean Gaussian distributions with variances \( A_0 \) and \( A_1 \). The problem of quickly identifying a fixed number of the sequences with variance \( A_1 \) is considered and an adaptive two-stage experimental design and testing procedure is proposed. The agility and reliability gains in comparison with the existing related methods for quick search over multiple sequences are quantified.

1. PRELIMINARIES

1.1. Model

Consider \( n \) observation sequences \( X_1, \ldots, X_n \), where for each \( i \in \{1, \ldots, n\} \), \( X_i \sim \{X_i(j); j = 1, 2, \ldots \} \). The observations of each sequence \( i \) are independent and identically distributed (i.i.d) and obey one of the two hypotheses:

\[
\begin{align*}
H_0 & : X_i(j) \sim Q_0 \sim N_C(0, A_0), \quad j = 1, 2, \ldots \quad (1) \\
H_1 & : X_i(j) \sim Q_1 \sim N_C(0, A_1), \quad j = 1, 2, \ldots 
\end{align*}
\]

where \( N_C(a, b) \) denotes the complex Gaussian distribution with mean \( a \) and variance \( b \), and \( A_0 \) and \( A_1 \) are specified positive real numbers. Each sequence \( X_i \) is distributed according to \( Q_0 \) or \( Q_1 \) independently of the rest. Also, we assume that the hypothesis \( H_0 \) occurs sparsely with the prior probability \( \epsilon_n \). To model the sparsity we also assume that \( \epsilon_n = o(1) \). Also, for \( m \in \{0, 1\} \) let us define the sets

\[
H_m \triangleq \{i \in \{1, \ldots, n\} : X_i \sim Q_m\}.
\]

1.2. Search Objective

The main objective is to use observations \( X_1, \ldots, X_n \) in order to identify \( T \in \mathbb{N} \) elements of \( H_0 \). The conventional non-adaptive sampling procedures use some pre-specified number of observations and locate \( T \) sequences of interest. This strategy is non-adaptive, in the sense that the measurement process is fixed \( a \ priori \) and does not change during the experiment. In contrast, we devise an adaptive procedure in which the measurement strategy is adjusted sequentially such that future measurements use information gathered from previous ones. We demonstrate that such measurement adaptation substantially improves the reliability and agility in identifying the \( T \) sequences of interest.

Clearly, for detecting multiple sequences drawn from \( Q_0 \) there exists a tradeoff between agility and reliability in the sense that achieving a higher level of detection reliability requires using more sampling resources, which in turn results in delay in reliable detection. We characterize this tradeoff in both non-adaptive and the proposed adaptive sampling procedures. Comparing these tradeoffs demonstrates the agility and reliability gains afforded by the adaptive procedure. The analysis provided is asymptotic with respect to a large number of sequences, \( n \).

2. NON-ADAPTIVE SAMPLING

The development of the non-adaptive sampling scheme serves a two-fold purpose. On one hand this detection scheme is also deployed in the detection phase of the adaptive procedure proposed in Section 3, and on the other hand, it offers a baseline for assessing the gain yielded by the adaptive procedure.

2.1. Non-Adaptive Sensing Procedure

Constructing a non-adaptive sampling procedure involves two issues. One pertains to the experimental design, which is the design of the information-gathering process. In our setup the experimental design elucidates the distribution of the sampling budget among the sequences \( X_1, \ldots, X_n \). The second issue is to design a detector based on some optimality criterion. Prior to the sampling procedure all the sequences are equally likely to be drawn from \( Q_0 \). Due to the inherent symmetry and the sparse occurrence of \( H_0 \) we assume that the experimental design measures all sequences equally.

Given this experimental design, it is straightforward to construct a detector that is optimal in the sense that it maximizes the \textit{a posteriori} probability of successfully detecting \( T \) elements of \( H_0 \). Suppose that the measurement budget is \( M \in \mathbb{N} \) per sequence. The observation vector \( X_i \sim \{X_i(j)\} \) for sequence \( i \) is a sample from a mixture Gaussian distribution. Given the observation set \( D_n \triangleq \{X_1, \ldots, X_n\} \), the maximum \textit{a posteriori} probability (MAP) rule for identifying \( T \) member of \( H_0 \) is formalized in the following remark.
Remark 1 The MAP rule for detecting $T$ elements of $\mathcal{H}_0$ is given by
\[
\hat{U}_{\text{MAP}} \triangleq \arg \max_{U : |U| = T} P(U \subseteq \mathcal{H}_0 \mid \mathcal{D}_n) = \arg \max_{U : |U| = T} \sum_{i \in U} \|X_i\|^2.
\]

Hence, in order to locate $T$ sequences of interest the MAP detector requires only the sufficient statistic $Y_i \triangleq \|X_i\|^2$ for $i = 1, \ldots, n$. Corresponding to the sequence of random variables $\{Y_1, \ldots, Y_n\}$ we define $\{Y_{(1)}, \ldots, Y_{(n)}\}$ as the sequence of order statistics in an increasing order, e.g., $Y_{(m)}$ represents the $m$th smallest element of $\{Y_1, \ldots, Y_n\}$.

2.2. Asymptotic Performance

Recalling the distribution of $X_i(j)$ given in (2), the sufficient statistics $Y_i$ under hypothesis $H_m$ for $m \in \{0, 1\}$ is distributed as
\[
Y_i \mid H_m \sim \text{Gamma} (\alpha, \beta) \quad \text{for} \quad i = 1, \ldots, n,
\]
where $\text{Gamma} (\alpha, \beta)$ denotes a Gamma distribution with parameters $\alpha$ and $\beta$. Clearly the detector makes a detection error if $\hat{U}_{\text{MAP}} \cap H_1 \neq \emptyset$. Let us define $u_i$ and $v_i$ as the indices of the $i^{th}$ smallest elements of the sets $\{Y_1 : i \in \{1, \ldots, n\}\}$ and $\{Y_i : i \in H_0\}$, respectively. From Remark 1 the detection error probability is given by
\[
P_{\text{err}}(n) \triangleq P \left( \hat{U}_{\text{MAP}} \cap H_1 \neq \emptyset \right)
= 1 - P \left( \hat{U}_{\text{MAP}} \cap H_1 = \emptyset \right)
= 1 - P(\{u_1, \ldots, u_T\} \subseteq H_0).
\]

Note that the event of having all the $T$ smallest measured $Y_i$ belonging to $H_0$ is equivalent to
\[
\{u_1, \ldots, u_T\} \subseteq H_0 \equiv \left\{ \max_{i \in \{1, \ldots, T\}} Y_{v_i} \leq \min_{i \in H_1} Y_i \right\}
\equiv \left\{ Y_{v_T} \leq \min_{i \in H_1} Y_i \right\}.
\]

Therefore,
\[
P_{\text{err}}(n) = 1 - P(Y_{v_T} \leq \min_{i \in H_1} Y_i). \tag{4}
\]

Assessing $P_{\text{err}}(n)$ as defined above relies on the properties of the order statistics of two sets of random variables. The following lemma is instrumental for characterizing the distributions of these sets of order statistics and evaluating $P_{\text{err}}(n)$. This is a generalization of a well-studied problem in the context of extreme value theory that considers the first order statistic [1]. In this lemma, we give the corresponding results for higher order statistics and at the same time we also allow for distribution evolution, i.e., the number and distribution of the involved random variables changes simultaneously, for which the existing results are not applicable.

Lemma 1 Let $\left\{ Y_{(i)} \right\}_{i=1}^{m}$ be a sequence of i.i.d. random variables distributed as Gamma $(M, \alpha_m)$ and denote its corresponding sequence of order statistics by $\left\{ Y_{(i)} \right\}_{i=1}^{m}$. Let $b_m \triangleq \alpha_m \left( \frac{\Gamma(M+1)}{m} \right)^{\frac{1}{m}}$ and for some $T \in \mathbb{N}$ define the sequence of random variables $W_{(i)} \triangleq \frac{Y_{(i)}}{b_m}$ for $i = 1, \ldots, T$. Then as $m \to \infty$, $W_{(i)}$ converges in distribution to a random variable $W_{(i)}$ with cumulative distribution function (CDF)
\[
Q_{(i)}(w; m) \triangleq P(W_{(i)} < w) \triangleq 1 - \exp(-w^M) \sum_{k=0}^{\infty} \frac{w^{kM}}{k!}.
\]

For the setting of Section 2.1, the following theorem characterizes the asymptotic performance of the MAP detector. It also establishes the tradeoffs among the prior probability $\epsilon_n$, per sequence sampling budget $M$, and the ratio of the variances of distributions $Q_0$ and $Q_1$.

Theorem 1 (Non-Adaptive Tradeoff) When $\epsilon_n = o(1)$ and $\alpha \epsilon_n = \omega(1)$, the error probability of the MAP for identifying $T$ elements of $\mathcal{H}_0$ is given by
\[
P_{\text{err}}(n) = P \left( \hat{U}_{\text{MAP}} \cap H_1 \neq \emptyset \right)
\triangleq 1 - (1 + [(A_1/A_0)^M \cdot \epsilon_n]^{-1})^{-T}. \tag{5}
\]

As expected, there exists a tension between reliability and agility. On one hand, increasing the sampling budget per sequence $M$ favors reliability, as according to (5) it improves the probability of successfully detection, and on the other hand, imposes more delay in detecting $T$ sequences distributed as $Q_0$. By using the result of Theorem 1, we offer a necessary and sufficient condition on the scaling of the ratios of the variances $A_1/A_0$ to guarantee asymptotically error-free detection in the non-adaptive sampling setting.

Corollary 1 (Non-Adaptive Variance Scaling) When $\epsilon_n = o(1)$ and $\alpha \epsilon_n = \omega(1)$, a necessary and sufficient condition for $P_{\text{err}}(n) \to 0$ as $n \to \infty$ is that $\frac{A_1}{A_0}$ scales with increasing $n$ as
\[
\frac{A_1}{A_0} = \omega \left( \frac{T}{\epsilon_n} \right).
\]

3. ADAPTIVE SAMPLING

3.1. Adaptive Sampling

Our proposed adaptive sampling procedure has two phases, namely the exploration phase and the detection phase. The exploration phase, being an iterative procedure, is intended to purify the set of the sequences to be observed carefully for detecting the sequences drawn from $Q_0$. This phase is accomplished by successively identifying and eliminating a group of sequences deemed to be drawn from $Q_1$. The detection phase is performed after the exploration phase in order to identify $T$ sequences of interest among the subset of candidate sequences retained after exploration. The detection scheme deployed is identical to the MAP detection scheme of Section 2.
The exploration phase proceeds in an iterative way. In each iteration it further monitors the sequences retained by the previous iteration and eliminates those deemed to be drawn from $Q_0$ least-likely. The core idea is that it is relatively easy to identify sequences drawn from $Q_1$ with low-quality measurements (recall that $\epsilon_n$ is small). Each iteration carries on by thresholding the observed energy on each sequence retained by the previous iteration. The threshold level depends only on $\frac{\lambda_k}{A_0}$, and is designed such that at each iteration roughly half of the existing sequences distributed as $Q_1$ are eliminated, while almost all of those distributed as $Q_0$ are preserved. The output of each exploration phase will have a more condensed proportion of the desired sequences to the non-desired ones. Subsequently, the detector developed for the non-adaptive procedure is applied on this refined set of sequences in order to identify the detector developed for the non-adaptive procedure is applied on this refined set of sequences in order to identify the desired sequences to the non-desired ones. Subsequently, the detector developed for the non-adaptive procedure is applied on this refined set of sequences in order to identify the desired sequences to the non-desired ones.

We show that the gains yielded by this adaptive procedure can be interpreted in two ways. First we demonstrate that when targeting at achieving the same level of detection reliability, the adaptive procedure requires substantially less sampling budget, or equivalently it is substantially more agile. Secondly, we show that under the same sampling budget, and targeting identical detection reliability, the adaptive procedure imposes less-stringent conditions on how fast the power of the active users must scale with increasing $n$. This essentially indicates that for some choices of $\gamma_n$, the adaptive procedure can guarantee successful detection while the most non-adaptive procedure fails to do so.

Let us define $K$ as the number of exploration cycles (iterations) in the exploration phase. Also denote the sampling budget per sequence in the $kth$ exploration cycle by $M_k$. The exploration phase is initialized by including all sequences for sampling and resumes as follows. In the first iteration all sequences are allocated the identical sampling budget of $M_1$. The energy levels of all sequences are compared against $\lambda_1(A_1/A_0)$, where $\lambda_1$ is the median of the distribution Gamma $(M_1, 1)$. The sequences for which the measured energy exceed this threshold are discarded and the rest are carried over to the second iteration for further sampling. The same procedure is repeated throughout all $K$ cycles. More specifically, in the $kth$ cycle all the sequences retained by the $(k-1)th$ iteration are allocated the identical sampling budget of $M_k$. The energy levels of these sequences are compared with $\lambda_k(A_k/A_0)$, where $\lambda_k$ is the median of the distribution Gamma $(M_k, 1)$ and the exploration is performed via thresholding as in the first iteration. Finally, after the exploration phase, each of the remaining sequences is allocated the sampling budget of $M_{K+1}$ and the MAP detection scheme provided in Remark 1 is applied in order to detect $T$ sequences distributed as $Q_0$.

We set $G_0 \triangleq \{1, \ldots, n\}$ and for $k = 1, \ldots, K$, we define $G_k$ as the set of the indices of the sequences that are retained by the $kth$ exploration cycle. Clearly we have $G_K \subseteq \ldots \subseteq G_1 \subseteq G_0$ and $G_K$ contains the set of the indices of the candidate sequences among which $T$ sequences distributed as $Q_0$ will be detected. The set of measurements defined for the non-adaptive scheme is extended for the proposed adaptive procedure as follows. We define the set of measurements in the $kth$ cycle as

$$
\mathcal{D}_n^k \triangleq \{X_i^k : \text{ for } i \in G_{k-1}\} \quad \text{for } k = 1, \ldots, K+1.
$$

The measurement sets $\mathcal{D}_n^1, \ldots, \mathcal{D}_n^K$ are processed in the exploration phase and the measurement set $\mathcal{D}_n^{K+1}$ is used in the detection phase. Under hypothesis $H_m$, the observation sample $X_i^k(j)$ for $k = 1, \ldots, K+1$ is distributed as

$$
X_i^k(j) \mid H_m \overset{iid}{\sim} \mathcal{N}_C(0, A_m), \quad \text{for } i \in G_{k-1}. \quad (7)
$$

We also define

$$
Y_i^k \triangleq \|X_i^k\|^2 \quad \text{for } i \in G_{k-1} \quad \text{and } k = 1, \ldots, K+1. \quad (8)
$$

Equations (7) and (8) provide that for $k = 1, \ldots, K+1$,

$$
Y_i^k \mid H_m \sim \text{Gamma} (M_k, A_m) \quad \text{for } i \in G_{k-1}. \quad (9)
$$

For each $k$, corresponding to the sequence $\{Y_i^k\}_{i \in G_{k-1}}$, we define the sequence of order statistics $\{Y_i^k\}_{i \in G_{k-1}}$ in an increasing order such that $Y_i^k$ represents the $i$th smallest element of this sequence. The adaptive sampling procedure is formally described in the table above.

### 3.2. Asymptotic Performance

We start by assessing the performance for any given value of the exploration cycles $K$. The analysis of the adaptive sampling procedure follows the approach of [2], albeit with the non-trivial modifications to deal with the different objective and the different observation model. The following lemmas shed light on how the adaptive procedure accomplishes the exploration cycles. Lemma 2 characterizes the proportion of the sequences distributed as $Q_0$ that are retained in each exploration cycle.

---

**Table: Exploration and Detection Phases**

<table>
<thead>
<tr>
<th>Exploration phase</th>
</tr>
</thead>
</table>
| 1: Input $K \in \mathbb{N}$ and $\{M_1, \ldots, M_{K+1}\}$. 
| 2: Initialize the index set $G_0 \leftarrow \{1, \ldots, n\}$. 
| 3: for $k = 1, \ldots, K$ do 
| 4: Set $Y_i^k = \{\|X_i^{k+1}\|^2 \text{ for } i \in G_{k-1}\}$ 
| 5: Obtain $G_k \leftarrow \{i \in G_{k-1} \mid Y_i^k < \lambda_k(A_1/A_0)\}$. 
| 6: end for |

<table>
<thead>
<tr>
<th>Detection phase</th>
</tr>
</thead>
</table>
| 7: Set $Y_i^{K+1} = \{\|X_i^{K+1}\|^2 \text{ for } i \in G_K\}$ 
| 8: Obtain $U_{\text{det}}$. 
| 9: Output $U_{\text{det}}$. |

---
Lemma 2 Let \( m_0 = |\mathcal{H}_0| \) and for \( k = 1, \ldots, K \) define \( m_k \) as the number of sequences distributed as \( Q_0 \) that are retained by the \( k \)-th exploration cycle. Conditionally on \( m_{k-1} \) and for sufficiently large \( n \) the event

\[
\left( \frac{A_1/A_0}{1 + A_1/A_0} \right)^{m_{k-1}} \leq m_k \leq m_{k-1}, \text{ for } k = 1, \ldots, K,
\]

holds with probability at least \( 1 - \exp\left(-\frac{m_{k-1}}{\alpha n^\alpha}\right) \) for any \( \alpha > 0 \).

The next lemma shows that during each exploration cycle almost half of the sequences distributed as \( Q_1 \) are eliminated.

Lemma 3 Let \( \ell_k = |\mathcal{H}_k| \) and for \( k = 1, \ldots, K \) define \( \ell_k \) as the number of the sequences distributed as \( Q_1 \) that are retained by the \( k \)-th exploration cycle. Conditionally on \( \ell_{k-1} \) and for sufficiently large \( n \), for all \( k = 1, \ldots, K \), the event

\[
\left( \frac{1}{2} - \frac{1}{\log n} \right) \ell_{k-1} \leq \ell_k \leq \left( \frac{1}{2} + \frac{1}{\log n} \right) \ell_{k-1},
\]

holds with probability at least \( 1 - 2 \exp\left(-\frac{2\ell_{k-1}}{(\log n)^2}\right) \).

A careful use of the above lemmas establishes the performance of the adaptive sampling in the following theorem.

Theorem 2 (Adaptive Tradeoff) When \( \epsilon_n = o(1) \) and \( \omega_n = \omega(1) \), the error probability of the adaptive sampling procedure for identifying \( T \) sequences distributed as \( Q_0 \) is given by

\[
P_{\text{adap}}(n) \geq P\left( \hat{\mathcal{U}}_{\text{adap}} \cap \mathcal{H}_1 \neq \emptyset \right)
\geq 1 - \left( 1 + [(A_1/A_0)^{M_k+1} \cdot 2^K\epsilon_n]^{-1} \right)^{-T}.
\]

In order to quantify the gains yielded by the adaptive procedure, we compare the results for the non-adaptive and adaptive schemes provided in Theorems 1 and 2. In particular we first characterize the agility factor, which we define as the ratio of the sampling budgets required by the adaptive procedure to that required by the non-adaptive scheme with the aim of attaining identical asymptotic reliability levels in the asymptotic and the non-asymptotic procedure, i.e., \( P_{\text{non}}(n) = P_{\text{adap}}(n) \).

Theorem 3 (Agility) When \( \epsilon_n = o(1) \) and \( \omega_n = \omega(1) \), the agility factor of the adaptive sampling approach with \( Mn \) sampling budget is asymptotically upper bounded by \( \left( \frac{2}{T^2} + \frac{3}{M} \right) \), where \( K \) is the number of exploration cycles.

It is noteworthy that while the number of exploration cycles \( K \) can be made arbitrarily large (but fixed as a function of \( n \)), increasing it beyond some point will affect the agility very insignificantly. More specifically, for large \( K \), the agility factor will be dominated by the term \( \frac{3}{M} \). This underlines the fundamental limit of the agility gain yielded by the adaptive procedure.

An analog of Corollary 1 can be derived for the adaptive procedure, providing a necessary and sufficient condition on the scaling of \( A_1/A_0 \) for guaranteeing asymptotic error-free detection of \( T \) sequences distributed as \( Q_0 \). For comparison purposes we assume that both adaptive and non-adaptive procedures are granted the same sampling budget.

Corollary 2 (Adaptive Power Scaling) When \( \epsilon_n = o(1) \) and \( \omega_n = \omega(1) \), given that the sampling budget is \( Mn \), a necessary and sufficient condition for \( P_{\text{adap}}(n) \to 0 \) as \( n \to \infty \) is that

\[
\frac{A_1}{A_0} = \omega\left( \frac{M}{\sqrt{T}} \right),
\]

where \( M' \geq 2^K(M - 2) + 2 \).

Comparing the result above with that of Corollary 1 shows that an adaptive scheme can cope with signals with much smaller variances. More specifically, by noting that \( M' \) is substantially larger than \( M \), the variance scaling requirement in the adaptive scenario, which is smaller than \( \frac{\omega}{\sqrt{T} \sqrt{\epsilon_n}} \) becomes substantially smaller than its counterpart in the non-adaptive scenario \( \omega\left( \frac{\sqrt{\epsilon_n}}{\sqrt{T}} \right) \). As a result, there are scenarios where non-adaptive schemes fail to successfully identify \( T \) sequences of interest, while the adaptive scheme succeeds.

4. CONCLUSION

In this paper we have presented an adaptive sampling methodology for quickly searching over finitely many sequences with the objective of identifying multiple sequences that are distributed according to a given distribution of interest. The core idea of the sampling procedure is to successively and gradually adjust the measurement process using information gleaned from the previous measurements. Compared to the non-adaptive procedures, dramatic gains in terms of reliability and agility are achieved.

5. REFERENCES

COGNITION AND INFERENCE IN AN ABSTRACT SETTING

Flemming Topsøe
University of Copenhagen
Department of Mathematical Sciences
Universitetsparken 5, dk-2100 Copenhagen, Denmark, topsoe@math.ku.dk

ABSTRACT

We continue the development of an abstract, though quantitative, theory of cognition which is rooted in philosophical considerations. Applications include classical Shannon theory and results from geometry. Special attention is paid to inference which is treated as the outcome of a situation of conflict between Nature and Observer, “you”.

1. INTRODUCTION

Last year, at WITMSE 2010, the author presented basic elements of an abstract theory of cognition, cf. [1]. Topics which we will focus on this year include those discussed in Section 3.

Emphasis will be on concrete results, especially concerning identification. From the point of view of applications, these are the most useful ones and also the technically simplest to establish.

Two key characteristics of the theory presented is that it is not tied to probabilistic notions and that it invokes game theoretical considerations. The desirability of a non-probabilistic approach has been advanced before, cf. [2], [3], [4], [5] and also the recent survey [6]. The relevance of games goes back to [7] and [8], cf. also [9], [10] and, as an example of a utility-based work, [11].

2. ELEMENTS OF COGNITION

In this section we outline parts of the abstract theory under development. Some passages are taken from [12].

2.1. Philosophical background

Observer is placed in a world and interacts with Nature when confronted with situations from the world. Nature does not have a mind and cannot act but is the holder of “truth”. Observer seeks the truth but is restricted to belief. Observer is guided by a creative mind which is exploited to obtain knowledge as effortlessly as possible through experiments and associated observations. Knowledge often comes in the form of perception of phenomena from the world.

“Belief is a tendency to act”\(^1\). Thus one should be aware of possibilities to transform belief to more action-oriented objects. Such objects we call controls. Description is the key to control through the design of experiments. An experiment involves a preparation which entails a limitation of the states – possible truth instances – available to Nature. Theoretically possible but unrealistic preparations should be distinguished from feasible preparations. Feasible preparations determine the knowable, thus provide limitations to what can be known, hence to obtainable information.

Description entails an effort which depends on the state as well as on Observers belief. This is the key to quantitative considerations. Insight into the knowable also comes from description: “what you can know depends on what you can describe”.

To be operational, description effort should satisfy the perfect match principle, viz. that effort, given the state, is the least under a perfect match, i.e. when belief equals truth. The minimal effort, given the state, is called entropy, and the excess effort, taking also belief into consideration, is called divergence\(^2\).

Interaction between Nature and Observer takes place as if they are players in a two-person zero-sum game with description effort as objective function, Nature as maximizer and Observer as minimizer. Ideally, one should not only aim at equilibrium but also at bi-optimality, i.e. the identification of optimal strategies which provides Observer not only with insight about what can be inferred but also on how.

2.2. Truth, belief and description

Given are sets \(X\), the state space, and \(Y \supseteq X\), the belief reservoir, as well as a relation \(X \otimes Y \subseteq X \times Y\), called visibility. A non-empty set \(Y_{det} \subseteq Y\) determines certain beliefs. We write \(y \succ x\) for \((x, y) \in X \otimes Y\) and say, either that “\(y\) can see \(x\)”, that “\(x\) is visible from \(y\)” or similar. By \(|y|\) we denote the outlook from \(y\), the set of \(x\) which are visible from \(y\), and by \([x]\) we denote the watchout for \(x\), the set of \(y\) from which \(x\) can be seen. We assume that \(x \succ x\) for all \(x\) and that there exists \(y\) from which all of \(X\) is visible, i.e. \(|y| = X\).

A preparation is a non-empty subset \(P\) of \(X\). A pair \((x, y) \in X \otimes Y\) is an atomic situation. The watchout for

\(^1\)a quotation from Good [13].

\(^2\)the term “divergence” appears justified as the quantity typically stands for the discrepancy, counted non-negative, between the “actual” and the “best possible” performance. Regarding “entropy”, terminology is less convincing and some other terminology for the abstract setting may be preferable.
a preparation $P$ is the set $|P| = \bigcap_{x \in P} [x]$, i.e. the set of y from which all of $P$ is visible. By assumption this set is non-empty. We may write $y > P$ in place of $y \in [P]$. For many applications, $X \otimes Y = X \times Y$.

Quantitative considerations are enabled through a function $\Phi : X \otimes Y \rightarrow [-\infty, \infty]$, the description, also called the effort function. This function determines the necessary effort by Observer in any atomic situation. We assume that $\Phi(x, y) = 0$ if $y \in Y_{\not\in P}$ and – the central assumption of our modelling – that $\Phi$ satisfies the perfect match principle or is proper, essentially that $\Phi(x, y) \geq \Phi(x, z)$. More precisely, we assume that there are functions $H : X \rightarrow [-\infty, \infty]$, called entropy, and $D : X \otimes Y \rightarrow [0, \infty]$, called divergence, such that, for all $(x, y) \in X \otimes Y$, firstly,

$$\Phi(x, y) = H(x) + D(x, y),$$ (1)

the linking identity, and, secondly, the fundamental inequality holds for D, i.e. $D(x, y) \geq 0$ with equality if and only if $y = x$.

The assumptions made are also expressed by saying that $(\Phi, H, D)$ is an effort-based information triple. A triple $(U, M, D)$ for which $-U - M, D)$ is an information triple after this definition is a utility-based information triple with $U$ as utility function and $M$ as maximal utility (as before, $D$ is the divergence).

Two descriptions which differ only by a positive scalar are equivalent. The choice among equivalent descriptions amounts to a choice of unit.

With a proper description $\Phi$, we define a (strict) feasible preparation as one of the form $\{\Phi^{\gamma} = h\}$ or a finite intersection of such sets. Here, $\Phi^{\gamma}$ denotes the marginal function $x \rightarrow \Phi(x, y)$ defined on $|y|$. This definition is sound on philosophical grounds. Further, it goes well with a definition of core, really an abstract notion of exponential families. For a family $\mathcal{P}$ of preparations (typically feasible ones), core($\mathcal{P}$) is the set of $y$ such that, for each $\mathcal{P} \in \mathcal{P}$, there exists $h$ such that $\Phi^{\gamma} = h$ on $\mathcal{P}$. See [14].

The choice of a proper effort function in concrete cases of interest is essential for the theory to render useful results. As examples of appropriate choices, we refer to [1], where cases of probabilistic modelling which lead to Tsallis entropy are discussed.

### 2.3. Inference

Consider partial information “$x \in \mathcal{P}$”.

The standard process of inference concerns the identification of a state in $\mathcal{P}$, the inferred state. This will be achieved by game theoretical methods involving the previously indicated game, $\gamma = \gamma(\mathcal{P})|\Phi|$, with $\Phi$ as objective function. For $\gamma$, also belief instances will be identified. An inferred belief instance $y^{\gamma}$ is, via the associated control, more of an instruction to Observer on how best to act regarding the set-up of experiments. Double inference gives Observer information both about what can be inferred about truth and how.

The value of $\gamma(\mathcal{P})$ for Nature is

$$\text{the MaxEnt-value, } H_{\text{max}}(\mathcal{P}).$$

Defining risk by

$$\text{Ri}(y|x) = \sup_{x \in \mathcal{P}} \Phi(x, y),$$

the value for Observer is the MinRisk-value of the game:

$$\text{Ri}_{\text{min}}(\mathcal{P}) = \inf_{y \not\in \mathcal{P}} \text{Ri}(y|x).$$ (3)

An optimal strategy for Nature is a strategy $x^{*} \in \mathcal{P}$ with $H(x^{*}) = H_{\text{max}}(\mathcal{P})$. An optimal strategy for Observer is a strategy $y^{*} \not\in \mathcal{P}$ with $\text{Ri}(y^{*}|x) = \text{Ri}_{\text{min}}(\mathcal{P})$.

The game is in equilibrium if $H_{\text{max}}(\mathcal{P}) = \text{Ri}_{\text{min}}(\mathcal{P}) < \infty$. By $\text{ctr}(\mathcal{P})$, the centre of $\mathcal{P}$, we denote the set $\mathcal{P} \cap \mathcal{P}$.

**Lemma 1** If $\gamma(\mathcal{P})$ is in equilibrium and both players have optimal strategies, then these strategies are unique, coincide and belong to the centre of $\mathcal{P}$.

**Proof** Let $x^{*} \in \mathcal{P}$ be any optimal strategy for Nature and $y^{*} \not\in \mathcal{P}$ any optimal strategy for Observer. By assumption, such strategies exist. Then $\Phi(x^{*}, y^{*}) \geq H(x^{*}) = H_{\text{max}}(\mathcal{P}) = \text{Ri}_{\text{min}}(\mathcal{P}) = \text{Ri}(y^{*}|x) \geq \Phi(x^{*}, y^{*})$, hence $\Phi(x^{*}, y^{*}) = H(x^{*})$ and we conclude that $y^{*} = x^{*}$ as desired. □

For a game in equilibrium with optimal strategies for both players, the common unique strategy is the bi-optimal strategy. In spite of the identity of the optimal strategies in such cases, we often use different notation, typically with $x^{*}$ when we focus on optimality for Nature and with $y^{*}$ when we focus on optimality for Observer.

**Theorem 1** Let $y^{*} = x^{*} \in \text{ctr}(\mathcal{P})$ with $H(x^{*}) < \infty$. Then $\gamma(\mathcal{P})$ is in equilibrium and has $x^{*}$ as bi-optimal strategy if and only if, for all $x \in \mathcal{P}$, $\Phi(x, y^{*}) \leq H(x^{*})$. When this condition is satisfied, the Pythagorean inequality as well as the dual Pythagorean inequality holds, i.e.

$$\forall x \in \mathcal{P} : H(x) + D(x, y^{*}) \leq H(x^{*}),$$

$$\forall y \not\in \mathcal{P} : \text{Ri}_{\text{min}}(\mathcal{P}) + D(x^{*}, y) \leq \text{Ri}(y|x).$$ (4) (5)

**Proof** In brief: In view of the assumptions imposed, the condition stated is one of the famous saddle-value inequalities often ascribed to Nash (but in the present simple case due to von Neumann), and the other saddle-value inequality is automatically fulfilled due to the perfect match principle. The result follows from these observations.

The Pythagorean inequality is a simple reformulation of the inequality $\Phi(x, y^{*}) \leq H(x^{*})$ and the dual Pythagorean inequality holds since, for $y \not\in \mathcal{P}$, $\text{Ri}_{\text{min}}(\mathcal{P}) + D(x^{*}, y) = H(x^{*}) + D(x^{*}, y) = \Phi(x^{*}, y) \leq \text{Ri}(y|x)$. □

The results above are developed for an effort-based information triple. Similar, or rather dual results apply to utility-based information triples. Then Nature is a minimizer, Observer a maximizer. We leave it to the reader to formulate appropriate concepts and results.
3. SPECIAL FEATURES

3.1. Adding a geometric flavour

Let us look specifically at models of updating. For this, \( D \) is a divergence function on \( X \otimes Y \). Suppose it satisfies the fundamental inequality, \( y_0 \) is a suitable prior and \( P \) a preparation such that \( D(0) < \infty \) on \( P \). We consider the utility-based information triple \( (U(y_0), D, P) \) with \( U(y_0) = D(x, y_0) - D(x, y) \), representing updating gain. The associated game is denoted \( \gamma = \gamma(P|U(y_0)) \). An optimal strategy \( x^* \) for Nature, if unique, is the \( D \)-projection of \( y_0 \) on \( P \), i.e., the unique element in \( P \) such that \( D(x^*, y_0) = D(0) \), the infimum of \( D(x, y_0) \) with \( x \in P \). Given \( y > P \), the guaranteed updating gain for Observer associated with the posterior \( y \) and the maximum guaranteed updating gain are given by

\[
G_{tu}(y|P, y_0) = \inf_{x \in P} U_{|y_0}(x, y) \tag{6}
\]

\[
G_{tu_{\max}}(P, y_0) = \sup_{y \in P} G_{tu}(y|P, y_0) \tag{7}
\]

Before introducing geometrical-like elements, note the following result which follows directly from Theorem 1:

**Theorem 2** A necessary and sufficient condition that \( \gamma \) is in equilibrium with \( x^* \in \text{ctr}(P) \) as bi-optimal strategy is that the Pythagorean inequality holds which, in this case means that, for \( x \in P \),

\[
D(x, y_0) \geq D(x, x^*) + D(x^*, y_0). \tag{8}
\]

If so, \( x^* \) is the \( D \)-projection of \( y_0 \) on \( P \).

Next, consider the open divergence ball with centre \( y_0 \) and radius \( r \), defined as the set

\[
B(y_0, r) = \{ D(0) < r \}. \tag{9}
\]

Also consider open half-spaces of size \( a \),

\[
\sigma^+(y, a|y_0) = \left\{ x \left| U_{|y_0} < a \right. \right\}. \tag{10}
\]

and, in particular, the open half-space

\[
\sigma^+(y|y_0) = \left\{ x \left| U_{|y_0} < D(y, y_0) \right. \right\}. \tag{11}
\]

We say that a set is external to \( P \) if it is contained in the complement of \( P \). The following result characterizes the values for the players in \( \gamma \) in geometrically flavoured terms, also in cases where \( \gamma \) is not in equilibrium:

**Proposition 1** (i) The value \( D_{\text{min}}(P, y_0) \) is the size of the largest ball \( B(y_0, r) \) which is external to \( P \), and the maximal guaranteed updating gain \( G_{tu_{\max}}(P, y_0) \) is the supremum of \( a \) for which there exists \( y > P \) such that the half-space \( \sigma^+(y, a|y_0) \) is external to \( P \).

(ii) The updating game \( \gamma(P|U(y_0)) \) is in equilibrium and has a bi-optimal strategy if and only if, for some \( y \in P \), \( \sigma^+(y|y_0) \) is external to \( P \). When this condition holds, \( y \) is the bi-optimal strategy, in particular, \( y \) is the \( D \)-projection of \( y_0 \) on \( P \).

If you consider the case where divergence is squared Euclidean distance, the geometric significance of this result becomes clear.

3.2. Adding convexity

For this subsection, \( X \) is a convex topological space, the marginals \( \Phi^\gamma \) are affine and the marginals \( D_\gamma : y \to D(x, y) \) are lower semi-continuous on \( X \).

Then, for every convex combination \( \pi = \sum a_i x_i \),

\[
H(\pi) = \sum a_i H(x_i) + \sum a_i D(x_i, \pi). \tag{12}
\]

In particular, \( H \) is strictly concave on \( X \).

Further, if \( H(\pi) < \infty \), then, for every \( y \in Y \), the compensation identity holds:

\[
\sum a_i D(x_i, y) = D(\pi, y) + \sum a_i D(x_i, \pi). \tag{13}
\]

In particular, for \( y \in Y \), the restriction of \( D_\gamma \) to convex preparations \( P \) with \( H_{\text{max}}(P) < \infty \) is strictly convex.

Let us look at a game \( \gamma(P) \). From Theorem 1 we realize the importance of the condition

\[
\forall x \in P : \Phi(x, y^*) \leq H(x^*). \tag{14}
\]

with \( y^* = x^* \in P \). It leads to equilibrium of \( \gamma(P) \) and bi-optimality of \( x^* \). In particular, it implies that \( H(x^*) = H_{\text{max}}(P) \).

**Proof** Consider an element \( x \in P \) and apply (12) to a convex combination of the form \( y_n = (1 - \frac{1}{n}) x^* + \frac{1}{n} x \). We find that \( H(x^*) \geq H(y_n) \geq (1 - \frac{1}{n}) H(x^*) + \frac{1}{n} H(x) + \frac{1}{n} D(x, y_n) \) from which we conclude that \( H(x^*) \geq H(x^*) + D(x, y_n) \). Exploiting the assumed lower semi-continuity, \( H(x) + D(x, x^*) \leq H(x^*) \) follows. As \( x \in P \) was arbitrary, (14) holds. Then apply Theorem 1. \( \square \)

We find it important that Theorem 3 also applies to the updating models of Theorem 2. Analyzing this it appears that this is indeed the case, provided you assume that the divergence function which Theorem 2 depends on satisfies the compensation identity. In this way one derives abstract versions of by now classical results of Shannon theory related to information projections and Pythagorean inequalities. These results go back to Čencov and Csiszár, cf. [15] and [16]. Also of relevance are [17] and [18].

3.3. Axiomatization

The key object which appears to be worth while axiomatizing is the information triples. Basic conditions are centred around the linking identity, the fundamental identity, convexity of \( X \) and affinity of the marginals \( \Phi^\gamma \). This may be supplied with topological conditions. Details may be found in [19]. One may start from atomic triples for which \( X \) and \( Y \) are the reals or the non-negative reals. A process of integration leads to more complicated triples,
often related to **Bregman divergencies**. Other processes involve **relativization and randomization**. A systematic study as indicated also helps in defining concrete triples of interest.

### 4. REFERENCES


Information Distance: New Developments

Paul M.B. Vitányi
CWI, Amsterdam, The Netherlands
(Invited Lecture)

Abstract

In pattern recognition, learning, and data mining one obtains information from information-carrying objects. This involves an objective definition of the information in a single object, the information to go from one object to another object in a pair of objects, the information to go from one object to any other object in a multiple of objects, and the shared information between objects. This is called “information distance.” We survey a selection of new developments in information distance.

I. The Case $n = 2$

The clustering we use is hierarchical clustering in dendrograms based on a new fast heuristic for the quartet method [5]. If we consider $n$ objects, then we find $n^2$ pairwise distances. These distances are between natural data. We let the data decide for themselves, and construct a hierarchical clustering of the $n$ objects concerned. For details see the cited reference. The method takes the $n \times n$ distance matrix as input, and yields a dendrogram with the $n$ objects as leaves (so the dendrogram contains $n$ external nodes or leaves and $n - 2$ internal nodes. We assume $n \geq 4$.

The method is available as an open-source software tool, [2].

Our aim is to capture, in a single similarity metric, every effective distance: effective versions of Hamming distance, Euclidean distance, edit distances, alignment distance, Lempel-Ziv distance, and so on. This metric should be so general that it works in every domain: music, text, literature, programs, genomes, executables, natural language determination, equally and simultaneously. It would be able to simultaneously detect all similarities between pieces that other effective distances can detect separately.

Such a “universal” metric was co-developed by us as a normalized version of the “information metric” of [1], [9]. There it was shown that the information metric minorizes up to a constant all effective distances satisfying a mild density requirement (excluding for example distances that are 1 for every pair $x, y$ such that $x \neq y$). This justifies the notion that the information distance is universal.

We may be interested what happens in terms of properties or features of the pair of objects analyzed, say $x$ and $y$. It can be shown that the information distance captures every property of which the Kolmogorov complexity is logarithmic in the length of $\min\{ |x|, |y| \}$. If those lengths go to infinity, then logarithm of those lengths go to infinity too. In this case the information distance captures every property.

This information distance (actually a metric up to minor additive terms) is normalized so that the resulting distances are in $[0, 1]$ and can be shown to retain the metric property, [8]. The result is the “normalized information distance” (actually a metric up to negligible terms). All this is in terms of Kolmogorov complexity [9].

It articulates the intuition that two objects are deemed close if we can significantly “compress” one given the information in the other, that is, two pieces are more similar if we can more succinctly describe one given the other. The normalized information distance discovers all effective similarities in the sense that if two objects are close according to some effective similarity, then they are also close according to the normalized information distance.

Put differently, the normalized information distance represents similarity according to the dominating shared feature between the two objects being compared. In comparisons of more than two objects, different pairs may have different dominating features. For every two objects,
this normalized information metric distance zooms in on the dominant similarity between those two objects out of a wide class of admissible similarity features. Since the normalized information distance also satisfies the metric (in)equality, and takes values in $[0, 1]$, it may be called “the” similarity metric.

Unfortunately, the universality of the normalized information distance comes at the price of noncomputability. Recently we have shown that the normalized information distance is not even semicomputable (this is weaker than computable) and there is no semicomputable function at a computable distance of it [13].

Since the Kolmogorov complexity of a string or file is the length of the ultimate compressed version of that file, we can use real data compression programs to approximate the Kolmogorov complexity. Therefore, to apply this idealization we can use real data compression programs to approximate the normalized distance of it [13].

Unfortunately, the universality of the normalized information distance comes at the price of noncomputability. Recently we have shown that the normalized information distance is not even semicomputable (this is weaker than computable) and there is no semicomputable function at a computable distance of it [13].

The rise of the Internet has enticed millions of users to type in trillions of characters to create billions of web pages of on average low quality contents. The sheer mass of the information available about almost every conceivable topic makes it likely that extremes will cancel and the majority or average is meaningful in a low-quality approximate sense. Below, we give a general method to tap the amorphous low-grade knowledge available for free on the Internet, typed in by local users aiming at personal gratification of diverse objectives, and yet globally achieving what is effectively the largest semantic electronic database in the world. Moreover, this database is available for all by using any search engine that can return aggregate page-count estimates like Google for a large range of search-queries.

While the previous NCD method that compares the objects themselves using (1) is particularly suited to obtain knowledge about the similarity of objects themselves, irrespective of common beliefs about such similarities, we now develop a method that uses only the name of an object and obtains knowledge about the similarity of objects by tapping available information generated by multitudes of web users. The new method is useful to extract knowledge from a given corpus of knowledge, in this case the pages on the Internet accessed by a search engine returning aggregate page counts, but not to obtain true facts that are not common knowledge in that database. For example, common viewpoints on the creation myths in different religions may be extracted by the web-based method, but contentious questions of fact concerning the phylogeny of species can be better approached by using the genomes of these species, rather than by opinion. This approach was proposed by [4]. We skip the theory.

In contrast to strings $x$ where the complexity $Z(x)$ represents the length of the compressed version of $x$ using compressor $Z$, for a search term $x$ (just the name for an object rather than the object itself), the code of length $G(x)$ represents the shortest expected prefix-code word length of the event $x$ (the number of pages of the Internet returned by a given search engine). The associated normalized web distance (NWD) is defined just as (1) with the search engine in the role of compressor yielding code lengths $G(x), G(y)$ for the singleton search terms $x, y$ being compared and a code length $G(x, y)$ for the doubleton pair $(x, y)$, by

$$NW D(x, y) = \frac{G(x,y) - \min(G(x), G(y))}{\max(G(x), G(y))}. \quad (2)$$

This NWD uses the background knowledge on the web as viewed by the search engine as conditional information.

The same formula as (2) can be written in terms of frequencies of the number of pages returned on a search query as

$$NW D(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log N - \min\{\log f(x), \log f(y)\}}, \quad (3)$$

where

$$NW D(x, y) = \frac{Z(x,y) - \min(Z(x), Z(y))}{\max(Z(x), Z(y))}. \quad (1)$$

where for convenience we have replaced the pair $(x, y)$ in the formula by the concatenation $xy$, and we ignore logarithmic terms in the numerator and denominator, see [8], [3]. In [3] we propose axioms to capture the real-world setting, and show that (1) approximates optimality. Actually, the NCD is a family of compression functions parameterized by the given data compressor $Z$.

A. Web-based Similarity

To make computers more intelligent one would like to represent meaning in computer-digestable form. Long-term and labor-intensive efforts like the Cyc project [7] and the WordNet project [11] try to establish semantic relations between common objects, or, more precisely, names for those objects. The idea is to create a semantic web of such vast proportions that rudimentary intelligence and knowledge about the real world spontaneously emerges. This comes at the great cost of designing structures capable of manipulating knowledge, and entering high quality contents in these structures by knowledgeable human experts. While the efforts are long-running and large scale, the overall information entered is minute compared to what is available on the Internet.

The rise of the Internet has enticed millions of users to type in trillions of characters to create billions of web pages of on average low quality contents. The sheer mass of the information available about almost every conceivable topic makes it likely that extremes will cancel and the majority or average is meaningful in a low-quality approximate sense. Below, we give a general method to tap the amorphous low-grade knowledge available for free on the Internet, typed in by local users aiming at personal gratification of diverse objectives, and yet globally achieving what is effectively the largest semantic electronic database in the world. Moreover, this database is available for all by using any search engine that can return aggregate page-count estimates like Google for a large range of search-queries.

While the previous NCD method that compares the objects themselves using (1) is particularly suited to obtain knowledge about the similarity of objects themselves, irrespective of common beliefs about such similarities, we now develop a method that uses only the name of an object and obtains knowledge about the similarity of objects by tapping available information generated by multitudes of web users. The new method is useful to extract knowledge from a given corpus of knowledge, in this case the pages on the Internet accessed by a search engine returning aggregate page counts, but not to obtain true facts that are not common knowledge in that database. For example, common viewpoints on the creation myths in different religions may be extracted by the web-based method, but contentious questions of fact concerning the phylogeny of species can be better approached by using the genomes of these species, rather than by opinion. This approach was proposed by [4]. We skip the theory.

In contrast to strings $x$ where the complexity $Z(x)$ represents the length of the compressed version of $x$ using compressor $Z$, for a search term $x$ (just the name for an object rather than the object itself), the code of length $G(x)$ represents the shortest expected prefix-code word length of the event $x$ (the number of pages of the Internet returned by a given search engine). The associated normalized web distance (NWD) is defined just as (1) with the search engine in the role of compressor yielding code lengths $G(x), G(y)$ for the singleton search terms $x, y$ being compared and a code length $G(x, y)$ for the doubleton pair $(x, y)$, by

$$NW D(x, y) = \frac{G(x,y) - \min(G(x), G(y))}{\max(G(x), G(y))}. \quad (2)$$

This NWD uses the background knowledge on the web as viewed by the search engine as conditional information.

The same formula as (2) can be written in terms of frequencies of the number of pages returned on a search query as

$$NW D(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log N - \min\{\log f(x), \log f(y)\}}, \quad (3)$$

where

$$NW D(x, y) = \frac{Z(x,y) - \min(Z(x), Z(y))}{\max(Z(x), Z(y))}. \quad (1)$$

where for convenience we have replaced the pair $(x, y)$ in the formula by the concatenation $xy$, and we ignore logarithmic terms in the numerator and denominator, see [8], [3]. In [3] we propose axioms to capture the real-world setting, and show that (1) approximates optimality. Actually, the NCD is a family of compression functions parameterized by the given data compressor $Z$.
and if $f(x), f(y) > 0$ and $f(x, y) = 0$ then $\text{NWD}(x, y) = \infty$. It is easy to see that

1) $\text{NWD}(x, y)$ is undefined for $f(x) = f(y) = 0$;
2) $\text{NWD}(x, y) = \infty$ for $f(x, y) = 0$ and either or both $f(x) > 0$ and $f(y) > 0$; and
3) $\text{NWD}(x, y) \geq 0$ otherwise.

The number $N$ is related to the number of pages $M$ indexed by the search engine we use. Our experimental results suggest that every reasonable (greater than any $f(x)$) value can be used for the normalizing factor $N$, and our results seem in general insensitive to this choice. In our software, this parameter $N$ can be adjusted as appropriate, and we often use $M$ for $N$. In the [4] we analyze the mathematical properties of NWD, and prove the universality of the search engine distribution. We show that the NWD is not a metric, in contrast to the NCD. The generic example showing the nonmetricity of semantics (and therefore the NWD) is that a man is close to a centaur, and a centaur is close to a horse, but a man is very different from a horse.

B. Question-Answer System

A typical procedure for finding an answer on the Internet consists in entering some terms regarding the question into a Web search engine and then browsing the search results in search for the answer. This is particularly inconvenient when one uses a mobile device with a slow internet connection and small display. Question-answer (QA) systems attempt to solve this problem. They allow the user to enter a question in natural language and generate an answer by searching the Web autonomously. The QA system QUANTA [15] that uses variants of the NCD and the NWD to identify the correct answer to a question out of several candidates for answers. QUANTA is remarkable in that it uses neither NCD nor NWD introduced so far, but a variation that is nevertheless based on the same theoretical principles. This variation is tuned to the particular needs of a QA system. Without going in too much detail it uses the maximal overlap of program $p$ going from file $x$ to file $y$, and program $q$ going from file $y$ to file $x$. The system QUANTA is 1.5 times better (according to generally used measures) than its competition.

II. $n > 2$

In many applications we are interested in shared information between many objects instead of just a pair of objects. For example, in customer reviews of gadgets, in blogs about public happenings, in newspaper articles about the same occurrence, we are interested in the most comprehensive one or the most specialized one. Thus, we want to extend the information distance measure from pairs to multiples. This approach was introduced in [10] while most of the theory is developed in [14].

Let $X$ denote a finite list of $m$ finite binary strings defined by $X = (x_1, \ldots, x_m)$, the constituting strings ordered length-increasing lexicographic. We use lists and not sets, since if $X$ is a set we cannot express simply the distance from a string to itself or between strings that are all equal. Let $U$ be the reference universal Turing machine. Given the string $x_i$ we define the information distance to any string in $X$ by $E_{\text{max}}(X) = \min \{|p| : U(x_i, p, j) = x_j \text{ for all } x_i, x_j \in X\}$. It is shown in [10], Theorem 2, that

$$E_{\text{max}}(X) = \max_{x \in X} K(X|x), \quad (4)$$

up to a logarithmic additive term. Define $E_{\text{min}}(X) = \min_{x \in X} K(X|x)$. Theorem 3 in [10] states that for every list $X = (x_1, \ldots, x_m)$ we have

$$E_{\text{min}}(X) \leq E_{\text{max}}(X) \leq \min_{1 \leq i \leq m} \sum_{x_i, x_k \in X \& k \neq i} E_{\text{max}}(x_i, x_k), \quad (5)$$

up to a logarithmic additive term. This is not a corollary of (4) as stated in [10], but both inequalities follow from the definitions. The left hand side is interpreted as the program length of the “most comprehensive object that contains the most information about all the others [all elements of $X$],” and the right hand side is interpreted as the program length of the “most specialized object that is similar to all the others.”

Information distance for multiples, that is, finite lists, appears both practically and theoretically promising. The results below appear in [14]. In all cases the results imply the corresponding ones for the pair-wise information distance defined as follows. The information distance in [1] between strings $x_1$ and $x_2$ is $E_{\text{max}}(x_1, x_2) = \max \{K(x_1|x_2), K(x_2|x_1)\}$. In the [14] $E_{\text{max}}(X) = \max_{x \in X} K(X|x)$. Two these definitions coincide for $|X| = 2$ since $K(x, y|x) = K(y|x)$ up to an additive constant term. The reference investigate the maximal overlap of information which for $|X| = 2$ specializes to Theorem 3.4 in [1]. A corollary in [14] shows (4) and another corollary shows that the left hand side of (5) can indeed be taken to correspond to a single program embodying the “most comprehensive object that contains the most information about all the others” as stated but not argued or proved in [10]. The reference proves metricity and universality which for $|XY| = 2$ (for metricity) and $|X| = 2$ (for universality) specialize to Theorem 4.2 in [1]; additivity; minimum overlap of information which for $|X| = 2$ specializes to Theorem 8.3.7 in [12]; and the nonmetricity of normalized information distance for lists of more than two elements and the failure of certain proposals of a normalizing factor (to achieve a normalized version). In contrast, for lists of two elements we can normalize the information distance as in Lemma V.4 and Theorem V.7 of
The definitions are of necessity new as are the proof ideas. Remarkably, the new notation and proofs for the general case are simpler than the mentioned existing proofs for the particular case of pairwise information distance.

III. Conclusion

By now applications abound. See the many references to the papers [8], [3], [4] in Google Scholar.

The methods turns out to be more-or-less robust under change of the underlying compressor-types: statistical (PPMZ), Lempel-Ziv based dictionary (gzip), block based (bzip2), or special purpose (Gencompress). Obviously the window size matters, as well as how good the compressor is. For example, PPMZ gives for mtDNA of the investigated species diagonal elements \( \text{NCD}(x,x) \) between 0.002 and 0.006. The compressor bzip2 does considerably worse, and gzip gives something in between 0.5 and 1 on the diagonal elements. Nonetheless, for texts like books gzip does fine in our experiments; the window size is sufficient and we do not use the diagonal elements. But for genomics gzip is no good.

References

MINIMUM VARIATIONAL STOCHASTIC COMPLEXITY AND AVERAGE GENERALIZATION ERROR IN LATENT VARIABLE MODELS

Kazuho Watanabe

Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5, Takayama-cho, Ikoma, Nara, 630-0192, JAPAN, wkazuho@is.naist.jp

ABSTRACT

Bayesian learning is often accomplished with approximation schemes because it requires intractable computation of the posterior distributions. In this paper, focusing on the approximation scheme, variational Bayes method, we investigate the relationship between the asymptotic behavior of variational stochastic complexity or free energy, which is the objective function to be minimized by variational Bayes, and the generalization ability of the variational Bayes approach. We show an inequality which implies a relationship between the minimum variational stochastic complexity of latent variable models and the generalization error of the approximate predictive distribution. This relationship is also examined by a numerical experiment of the variational Bayesian learning of the GMM.

1. INTRODUCTION

Bayesian estimation provides a powerful framework for learning from data. Recently, its asymptotic theory has been established, which supports its effectiveness for latent variable models such as the Gaussian mixture model (GMM) and hidden Markov model (HMM). More specifically, a formula for evaluating asymptotic forms of Bayesian mixture-type stochastic complexity or free energy was obtained and the generalization errors of statistical models have been intensively analyzed [1, 2].

Practically, however, Bayesian estimation requires some approximation method since computing the Bayesian posterior distribution is intractable in general. In this study, we focus on the approximation method, variational Bayes for Bayesian estimation. This method has been successfully applied to latent variable models such as mixture models and HMMs [3, 4]. Furthermore, its asymptotic analysis has progressed in several statistical models [5, 6]. More specifically, a formula for evaluating the asymptotic form of the minimum variational free energy was derived [6]. The variational free energy, what we call variational stochastic complexity in this paper, is the objective function to be minimized by variational Bayes and provides an upper bound for the Bayesian mixture-type stochastic complexity.

In this paper, we provide as a byproduct of this analysis, a quantity which is related to the generalization ability of the variational Bayesian approach. Analysis of generalization ability of a learning machine when it is used with the variational Bayesian approximation has been successful in quite limited cases [7]. We show an inequality which implies a relationship between the minimum variational stochastic complexity of latent variable models and the generalization error of the approximate predictive distribution. This relationship is also examined by a numerical experiment of the variational Bayesian learning of the GMM.

2. VARIATIONAL BAYES FOR LATENT VARIABLE MODELS

Let $y^n_i = \{y_i\}_{i=1}^n$ be the latent (unobserved) variables corresponding to the i.i.d. observations $x^n_i = \{x_i\}_{i=1}^n$ and consider the latent variable model with parameter $w$,

$$p(x^n_i|w) = \sum_{y^n_i} p(x^n_i, y^n_i|w) = \prod_{i=1}^n p(x_i, y_i|w),$$

where $\sum_{y^n_i}$ denotes the summation over all possible realizations of the latent variables.

The Bayesian posterior distribution of the latent variables and parameter $w$ is

$$p(y^n_i, w|x^n_i) = \frac{p(x^n_i, y^n_i|w)p_0(w)}{\sum_{y^n_i} \int p(x^n_i, y^n_i|w)p_0(w)dw},$$

where $p_0(w)$ is the prior distribution of $w$. The posterior distribution is intractable when the marginal likelihood $Z(x^n_i) = \sum_{y^n_i} \int p(x^n_i, y^n_i|w)p_0(w)dw$ requires the sum over exponentially many terms as in the Gaussian mixture model (GMM) and the hidden Markov model (HMM). In this article,

$$F(x^n_i) = -\log Z(x^n_i)$$

is referred to as the Bayesian mixture-type stochastic complexity [8].

The variational Bayesian framework approximates the Bayesian posterior distribution $p(y^n_i, w|x^n_i)$ of the hidden variables and the parameters by the variational posterior distribution $q(y^n_i, w|x^n_i)$, which factorizes as

$$q(y^n_i, w|x^n_i) = q(y^n_i|x^n_i)q(w|x^n_i),$$

(1)

where $q(y^n_i|x^n_i)$ and $q(w|x^n_i)$ are probability distributions on the hidden variables and the parameters respectively. The variational posterior $q(y^n_i, w|x^n_i)$ is chosen so that
it minimizes the functional $\mathcal{F}[q]$, referred to as the variational stochastic complexity or variational free energy,

$$
\mathcal{F}[q] = F(x^n_1) + K(q(y^n_0, w|x^n_1)||p(y^n_0, w|x^n_1)),
$$

(2)

where $K(q(y^n_0, w|x^n_1)||p(y^n_0, w|x^n_1))$ is the Kullback Leibler information from the variational posterior $q(y^n_0, w|x^n_1)$ to the Bayesian posterior $p(y^n_0, w|x^n_1)$. This reduces to the following alternating optimization of $q(y^n_0|x^n_1)$ and $q(w|x^n_1)$,

$$
q(w|x^n_1) \propto p_0(w) \exp \left( \log p(x^n_1, y^n_0|w) \right) q(y^n_0|x^n_1),
$$

(3)

and

$$
q(y^n_0|x^n_1) \propto \exp \left( \log p(x^n_1, y^n_0|w) \right) q(w|x^n_1),
$$

(4)

where $\langle \cdot \rangle_p$ denotes the expectation with respect to $p$ [3, 4].

Let

$$
\mathcal{F}_\text{min}(x^n_1) = \min_{q(y^n_0|x^n_1), q(w|x^n_1)} \mathcal{F}[q]
$$

be the minimum variational stochastic complexity. We assume that $p(x|w^*)$ with the parameter $w^*$ is the underlying distribution generating the data $x^n_1$ independently and identically. Because of the non-identifiability of the latent variable model, the set of true parameters $x^n_1$ is not generally a point but can be a union of several manifolds with singularities [1].

For arbitrary $\tilde{w}^* \in W^*$,

$$
\mathcal{F}^\dagger(n) \equiv \langle \mathcal{F}_\text{min}(x^n_1) + \log p(x^n_1|w^*) \rangle_{p(x^n_1|w^*)}
$$

is bounded from above by

$$
U^*(n) = \langle U^*(x^n_1) \rangle_{p(x^n_1|w^*)},
$$

(5)

where $U^*(x^n_1)$ is given by

$$
- \log \int \exp \left\{ \log \frac{p(x^n_1, y^n_0|w)}{p(x^n_1, y^n_0|\tilde{w}^*)} \right\} p_0(w) dw.
$$

Asymptotic evaluation of $U^*(n)$ is elaborated in [6] with an alternative view of variational Bayes as a local variational approximation [9].

3. VARIATIONAL STOCHASTIC COMPLEXITY AND GENERALIZATION ERROR

Let $p(x, y|\tilde{w}^*)$ be the true distribution of the observed variable $x$ and the latent variable $y$ which has the marginal distribution $p(x|w^*)$. We define by

$$
\bar{G}^\dagger(x^n_1) = K(p(x, y|\tilde{w}^*)||\bar{p}^\dagger(x, y|x^n_1)),
$$

(6)

the generalization error of the predictive distribution,

$$
\bar{p}^\dagger(x, y|x^n_1) = \langle p(x, y|w^*) \rangle_{q^\dagger(w|x^n_1)} = \int p(x, y|w^*) q^\dagger(w|x^n_1) dw,
$$

(7)

where $q^\dagger(w|x^n_1)$ is the optimal approximating posterior distribution (3) for $q(y_0^n|x^n_1) = p(y_0^n|x^n_1, \tilde{w}^*)$. We denote its mean by

$$
\bar{G}^\dagger(n) = \langle \bar{G}^\dagger(x^n_1) \rangle_{\prod_{i=1}^n p(x_i|w^*)}.
$$

Then, the following inequality holds,

$$
U^*(n+1) - U^*(n) \geq \bar{G}^\dagger(n),
$$

(8)

where $U^*(n)$ is the upper bound (5) of the minimum variational stochastic complexity.

(Proof of the inequality (8)) Let

$$
\tilde{p}_n(y) = p(y|x_i, \tilde{w}^*).
$$

Then it follows that

$$
U^*(x^n_1) - U^*(x^n_1) = \int \prod_{i=1}^n p(x_i, y_i|w^*) \log \frac{p(x_i, y_i|w^*)}{\tilde{p}_n(y)} p_0(w) dw
$$

$$
= \int \prod_{i=1}^n p(x_i, y_i|w^*) \log \frac{p(x_i, y_i|w^*)}{\tilde{p}_n(y)} p_0(w) dw
$$

$$
= \int \prod_{i=1}^n \log \left\{ \frac{p(x_{i+1}, y|w^*)}{p(x_{i+1}, y|\tilde{w}^*)} \right\} \tilde{p}_n(y) q^\dagger(w|x^n_1)
$$

$$
\geq \int \prod_{i=1}^n p(y|x_{i+1}, \tilde{w}^*) \log \frac{p(x_{i+1}, y|\tilde{w}^*)}{\tilde{p}_n(y)} q^\dagger(w|x^n_1),
$$

(9)

In the last inequality, we have applied Jensen’s inequality due to the convexity of the function log $\int \exp \langle \cdot \rangle p(w) dw$.

Taking expectation with respect to $\prod_{i=1}^n p(x_i|w^*)$ in both sides of the above inequality yields the inequality (8). (Q.E.D)

The inequality (8) is analogous to the equality,

$$
F^*(n+1) - F^*(n) = G(n),
$$

which holds for the average mixture-type stochastic complexity,

$$
F^*(n) = \langle F(x^n_1) + \log p(x^n_1|w^*) \rangle_{p(x^n_1|w^*)},
$$

and the generalization error of the Bayesian predictive distribution,

$$
G(n) = \langle K(p(x|w^*)||p(x^n_1|w^*)) \prod_{i=1}^n p(x_i|w^*) \rangle,
$$

where $p(x|w^*) = \langle p(x|w^*) \rangle_{p(w|x^n_1)}$.

If $U^*(n)$ has the asymptotic form $U^*(n) \simeq \chi \log n + O(1)$ as in eq. (14), the inequality (8) suggests that

$$
\bar{G}^\dagger(n) \leq \frac{\chi}{n} + o \left( \frac{1}{n} \right).
$$

(11)

This means that the coefficient $\chi$ of the leading term of $U^*(n)$ is directly related to the generalization error of the variational Bayes approach measured by eq. (6).

By applying Jensen’s inequality with respect to $\langle \cdot \rangle_{q^\dagger(w|x^n_1)}$ and the convexity of the negative logarithmic function in eq. (9), we further obtain,

$$
U^*(n+1) - U^*(n) \leq \tilde{G}^\dagger(n),
$$

76
where $\tilde{G}^*(n)$ is the expectation of the Gibbs generalization error, 
\[
\langle K(p(x, y|\hat{\theta}^*))[p(x, y|\theta^*)]\rangle_q^*(w|\alpha_k^*) .
\]

4. GAUSSIAN MIXTURE MODEL

Let \( g(x|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{||x-\mu||^2}{2}\right\} \) be the \( M \)-dimensional Gaussian density and consider the GMM with \( K \) components, 
\[
p(x|w) = \sum_y p(x, y|w),
\]
where 
\[
p(x, y|w) = \prod_{k=1}^K \{a_k g(x|\mu_k)\}^{y_k(k)} . (12)
\]

where \( x \in \mathbb{R}^M \) and the parameter vector \( w \) consists of the mean vectors \( \{\mu_k\}_{k=1}^K \) and the mixing proportions \( a = \{a_k\}_{k=1}^K \) that satisfy \( 0 \leq a_k \leq 1 \) for \( k = 1, \ldots, K \) and \( \sum_{k=1}^K a_k = 1 \). The latent variable \( y = (y^{(1)}, y^{(2)}, \ldots, y^{(K)}) \) indicates the component from which the datum \( x \) is generated, that is, \( y^{(k)} = 1 \) if \( x \) is from the \( k \)th component and \( y^{(k)} = 0 \) otherwise. The variational Bayes framework is successfully applied to this model using the prior distribution,
\[
p_0(w) = p_0(a) \prod_{k=1}^K p_0(\mu_k) . (13)
\]
where 
\[
p_0(a) = \frac{\Gamma(Ka_0)}{\Gamma(a_0)^K} \prod_{k=1}^K a_k^{a_0-1}
\]
is the Dirichlet distribution with hyperparameter \( a_0 > 0 \) and 
\[
p_0(\mu_k) = \sqrt{\frac{\beta_0}{2\pi}} \exp\left\{-\frac{\beta_0||\mu_k - \nu_0||^2}{2}\right\}
\]
is the Gaussian distribution with hyperparameters \( \beta_0 > 0 \) and \( \nu_0 \in \mathbb{R}^M \). They are the conjugate prior distributions for the mixing proportions and each mean vector respectively.

Let the true distribution \( p(x|\theta^*) \) be the GMM with \( K_0 \) components, that is, realizable by the model. Then it can be shown that the upper bound of the minimum variational stochastic complexity is asymptotically bounded as 
\[
U^*(n) \leq \overline{\lambda} \log n + O(1), (14)
\]
where 
\[
\overline{\lambda} = \left\{ \begin{array}{ll} 
(K - K_0)a_0 + \frac{MK_0 + K_0 - 1}{2} & (a_0 \leq \frac{M+1}{2}), \\
(M+1)(K - K_0)a_0 - (a_0 > \frac{M+1}{2}). 
\end{array} \right.
\]
The proof is given in [5, 6].

5. NUMERICAL EXPERIMENT

We implemented the variational Bayesian learning of the GMM with \( K \) components (12). For simplicity, we chose the true distribution to be the standard normal distribution in \( \mathbb{R}^2 \), \( g(x|[0,0]^T) \). According to the choice of \( \theta^* \) for evaluating \( \overline{\lambda} \) in eq.(14) [6], we consider this distribution as the choice, \( \tilde{\theta}^* = \{\tilde{\mu}_k\}, \tilde{\nu}_k \) for \( k = 0 \) for \( k = 2, \ldots, K \), \( \tilde{\mu}_k = (0,0)^T \) for \( k = 1, 2, \ldots, K \) and focus on the case where \( a_0 < (M+1)/2 = 1.5 \).

Samples of the size \( n = 100 \) were generated by the true distribution. The variational Bayes algorithm was executed 21 times with 20 different random initializations and the one from the true parameter \( \tilde{\theta}^* \). We adopted the estimate \( \hat{\theta}(w|x^n) \) that attained the minimum of the variational stochastic complexity and evaluated the generalization error,
\[
\overline{\theta}(x^n) = K(p(x, y|\theta^*)|p(x, y|x^n)), (15)
\]
where \( \hat{\theta}(x, y|x^n) = \langle p(x, y|\theta^*)|\tilde{\theta}(w|x^n) \rangle \) is the (approximate) predictive distribution.

To investigate the difference between \( \overline{\theta}(x^n) \) and \( \overline{\theta}(x^n) \) introduced in Section 5, we also evaluated \( \overline{\theta}(x^n) \), on the expectation of which we can show that 
\[
\overline{\theta}(n) \approx \left\{ \frac{M}{2} + (K - 1)a_0 \right\} \frac{1}{n} + o\left(\frac{1}{n}\right) . (16)
\]

Note that the coefficient \( \frac{M}{2} + (K - 1)a_0 \) is exactly equal to \( \overline{\lambda} \) in the inequality (14) for the case where \( K_0 = 1 \) and \( a_0 < \frac{M+1}{2} \). This means that the inequality (11) is tight in this case.

Additionally, we calculated the generalization error of the marginal distribution,
\[
\overline{\theta}(x^n) = K(p(x|\theta^*)|\tilde{\theta}(w|x^n)) (17)
\]
where \( \tilde{\theta}(x|\theta^*) = \langle p(x|\theta^*)|\tilde{\theta}(w|x^n) \rangle \) is the marginal predictive distribution.

Fig.1 and Fig.2 show the generalization errors for \( n = 100 \) and \( K_0 = 1 \) averaged over 100 trials with different data sets. Fig.1 is for the case of \( K = 2 \) with different values of the hyperparameter \( a_0 \). We can see that for small \( a_0 \), the behavior of the generalization error of the joint predictive distribution is well described by that of \( \overline{\theta}(n) \) and hence by the coefficient \( \overline{\lambda} \) in the upper bound (14). As \( a_0 \) tends larger, the average of \( \overline{\theta}(x^n) \) also increases, as does that of the generalization error \( \overline{\theta}(x^n) \) of the marginal distribution, although only slightly. This may be caused by overfitting. Fig.2 shows the average of the generalization errors for the case of \( a_0 = 0.2 \) with different number \( K \) of components. Again, we can see that for small \( a_0 \) the generalization error of the joint predictive distribution is described by \( \overline{\lambda} \) in eq.(14) while the generalization error of the marginal distribution stays constant even when the model becomes more redundant.

6. CONCLUSION

In this paper, we have investigated the average generalization error of the variational Bayesian approach for latent
variable models by deriving inequalities on the difference of the minimum variational stochastic complexity. We have demonstrated that the coefficient of the asymptotic minimum variational stochastic complexity partly describes the behavior of the generalization error. Thorough investigation of the generalization ability of the variational Bayes algorithm including the case for large $\alpha_0$ and for the marginal predictive distribution will be left for future work.

In the original (not approximate) Bayesian estimation, the universal relation among the quartet, Bayes and Gibbs generalization errors and Bayes and Gibbs training errors, was proved [1]. It is an important undertaking to explore such relationships among the quantities introduced in this paper for the approximate Bayesian estimation.

7. ACKNOWLEDGMENTS

This work was partly supported by KAKENHI (23700175) from the Japan Society for the Promotion of Science.

8. REFERENCES


MDL-based Modeling of Etymological Sound Change in the Uralic Language Family

Hannes Wettig, Suvi Hiltunen, Roman Yangarber
Department of Computer Science
University of Helsinki, Finland

Extended Abstract

We present work on the UraLink Project, jointly funded by the Academy of Finland and by the Russian Fund for the Humanities, aiming to study computational methods for analyzing etymological data in the Uralic Language Family. The Project is conducted jointly with partners who provide us with Uralic etymological databases.

One striking feature of the databases is a large amount of uncertainty, marked explicitly inside entries in the databases, and statements, or theories, that conflict between the databases. This alone indicates an urgent need to attempt to quantify the uncertainty in etymological data. Uncertainty is a natural and inescapable aspect of the complex task of establishing genetic relationships among languages by positing cognate sets. However, having rigorous methods for analyzing and quantifying the uncertainty can provide essential insights into the quality of the databases. Among other results, we seek measures of internal consistency of the datasets that have been compiled by linguists.

We approach the problem as follows. Our starting point is two databases—different compilations of cognate sets, done at different times by different scholars. We treat each data set in turn and simultaneously try to:

1. establish a complete alignment of all lexical items contained in the data set—an alignment that is globally optimal, and
2. discover the rules of correspondence that govern the alignments.

There is a duality to the problem posed in this way: on one hand, if we had a complete and accurate alignment of the entire data, we could simply read off the rules of correspondence from the aligned data. On the other hand, if we had a complete set of rules of correspondence, we could construct the alignment, simply by following the rules. We have neither. This suggests an iterative method: we begin with some random alignment, and proceed to discover incrementally better rules from the aligned data, while at the same time obtaining an improved alignment. Our methods are automatic, using the non-aligned data only, with no additional parameters and no supervision.

We make the assumption that the “true” rules of regular sound correspondence that underly the relationships within the family are violated very infrequently; however, we admit the possibility in principle that they may be violated. This suggests a probabilistic (rather than deterministic) approach, since that allows us to model in a continuous fashion the probability that some rule will hold. It also suggests that the information-theoretic Minimum Description Length Principle (MDL) is a natural way to model our data: we try to encode as much of the information in the dataset as possible via the rules, and then explicitly code the exceptions, which are hopefully infrequent. MDL establishes a link between the quality of our “theory” and the code-length of the data: if we can find very crisp rules that describe the sound correspondences among the languages, we will be able to encode the data compactly (short code-length) by writing down the rules. If the rules are too weak—not sufficiently descriptive—we will need to spend a lot of code length on writing down the exceptions. If the rules are overly strong—with conditions that are too rigid, and therefore apply in very few instances—we will spend a lot of code-length on writing down the rules themselves. MDL finds the optimal balance between these two undesirable extremes.

Our presentation is structured as follows. First we introduce a group of models, starting with a simplified baseline to facilitate the presentation of the key ideas, as well as to provide a performance benchmark. Second, we introduce several means for evaluating the performance of the models, discuss what they help us to prove and achieve, and indicate directions for future enhancements.

The Baseline Model makes the following initial simplifications:
i Pairwise alignment: we align words from only two languages at a time. We provisionally call them the “source” and “target” languages, although the Baseline Model is actually symmetric. The more advanced models are directional.

ii 1-1 symbols alignment: one source symbol may correspond only to one target symbol—or to the empty symbol to model insertions and deletions.

iii We ignore the context of the aligned symbol pair.

iv Symbols are treated as atoms, i.e., they are not analyzed in terms of their distinctive features.

We make the assumption that the alphabet symbols are identified with sounds. In Uralic data, it is fairly simple to transform written form to phonetic form.

The cost of coding the complete data—i.e., the observed data plus the alignments—is computed using Bayesian marginal likelihood, or prequential coding:

\[
L_{\text{base}}(D) = -\sum_{e \in E} \log \Gamma(c(e) + \alpha(e)) + \sum_{e \in E} \log \Gamma(\alpha(e)) + \log \Gamma \left( \sum_{e \in E} (c(e) + \alpha(e)) \right) - \log \Gamma \left( \sum_{e \in E} \alpha(e) \right)
\]

where \( L_{\text{base}} \) is a kind of entropy term, of the distribution over \( E \), the set of all possible alignment “events”. Each event \( e \in E \) is a pair \((\sigma : \tau)\), with \( \sigma \) and \( \tau \) symbols in the alphabets of the source and target language. \( c(e) \) is the count of occurrences of \( e \) in the complete (aligned) data, and \( \alpha(e) \) is a prior, which is set to 1 in the Baseline model. The use of this objective function in the models is described in detail in related published work, (Wettig and Yangarber, 2011; Wettig et al., 2011). A greedy iterative algorithm searches for a set of alignments that optimize this objective function. It starts with a random alignment of all word pairs in the two languages, and then re-aligns the word pairs one by one, using dynamic programming.

The “rules” obtained by this model are very simplistic: they are embodied in a probability distribution over all possible alignments. To make sure that there is a small number of frequent symbol alignments, the search algorithm tries to reduce the entropy of this distribution, by reducing the value of the cost function as far as possible.

The baseline model has two main problems: 1. it appears to get stuck in local optima, and 2. it produces many alignment events with very low counts, occurring only once or twice.

To address problem 1 we introduce simulated annealing. This yields a reduction in cost, and a better—more sparse—alignment count matrix. To address problem 2, we move to using a two-part code. We expect the data to be sparse—so only a small proportion of all possible events in \( E \) will actually ever be observed. The first part of the code encodes the model: which alignment events in \( E \) are actually observed in the complete data—have non-zero counts. The second part is again the marginal likelihood, but only over the observed events, which is closely related to the entropy of their distribution.

Figure 1 shows the kind of “rules” that the two-part model learns, from Finnish/Estonian data. Finnish and Estonian are closely related, and the model finds that a sound in one often corresponds to the same sound in the other. We also see that Finnish a, i, ä are frequently deleted—align to the empty symbol, marked “.” in the matrix. This happens because final vowels are often lost in Estonian, but the model leaves this and many other phenomena unexplained, which contributes to entropy. Subsequent models discover rules that explain exactly when and why, e.g., these vowels are preserved vs. deleted in Estonian—thereby reducing the entropy.

We introduce improvements over the Baseline Model by relaxing the simplifying assumptions one by one. All our subsequent models use two-part coding as well. The first part codes the model itself, and the second part codes the complete aligned data using the model.

First we extend the model to align multiple symbols, not just 1-1 alignment. We allow up to two symbols in either language as part of a single alignment event. This model therefore captures a simplified notion of context, by learning that certain pairs of consecutive symbols may function as a unit. Also, in this model, the cost function codes separately the different kinds of events—1-1, 2-1, 2-0, etc.—because we expect them to behave differently, i.e., to have different distributions. This model produces lower cost, better alignments, and discovers some cor-

Figure 1: Alignment matrix using the two-part code
rect rules. For example, the model learns that Finnish diphthongs -ie-, -uo-, -yö- always correspond to Estonian long vowels -ee-, -oo-, -õõ-, (which is also where they historically originate). Comparing with Figure 1, this reduces the entropy by explaining away all instances where Finnish i mapped to Estonian e, u to o, and so on.

This removes simplifying assumption ii. We remove assumption i of pairwise language alignment, by aligning multiple languages simultaneously, since our ultimate goal is to model the entire family tree. The baseline model has a natural extension to 3 dimensions, by considering triplets of symbols, rather than pairs. For example, the words meaning “9” in Finnish, Estonian and Mordva, can be aligned simultaneously as:

\[
\begin{align*}
y & . h & d & e & k & s & å & n \\
| & | & | & | & | & | & | \\
ü & . h & e & k & s & a \\
| & | & | & | & | & | & | \\
v & e & χ & . & k & s & a \\
\end{align*}
\]

The cost of each three-way symbol alignment is defined as the sum of three pairwise alignments. This method allows us to easily extend any model for aligning a pair of languages—the baseline, the multi-symbol model, or the models that follow below—to higher dimensions.

Lastly, we introduce a context-based model (which relaxes assumptions iii and iv). The context model takes information from the environment of each symbol into account to predict the symbol with higher probability, to reduce code length. It also treats each symbol as a vector of its phonetic features. A consonant, e.g., has the voiced feature—it can be voiced (b, v, ...) or voiceless (f, v, h, ...); another feature is the place of articulation—where in the mouth it is produced—labial, dental, ..., velar; manner of articulation—stop (p, b, k, ...) vs. fricative (f, v, h, ...); etc. A vowel has a feature for the horizontal position of the tongue—with values from front to back; vertical position—from high to low; and so on.

We have several variants of the context model. The common idea is that the model learns one decision tree for each feature, and the branches of the tree correspond to queries about the context of the symbol we are coding next. The root of the tree simply contains the unconditional distribution of the observed data over the feature’s values. Then we try to split the root. For example, to predict the voiced feature of a consonant symbol on the source level we may consider querying the horizontal feature of the preceding vowel on the target level. There is a fixed number of such candidate queries about the context that the model allows. The candidate that yields the best split—largest drop in cost—is added to the tree, and we proceed down each branch recursively, until the cost cannot be reduced any further. To compute the cost at each node we use the same prequential coding as the baseline model, but over the possible feature values (rather than over all possible symbols, as is done in the baseline).

Each successive extension yields an improvement in code length over the previous models. However, we need to address the question: what do the models tell us about the quality of the data? We first verify that the models behave sensibly by showing that they outperform standard data compressors; this uses the principle that we can claim to have discovered the regularities in the data if and only if we can compress the data effectively.

We next check that the pairwise alignment costs (code-length) can be turned into good measures of language distance. We align all languages in our data set pairwise, e.g., using the two-part 1-1 model. We can then measure the Normalized Compression Distance, introduced in (Cilibrasi and Vitanyi, 2005):

\[
NCD(a, b) = \frac{C(a, b) - \min(C(a, a), C(b, b))}{\max(C(a, a), C(b, b))}
\]

where \(0 < NCD < 1\), and \(C(a, b)\) is the compression cost—i.e., the cost of the complete aligned data for languages a and b. Using algorithms for building genealogical trees from a pairwise distance matrix—for example, UPGMA—we can show that the NCD yields family trees that strongly resemble trees constructed by linguists using traditional methods, see Figure 2.

We examine in detail some of the complex rules of correspondence among different languages, that the algorithm discovers from data, and confirm that they agree with rules posited by linguists. For example, Figure 1 shows that Finnish t corresponds well to Estonian t, but also d; Finnish k corresponds to both k or g, etc. The context model discovers an explicit rule that explains this variation: a voiceless stop consonant in Finnish becomes voiced in Estonian, if it is in the middle of the word and is preceded by a voiced consonant or a vowel. (Finnish preserves the historically original voiceless consonants.)
We wish to demonstrate that the acquired rules are objectively sound. However, direct evaluation of alignments produced by the models is difficult, since we have no gold-standard alignment to compare against, and even if one existed, it would be subjective. We evaluate the quality of the model indirectly by data imputation. The idea of this technique is: given a data set, say, Finnish-Estonian, 

a. withhold one word-pair \((w_F, w_E)\) from the data set;

b. train a model on the remainder of the data; 
c. show the model the hidden Finnish word \(w_F\), and ask it to predict the corresponding \(w_E\). (This is done by selecting the best \(w_E\) over all possible Estonian strings, via dynamic programming.) Repeat for all word pairs in the data set.

The Levenshtein distance from the predicted \(\hat{w}_E\) to \(w_E\), averaged over the entire data set, gives the imputation power of the model: how well the model predicts missing words, given the rest of the data. We then demonstrate that the successive improvements in the models indeed yield improvements in the imputation power. Figure 3 compares the imputation power of the baseline model vs. a context model; each point in the scatter plot compares the models on one of the 10x9 language pairs in our data set. (Lower imputation costs—toward the bottom-right—are obtained for closely related languages.)

The main contribution of this work is a set of methods that allow us to obtain quantitative measures of the goodness of etymological data sets, and to compare different sets by using these metrics. The methods are objective in the sense that they rely only on the data provided—the cognate relationships posited in the dataset—and not on any additional assumptions. We do not aim to construct etymological databases from scratch, nor from raw linguistic data. Rather we start with a given dataset, and study what rules it already inherently contains, what philogenetic trees it inherently induces, and how consistent it is. Although the methods are explored using Uralic data, they can be easily applied to many other language families as well.

We discuss the implications of these methods and how they can support etymological research. One way to view our approach is as an attempt to combine the best of both worlds: allow humans to do what they do best—posit semantic links and propose hypotheses, and allow machines to do what they do best—check the consistency of these assumptions over large data sets.

Our current research includes extension of the methods to the problem of reconstruction. This involves imputation of hidden data in the internal nodes in the family tree, as well as modeling of the temporal aspects of the evolution of the language family, where methods from population genetics may be potentially useful.

Acknowledgments

This research is supported by the Uralink Project, through the Academy of Finland grant 129185.

References


Author Index

Anantharam, Venkat, 51
Baron, Dror, 7
Benatti, Fabio, 11
Bickel, David R., 15
Danielyan, Aram, 41
Egiazarian, Karen, 41
van Erven, Tim, 19
Forchhammer, Søren, 21
García, Jesús E., 25
Giurcăneanu, Ciprian Doru, 29
González-López, Veronica A., 25
Haghverdi, Esfandiar, 31
Harremoës, Peter, 35
Hiltunen, Suvi, 79
Huang, Xin, 21
Hukkanen, Jenni, 39
Katkovnik, Vladimir, 41
Koolen, Wouter, 49
Liski, Antti, 29
Luong, Huynh Van, 21
Makalic, Enes, 45
Poor, H. Vincent, 63
Razavi, Seyed Alireza, 29
Reid, Mark D., 19
de Rooij, Steven, 49
Sabo, Edmond, 39
Santhanam, Narayana, 51
Şchiopu, Ionuţ, 55
Schmidt, Daniel, 45, 59
Tăbuş, Ioan, 39, 55
Tajer, Ali, 63
Topsoe, Flemming, 67
Vitányi, Paul, 71
Watanabe, Kazuho, 75
Wettig, Hannes, 79
Williamson, Robert C., 19
Yangarber, Roman, 79